ACS Sustainable Chemistry & Engineering

pubs.acs.org/journal/ascecg

Research Article

APPROPRIATE Life Cycle Assessment: A PROcess-Specific, PRedictive Impact AssessmenT Method for Emerging Chemical Processes

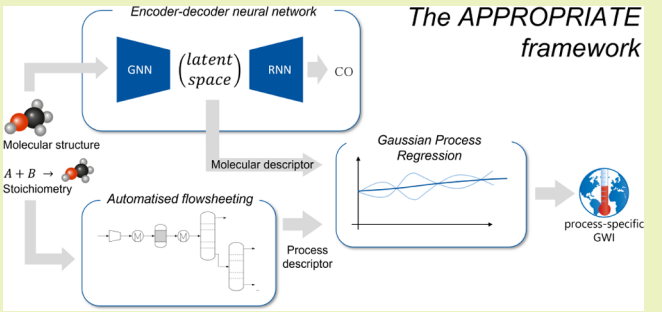
Johanna Kleinekorte, Jonas Kleppich, Lorenz Fleitmann, Verena Beckert, Luise Blodau, and André Bardow*

Cite This: *ACS Sustainable Chem. Eng.* 2023, 11, 9303−9319

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A sustainable chemical industry needs to quantify its emissions and resource consumption by life cycle assessment (LCA). However, LCA requires detailed mass and energy balances, which are usually not available at early process development stages. Here, we introduce a framework (**A PRO**cess-specific, **PR**edictive impact **A**ssessmen**T** method for **E**merging chemical processes, APPROPRIATE) to provide a fully automated, predictive LCA framework for the early phases of process development. Based on Gaussian Process Regression, the framework is already applicable at Technology Readiness Level 2. To overcome the limited LCA data availability, we employ an encoder−decoder network in combination with transfer learning to achieve a latent representation as a condensed molecular descriptor. We further propose to integrate not only molecular but also process descriptors, e.g., the stoichiometric sum of the reactants' impacts. Thereby, we can distinguish between process alternatives and incorporate changes in the background systems. The framework is compared to state-of-the-art predictive LCA approaches and shows increased prediction accuracy in terms of the coefficient of determination of $R^2 = 0.61$ for the global warming impact compared to an $R^2 = 0.3$ in former studies. Highly relevant features are the stoichiometric sum of the reactants' impacts and the condensed molecular descriptors. APPROPRIATE supports decision making in early process development stages by allowing the distinction between process alternatives and quantifying predictions' uncertainty.



KEYWORDS: *Graph neural networks, Latent space, Autoencoder, Gaussian process regression, Automatized flowsheeting*

INTRODUCTION

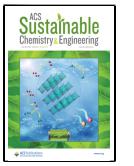
The integration of environmental objectives into process development is key to a sustainable chemical industry. To maximize environmental savings at low costs, the environmental assessment must be integrated in early process development stages.[1]

A holistic and standardized method for assessing environmental impacts is life cycle assessment (LCA).[2] LCA is holistic in two aspects: First, LCA takes into account the entire life cycle of a process or a product starting with the provision of feedstocks and energy, through the production and use of the product, to its recycling or final disposal at the end of its life. Second, LCA includes several types of environmental impacts. Therefore, LCA prevents problem shifting between life cycle phases as well as between environmental impacts.

However, due to its holistic nature, LCA requires a significant amount of data, which is usually not available in the early process development stages.[3] Furthermore, when conducting an LCA for an emerging technology, the time of modeling differs from the modeled point in time, i.e., the technology is usually considered to be scaled up on an industrial scale at a future date. As a consequence, external influences in the background system, i.e., supply processes outside the considered system boundary, that could change in the future must be considered accordingly.[4]

To overcome these challenges, predictive LCA approaches have been presented. These predictive LCA methods use a regression model to predict the environmental impact of a chemical of interest based on easily available descriptors, e.g., physical properties such as the molar mass or the number of various functional groups. Common regression models used are multilinear regression,[5,6] Artificial Neural Networks (ANN),[7−10] or decision trees.[11] In the following, we briefly summarize the employed models, before discussing the prediction accuracies for the example of global warming impacts (GWIs).

Received: December 27, 2022
Revised: May 17, 2023
Published: June 9, 2023



In pioneering work, Wernet et al.[7] compared the prediction accuracy of an ANN with linear regression predicting cradle-to-gate environmental impacts of organic chemical production. The multilinear regression model proposed by Calvo-Serrano et al.[5,6] includes molecular and thermodynamic descriptors of the product to predict environmental impacts. Song et al.[9] also developed a multilayer ANN to estimate environmental impacts for chemicals based on molecular structure descriptors.[9] Baxevanidis et al.[12] recently compared six linear and nonlinear regression models for predictive LCA: multiple linear regression, principal component regression, partial least-squares, Kriging (which equals Gaussian Process Regression), radial basis functions, and a combination of radial basis functions with principal component analysis. Their radial basis function approach corresponds to a single-layer neural network with a radial basis function as activation function in the hidden layer.[13]

The discussed predictive LCA models can be compared for the example of GWI, which is reported in all studies. The comparison reveals an overall poor prediction performance: Wernet et al.[7] report a squared Pearson's coefficient of approximately $\rho^2 = 0.6$. (Wernet et al. name the error measure "Pearson's correlation coefficient (R^2)". However, Pearson's correlation coefficient is usually referred to as ρ and used for evaluating correlations between targets and descriptors of regression models rather than for prediction accuracy. Therefore, the authors of the present work assume that Wernet et al. used the squared Pearson coefficient ρ^2, which is common for the evaluation of regression models. However, the squared Pearson's coefficient ρ^2 is not equal to, and cannot be compared to, the more generally valid coefficient of determination (R^2).[14] For the evaluation of prediction accuracy on test data, the squared Pearson's coefficient ρ^2 yields higher values than the coefficient of determination R^2 and thus, should not be used for assessing prediction accuracy.)[14] The improved model published in Wernet et al.[8] achieves a coefficient of determination of $R^2 = 0.41$. Similarly, Song et al.[9] state a coefficient of determination of $R^2 = 0.48$ on the test set. Calvo-Serrano et al.[15] reported relative errors in the range of 20−44% using leave-one-out cross-validation. However, a high predictive performance in a leave-one-out cross-validation is a necessary but not sufficient condition for high predictive power in general.[16] Thus, evaluating the prediction quality requires an external test set.[16] Baxevanidis et al.[12] published coefficients of determination ranging from 0.1 to 0.26 on their GWI test set. The highest prediction performance is achieved using partial least-squares. In contrast, the Kriging model achieves the lowest accuracy. However, the high R^2 of 0.97 on the training set suggests that the model was heavily overfitted.

All these studies have in common that they use molecular descriptors as input and thus provide component-specific predictions. However, to be able to compare and, if necessary, exclude process alternatives at an early stage of development, a process-specific prediction is required. A first approach for process-specific predictions was presented in our earlier conference publication.[10] In that study, we introduced process descriptors that can encode the reaction equation of the considered technology and thus allow us to distinguish between process alternatives. However, the model accuracy was very low due to limited training data.[10]

Karka et al.[11] propose decision trees to classify the expected environmental impact of a chemicals' production into low, medium, or high. This approach is based on if−then rules, which use a set of critical parameters of the process chain, e.g., the product's molecular structure and process chain-related variables corresponding to chemistry, complexity, and generic process conditions, e.g., the solvent used or the percentage of energy integration. Unfortunately, this process information is not available at early process development stages.

Karka et al.[17] recently compared their decision trees[11] with a process-specific ANN. The authors use 91 impacts of biobased processes as training data and up to 30 molecular and process descriptors as input. Despite the limited training data, the ANN achieves high prediction accuracies of 0.55 to 0.72 as coefficient of determination for GWI, depending on the allocation method and descriptors used. The model still uses the descriptors published in Karka et al.,[11] which are usually not available at early process development stages. Furthermore, the descriptors used do not consider changes in the background system.

