

# From Reaction Stoichiometry to Life Cycle Assessment: Decision Tree-Based Estimation Tool

Tim Langhorst, Benedikt Winter, Moritz Tuchschnid, Dennis Roskosch, and André Bardow\*



Cite This: <https://doi.org/10.1021/acsenvironau.4c00065>



Read Online

ACCESS |



Metrics & More



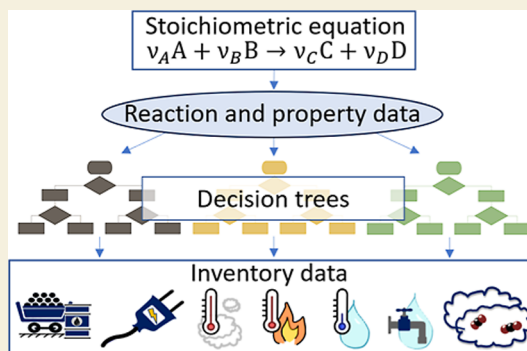
Article Recommendations



Supporting Information

**ABSTRACT:** Decision-making during the early stages of research and development (R&D) should be informed by both economic and ecological perspectives. While early stage cost assessments are well established, life cycle assessment (LCA) is still largely descriptive but should expand to a more prospective tool for early assessing the ecological effects of future processes. Chemical processes should be first assessed as early as when only the reaction equation is known. Our previous comparison of estimation methods based on the reaction equation identified three requirements to foster early stage LCA: (1) estimate inventories rather than final impacts to ensure flexibility, (2) distinguish between processes, as single values cannot reflect the variety of chemical processes, (3) provide a measure of uncertainty. In this publication, we propose regression trees to estimate key inputs for industry-scale life-cycle inventories of chemical processes directly from the underlying reaction equation. In detail, the regression trees yield the raw materials' impact, the direct greenhouse gas (GHG) emissions in CO<sub>2</sub>eq, and the demands for electricity, steam, natural gas, cooling water, and process water. The regression trees outperform the current best available proxy values and provide inventory information that is as accurate as cost estimates. Thus, our work enables decision-makers to consider environmental aspects with the same level of accuracy as costs projections.

**KEYWORDS:** chemical process design, prospective LCA tool, life cycle assessment (LCA), life cycle inventory (LCI), decision trees



## INTRODUCTION

Prospective life cycle assessment (LCA) plays a key role in informing early stage decision-making by evaluating the environmental impacts of future chemical processes. The prospective case of the standardized and holistic LCA approach models and assesses an emerging technology at an early stage of development (e.g., at lab scale) at a later stage of development in the future (e.g., industry scale).<sup>1–4</sup> Conducting LCAs as early as possible is crucial to incorporate sustainability considerations into the design and development of new chemical processes, because at later stages changes become very costly.<sup>5</sup> However, at this early stage, the information available about a process is often limited to reaction equations, restricting the quality of the LCA and its usefulness for decision-making.<sup>6,7</sup> Furthermore, the environmental impacts of background processes connected to the assessed process are often of low quality or unavailable, as only a small fraction of all chemicals produced worldwide is covered by LCA databases.<sup>8</sup> In both early stage assessments and missing background data situations, LCA primarily relies on stoichiometry-based estimation methods and proxies to estimate potential environmental impacts.<sup>9</sup> Such proxy values are usually derived from averages for a set of chemical processes and are commonly applied in life cycle inventory databases.<sup>8,10</sup>

In a previous study, we compared existing stoichiometry-based methods to estimate the inventory data of chemical processes.<sup>11</sup> These methods base estimations of raw material demands on the stoichiometric equation<sup>9,12</sup> and optionally combine them with default proxies for the yield and energy demands.<sup>13–15</sup> We highlighted the potential for enhancing those stoichiometry-based estimation methods and suggested an improved method by combining the yield estimation suggested by Geisler et al.<sup>14</sup> with the average of the process energy demands calculated by Kim and Overcash.<sup>13</sup> However, we found that single proxy values cannot adequately capture the variety within chemical processes.<sup>11</sup> A differentiation of the processes is expected to enhance the prediction accuracy of estimation methods. A promising avenue to realize this differentiation are machine learning models.

Available machine learning models for LCA use mixed integer programming,<sup>16,17</sup> group contribution methods,<sup>18</sup> or artificial neural networks to directly predict the environmental

**Received:** July 30, 2024

**Revised:** June 29, 2025

**Accepted:** June 30, 2025

impacts of target molecules.<sup>19–21</sup> Since these molecule-based approaches inherently neglect all process and pathway information, process-specific machine-learning models have recently been developed using both molecular and process descriptors to predict environmental impacts.<sup>22–24</sup>

These machine learning models quickly estimate the environmental impacts of new processes and thus are a promising addition for the LCA practitioner's toolbox. While the current models focus on aggregated cradle-to-gate impacts,<sup>25</sup> directly estimating the inventory data can increase transparency and interpretability.<sup>9,26,27</sup> In contrast to estimating environmental impacts, estimating the inventory data, at the level of technical flows, improves the understanding and adaptability of the assessment results:

- 1 Data is easy to interpret and cross-check with process experts. Consequently, energy scenario analysis and sensitivity checks can be performed based on the estimated technical flows.
- 2 Technical flow data provides the basis for generating gate-to-gate inventory data which can be used independently from the definition of system boundaries, approaches for solving multifunctionality, the choice of attributional or consequential LCA, Life Cycle Impact Assessment models, and other methodological choices.
- 3 Technical flow data can be used modularly, e.g., if the raw material demands of a process are known, but heat or electricity demands are unavailable to the LCA practitioner.
- 4 Input data can easily be refined, e.g., when new information becomes available for single technical flows.

With these advantages, estimated inventories can also enhance the quality of life-cycle inventory (LCI) databases by addressing data gaps resulting from confidentiality concerns or other limitations.<sup>8,11</sup>

Currently, methods are missing to estimate process-specific inventory data from information that is available at early stages of research. The methods should be transparent and easy to apply for practitioners. Decision trees offer this transparency by grouping similar processes by sequential splits of the data.<sup>28</sup> The sequential splits capture complex interdependencies between input features, while each split is performed based on a simple “yes” or “no” decision.

The concept of decision trees has already been applied in prospective LCA studies. For instance, Karka et al.<sup>29</sup> employed decision trees to estimate environmental impacts for biobased processes, requiring information from laboratory and early conceptual design stages. However, even the reduced set of input features requires information about the number of solvent types, maximum temperature, and the number of processing steps, which may not be available at an early research stage.

Zhao et al.<sup>30</sup> applied decision trees to estimate missing elementary flow data for life cycle assessment. Their method identifies and estimates nonzero elementary flows if less than 10% of all elementary flows are missing. However, the authors acknowledge that substantial research is still needed for practical application as the use case of this study was rather artificial, and the exact number of nonzero flows remains unknown in practice.<sup>30</sup> Pereira et al.<sup>31</sup> used separate decision trees for each design stage to predict steam demands for batch processes. The classification achieved by both studies (i.e., “low”, “mid”, and “high”) is helpful for initial comparisons but

lacks the quantification required for effective early stage decision-making. Decision trees can be used for regression tasks instead of classification to overcome this limitation.

This paper introduces decision trees to estimate important information for chemical processes required to create Life Cycle Inventories: raw materials' global warming impact, direct GHG emissions in CO<sub>2</sub>eq, and demands for electricity, steam, natural gas, cooling water, and process water. The trees are trained for regression and thus provide specific values and the mean absolute error (MAE) as a measure of uncertainty.

All input data for the trees is derived from the reaction stoichiometry and simple thermophysical data available at a very early research stage, even before first lab experiments are conducted. The trees are trained on inventory data of 409 organic chemical processes based on the Process Economics Programme (PEP) yearbook.<sup>32</sup> We share the decision trees in a toolbox, including input and output files and a Python code to use the trees. Additionally, we provide short versions of the decision trees to enable manual calculations.

## METHODOLOGY

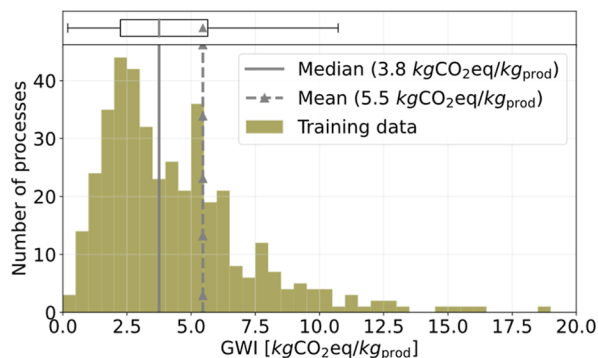
We train decision trees to estimate process-specific energy and utility demands from information available at early research stages. Furthermore, the decision trees provide a coefficient to scale the raw material demands based on the reaction stoichiometry and an estimate for the direct emissions of greenhouse gases. All output data can be estimated based on data available from the stoichiometric equation without knowledge about the process design or reaction conditions. Decision trees are interpretable supervised learning algorithms used for classification and regression tasks.<sup>33</sup> When predicting LCI data on the level of technical flows, the target values are continuous and thus enable regression. The regression tree algorithm builds “if-then-else” rules to split data into separate nodes, based on the provided input features. Multiple splits in sequence result in a tree-like structure. Features can be categorical, such as whether a side reaction exists, or continuous values, such as boiling points.

In the following, we discuss (1) the training data used for this study, (2) the LCA method and assumptions for this study, (3) the features selected as inputs for the decision trees, and (4) the design and fitting approach of the trees. Subsequently, we discuss (5) the performance indicators and model validation approach.

### Training Data

A large set of detailed chemical gate-to-gate inventory data (e.g., energy, utility, and raw materials) derived from the Process Economics Programme (PEP) yearbook<sup>32</sup> is available from our previous studies.<sup>11,34</sup> The PEP yearbook contains validated process reports with information about the required amounts of raw materials and utilities, such as steam, natural gas, electricity, and water. It covers a wide range of chemicals, including platform chemicals, intermediates, and consumer chemicals. The data set is much larger than the data used for calculating the proxy values currently used in Life Cycle Inventory databases. The PEP data is developed for techno-economic assessments and is based on company reports, patents, process analyses, and simulations and is a well-established source for techno-economic process information in industry.<sup>35</sup> We adjusted the data to reflect the requirements of LCA, e.g., by closing the mass balances, and added the related stoichiometric equations to enable the use as training data.<sup>11,34</sup>

For this work, we focus on processes for organic reactions with chemically well-defined inputs and exclude biochemical and electrochemical reactions. This focus is necessary as the number of processes for inorganic (33), biochemical (15) and electrochemical (9) reactions does not allow accurate training of decision trees for these groups of reactions. The remaining data set contains 409 processes producing 148 unique main products, ranging from platform chemicals such as butadiene, ethylene, or methanol to intermediates such as cyclo-dodecatriene or isopropanol and finally to consumer chemicals such as butylated hydroxytoluene or pyromellitic dianhydride (see Supporting Information Table S1). The data set covers a wide range of global warming impacts (GWIs) (Figure 1) and



**Figure 1.** Distribution of the overall GWI (cradle-to-gate, system expansion for valuable byproducts) of the training data as boxplot and histogram (bin width: 0.5), with median (gray line), mean (dashed line and triangle). The GWIs of the processes for the following products are out of the shown range but are included in the training data: hexamethylenediamine, vinylidene fluoride, tetrafluoroethylene, and hexafluoropropylene.

molar masses, from 26  $\frac{\text{g}}{\text{mol}}$  for acetylene to 270  $\frac{\text{g}}{\text{mol}}$  for alachlor, a herbicide banned in the European Union. The GWI reflects the global warming caused by all greenhouse gases emitted during the production of a chemical and is calculated by multiplying the overall emitted amount of each greenhouse gas with its distinct global warming potential over a time horizon of 100 years (GWP100).

### LCA Methodology

We trained seven separate decision trees for estimating the following gate-to-gate outputs: the demands of (1) electricity, (2) steam, and (3) natural gas in MJ, the demands of (4) cooling water and (5) process water in m<sup>3</sup>, (6) the direct greenhouse gas emissions in CO<sub>2</sub>eq, and (7) the raw materials coefficient ( $C_{\text{rm}}$ ). The outputs (1) to (5) reflect technical flows and thus can be used to build life cycle inventories applicable to all impact categories. Here, we use the global warming impact (GWI) as an example to enable comparison across all outputs and show how the single outputs can be combined to assess environmental impacts. The GWI was chosen as it is usually a key criterion for early stage decision-making and is categorized in the highest recommendation level for impact factors in the ILCD handbook.<sup>36–38</sup>

The GWI impact category is also the most comprehensive since the direct greenhouse gas (GHG) emissions from the processes are well-represented in the training data, making them suitable for model training (output (6)). In contrast, data on emissions of other chemicals are scarce in the training set. Thus, we decided to only estimate the direct emissions of

GHGs. For all other direct chemical emissions, decision trees could theoretically be trained. However, a larger and more detailed data set would be required to achieve predictions of comparable quality to those for GHG emissions. Thus, we recommend using generic emission factors as suggested by Jiménez-González et al., Geisler et al., or Hischer et al. to reflect direct emissions relevant to other environmental impacts.<sup>14,15,39</sup> Langhorst et al. provide an overview of those emission factors and give recommendations depending on the intended use case.<sup>27</sup>

Furthermore, we estimate the required amount of raw materials (7) to complement the inventories and estimate the overall GWI of the processes. The raw materials coefficient ( $C_{\text{rm}}$ ) is a dimensionless factor that quantifies how much of the GWI of the raw materials is accounted for through the stoichiometric use of reactants in the process (eq 1). This factor accounts for reaction yield, potential excess of reactants, and the potential use of additional auxiliaries, such as solvents. Thus, this factor streamlines the calculation of the overall raw materials' GWI from the reaction equation without the need to know the yield for each reactant and the reaction and separation conditions. While this factor is named  $C_{\text{rm}}$  to avoid confusion with the yield, it serves as an alternative to a generic proxy yield for scaling the stoichiometric material demands to generate inventories. However, similar to a generic proxy yield, raw material demands calculated from  $C_{\text{rm}}$  should be used with caution, particularly when calculating other impact categories, because generic proxies and the  $C_{\text{rm}}$  provide only one value for all reactants in a given reaction.

$$C_{\text{rm}} = \frac{\sum \text{GWI}_{\text{reactants, stoichiometry}}}{\sum \text{GWI}_{\text{raw materials, final process}}} = \frac{\sum \left( \text{GWI}_i \times \frac{m_{i, \text{stoich}}}{m_{p, \text{stoich}}} \right)}{\sum \left( \text{GWI}_i \times \frac{m_i}{m_p} \right) + \sum \left( \text{GWI}_j \times \frac{m_j}{m_p} \right)} \quad (1)$$

$\text{GWI}_{i/j}$  is the mass-specific global warming impact of reactant  $i$  respective auxiliary  $j$ ,  $m_{i, \text{stoich}}$  is the mass of reactant  $i$  as calculated from the stoichiometric coefficient,  $m_i$  is the mass of reactant  $i$  as required in the real process,  $m_j$  is the mass of auxiliary  $j$  required in the real process. All masses are normalized to the mass of the main product ( $m_p$ ,  $m_{p, \text{stoich}}$ ).

All estimates reflect the material and energy requirements of the chemical process described by the reaction equation. The results of our comparative study are shown on a cradle-to-gate basis. We apply the functional unit '1 kg of main product + side products' to enable users to use their own allocation approach. We credit produced steam, methane, and electricity via substitution if they are labeled as recovered energy in the PEP yearbook to determine the net energy demands of the process.