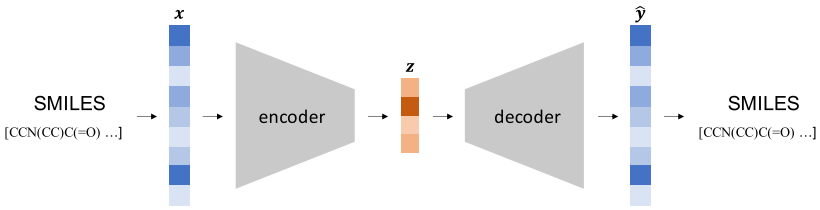
A general drawback of predictive LCA approaches is the dependence on the data basis used to train the models. Data-inherent errors and uncertainties due to noise, false information, or oversimplifying assumptions made in the database setup are propagated in the learning model. Common LCA databases contain approximately 500 chemicals.[18] However, the data underlying these chemicals is often not based on real manufacturing plants but relies on proxies.[19] In addition, the reported product-specific impacts are not based on a consistent methodology. For example, different background systems or allocation methods are used for different chemicals. As a result, the regression model is often trained on inconsistent data, thereby worsening prediction accuracy. Consequently, the obtained predictions are less meaningful. In addition, the low data availability requires limiting the input size of the trained regression model to prevent overfitting.

In conclusion, a predictive LCA model is required which (1) provides a high information density in the used inputs to tackle the challenge of limited LCA data sets available and (2) allows for process-specific prediction to distinguish between process alternatives.

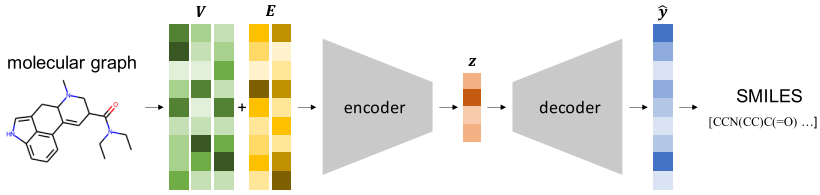
Thus, in this work, we propose a predictive LCA framework consisting of an encoder−decoder neural network and a Gaussian Process regression (GPR), allowing for process-specific prediction of the GWI of organic molecules with high information density in the input. We use transfer learning to train an encoder−decoder neural network translating the molecular structure given as a graph to SMILES.[20] The resulting latent representation captures the most relevant information on the molecular structure and thus provides a condensed molecular representation. Due to the transfer learning on a second chemical database without LCA data, the limited data availability of the LCA data is overcome. The process-specific prediction is enabled through the use of newly introduced process descriptors. As a result, our predictive LCA framework allows us to distinguish between process alternatives and to incorporate changes in the background system with acceptable prediction accuracy.

This work is structured as follows: First, a set of suitable descriptors is developed including the latent representation and fully automatically derived process descriptors. Next, the overall predictive LCA framework is explained. In the results, we first compare the prediction performance of our framework to current predictive LCA approaches from literature. We discuss the influences of specific descriptors on the prediction

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319



(a) Autoencoding framework of the CNN2RNN set-up.



(b) Translating framework of the GNN2RNN set-up.

Figure 1. Schematic illustration of the encoder−decoder neural networks. z and $\hat{y}$ denote the latent representation vector and the predicted sequence, i.e., the SMILES, respectively. (a) Autoencoding structure, which reproduces the input SMILES x. (b) Translational framework, wherein the molecular graph consisting of nodes V and edges E is translated into a latent representation. The latent representation is subsequently translated into a SMILES sequence.

performance and uncover correlations between the most influential descriptors and contributions to the GWI. We then discuss the particular characteristics of the process descriptors in more detail. Finally, we draw conclusions on our findings and suggest further steps.

DEVELOPMENT OF A SUITABLE SET OF DESCRIPTORS

We define the term *early process development stage* as referring to the Technology Readiness Level (TRL) 2.[21] According to Buchner et al.,[21] who adapted the TRL scale to chemical processes, the chemical reaction is selected at TRL 2. Therefore, at this level, only the molecular structure of the main product and the gross reaction equation are known. Subsequently, at TRL 3, the effort for process development increases substantially, as the reaction kinetics have to be determined, physical properties and catalyst synthesis have to be carried out, and first process concepts have to be tested.[21] Thus, the first environmental assessment of the emerging technology should be carried out before TRL 3 is reached. As a result, only available models inputs are the molecular structure of the main product, e.g., given by the SMILES code,[20] and a gross reaction equation.

Thus, all derived features must be fully obtained by predictive models, e.g., by quantum mechanics and statistical thermodynamics. As molecular descriptors, we consider several functional groups, e.g., UNIFAC main and subgroups,[22] and the number of specific atoms and bonds. Additionally, we include thermodynamic properties obtained from quantum mechanical calculations in Gaussian,[23] using geometries and frequencies from B3LYP/TZVP, and energies from the b2plyp/aug-cc-pVQZ level of theory. We use statistic thermodynamics from COSMO-RS[24] on the TZVP level to predict further thermodynamic properties such as boiling points or aqueous solubility. These commonly used softwares are selected since they allow for full automation of the predictions.

Additionally, the predictive LCA framework should be capable of distinguishing between process alternatives. Thus, in this paper, process descriptors are introduced, which can be also obtained fully predictively. To provide a condensed molecular representation and to overcome data limitations, an encoder−decoder approach in combination with transfer learning is introduced afterward. A list of all considered descriptors is given in the Supporting Information.

Condensed Molecular Descriptors: Latent Representations Using Encoder−Decoder Neural Networks. Encoder−decoder neural networks (EDNN) consist of two neural networks:[25] An encoder network encodes a source into a fixed-length continuous vector, i.e., the latent representation. Afterward, a decoder network translates the latent representation into an output. However, in contrast to autoencoders, the output is not necessarily a copy of the input but a translation into an alternative format, e.g., the trivial name of a chemical is translated into the SMILES representation.[26]

The SMILES code is employed to describe the molecule and thus generate molecular features as input to the machine learning model. However, while the SMILES code is a string representation that follows a specific grammar, this grammar gets completely lost during the classical translation of the string into a numeric representation, e.g., using one-hot encoding, to be used as a feature for a predictive model.[27] In contrast, we generate a numerical representation that contains more information about the underlying grammar of the molecule. For this purpose, we connect two neural networks in series (cf. Figure 1), which learn to generate a condensed latent representation z from the generic string representation, i.e., the encoder, and to predict the original string again from this latent space, i.e., the decoder. The final framework then uses only the encoder to generate the latent representation given the molecular structure as SMILES. This latent representation can then be used as a feature in the second step for the predictive LCA model.

We consider two EDNN frameworks in this work: (1) an autoencoder using a convolutional neural network as encoder and a recurrent neural network as decoder (CNN2RNN) as proposed by Gómez-Bombarelli et al.[28] (cf. Figure 1a) and (2) and a translating encoder−decoder neural network consisting of a graph neural network as encoder and a recurrent neural network as decoder (GNN2RNN) (Figure 1b). A CNN2RNN autoencoder has already been successfully used to translate the SMILES of a molecule into a latent representation, which was

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

subsequently used to predict chemical properties.[26,28] Thus, in this work, the CNN2RNN autoencoder is also trained to copy the SMILES. The second framework is based on a translation from a molecular graph representation to a SMILES encoding. This approach is motivated by the recent success of GNN in machine learning applications for chemicals.[29−32] Afterward, the obtained latent representations are used as molecular features for the predictive LCA framework.

Using an encoder−decoder neural network to generate a latent representation from the molecular structure as a feature for predictive LCA has two advantages: First, the training of the EDNN frameworks is not limited by the limited data availability of the LCA database. Instead, the latent representation can be trained on numerous data from open-source databases, e.g., the QM9[33,34] or the ZINC database,[35] which contain several thousand molecules. The QM9 database focus on small organic molecules with up to 9 heavy atoms (i.e., any atoms other than hydrogen), while the ZINC database consists of drug-like chemicals. Both databases are commonly used databases in the context of cheminformatics.[36] Second, the latent size can be used to adjust the compression of the features. Due to the limited availability of LCA training data, the input space of the regression models in the predictive LCA framework has to be significantly reduced to avoid overfitting. Thus, a compressed latent representation increases the information density in this input for the regression model. However, with decreasing latent size, the information loss in the latent representation increases (cf. Figure S1). In the following, the latent size is first fixed to $d = 20$. Afterward, this assumption is confirmed in the results by assessing the prediction performance of the predictive LCA framework for varying latent sizes.