The decision trees for the raw materials coefficient and the direct emissions have been trained on the GWI calculated with the life cycle impact assessment method ReCiPe(H)V1.13.<sup>40</sup> We chose ReCiPe(H)V1.13 to stay consistent with previous studies using the same data set.<sup>11,41</sup> The impacts of the raw materials and energy carriers are taken from the ecoinvent database 3.5<sup>42,43</sup> using "cut-off by classification" and global market mixes, if available, to ensure comparability to our previous study on estimation methods.<sup>11</sup> When calculating the raw materials coefficient, we assume that (i) oxygen, (ii) methane, and (iii) the combination of carbon monoxide and

Table 1. List of Features Used for the Final Decision Trees<sup>a</sup>

| feature (group)            | feature name           | description   | minimum value in training set | maximum value in training set | median of training set |
|----------------------------|------------------------|---|-------------------------------|-------------------------------|------------------------|
| molecular features         |                        |   |                               |                               |                        |
| boiling points             | BP <sub>maxE</sub>     | maximum boiling point of all involved reactants   | 20 K                          | 844 K                         | 351 K                  |
|                            | BP <sub>minE</sub>     | minimum boiling point of all involved reactants   | 20 K                          | 605 K                         | 90 K                   |
|                            | BP <sub>maxP</sub>     | maximum boiling point of all involved products  | 20 K                          | 762 K                         | 402 K                  |
|                            | BP <sub>minP</sub>     | minimum boiling point of all involved products  | 20 K                          | 538 K                         | 370 K                  |
| molecular weight           | M <sub>WmainP</sub>    | molecular weight of main product  | 26 g/mol                      | 270 g/mol                     | 89 g/mol               |
| key atoms in the reactants | Cl                     | mol of chlorine atoms in the reactants per mol of product                                       | 0 (361)                       | 8                             | 0                      |
|                            | F                      | mol of fluorine atoms in the reactants per mol of product                                       | 0 (403)                       | 6                             | 0                      |
|                            | N                      | mol of nitrogen atoms in the reactants per mol of product                                       | 0 (341)                       | 12                            | 0                      |
|                            | C                      | mol of nonpi/noncyclic carbon atoms in the reactants per mol of product                         | 0 (22)                        | 16                            | 3                      |
|                            | c                      | mol of cyclic carbon atoms in the reactants per mol of product                                  | 0 (297)                       | 18                            | 0                      |
| reaction-based features    |                        |   |                               |                               |                        |
| reactants                  | countReac              | number of reactants in the reaction stoichiometry   | 1                             | 5                             | 2                      |
|                            | stoichioH <sub>2</sub> | mol of H <sub>2</sub> required as reactants   | 0 (334)                       | 10                            | 0                      |
| products                   | countPro               | number of products in the reaction stoichiometry  | 1                             | 5                             | 2                      |
|                            | AddSidePro             | expected occurrence and separation of additional side products from possible side reactions     | 0 (=no) (303)                 | 1 (=yes)                      | 0                      |
|                            | water                  | mass of water [ $\frac{\text{kg}}{\text{kg}_{\text{main product}}}$ ] formed stoichiometrically | 0 (217)                       | 2.1                           | 0                      |
|                            | X <sub>mainP</sub>     | molar fraction of the main product assuming 100% yield  | 0.1                           | 1                             | 0.5                    |

<sup>a</sup>The minimum and maximum values for each feature as used in the final training set indicate the applicability range of this model. In case the minimum value is “0”, the number of “0” in this data set is provided in brackets. Table S1b provides additional information on the 1st and 3rd quartile.

hydrogen will be provided to the final process as (i) air, (ii) natural gas and (iii) synthesis gas. The direct emissions are calculated from the PEP yearbook by closing the mass balances and assuming losses are treated according to Doka and Kätelhön et al.<sup>34,44,45</sup>

To calculate the overall cradle-to-gate environmental impacts, here represented by the GWI, from the decision trees, the stoichiometric impact of the reactants is divided by the raw materials coefficient ( $C_{\text{rm}}$ ) and the estimated demands  $D_x$  are multiplied with their corresponding impacts, here  $\text{GWI}_x$  (eq 2). This calculation allows practitioners to choose their own scenarios, e.g., for energy and water.

$$\text{GWI}_{\text{estimated}} = \frac{\sum \left( \text{GWI}_i \times \frac{m_{i,\text{stoich}}}{m_{\text{P},\text{stoich}}} \right)}{C_{\text{rm}}} + \text{GWI}_{\text{directemissions}} + D_{\text{electr}} \times \text{GWI}_{\text{electr}} + D_{\text{heat,CH}_4} \times \text{GWI}_{\text{heat,CH}_4} + D_{\text{heat,steam}} \times \text{GWI}_{\text{heat,steam}} + D_{\text{cooling}} \times \text{GWI}_{\text{cooling}} + D_{\text{water}} \times \text{GWI}_{\text{water}} \quad (2)$$

The overall quality of the GWI prediction improves with a cleaner energy scenario, as the contribution of the energy demands decreases compared to the raw material demands. The raw material demands can be predicted more accurately based on the reaction equation than the energy demands, which are largely affected by later process design.

Please note that the technical flows estimated by the decision trees are agnostic to the energy scenario. The choice of an energy scenario only becomes relevant when combining and comparing the individual outputs of the decision trees on the level of environmental impacts. In that case, the contribution of the energy demands to the overall GWI

depends on the GWI assumed for the energy inputs. For example, a supply of low GHG emission electricity, as expected in the future, might compensate for a poor estimate for electricity.<sup>11</sup> To prevent any overstatement relating to the accuracy of the decision trees, we use an energy scenario with high GWIs, leading to the least favorable accuracy scenario. For this purpose, electricity is assumed to be provided by combustion of lignite in a power plant.<sup>42,43</sup>

### Feature Selection

Our goal is to ensure the applicability of the decision trees in the early stages of research and development (R&D). Thus, only features about the processes that can be derived directly from the reaction equation or from expert opinions are applicable. However, the features must relate to the property predicted by the trees. We identified the most useful features in a prestudy by calculating the feature importance and systematically deleting features that do not improve the overall model performance based on a leave-one-out cross-validation (see Supporting Information Figure S1). The final decision trees are trained using ten molecular features and six reaction-based features (Table 1). The boiling points have been calculated using statistical thermodynamics from COSMO-RS.<sup>46</sup>

Machine learning models generally have a limited capability to extrapolate, particularly in case the unseen data strongly differs from the training data. Thus, we recommend using the decision trees only for organic reactions within the ranges of the training data provided in Table 1. The trees are not trained for electrochemical or biochemical routes.

### Tree Design and Fitting Approach

We used the decision tree regression model from the scikit-learn module (version 0.23.2) in Python 3.8 to train the decision trees. This model uses a modified version of the



CART algorithm to find the best split.<sup>47,48</sup> Each split partitions the data into two subsets, called nodes, with similar target values by minimizing the variability within both resulting nodes. To determine the best feature and location for the next splits, we let the model minimize the mean absolute error (MAE). Based on this criterion, the predicted values are less prone to outliers because they are based on the median of their node, while the algorithms for all other minimization criteria use the mean.<sup>48</sup> For each possible node, the  $MAE_{\text{node}}$  is calculated (eq 3). The splitting algorithm searches for the split minimizing the weighted sum of both nodes resulting from a split.

$$MAE_{\text{node}} = \frac{1}{n_{\text{node}}} \sum_{y \in \text{node}} |y - \text{median}(y_{\text{node}})| \quad (3)$$

To avoid overfitting, the two most efficient and common approaches for decision trees are (1) to impose a minimum number of samples in a single node and (2) to limit the maximum depth of the decision trees.<sup>47,49</sup> The depth describes the maximum number of splits in sequence within the decision tree. Based on a prestudy, we limited the minimum number of samples in each node to 4, which is approximately 1% of all samples. The decision tree should be as deep as necessary to distinguish between data points and as short as possible to avoid overfitting. The depth was, therefore, set to 7, which allows the tree to grow to a maximum of 128 final nodes. With a fixed minimum of 4 samples per node, the possible nodes could cover all processes within the training data set. The tree algorithm does not necessarily result in equal sample sizes in each node. Thus, the actual number of final nodes is expected to be lower. We checked the performance in a prestudy with a reduced training set and found the depth to be appropriate (see Supporting Information “Prestudy”).