Comparison of the Encoding−Decoding Performance. We trained the CNN2RNN and the GNN2RNN framework on a training set combining data from the QM9 and the ZINC database and measures the prediction performance on a test set consisting of 166 unique chemicals. This test set consists of all chemicals, which are also included in the LCA training set, and can thus be used afterward to train the predictive LCA framework. The prediction performances are compared in terms of the cross-entropy (cf. Supporting Information for details) on the predicted SMILES from the decoder (Figure 2).

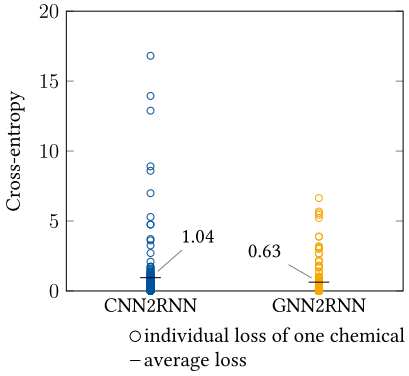


Figure 2. Comparison of the prediction accuracy for the CNN2RNN and a GNN2RNN framework in terms of the cross-entropy on the reproduced SMILES. The black bars indicate the mean cross-entropy over all 166 unique chemicals, while individual losses are plotted by the circles.

The CNN2RNN framework achieves an averaged cross-entropy of 1.04 (black bar), while the GNN2RNN achieves an averaged cross-entropy of 0.63. Generally, a cross-entropy smaller than 0.7 is preferred, since this value refers to an averaged probability greater than 50% for the correct decoding:

$$L_{\text{Cross-entropy}} = -\ln(\hat{y}_i) = -\ln(0.5) = 0.7 \quad (1)$$

Since all probabilities over the possible tokens sum up to 100%, a probability greater than 50% already indicates the correct prediction. Thus, the GNN2RNN framework is beneficial for two reasons: The averaged cross-entropy is lower than the averaged cross-entropy of the CNN framework, and the averaged cross-entropy is lower than 0.7. Additionally, the spread of the cross-entropy over the test set is decreased with a maximum cross-entropy of 6.64 compared to the maximum cross-entropy of 16.8 for the CNN2RNN framework. In both frameworks, the highest cross-entropies are observed on small molecules consisting of two heavy atoms, e.g., methyl chloride, methanol, formaldehyde, hydrogen cyanide, or hydrogen peroxide. However, this observation is expected due to the increased statistical significance of an incorrect prediction of a character in a small molecule than in a large one.

A potential explanation for the success of the GNN2RNN framework could be drawn from the comparison to high-order group-contribution methods. Marrero and Gani[37] introduce groups of second- and third-order to incorporate the spatial positioning of the groups and state an increased prediction performance for physical and thermodynamic properties. Similarly, the CNN2RNN framework only considers certain groups in the SMILES, whereas the GNN2RNN framework can also consider the relationships between the groups through the molecular graph as input.

In conclusion, the GNN2RNN framework outperforms the CNN2RNN framework, and thus, the subsequent evaluations are done with the GNN2RNN framework.

Predictively Available Process Descriptors. The latent representation is a condensed molecular representation that describes the main product of a process. In contrast, process descriptors contain information about the process itself and thus, aim to estimate process-related emissions more directly. A chemical process has four emission sources in a cradle-to-gate system boundary:[19] (1) the emissions related to the feedstock supply, (2) emissions caused by the energy supply, (3) emissions caused by the auxiliary supply, and (4) direct process emissions. Thus, suitable process descriptors should also be able to describe all four sources of emissions. The collection used in this work is introduced in the following sections.

1. Feedstock-Related Emissions. The feedstock-related emissions are encoded by the stoichiometric sum of the reactants' impacts:[38]

$$\text{EI}_{\text{stoichiometric},i} = \sum_k \left(\frac{\mu_k M_k \text{EI}_k}{\mu_i M_i \cdot X \cdot S} \right) \cdot \lambda_i \quad (2)$$

wherein k denotes the running variable over all reactants and i is the regarded product. μ_i and μ_k denote the stoichiometric coefficients of the reactant and the product, respectively, and M_i, M_k describe the molar masses. EI_k describes the environmental impact of reactant k. Following Patel et al.,[38] the allocation factor λ_i is calculated economically based on the product prices if required to account for valuable byproducts:

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

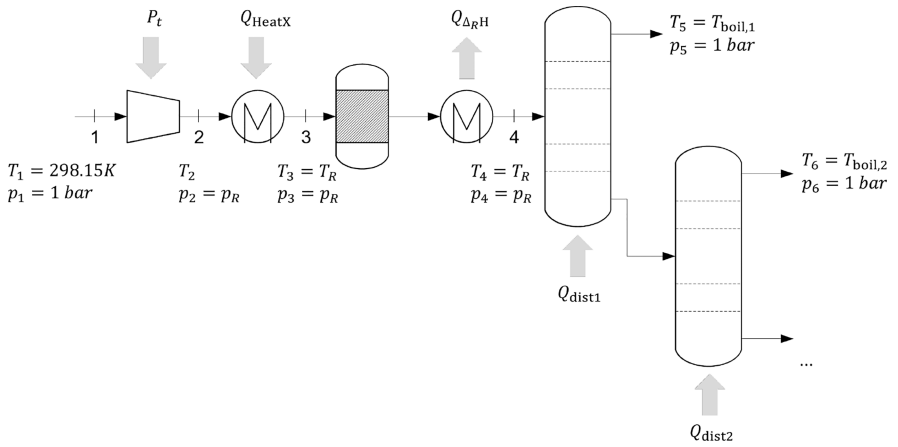


Figure 3. Flowsheet considered for automatized flowsheeting according to the Douglas hierarchy.[42] The electricity demand, power P_t, is considered as the value required to pressurize the reactants from standard pressure to reaction pressure p_R. The heat demand is summed from the heat demand Q_{HeatX} required to heat the mixture from standard conditions to reaction temperature T_R, the heat supply or removal to keep the reactor isothermal $Q_{\Delta_R\mathrm{H}}$ and the distillation energies $Q_{\mathrm{dist},i}$. In addition, a flash is implemented in accordance with ref 42 to separate gaseous from liquid phases before entering the distillation section. However, the flash is neglected in this illustration since it requires no additional energy. Furthermore, we neglected recycling, since otherwise the stoichiometric sum of the reactant's impacts is always determined for 100% yield.

For example, if a reaction produces chemicals A and B with market prices of 1000 and 2000 $ per ton, then we assign 2/3 of the environmental impacts to chemical B. For further information on the concept of allocation factors to handle multifunctional processes in LCA, the reader is referred to ISO 14040,[2] Jung et al.,[39] Heijungs and Guinée,[40] and Heijungs and Frischknecht.[41] For the calculation, a conversion X and selectivity S are assumed to be 85% and 100%, respectively, i.e., side-products caused by a second, simultaneous reaction are neglected. This approach allows us to keep a fixed flowsheet for all reactions. The resulting errors need to be corrected by the data-driven model.

2. Energy-Related Emissions. The energy-related emissions are encoded by descriptors that are a further development of the indicators proposed by Patel et al.[38] Based on the gross reaction equation and the thermodynamic properties of all involved components calculated by quantum mechanics, the Gibbs free energy of the reaction $\Delta_R G$ and the reaction enthalpy $\Delta_R H$ are calculated as additional process descriptors. The reaction enthalpy $\Delta_R H$ indicates whether the reaction is endothermic or exothermic and could therefore correlate with the reaction section's energy requirements.

Furthermore, the number of byproducts of the regarded process and the minimum difference in the boiling points are used as descriptors since a correlation between these indicators and the separation effort of the process is assumed.

To obtain further quantitative descriptors encoding the process energy demand and thus correlating with the energy-related emissions, an automatized flowsheeting approach according to the Douglas Hierarchy[42] is used. Following Douglas,[42] we consider only reactions in the gas and the liquid phase. The generalized flowsheet consists of a compressor/a pump, a heat exchanger, a reactor, and several distillation columns (Figure 3).

In a first step, the reactants are brought from ambient pressure to reaction pressure p_R. The reaction temperature and pressure can be obtained, for example, from the machine learning approach proposed by Gao et al.,[43] which is publicly available. However, the integration of this model into the presented framework of this work is out of scope. Thus, the reaction conditions are provided manually as additional input. When no reaction conditions are available, standard conditions are assumed. The pressure increase is assumed as adiabatic compression of an ideal gas, including interstage cooling, or as a pump of ideal liquids.

The compressed mixture is brought to reaction temperature T_R in the subsequent heat exchanger. The temperature-dependent enthalpies of formation for each component and the enthalpies of vaporization are obtained by quantum mechanical calculations and COSMO-RS.

In the third step of the flowsheet, the reactor, a simultaneous chemical and phase equilibrium, is calculated as proposed by Scheffczyk.[44] The required NRTL parameters are predicted using COSMO-RS. The reaction enthalpy is assumed as the heating or cooling demand of the reactor. Additionally, we consider a stoichiometric reactor with fixed yield, here assumed to be 85%, as an alternative reactor model.

The choice of the reactor model significantly influences the overall estimated energy demand. A comparison of the reactor models and their influence on the energy demand of the process is given in the Supporting Information. In conclusion, the stoichiometric reactor outperforms the equilibrium-based reactor because the equilibrium reactor calculates low conversion rates (<50%) for 41 of 301 processes. These low conversion rates lead to a strong increase of the separation energies resulting in largely overestimated process energies in total (cf. Figure 5a).

As the last step, the reactor output is separated into pure product streams using a sequence of distillation columns. The sequence is designed following the design procedure proposed by Douglas,[42] starting with the sorting of all components in the reactor outlet according to their boiling points. Afterward, the components are classified as product, recycle, or waste stream. Recycle and waste streams with neighboring boiling points are not further separated as these streams can be recycled or disregarded in a none-pure state. Product streams are always fully purified (sharp splits). The number of output streams then determines the number of columns. The sequence order of columns is defined by the boiling points of the head products, starting with the separation of the lowest boiling

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

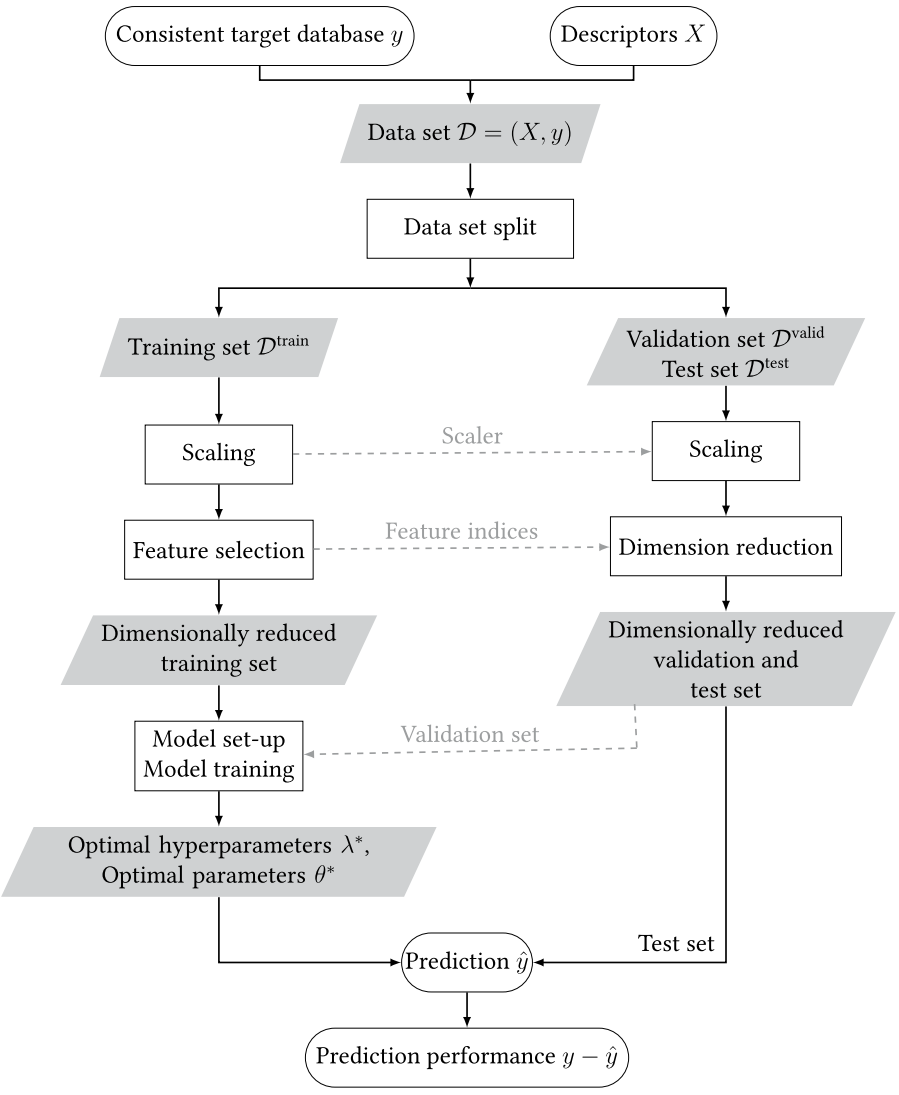


Figure 4. Algorithm of the proposed framework for predictive life cycle assessment.

component. The component to be purified is assumed to be always separated as the head product. Thus, the distillation sequence also specifies the mixture compositions in each distillation column. The required separation energy for each distillation column is calculated with a pinch-based process model[45,46] or using process shortcut equations.[47] The required NRTL and Antoine parameters for the pinch-based process model are estimated using COSMO-RS and are thus obtained fully predictively. Similarly, the Antoine parameters required for the shortcut equations are obtained from COSMO-RS.

In contrast to the reactor models, the choice of the distillation model has little influence on the overall predicted energy demand of the process (cf. Figure S6). Thus, a stoichiometric reactor is used in combination with the shortcut equation describing the distillation columns for subsequent analyses.

The overall generalized flowsheet is used to estimate the heat and power demand for the process as additional descriptors. Details on the modeling assumptions and the estimated energies are given in the Supporting Information.

Furthermore, an approximated impact $\mathrm{EI}_{\mathrm{heuristics}}$ is calculated as the sum of the stoichiometric sum of the reactants' impacts $\mathrm{EI}_{\mathrm{stoichiometric}}$ (eq 2) and the impact due to the estimated energy supply according to

$$\mathrm{EI}_{\mathrm{heuristics}} = \mathrm{EI}_{\mathrm{stoichiometric}} + \sum_j \mathrm{IS}_j \cdot E_j \quad \text{with } j \in \{\text{heating, cooling, electricity}\} \tag{3}$$

where IS denotes the specific impacts per MJ energy E_j, including heat, cooling energy, and electricity. Additionally, the process shortcuts are used to quantify further descriptors proposed by Patel et al.,[48] e.g., the mass fraction of the product or water at reactor outlet.

3. Auxiliary-Related Emissions. The required auxiliary materials, e.g., solvents or catalysts, of a process are not yet known in TRL 2 and are highly dependent on the process design. In addition, in conventional LCA studies, the influences of these auxiliary materials are often neglected due to a lack of data.[49] Therefore, no descriptors for auxiliary materials are proposed in this work.

4. Direct Process Emissions. Direct emissions occur as an undesired side product during the production of the actual main product. The emissions therefore do not necessarily consist of the same elements as the main product but also may

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

contain elements of the byproducts produced according to the reaction equation. While the elements of the main product are already contained in the molecular descriptors describing the main product, further descriptors are needed to describe the elements in the byproducts. Thus, a measure inspired by the carbon efficiency of the Green Chemistry metrics[50] is the number of specific elements in the byproducts. The number is normalized by the products' mass according to the functional unit of 1 kg product of the predictive LCA yielding the direct emissions coefficient:

$$f_i = \frac{n_{i,\mathrm{by-product}}}{m_{\mathrm{product}}} \quad i \in \{\mathrm{C, F, Cl, Br, N}\} \tag{4}$$

The considered elements were chosen to represent all relevant emissions that have significant influences on the ReCiPe 2016 v1.1 impact categories.[51]

PREDICTIVE LCA FRAMEWORK

We propose a fully automatized framework (Figure 4) to predict cradle-to-gate environmental impacts using only the molecular structure of the main product and the gross reaction equation as input, which is thus already applicable at TRL 2. The framework consists of 6 steps: (1) the data collection, i.e., the development of a consistent target database and suitable descriptors, (2) the splitting of the data set into three subsets, (3) the scaling, (4) the feature selection, (5) the automatized model setup, and finally (6) the prediction.