### Performance Indicators

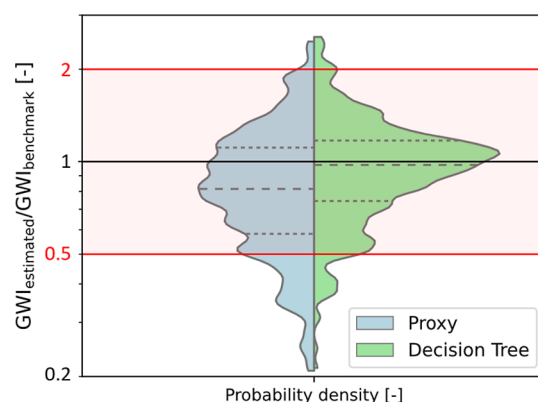
Our goal is to enable early stage LCA with a quality comparable to cost estimates, as both costs and environmental impacts should be used in decision-making during R&D. LCA can benefit from decades of experience in quantifying and communicating uncertainties of cost estimates as collected and published by the Association for the Advancement of Cost Engineering (AACE) and used for techno-economic assessments. Thus, we use established performance indicators from cost engineering for our study: the mean absolute error (MAE) and the percentage of processes outside the target range published by the AACE.<sup>50,51</sup> This comparison basically answers the question: “Can we predict LCI data within the same accuracy ranges as cost data?” The mean absolute error is also used to train the decision tree algorithm and is described in eq 3. The AACE proposes an accuracy range of −50% to +100% for cost estimates at this early research stage.<sup>50,51</sup> For costing, this accuracy range covers approximately 90% of all processes.<sup>50,51</sup> We aim to achieve comparable accuracy for LCA to enable consistent decision-making in cost estimates and LCA.

The results shown in the following are based on unseen data from a leave-one-out cross-validation. A prestudy showed that this approach reflects the performance of the final decision trees on an independent test set (see Supporting Information Table S2). This prestudy showed that 369 processes were sufficient for training decision trees that outperformed generic proxy values on an independent test set of 40 processes.

## RESULTS

Performance of the decision tree regression model.

Our goal is to predict the most common technical flows for life cycle inventories from a reaction equation and improve the accuracy compared to existing proxy-based approaches. To test the overall performance of our prediction, we combine the results of the individual decision trees for the raw materials' impact, the direct emissions, and the utilities (heat from steam and natural gas, electricity, cooling water, and process water) according to eq 2. We use the data from the PEP yearbook to benchmark the estimated results, as the best data available to us. However, we want to highlight that all LCA data can have high degrees of uncertainty and variability and that our results can only be as accurate as the process data provided in the PEP yearbook. Figure 2 compares the accuracy of the decision trees



**Figure 2.** Violin plot comparing the overall GWI estimation performance when using the best available proxy values<sup>11</sup> for each technical flow (left) with the new decision trees (right). Results reflect cradle-to-gate system boundaries in both cases. The performance is measured using a leave-one-out cross-validation. The sample lies on the black horizontal line if the estimated GWI equals the benchmark GWI (perfect prediction). Below this line, the GWI is underestimated; above, it is overestimated. The logarithmic scale shows the equal relevance of under and overestimating by a factor of 2 (red lines). The red area between those red lines (0.5 to 2) provides orientation between Figures 2 and 3. The dashed lines in the violin plots mark the quartile ranges and the median of the data sets.

in predicting the GWI to the best performance achieved by default proxy values.<sup>11</sup> The proxy approach applies a generic demand for heat and electricity to all processes.<sup>13</sup> Furthermore, we assume one of two generic yields, depending on whether side products occur or not.<sup>14</sup> The proxy values used for this comparison have been identified as the most suitable in our previous study.<sup>11</sup> Thus, these values reflect the highest accuracy that can be reached using established default values. For a fair comparison, we used the same LCA method and assumptions for both estimation methods.

Default values cannot reflect the full range of chemical processes due to the high variability of chemical reactions and process setups. Such variability can lead to substantially different demands for utilities like heat and electricity.<sup>11</sup> Decision trees can reflect such variability by offering distinct values instead of a single proxy. Combining the results of the decision trees increases the accuracy of prediction compared to combining default proxies. The increase in accuracy is visible in two effects: First, we see a shift of the mode and median toward 1, reflecting a perfect prediction (from 0.82 to 1.06 and

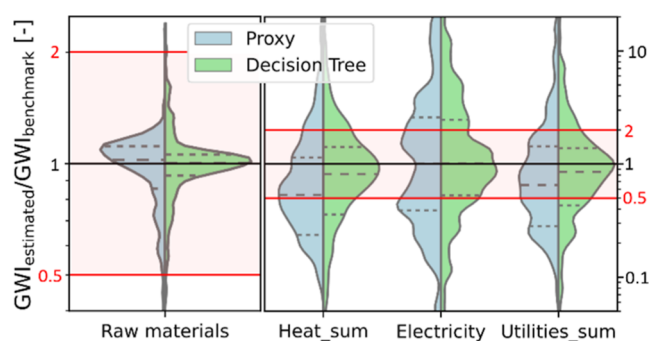
from 0.81 to 0.98). Second, the more narrow distribution indicates an improved precision. Using the decision trees, the second and third quartile of all processes are estimated within  $-26\%$  to  $+18\%$  to a perfect prediction, compared to the method using default proxies where the second and third quartile are predicted in a range of  $-43\%$  to  $+12\%$  to a perfect prediction. Furthermore, only 11% of all processes are estimated outside the target range from  $-50\%$  to  $+100\%$  (see Supporting Information Table S3).

The mean absolute error of the separate utility estimates reduced by 8–20%, resulting in an improvement of 15% for the sum of all utilities (see Supporting Information Table S4). By reducing the mean absolute error, the decision trees also reduce the number of processes estimated outside the AACE target range by 17–31% resulting in an improvement of 19% for the sum of all utilities (see Supporting Information Table S4).

The overall GWI of the raw materials can be predicted with a high level of accuracy, reflecting the strong correlation of the reaction equation and the final raw material demands. The best available proxy-based method distinguishes between processes where side reactions occur and those without side reactions. This distinction already leads to a high level of accuracy with the median at  $1.022 \left[ \frac{GWI_{\text{estimated}}}{GWI_{\text{benchmark}}} \right]$  and the second and third quartile of the processes estimated between  $-15\%$  and  $+11\%$  above or below 1 (perfect prediction). The decision trees further increase the accuracy of the prediction with the median at  $1.003 \left[ \frac{GWI_{\text{estimated}}}{GWI_{\text{benchmark}}} \right]$  and the second and third quartile estimated between  $-7\%$  and  $+6\%$  to a perfect prediction. This increase in accuracy becomes particularly relevant as the GWI of raw materials is often the major contributor to the overall GWI.<sup>11</sup> The required raw materials were calculated using the raw materials coefficient  $C_{\text{rm}}$  and thus may not present other impact categories as good as the GWI. Nevertheless, the  $C_{\text{rm}}$  may still better represent the required amount of raw materials than a generic yield.

The utility demands are affected by the whole process design and become known only at a late stage of process development.<sup>52</sup> Therefore, the information derived solely from the reaction equation cannot fully reflect the utility demands. Consequently, the distribution remains relatively wide compared to the raw materials' GWI (Figure 3). Thus, the decision trees improve the predictions even more. By providing estimates specific to reactions, the decision trees shift the mode and median toward 1 and reduce the variance, indicating they effectively capture the diversity of processes.

In particular, we observe a strong shift of mean (0.66 to 0.88) and mode (0.59 to 0.89) from the lower end of the target range to a perfect prediction for the heating demands (steam and natural gas combined, see Figure 3 "Heat\_sum"). In the case of electricity demand, the median resulting from the proxy use was already close to a perfect prediction (1.00). However, we observe a bimodal distribution characterized by peaks at both ends of the target range (0.44–1.5). This bimodality is due to two clusters of processes that a single proxy cannot reflect. In contrast, the decision trees result in a single mode (0.86), indicating they can distinguish those clusters. For all technical flows, the trees reduce the interquartile ranges, meaning the prediction accuracies of the middle 50% of all processes are less dispersed compared to the estimates by single proxies. This reduced dispersion is also reflected in the

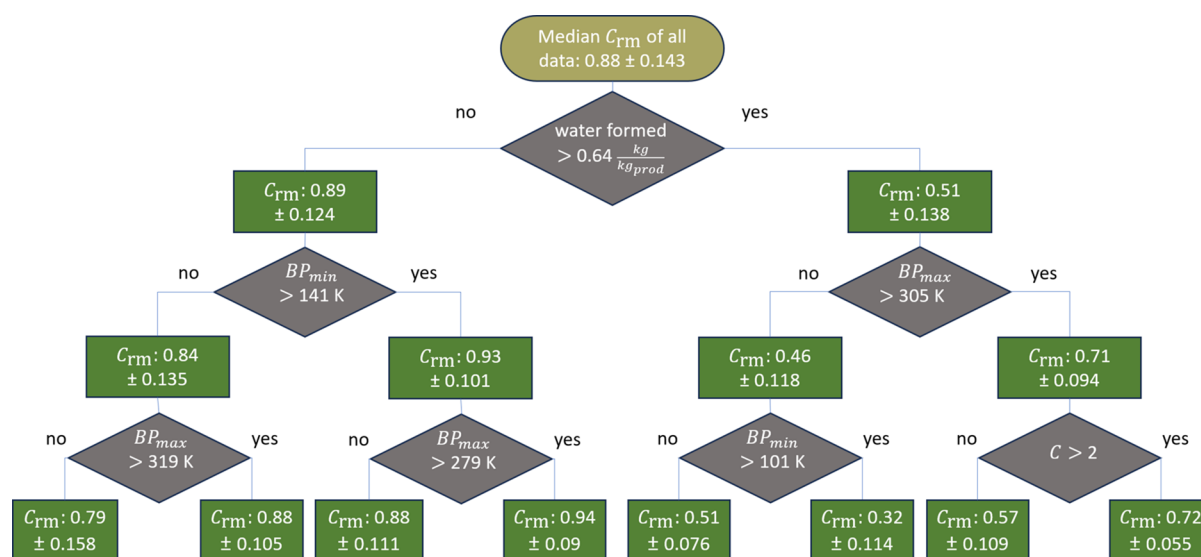


**Figure 3.** Violin plots comparing the estimation performance of the best available proxies (left) with the new decision trees (right) for estimating the aggregate GWI of the raw materials, the total heating demands in MJ, electricity demand in MJ and the aggregate GWI of all utilities. Please note the different ranges on the y-axis. For each utility, the results are equal on the inventory flow level and on the impact level, as both the estimated value and the benchmark value are multiplied with the same characterization factor. The performance is measured using a leave-one-out cross-validation. The sample lies on the black horizontal line if the estimated value equals the benchmark process data (perfect prediction). Below this line, the technical flow is underestimated; above, it is overestimated. The logarithmic scale shows the equal relevance of under and overestimating by a factor of 2 (red lines). The red area between those red lines (0.5 to 2) provides orientation between the two y-axis scales and simplifies the comparison to Figure 2. The dashed lines in the violin plots mark the quartile ranges and the median of the data sets.

last violin plot, which shows the overall accuracy of all utilities combined with regard to their GWI (see Figure 3 "Utilities\_sum", mode shifted from 0.78 to 0.98). Table S4 lists the mean absolute errors (MEA) and the percentage of processes outside the target range for all estimated technical flows and the direct emissions of GHGs based on the units estimated by the decision trees, e.g., energy in MJ and water in  $\text{m}^3$ .

In summary, the decision trees can reflect differences between chemical reactions and thus better reflect the variability in chemical processes. For early stage assessments, the estimated yield (via the raw materials coefficient  $C_{\text{rm}}$ ), direct GHG emissions, and utility demands enable a distinction between chemical reactions during research and development. Hotspots like a high estimated demand for heat can be discussed with process engineers to raise concerns early on and check potential issues.

However, the uncertainty of the estimated results remains high due to the limited information required for their use and the variability in final process designs, even for the same chemical reactions. Process setups can vary significantly even for existing production routes, including our training data. The mean absolute error provided by the decision trees helps assess the uncertainty of the decision trees but misses further uncertainty arising from the limitations of the data set. Thus, estimates should be regarded critically and improved once better data becomes available during process design. If such more detailed data becomes available, the analysis can and should be refined, e.g., using methods from Parvatker and Eckelman or El-Halwagi.<sup>9,53</sup> The Low-TRL Guidelines and the ESTIMATE tool summarize and recommend specific methods for each research and development stage.<sup>27,54</sup> The guidelines are supplemented by multiple worked examples, YouTube videos, and a Web site offering further tools.<sup>55,56</sup>



**Figure 4.** First three layers of the decision tree for estimating the raw materials coefficient  $C_{rm}$ . Without any decision taken (pill shaped box), the median of all data points is at 0.88  $\left[\frac{GWI_{stoichiometric}}{GWI_{benchmark}}\right]$  with a mean absolute error of 0.143. The data is split with each decision (diamonds) to minimize the overall mean absolute error. The rectangles represent the median and the mean absolute error of the resulting split groups (nodes). BP represents the normal boiling point of the reactants. Here, the minimum or the maximum of all reactants are used as splitting criteria. C represents moles of (aliphatic) carbon atoms involved in the reaction per mole of the main product.

### Interpretation of the Decision Trees

Decision trees can visualize the decision path and thus enable an understanding of the learned decisions. The trees were trained using a decision tree regression model without prior chemistry or process engineering knowledge. Thus, we want to understand whether the decisions align with such knowledge. In the following, we exemplarily discuss the first layers of the decision tree estimating the raw materials coefficient  $C_{rm}$  (Figure 4).  $C_{rm}$  reflects the average yield for all reactants and further accounts for auxiliaries (eq 1). The yield for single reactants may vary from this average depending on which reactants may be provided in excess (see Methodology).

The first decision divides the data set into those processes with a maximum of  $0.64 \frac{\text{kg water}}{\text{kg product}}$  formed, and those processes with higher amounts of water formed. If a high amount of water is formed during the reaction,  $C_{rm}$  is expected to be only 0.51. This first decision is reasonable, as there is generally a negative correlation in the training data between the amount of water formed and the required material inputs. Water is an additional byproduct increasing the purification efforts and decreasing the overall yield of the final purified product.

If more than  $0.64 \frac{\text{kg water}}{\text{kg product}}$  is formed, the next decision is whether the maximum normal boiling point of the reactants is above 305 K or not. This split roughly divides the reactions into those with oxygen included in the organic reactant (here, 10 out of 11 reactions contain alcohols or carboxylic acids) and those in which pure oxygen is provided (true for all 21 processes). Thus, the decision tree distinguishes between oxidative dehydrogenation and dehydration reactions as the reason for water formation. Oxidative dehydrogenation processes are often unspecific, with unwanted side reactions likely to reduce the yield, particularly for alkanes.<sup>57,58</sup> Furthermore, often an oxidative dehydrogenation, as indicated in the reaction equation, cannot be realized at a large scale, thereby requiring the use of even more unspecific dehydrogen-

ation routes with low equilibrium concentrations and auxiliaries required for catalyst regeneration.<sup>57,58</sup> Thus, a higher boiling point indicating dehydration processes being used, instead of dehydrogenation, correlates with a lower raw material demand (higher  $C_{rm}$ ). This correlation is reasonable, as a high mass ratio of water indicates that the product molecules are relatively small, compared to the number of functional groups. In our case, the average molecular weight of the main product decreases from  $98 \frac{\text{g}}{\text{mol}}$  to  $61 \frac{\text{g}}{\text{mol}}$  if more than  $0.64 \frac{\text{kg water}}{\text{kg product}}$  is formed. In this case, a high boiling point usually results from functional groups increasing the molecule's polarity. A possible exception to this rule are ammoxidation processes, as those also result in a high amount of water as a byproduct and can be used for small and large molecules to achieve nitriles with a medium  $C_{rm}$  (44–72% in our training data).<sup>59</sup>

If less than  $0.64 \frac{\text{kg water}}{\text{kg product}}$  is formed during the reaction, the raw materials coefficient  $C_{rm}$  remains at 0.89. In this case, the next decision is whether the minimum normal boiling point of the reactants is above 141 K or not. By choosing this temperature to split the data, the model incorporates important information: For our data set, the reactants below this boiling point are hydrogen, carbon monoxide, or methane, which are gases used for C1 chemistry. Thus, if the question is answered with “no”, the routes involve C1 chemistry or the reaction with hydrogen. Those reactions often use the gases in excess and are more likely to require auxiliaries (1.16 auxiliaries in average, compared to 0.75 auxiliaries for the other reactions). Furthermore, the reactants for C1 chemistry have a rather low GWI (average: 2.1) compared to the reactants for other reactions (average: 5.1) increasing the effect of additional auxiliaries on the overall GWI. Thus, a decrease of  $C_{rm}$ , as suggested by the decision tree, seems well aligned with industrial practice.