In a first step, we develop a consistent data set of GWIs of chemical processes, which is crucial for the training of machine learning models.[52] The GWI is studied as an exemplary impact category. However, the proposed framework can be applied accordingly to any other impact category.

Parvatker and Eckelman[53] have shown that process simulations are one of the best methods to generate Life Cycle Inventories when primary plant data is not available. Therefore, the LCA data set is developed based on the Process Economics Yearbook by IHS Markit,[54] using all contained processes from Germany (1692 processes) due to good data availability. This economic database is based on press releases, patents and advanced process simulations, providing the highest quality inventory data available for a wide range of processes.[54] Since the proposed LCA framework aims at predicting the production emissions of chemicals, processes not directly describable with a reaction equation are excluded. An example would be a purification process in which no reaction occurs and only a stream is concentrated. As a result, 1178 processes are removed resulting in 514 remaining processes for further preprocessing. In a next step, all processes with the same gross reaction equation but different impacts, e.g., due to different process flowsheets for the same reaction, are averaged to obtain one impact per gross reaction equation (310 remaining processes with unique reaction equation). Furthermore, a k-means cluster algorithm[55] is applied on the descriptors to identify similarities in the represented processes and to define outliers. In an iterative process, first, the number of clusters is optimized using the elbow method.[56,57] Then, clusters containing less than 3% of the processes are identified as outliers and removed from the data set. Afterward, the steps are repeated until each cluster contains less than 80% of the processes. In total, 3 processes are identified as outliers and removed. Last, all processes with a GWI greater than 20 CO_2-eq are removed to ensure a well-balanced training set (see Supporting Information for a detailed discussion). The final data set $\mathcal{D}$ includes 304 unique processes producing 166 unique chemicals (50 aromatic and 116 aliphatic compounds). This data set contains the targets to be predicted by the machine learning model, i.e., the GWIs y, as well as the descriptors describing the processes X.

This data set is then split into three subsets to avoid overfitting and to finally allow for an unbiased evaluation of the overall framework: the training set $\mathcal{D}_{\mathrm{train}}$, the validation set $\mathcal{D}_{\mathrm{valid}}$, and the test set $\mathcal{D}_{\mathrm{test}}$. The validation set of 10% of the data samples is used to set up the chosen regression model. A test set of 20% of the data samples is withheld from the framework to test the optimized model. Thus, an unbiased performance evaluation of the overall framework is possible.

To ensure a complete representation of the data set domain and characteristics in each subset and thus reduce the model's uncertainty, the distribution in training and validation or test set has to be similar. Therefore, 10^6 data splits are randomly generated, and the divergence between the set distributions is measured using the Kullback–Leibler (KL) divergence.[58] Then, the set split with minimum divergence is selected, indicating maximum similarity between the training and validation or test set, respectively.

Based on the training set, the input matrix is scaled using the Robust Scaler from Scikit-Learn.[59] Next, the data sets' dimension is reduced by selecting a suitable subset of features from all collected descriptors using sequential forward selection.[60] An overview of all considered descriptors and the finally used features for the regression models is given in the Supporting Information. The scaler and the indices of the selected features are then passed to the validation and the test set to process those input matrices similarly to the training set. However, no information from the validation and test set is incorporated in the scaling range or the selected features. As a result, these sets still allow for a sound evaluation of the prediction performance during the model setup or the final performance of the overall framework.

The dimensionally reduced training set is then used for training the regression model, while the dimensionally reduced validation set is used for the automated model setup, i.e., the hyperparameter optimization. The final model is defined by optimal parameters θ^* and optimal hyperparameters λ^*. This final model is then applied to the test set $\mathcal{D}_{\mathrm{test}}$ to predict the GWIs $\hat{y}_i$ and to quantify the overall prediction performance of the framework.

Considered Regression Models. For predictive LCA to be used as a decision support tool in early stages of process development, the regression model must not only provide predictions with sufficient accuracy but also be able to quantify the uncertainties of its predictions. A promising model is therefore Gaussian process regression (GPR). As a Bayesian modeling approach, the GPR does not only learn a single parameter but a probability distribution of parameters.[27] Gaussian Process Regression can thus be seen as a Gaussian distribution over functions.[61] Based on the probability distribution, a single-valued estimation can be obtained by the mean of the distribution. The distribution itself reflects the uncertainty of the prediction. In the following, we therefore use a GPR as regression model in our predictive LCA framework. To compare the predictive performance with current literature predictive LCAs, we set up an additional framework using an ANN as regression model.

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

The GPR is implemented using the Scikit-Learn package.[59] The kernels are optimized using the Automatic Model Construction approach proposed by Duvenaud.[62] The maximum number of basic kernels is set to 5 to avoid overfitting. The prior mean is assumed to be zero since deviations from this mean can be modeled by an additional kernel and are thus captured in the automated setup. Overall, 100 iterations of a greedy search are performed wherein base kernels are combined by addition, multiplication or replacement. As base kernels, we consider the following kernels: squared-exponential, rational quadratic, periodic, linear and constant kernels. A detailed explanation of these base kernels and how to express structures with a series of base kernels can be found in Duvenaud.[62]

Applied to the training set used in this work, the compositional kernel search proposed a combination of linear (Lin) and squared exponential (SE) kernels to represent the underlying covariances in the LCA data. The multiplication with a squared exponential kernel converts any global correlation into a local correlation since the squared exponential function decreases monotonically to 0 with increasing distance between two inputs. In contrast, the multiplication with a linear kernel only causes the standard deviation of the model to vary linearly and thus does not affect the correlation of the inputs.

Two kernel combinations are optimized, once for a component-specific and once for a process-specific framework. The resulting kernel combinations are summarized in Table 1.

Table 1. Optimized Kernel Structure for the Gaussian Process Regression[a]

data set	kernel structure
component-specific	SE·Lin + Lin^2
process-specific	SE·Lin^2 + Lin^2

[a]The compositional kernel search from Duvenaud found a combination of squared exponential (SE) and linear (Lin) kernel for the predictive LCA model to be most promising to model the GWI based on molecular descriptors only (component-specific) and the combination of molecular and process descriptors (process-specific).

Applying the rules of the automated description generation from Duvenaud[62] on the created kernel compositions, the kernel search uncovered a smooth function with linearly varying amplitude and a quadratic function between the molecular descriptors and the GWI. Similarly, the kernel combination for the process descriptors can be interpreted as smooth function with polynomially varying amplitude and a quadratic function.

The automated setup of the ANN includes two steps: the hyperparameter optimization providing optimal hyperparameters λ^*, which describe the architecture of the ANN, and the training of the final model leading to optimal model parameters θ^*. The ANN is implemented via the *tensorflow* package,[63] and the hyperparameter optimization is undertaken using the *optuna* package.[64]

The following hyperparameters are optimized: the activation function, the number of layers and neurons per layer, the regularization and learning rates, the number of epochs, and the batch size. As activation functions, the four most common functions are considered: hyperbolic tangent, sigmoid, rectified linear unit, and exponential linear unit. The ranges for the number of layers and neurons per layer are chosen to be [2; 5] and [4; 512], respectively. The limits of the regularization and the learning rate are set to $[10^{-5}; 10^{-2}]$. The number of epochs is limited to the range [50; 1000], and the batch size is chosen to be within [2; 64]. The bounds for the hyperparameters were chosen based on the expert knowledge of the author and were never reached during optimization.

The results of the hyperparameter optimization are summarized in Table 2. Similar to the setup of the GPR, the hyperparameters are optimized once for the component-specific data set and once for the process-specific data set. For both sets, the sigmoid activation function outperforms the alternative activation functions. Both ANNs are optimized to be small with only one hidden layer. However, the process-specific ANN contains approximately twice the number of neurons in the first layer with 308 neurons compared to the 147 neurons in the component-specific ANN. Since a high number of neurons in the process-specific ANN tend to overfit, the batch size is optimized to be 7 times higher, and the learning rate is decreased to approximately half. However, the regularization rate, which is usually increased for architectures that tend to overfit, is approximately half the size of the component-specific ANN. Overall, the optimized architecture of the process-specific ANN implies a higher complexity of the functional relationships between process descriptors and targets compared to the component-specific ANN. This observation is consistent with the results from the kernel search for the GPR.

PREDICTION ACCURACY AND FEATURE IMPORTANCE

The prediction performance of the proposed framework is compared in terms of the root-mean squared error and the coefficient of determination (cf. the Supporting Information for further information on the error measures). As a test set, 20% of the data set is split randomly using the KL divergence approach described above. To ensure comparability of all approaches, this test set is kept constant for all comparisons regarding the prediction performance. In the following, this set is referred to as the *randomly generated LCA test set*. An overview of all considered chemicals and processes is given in the Supporting Information.

Prediction Performance on the Randomly Generated LCA Test Set. The prediction performance is discussed for predictive LCA models on the randomly generated LCA test set using process descriptors and the latent representation as

Table 2. Optimized Hyperparameters for the Artificial Neural Network Modeling the GWI, Once Based on Molecular Descriptors Only (Component-Specific) and Once Using the Combination of Molecular and Process Descriptors (Process-Specific)

data set	activation function	no. hidden layers	no. neurons in hidden layer	regularization	epochs	batch size	learning rate
component-specific	sigmoid	1	147	0.00471	440	2	0.01
process-specific	sigmoid	1	308	0.00218	410	14	0.0052

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

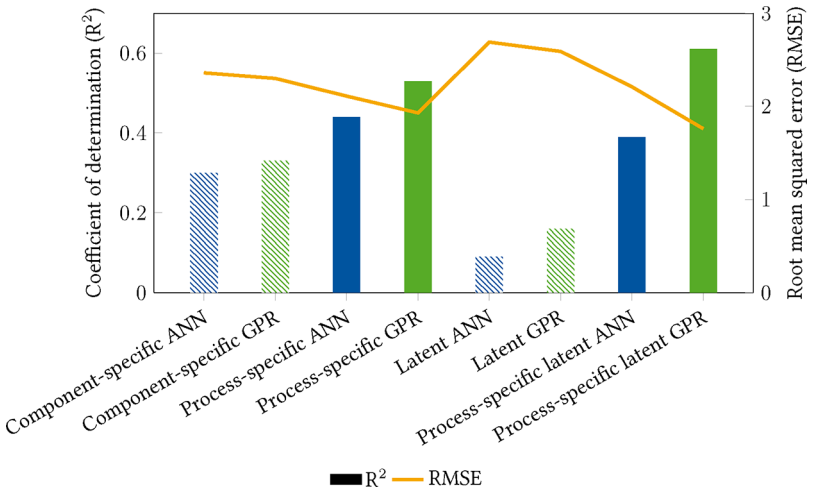


Figure 5. Prediction accuracy of the predictive LCA models using molecular descriptors only (component-specific), molecular and process descriptors (process-specific), only the latent representation (latent), and the combination of latent representation and process descriptors (process-specific latent) as features. The prediction performance of all four combinations is given once for the ANN-based (blue bars) and once for the GPR-based predictive LCA (green bars). Shaded bars indicate component-specific models, while filled bars refer to process-specific models.

features. In total, four combinations of features are considered for the assessment: (1) using molecular descriptors only, (2) using molecular and process descriptors, (3) using the latent representation only, and (4) complementing the latent representation with molecular and process features (Figure 5). The complementing features are chosen using the sequential forward selection as a feature selection method, whereby the latent representation was fixed as input, and the remaining features were selected complementarily. The latent size is kept constant on $d = 20$.

The component-specific ANN represents the state-of-the-art from the literature and achieves an R^2 of 0.3 and an root mean squared error (RMSE) of 2.36 kg CO_2-equiv. In contrast, Song et al.[9] reports a coefficient of determination factor 1.6 higher ($R^2 = 0.48$) for their component-specific ANN. However, Song et al.[9] used leave-one-out cross-validation to determine the R^2, and thus, the results are not directly comparable. Furthermore, the deviation in the prediction performances might result from the different training sets used. Song et al.[9] trained on a training set including only one production process per component. In contrast, the training set used in this work contains on average at least 2 process alternatives producing the same compound. As a result, the component-specific model presented in this work averages between process alternatives' impacts and thus, the prediction performance decreases.

To estimate the influence of the different data sets on prediction accuracy, we can approximate the target values as predictions and determine the maximum possible R^2. Using a product-specific data set, i.e., one like Song et al.[9] used, the R^2 would be 1. In contrast, for the process-specific data set, assuming the mean values from all processes with the same product as the prediction, the R^2 drops to 0.81. In other words, our model can reach a maximum of 0.81, whereas the model from Song et al.[9] can reach 1, since the training data, i.e., what they later have to compare against, is already averaged.

When a process-specific ANN is trained using the process features developed in this work, the prediction performance increases by 46%, resulting in a coefficient of determination $R^2 = 0.44$ (RMSE = 2.1 kg CO_2-equiv). Similarly, the R^2 of the GPR-based model increases significantly by 60%, from 0.33 to 0.53, when extending the molecular descriptors with process features. The RMSE is reduced by about 16% from 2.3 to 1.9 kg CO_2-equiv.

Compared to the R^2 of 0.44 for the process-specific ANN-based model, training the ANN solely with the latent representation reduces the coefficient of determination by a factor of approximately 4.8 to $R^2 = 0.1$. The RMSE is increased to 2.7 kg CO_2-equiv, which equals an increase of 28%. Combining the latent representation with additional process features as input for the ANN, the prediction performance increases up to $R^2 = 0.39$ and an RMSE of 2.21 kg CO_2-equiv compared to the ANN solely trained with the latent representation. Nevertheless, the ANN trained only with molecular and process descriptors still achieves the highest performance when comparing all ANN-based methods.

Similar to the ANN, the prediction performance of the GPR-based model decreases when trained with the latent representation only compared to the process-specific GPR. However, the coefficient of determination only decreases by a factor of 3.4 from 0.53 to 0.16, and thus, the GPR-based model outperforms the ANN-based model solely trained with the latent representation. In contrast, the GPR-based model can nearly quadruple its prediction performance in terms of R^2 from 0.16 to 0.6, when comparing the training with latent representation only and the combination of latent representation and additional process features. Thus, the GPR-based predictive LCA model trained on the latent representation and process features also outperforms the GPR trained with molecular and process descriptors only ($R^2 = 0.53$). In terms of RMSE, the prediction performance of the combined input of latent representation and process features reduces the RMSE by 0.36 kg CO_2-equiv from 2.59 to 1.76 kg CO_2-equiv, which equals a reduction of 32%.

In conclusion, utilizing the process features as input always increases the prediction performance. In contrast, utilizing the latent representation as input does not necessarily improve the prediction performance. Furthermore, the GPR-based frameworks outperform the ANN-based frameworks in all considered input combinations. The GPR-based predictive LCA framework which is trained on the latent representation and process features as input achieves the highest prediction

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

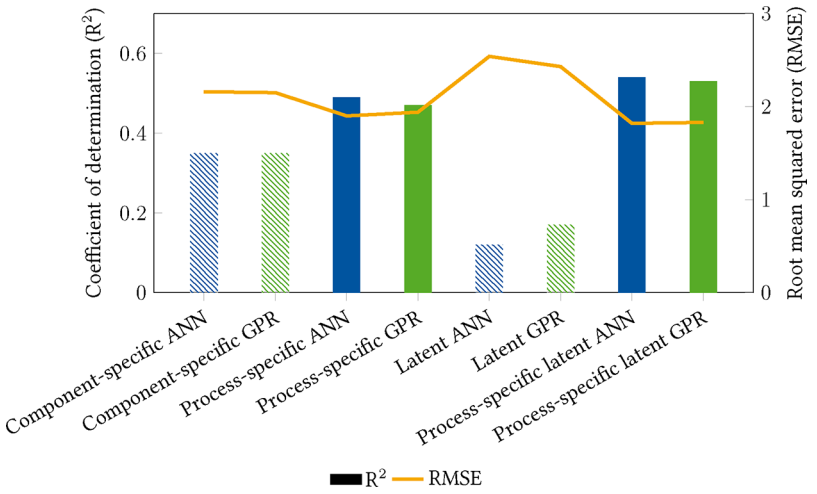


Figure 6. Prediction accuracy of the predictive LCA models averaged over 30 data set splits according to Wernet et al.[8] using molecular descriptors only (component-specific), molecular and process descriptors (process-specific), only the latent representation (latent), and the combination of latent representation and process descriptors (process-specific latent) as feature. The prediction performance of all four combinations is given once for the ANN-based (blue bars) and once for the GPR-based predictive LCA (green bars). Shaded bars indicate component-specific models, while filled bars refer to process-specific models.

performance. The influences of single predictions on the overall models' prediction performance are discussed in further detail in the Supporting Information.