The third decision layer can be summarized by one general rule: A decrease in the raw materials coefficient  $C_{rm}$  is expected when the boiling points of the reactants are closely aligned. This comparison between highest and lowest boiling point is possible, as the previous decision was based either on the lowest or highest boiling point of the reactants and the current decision layer is based on the remaining. If the boiling points are closely aligned,  $C_{rm}$  is expected to be lower. This expectation is reasonable, as we can expect the separation to be more challenging. Thus, the separation may require additional auxiliaries and result in product losses if their recovery is not feasible.

One decision does not focus on the boiling points but rather checks whether the reactants have more than two carbon atoms in total, which usually means that the final product consists of more than two carbon atoms. If this is not the case, a lower  $C_{rm}$  is estimated. Our data shows that the related processes share a low selectivity of the catalysts, resulting in a smaller  $C_{rm}$  for the process. The side products are recovered and sold for those processes. The small  $C_{rm}$  reflects a high raw materials demand of the overall process, including all side products within the system boundaries. This example shows the general importance of considering the system expansion approach used with the presented decision trees to estimate the utility and raw materials demands of the overall processes. Product-specific assessments could be achieved by crediting the side products in a subsequent step.

The trees created by the decision tree regression model make decisions that can be explained from a chemical and process engineering perspective even though the required knowledge was not explicitly provided to the model, e.g., the reaction mechanism. This knowledge was derived from other related features combining sequential decisions. Furthermore, the decision trees provide the mean absolute error, reflecting the uncertainty of each node and thus allowing for individual uncertainty assessments.

### Accessibility of Decision Trees as a Toolbox

With this study, we aim to provide a user-friendly approach to estimating missing inventory data in the form of important technical flows. Therefore, the final decision trees are made available in a toolbox, which includes a Python script to use the decision trees and a file for manual user inputs for all feature data (further details in readme file). The required reaction data can be derived directly from the stoichiometric coefficients:

- Amount of water formed during the reaction,
- Amount of hydrogen required for the reaction,
- Total number of products and reactants
- Molar ratio of the main product compared to all products.

The required property data are:

- The boiling points of reactants and products,
- The amount of Cl, F, N, and C atoms in the reactants
- The molecular weight of the main product.

Furthermore, users may indicate whether side-products from side reactions are expected and should be separated during the downstream process. Often, this information is not available at the early stages of research. Thus, we recommend setting this value to 0 (no side-product) per default (see [Supporting Information](#) “Additional side products”). In addition to the manual entry of all feature data, the toolbox allows the user to alternatively derive all required feature data directly from the

stoichiometric coefficients and the CAS Registry Numbers. To offer this alternative, the script contains a workflow to calculate the reaction-specific features from the stoichiometric coefficients and search property data in the DIPPR 801 database.<sup>60</sup> This automated search requires access to the DIPPR 801 database.<sup>60</sup> However, the open-source script in the modular toolbox allows practitioners to add further databases, if required.

In addition to the full decision trees in the toolbox, we offer short visual versions of the trees that can be used manually. The short trees represent the first four decisions of the full trees (see [Supporting Information Figures S5–S11](#)). This depth is sufficient to outperform default proxies (see [Supporting Information](#) “Additional side products”, and [Figures S3 and S4](#)).

The decision trees can be used modularly, e.g., only using those outputs that are not available for the studied process. This modularity allows training additional trees for further outputs or improving specific trees when additional training data becomes available. The toolbox, therefore, allows adding further decision trees in the future. Thus, we want to highlight the limitations of the decision trees that users should consider, and that may be closed in the future:

First, the decision tree for direct emission accounts only for greenhouse gases due to a lack of other information on direct emissions (see LCA methodology). In principle, decision trees can and should be trained for all types of direct emissions once sufficient training data is available.

Second, decision trees may be trained to separately predict the yield and demands for solvents and other auxiliary materials. However, predicting solvents and auxiliary materials in detail is challenging, particularly when only the reaction equation is available. Thus, the raw material coefficient introduced here seems to provide an easy-to-use alternative at early stages of research and development.

Third, the decision trees currently do not capture catalysts due to the lack of training data and confidentiality issues. While involved researchers may provide information on catalysts, the challenge of missing background data on catalysts in databases remains.

## CONCLUSIONS AND OUTLOOK

Our study introduces accessible and transparent decision trees that enable the estimation of inventories for organic chemical processes at a very early stage of process development by relying solely on information derived from the reaction equation and thermo-physical properties of the reactants and products. The decision trees can be used in screening studies without detailed process design or operating conditions. Seven individual trees were trained to estimate the raw materials coefficient, the direct emissions in CO<sub>2</sub>eq, and the demands for electricity, steam, natural gas, cooling water, and process water.

Compared to the best available proxy values, the decision trees reduce the mean absolute error of the estimates for the overall GWI by more than 20% and the number of processes estimated worse than the target accuracy by more than 40%. The decision trees outperformed each of the default proxies for the technical flows, with the mean absolute error reduced by 8–24% and the number of processes out of the target range reduced by 17–31%. With the trees, LCA can reach similar confidence levels as expected for cost estimates at the same early R&D stage.<sup>51</sup>



The decisions of the trees align with basic chemistry intuition and process design knowledge. Thus, on the one hand, the decisions made can be discussed with process designers. On the other hand, the decision criteria can be useful for chemists when searching for promising reaction equations. In addition to the estimated value, the trees provide the mean absolute error and thus reflect the uncertainty of the estimate. Both, the provided uncertainty measures and the transparency of the decision paths, enable more comprehensible assessments and more thorough discussion of the limitations than default proxies. Still, the decision trees are also limited by the accuracy of their training data, such that stoichiometry-based estimation methods can only provide initial LCA results. Thus, we recommend discussing the uncertainties and limitations with stakeholders early on and checking the decision path provided by the trees. The decision trees are available in a ready-to-use toolbox together with input and output Excel files, and a Python code to eventually fill in missing inputs and run the decision trees. As an alternative to the manual input of the reaction-based and property data, we offer an autofill option requiring access to the DIPPR database.<sup>60</sup> We thereby reduced the required user input to the CAS Registry Numbers and stoichiometric coefficients. Thus, the toolbox enables the broad use of our decision trees.

Our toolbox allows for the rapid construction of preliminary inventories for industry-scale processes solely from the stoichiometric equation and can be used for processes at the earliest R&D stage as well as for gaps in inventory databases. As the tool calculates each technical flow separately, the flows can and should be adjusted as soon as more detailed information becomes available during the R&D process.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The trained decision trees and the supporting files are openly available on GitHub at: <https://tlanghorst.github.io>.

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsenvironau.4c00065>.

Including a list of included chemicals, information about the Prestudy, and short decision trees (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**André Bardow** – Energy and Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Zurich 8092, Switzerland; [orcid.org/0000-0002-3831-0691](https://orcid.org/0000-0002-3831-0691); Email: [abardow@ethz.ch](mailto:abardow@ethz.ch)

### Authors

**Tim Langhorst** – Energy and Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Zurich 8092, Switzerland; [orcid.org/0000-0002-6362-1192](https://orcid.org/0000-0002-6362-1192)

**Benedikt Winter** – Energy and Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Zurich 8092, Switzerland; [orcid.org/0000-0001-5764-9757](https://orcid.org/0000-0001-5764-9757)

**Moritz Tuchschnid** – Energy and Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Zurich 8092, Switzerland

**Dennis Roskosch** – Energy and Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Zurich 8092, Switzerland; [orcid.org/0000-0001-5144-8348](https://orcid.org/0000-0001-5144-8348)

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsenvironau.4c00065>

## ■ Author Contributions

CRediT: **Tim Langhorst** conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing - original draft, writing - review & editing; **Benedikt Winter** conceptualization, methodology, supervision, visualization, writing - original draft; **Moritz Tuchschnid** formal analysis, software, validation, writing - original draft; **Dennis Roskosch** conceptualization, supervision, visualization, writing - original draft, writing - review & editing; **André Bardow** conceptualization, resources, supervision, visualization, writing - original draft, writing - review & editing.