Averaged Prediction Performance on 30 Randomly Generated Test Sets. Wernet et al.[8] evaluated their prediction performances on average over 30 models with varying training/test set splits. Thus, to compare the achieved prediction performance of our models with the results of Wernet et al.,[8] an averaged prediction performance over 30 data sets is required. We therefore trained 30 predictive models for each input configuration not only using the random generated LCA test set with lowest KL divergence but 30 data set splits with the 30 lowest KL divergences in the next step. The achieved prediction performances are then averaged (Figure 6). Generally, the trend in the averaged performances is identical to the trends discussed on the randomly generated test set comparing the various input combinations. However, the differences in the prediction performance between the ANN- and GPR-based models decrease when compared for similar inputs.

Both component-specific models achieve an average coefficient of determination of R^2 = 0.35 and an RMSE of 2.15 kg CO_2-equiv. In contrast, Wernet et al.[8] report an increased R^2 of 0.41, which can be explained again by the different training set including only one production process per component.

When the process features are used as input, the prediction performance increases by up to 40% and 11% in terms of the coefficient of determination (R^2 = 0.47−0.49) and the root mean squared error (RMSE = 1.9−1.93 kg CO_2-equiv), respectively, compared to the component-specific models. However, on average, the GPR-based model deteriorates its prediction performance compared to the randomly generated LCA test set by 20 percentage points on the coefficient of determination. In contrast, the ANN-based model improves its prediction performance by 6 percentage points. As a result, the prediction performance no longer substantially depends on the choice of regression model. In contrast, the choice of the input features gains relevance.

The predictive LCA models trained solely on the latent representation as input perform on average similarly poorly as observed on the randomly generated LCA test set. The ANN-based model achieves an R^2 = 0.13, while the GPR-based model achieves a slightly increased R^2 = 0.18.

The highest prediction performance is achieved with the predictive LCA models using process features and the latent representation as input (R^2 = 0.53, RMSE= 1.82). Similar to the process-specific models trained without the latent representation, the choice of the regression model does not influence the prediction performance.

In conclusion, the selection of the regression model is only less relevant than the selected features. Utilizing the process features as input increases the prediction performance substantially. The prediction performance can be further increased when the process features are extended by the latent representation. Thus, in the following section, the influence of the features on the prediction performance is discussed in detail.

Influence of the Latent Size on the Prediction Performance. In Figure 7, the coefficient of determination is plotted against the size of the latent representation, which is then used as a feature for the predictive LCA model. As a predictive LCA model, the GPR-based model is used, which is identified as most promising for the randomly generated data set (Figure 5). Since the feature space of the LCA regression model is limited to 35 features, the maximum latent size is set to 30, allowing for including at least five additional features, e.g., process descriptors. For each latent size and each feature combination, the training data for the encoder−decoder neural network is split ten times into a training set and a validation set, allowing for the training of ten models for each combination of features and latent sizes. To limit the computational effort, each model is only trained for one epoch. The test set is always kept constant, using the same chemicals contained in the LCA database. Afterward, the prediction performance on this test set is averaged over the ten models for each combination.

Overall, the averaged coefficients of determination achieved by the latent representation-using models are substantially

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

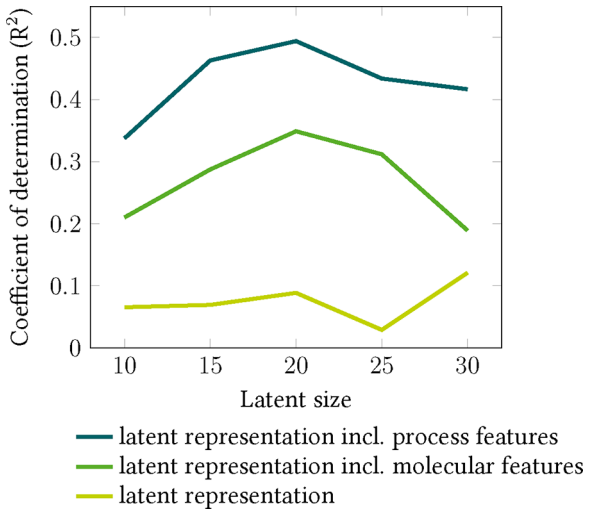


Figure 7. Trade-off between the latent size and the prediction accuracy of the GWI using the GPR-based predictive LCA framework.

lower than the ones presented in Figure 5 due to the limited training of one epoch only. The predictive LCA models trained with the latent representation and additional molecular features (green line) and with the latent representation and additional molecular and process features (petrol blue line) clearly show a maximum at a latent size of $d = 20$ and an R^2 of 0.35 and 0.49, respectively. Therefore, the assumption of fixing the latent size to $d = 20$ can be confirmed.

Influence of the Selected Features on the Prediction Performance. As demonstrated in the previous section, the chosen features influence the prediction performance substantially. Thus, the importance of each feature is analyzed according to Breiman:[65] For one feature at a time, the value is randomly shuffled from a uniform distribution over the entire value range, while all other features are held constant. The feature importance is then measured by the introduced degradation of the predictive accuracy on the targets normalized on 100%.

The feature importance analysis is conducted for the process-specific latent GPR (Figure 8), which uses 34 features in total as input. 20 of these 34 features refer to the latent representation, while the remaining 14 features are composed of 10 additional molecular and 4 process features (cf. Table S15). The stoichiometric sum of the reactants' impact $EI_{stoichiometric}$ is the most important feature with an importance of 18%. However, this feature is also the only process feature achieving a high importance for the prediction. Further utilized process features such as the estimated heat demand or the approximated environmental impact $EI_{heuristics}$ achieve importance scores lower than 3%. Entries of the latent representation are ranked 2−6, and at rank 8 with a feature importance of 4−17% thereby explaining the increased prediction performance of the process-specific latent predictive LCA frameworks compared to solely process-specific frameworks. 24 features achieve an importance less than 3% and are thus of only low importance.

In particular, the high importance of the seventh dimension of the latent representation is remarkable. Although a direct physical interpretation is not possible, an apparent correlation exists between the values of this dimension and the structure of the main product: all chemicals with low values in dimension 7 contain a benzene ring, while high values describe aliphatic compounds, mostly acids or acetates. As a result, the aromaticity of the main product seems to influence the GWI of the production process. Nevertheless, this influence is not apparent since dimension 7 of the latent representation is only poorly linearly correlated with the target GWI, with a negligible correlation coefficient of $\rho^2 = 0.01$. The nonlinear machine learning model is still able to identify this relation.

The scoring of the process features $EI_{stoichiometric}$, $EI_{heuristics}$, and the estimated heat demand can be explained by a closer look at the contributions to the target process impact. The GWI of a process is composed of the feedstock- and the energy-related emissions, as well as emissions caused by auxiliaries and direct process emissions. The stoichiometric sum of the reactants' impact $EI_{stoichiometric}$ is intended to provide information on the feedstock-related emissions as a feature. Comparing $EI_{stoichiometric}$ with the feedstock-related contributions to the target GWIs (Figure 9) reveals a strong correlation ($\rho^2 = 0.75$) and thus, a high informative value of the feature.