## ■ Notes

The authors declare the following competing financial interest(s): A. B. has served on review committees for research and development at ExxonMobil and TotalEnergies SE, oil and gas companies that are also active in chemicals production. A.B. serves as a scientific advisor at Carbon Minds GmbH, a company that is offering life cycle assessment databases, training and consulting to clients in the chemical industry. All other authors declare no competing financial conflict of interest.

## ■ ACKNOWLEDGMENTS

T.L., B.W., and A.B. acknowledge funding by NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation, grant number 180544. The authors would like to thank Johanna Kleinekorte, Lorenz Fleitmann and Raoul Meys for providing property and reaction data for most chemicals, Frederike Kuperjans for her support in curating the data and Shaohan Chen, Martin Pillich and Christopher Oberschelp for their valuable contribution to discussions.

## ■ ABBREVIATIONS

|       |   |
|-------|---|
| AACE  | association for the advancement of cost engineering |
| CART  | classification and regression trees                 |
| CAS   | chemical abstracts service                          |
| DIPPR | design institute for physical properties            |
| GW    | global warming impact                               |
| LCA   | life cycle assessment                               |
| LCI   | life cycle inventory                                |
| MAE   | mean absolute error                                 |
| PEP   | process economics program                           |
| R&D   | research and development                            |

## ■ REFERENCES

- (1) *Environmental Management – Life Cycle Assessment – Requirements and Guidelines*, 14044; International Organization for Standardization: Geneva, 2006, Amd2:2020.
- (2) Arvidsson, R.; et al. Environmental Assessment of Emerging Technologies: Recommendations for Prospective LCA. *J. Ind. Ecol.* **2018**, *22* (6), 1286–1294.
- (3) Wang, P.; Robinson, A. J.; Papadokonstantakis, S. Prospective techno-economic and life cycle assessment: a review across

established and emerging carbon capture, storage and utilization (CCS/CCU) technologies. *Front. Energy Res.* **2024**, *12*, 1412770.

(4) Arvidsson, R.; Svanström, M.; Sandén, B. A.; Thonemann, N.; Steubing, B.; Cucurachi, S. Terminology for future-oriented life cycle assessment: review and recommendations. *Int. J. Life Cycle Assess.* **2024**, *29* (4), 607–613.

(5) Ruiz-Mercado, G. J.; Smith, R. L.; Gonzalez, M. A. Sustainability Indicators for Chemical Processes: I. Taxonomy. *Ind. Eng. Chem. Res.* **2012**, *51* (5), 2309–2328.

(6) Zimmermann, A. W.; Langhorst, T.; Moni, S. M.; Schaidle, J.; Bensebaa, F.; Bardow, A. Life-cycle and Techno-Economic Assessment of Early-Stage Carbon Capture and Utilization Technologies – a Discussion of Current Challenges and Best Practices. *Front. Clim.* **2022**, *4*, 841907.

(7) Moni, S. M.; Mahmud, R.; High, K.; Carbajales-Dale, M. Life cycle assessment of emerging technologies: A review. *J. Ind. Ecol.* **2020**, *24* (1), 52–63.

(8) Oberschelp, C.; Hellweg, S.; Bradford, E.; Pfister, S.; Huo, J.; Wang, Z. Poor data and outdated methods sabotage the decarbonization efforts of the chemical industry. *ChemRxiv* **2023**, chemrxiv-2023-8c86t.

(9) Parvatker, A. G.; Eckelman, M. J. Comparative Evaluation of Chemical Life Cycle Inventory Generation Methods and Implications for Life Cycle Assessment Results. *ACS Sustainable Chem. Eng.* **2019**, *7* (1), 350–367.

(10) Althaus, H.-J.; Hirsch, R.; Osses, M.; Primas, A.; Jungbluth, N.; Chudacoff, M. *Life Cycle Inventories of Chemicals: Final Report Ecoinvent Data No. 8 v2.0, EMPA*; Swiss Centre for Life Cycle Inventories: Dübendorf, CH, 2007.

(11) Langhorst, T.; Winter, B.; Roskosch, D.; Bardow, A. Stoichiometry-Based Estimation of Climate Impacts of Emerging Chemical Processes: Method Benchmarking and Recommendations. *ACS Sustainable Chem. Eng.* **2023**, *11* (17), 6600–6609.

(12) Roh, K.; Bardow, A.; Bongartz, D.; Burre, J.; Chung, W.; Deutz, S.; Han, D.; Heßelmann, M.; Kohlhaas, Y.; König, A.; et al. Early-stage evaluation of emerging CO<sub>2</sub> utilization technologies at low technology readiness levels. *Green Chem.* **2020**, *22* (12), 3842–3859.

(13) Kim, S.; Overcash, M. Energy in chemical manufacturing processes: gate-to-gate information for life cycle assessment. *J. Chem. Technol. Biotechnol.* **2003**, *78* (9), 995–1005.

(14) Geisler, G.; Hofstetter, T. B.; Hungerbühler, K. Production of fine and speciality chemicals: procedure for the estimation of LCIs. *Int. J. Metalcast.* **2004**, *9* (2), 101–113.

(15) Hirsch, R.; Hellweg, S.; Capello, C.; Primas, A. Establishing Life Cycle Inventories of Chemicals Based on Differing Data Availability (9 pp). *Int. J. Life Cycle Assess.* **2005**, *10* (1), 59–67.

(16) Calvo-Serrano, R.; González-Miquel, M.; Papadokostantakis, S.; Guillén-Gosálbez, G. Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming. *Comput. Chem. Eng.* **2018**, *108*, 179–193.

(17) Calvo-Serrano, R.; González-Miquel, M.; Guillén-Gosálbez, G. Integrating COSMO-Based  $\sigma$ -Profiles with Molecular and Thermodynamic Attributes to Predict the Life Cycle Environmental Impact of Chemicals. *ACS Sustainable Chem. Eng.* **2019**, *7* (3), 3575–3583.

(18) Baxevanidis, P.; Papadokostantakis, S.; Kokossis, A.; Marcoulaki, E. Group contribution-based LCA models to enable screening for environmentally benign novel chemicals in CAMD applications. *AIChE J.* **2022**, *68* (3), No. e17544.

(19) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokostantakis, S.; Hungerbühler, K. Molecular-structure-based models of chemical inventories using neural networks. *Environ. Sci. Technol.* **2008**, *42* (17), 6717–6722.

(20) Wernet, G.; Papadokostantakis, S.; Hellweg, S.; Hungerbühler, K. Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. *Green Chem.* **2009**, *11* (11), 1826.

(21) Song, R.; Keller, A. A.; Suh, S. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environ. Sci. Technol.* **2017**, *51* (18), 10777–10785.

(22) Kleinekorte, J.; Kröger, L.; Leonhard, K.; Bardow, A. A Neural Network-Based Framework to Predict Process-Specific Environmental Impacts. In *Computer Aided Chemical Engineering, 29th European Symposium on Computer Aided Process Engineering*; Elsevier, 2019; pp 1447–1452.

(23) Kleinekorte, J.; Kleppich, J.; Fleitmann, L.; Beckert, V.; Blodau, L.; Bardow, A. APPROPRIATE Life Cycle Assessment: A PROCESS-Specific, PREDICTIVE Impact Assessment Method for Emerging Chemical Processes. *ACS Sustainable Chem. Eng.* **2023**, *11* (25), 9303–9319.

(24) Karka, P.; Papadokostantakis, S.; Kokossis, A. Digitizing sustainable process development: From ex-post to ex-ante LCA using machine-learning to evaluate bio-based process technologies ahead of detailed design. *Chem. Eng. Sci.* **2022**, *250*, 117339.

(25) Ghoroghi, A.; Rezgui, Y.; Petri, I.; Beach, T. Advances in application of machine learning to life cycle assessment: a literature review. *Int. J. Metalcast.* **2022**, *27* (3), 433–456.

(26) Kleinekorte, J.; et al. Life Cycle Assessment for the Design of Chemical Processes, Products, and Supply Chains. *Annu. Rev. Chem. Biomol. Eng.* **2020**, *11*, 203–233.

(27) Langhorst, T.; et al. *Techno-Economic Assessment & Life Cycle Assessment Guidelines for CO<sub>2</sub> Utilization*. Version 2.0; Michigan, 2022.

(28) Rokach, L.; Maimon, O. Decision trees. *Data mining and knowledge discovery handbook*; Springer Cham, 2005, pp 165–192.

(29) Karka, P.; Papadokostantakis, S.; Kokossis, A. Environmental impact assessment of biomass process chains at early design stages using decision trees. *Int. J. Life Cycle Assess.* **2019**, *24* (9), 1675–1700.

(30) Zhao, B.; Shuai, C.; Hou, P.; Qu, S.; Xu, M. Estimation of Unit Process Data for Life Cycle Assessment Using a Decision Tree-Based Approach. *Environ. Sci. Technol.* **2021**, *55* (12), 8439–8446.

(31) Pereira, C.; Hauner, I.; Hungerbühler, K.; Papadokostantakis, S. Gate-to-Gate Energy Consumption in Chemical Batch Plants: Statistical Models Based on Reaction Synthesis Type. *ACS Sustainable Chem. Eng.* **2018**, *6* (5), 5784–5796.

(32) IHS Markit *Process Economics Program (PEP) Yearbook*; IHS Markit: London.

(33) Loh, W.-Y. Classification and regression trees. *WIREs Data Mining. Knowl. Discov.* **2011**, *1* (1), 14–23.

(34) Käthelhön, A.; Meys, R.; Deutz, S.; Suh, S.; Bardow, A. Climate change mitigation potential of carbon capture and utilization in the chemical industry. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116* (23), 11187–11194.

(35) IHS Chemical *PEP Yearbook Introduction and Appendices*; IHS Markit: London.

(36) European Commission-Joint Research Center *ILCD Handbook - Framework and Requirements for Life Cycle Impact Assessment Models and Indicators*; Publications office: Luxembourg, 2010.

(37) Fazio, S.; Biganzoli, F.; De Laurentiis, V.; Zampori, L.; Sala, S.; Diaconu, E. *Supporting Information to the Characterisation Factors of Recommended EF Life Cycle Impact Assessment Methods*, 2018 1831–9424 KJ-NA-29600-EN-N (online), KJ-NA-29600-EN-C (print).

(38) European Commission-Joint Research Centre-Institute for Environment and Sustainability. *International Reference Life Cycle Data System (ILCD) Handbook: General Guide for Life Cycle Assessment: Provisions and Action Steps*; Publications Office of the European Union: Luxembourg, 2011.

(39) Jiménez-González, C.; Kim, S.; Overcash, M. R. Methodology for developing gate-to-gate Life cycle inventory information. *Int. J. Metalcast.* **2000**, *5* (3), 153–159.

(40) Huijbregts, M. A. J.; et al. ReCiPe2016: A harmonised life cycle impact assessment method at midpoint and endpoint level. *Int. J. Metalcast.* **2017**, *22*, 138–147.

(41) Kleinekorte, J.; Kleppich, J.; Fleitmann, L.; Beckert, V.; Blodau, L.; Bardow, A. APPROPRIATE Life Cycle Assessment: A PROCESS-Specific, PREDICTIVE Impact Assessment Method for Emerging

Chemical Processes. *ACS Sustainable Chem. Eng.* **2023**, *11* (25), 9303–9319.

(42) Frischknecht, R.; et al. The ecoinvent Database: Overview and Methodological Framework (7 pp). *Int. J. Life Cycle Assess.* **2005**, *10* (1), 3–9.

(43) Weidema, B. P.; et al. *Overview and Methodology: Data Quality Guideline for Ecoinvent Database Version 3*; Swiss Centre for Life Cycle Inventories, 2013.

(44) Doka, G. *Life Cycle Inventories of Waste Treatment Services: Ecoinvent Report No. 13*; Swiss Centre for Life Cycle Inventories: Dübendorf, CH, 2003.

(45) Doka, G. *Updates to Life Cycle Inventories of Waste Treatment Services - Part II: Waste Incineration*; Doka Life Cycle Assessments; Zurich, 2013.

(46) Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: A novel view to physiological solvation and partition questions. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 355–365.

(47) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, 2017.

(48) Pedregosa, F.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(49) Bramer, M. *Avoiding Overfitting of Decision Trees*; Principles of data mining, 2007; pp 119–134.

(50) Christensen, P.; Dysert, L. R.; Bates, J.; Burton, D.; Creese, R. C.; Hollmann, J. *Cost Estimate Classification System—As Applied in Engineering, Procurement, and Construction for the Process Industries*; AACE, Inc, 2005; Vol. 2011.

(51) Dysert, L. R.; Christesen, P. *AACE International Recommended Practice No. 18R-97, Cost Estimate Classification System—As Applied in Engineering, Procurement, and Construction for the Process Industries*; AACE, Inc: New York, NY, USA, 2016.

(52) Seider, W. D.; Lewin, D. R.; Seader, J. D.; Widagdo, S.; Gani, R.; Ng, K. M. *Product and Process Design Principles: Synthesis, Analysis, and Evaluation*; John Wiley & Sons, 2017.

(53) El-Halwagi, M. M. *Sustainable Design through Process Integration: Fundamentals and Applications to Industrial Pollution Prevention, Resource Conservation, and Profitability Enhancement*; Butterworth-Heinemann, 2017.

(54) Minten, H.; Vandegehuchte, B. D.; Jaumard, B.; Meys, R.; Reinert, C.; Bardow, A. Early-stage impact assessment tool (ESTIMATE) for the life cycle assessment of CO<sub>2</sub>-based chemicals. *Green Chem.* **2024**, *26* (15), 8728–8743.

(55) Global CO<sub>2</sub> Initiative. The Techno-Economic Assessment and Life Cycle Assessment Toolkit. <https://www.globalco2initiative.org/research/techno-economic-assessment-and-life-cycle-assessment-toolkit/> (accessed Jan 5, 2025).

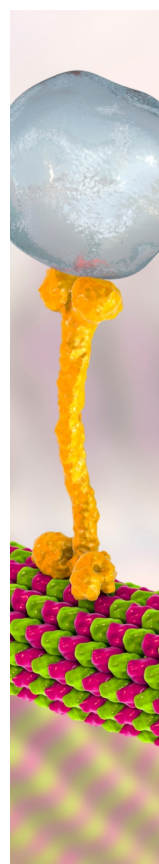
(56) Faber, G.; Mangin, C.; Olfe-Kräutlein, B.; Schaidle, J. A. AssessCCUS: An Integrated Approach for Aggregating Resources to Enable Techno-Economic and Life Cycle Assessment of Carbon Management Technologies. *Front. Clim.* **2022**, *4*, 817211.

(57) Likholobov, V. A.; Moroz, B. L.; van Santen, R. A. Organic Reactions: Sections 4.5 – 4.6:4. In *Handbook of Heterogeneous Catalysis*; John Wiley & Sons, Ltd, 1997; pp 2231–2334.

(58) Stock, C. Hydrogenation and Dehydrogenation. In *Ullmann's Encyclopedia of Industrial Chemistry*; Ley, C., Elvers, B., Eds., 7th ed., Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2023; Vol. 40; pp 1–20.

(59) Grasselli, R. K.; Tenhover, M. A. Ammoxidation. In *Handbook of Heterogeneous Catalysis*; Ertl, G., Knözinger, H., Schüth, F., Weitkamp, J., Eds.: Wiley, 2008; pp 3489–3517.

(60) Wilding, W. V.; Knotts, T. A.; Giles, N. F.; Rowley, R. L. *DIPPR Data Compilation of Pure Chemical Properties*; Design Institute for Physical Properties-AIChE: New York, NY, 2020.



CAS BIOFINDER DISCOVERY PLATFORM™

## BRIDGE BIOLOGY AND CHEMISTRY FOR FASTER ANSWERS

Analyze target relationships,  
compound effects, and disease  
pathways

Explore the platform

**CAS**  
A Division of the  
American Chemical Society