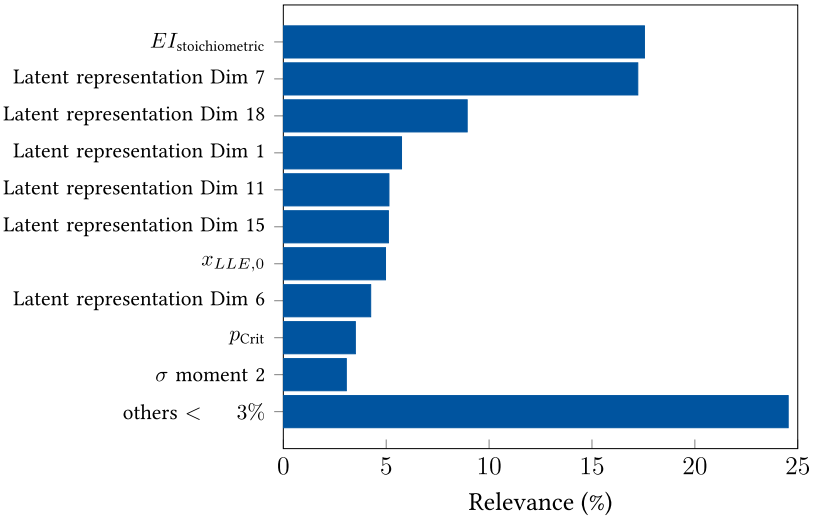


Figure 8. Feature relevance for the life cycle assessment prediction using the latent representation as feature. In total, 34 features are used as input, of which 24 each achieved a relevance of less than 3%.

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

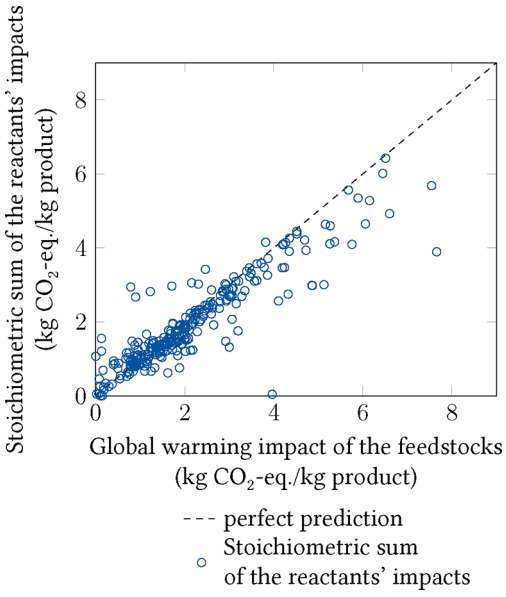


Figure 9. Parity plot of the stoichiometric sum of the reactants' impacts against the feedstock-related impacts of the targets. In general, the stoichiometric sum of the reactants' impacts represents a lower bound for the impact of a process. However, suppose a process produces a valuable byproduct that is not included in the stoichiometric equation but is produced by a second, parallel side reaction. In that scenario, the stoichiometric impact can also take on a higher value than the actual process impact due to allocation.

Similarly, the estimated heat demand is intended to describe the energy-related emissions. However, the resulting impact based on the estimated heat demand correlates poorly (R^2 = 0.03) with the energy-related contributions to the target GWIs (Figure S8, orange points), and thus, this feature contains only low informative value for the predictive LCA model. The low correlation is caused by a strong underestimation of the process energies. A detailed discussion of the accuracy of the process models is given in the Supporting Information.

Since the feature $EI_{heuristics}$ corresponds to the sum of the estimated feedstock and energy-related impacts, this feature combines the information content of $EI_{stoichiometric}$ and the estimated energy demand. In addition, however, this feature further contains the information of the selected energy scenario, since the specific impact per MJ heat is also passed on to the predictive LCA model. Nevertheless, the feature $EI_{heuristics}$ achieves only a low importance of 0.7% since it is strongly correlated to the $EI_{stoichiometric}$ and the heat demand. As a result, the predictive LCA model still has access to the required input information by the correlated features although the $EI_{heuristics}$ is varied. The low feature importance is thus not reliably interpretable.

To gain further insights into the importance of the latent representation and its influence on the prediction performance, the feature importance analysis of the process-specific latent GPR can be compared to the analysis conducted for the process-specific GPR trained without latent representation (Figure 10). The latter model uses 18 features as input from which 7 features are process features and the remaining 11 are molecular features (cf. Table S15). Similar to the process-specific latent GPR, the stoichiometric sum of the reactants' impacts achieves the highest importance. However, the importance increases substantially, up to 33%, while the molecular features' importance decreases compared to the

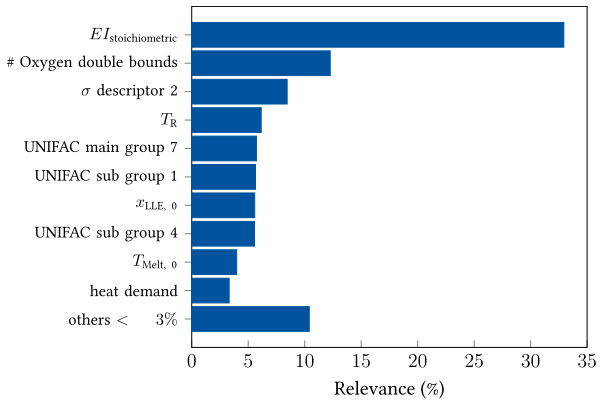


Figure 10. Feature relevance for the life cycle assessment prediction without the latent representation. In total, 18 features are used as input, of which 8 each achieved a relevance of less than 3%.

latent representation dimensions. Thus, the molecular features contain less informative value compared to the latent representation.

This observation can be explained exemplarily by the second most important feature, the number of oxygen double bonds. The number of oxygen double bonds contributes 12% to the prediction of the process-specific GPR without latent representation. Thus, the importance is 5% percentage points lower than the seventh dimension of the latent representation, the second most important feature of the process-specific latent GPR. Both features contain the information regarding whether the molecule considered is an acid or acetaldehyde. However, the seventh dimension of the latent representation contains additional information about the aromaticity of the molecule and has thus overall an increased informative value.

In summary, the feature importance analysis indicates that the process-specific predictive LCA models use the stoichiometric sum of the impacts of the reactants $EI_{stoichiometric}$ as the basis for the prediction. A component-specific offset is then added based on the molecular features to approximate the energy-related impact.

PROCESS DESCRIPTORS AS KEY FEATURE

To further investigate the benefits of process descriptors, a second data set split is performed including at least two process alternatives for each chemical in the test set. Therefore, 40 samples (13% of the overall LCA data samples) are chosen manually as a second test set. This test set is referred to as the *manually selected LCA test set* in the following section.

Process Descriptors Allow Us to Distinguish between Process Alternatives. The process-specific latent GPR model is used to predict the GWIs of producing each chemical via two alternative process routes. In Figure 11, the difference between these two process alternatives' predictions is given. If the trend between the process impacts is represented correctly by the predicted impacts, then the differences for both the targets and the predictions have the same sign, and the data points end up in the first or third quadrant. To simplify the diagram, the difference between two process alternatives is always defined such that the representation ends in the first or second quadrant, i.e., $\Delta = \hat{y}_1 - \hat{y}_2$ with $y_1 > y_2$, where $\hat{y}$ and y describe the prediction and the target impacts, respectively.

Overall, the process-specific latent GPR can correctly capture the trend in 18 out of 21 process comparisons

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

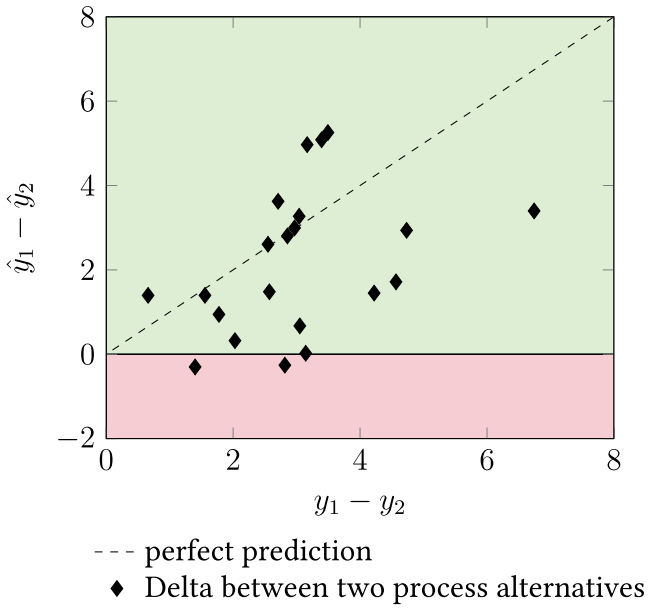


Figure 11. Parity plot of the GWI differences between process alternatives yielding the same product. The difference $\hat{y}_1 - \hat{y}_2$ refers to the difference of the predicted impacts, while the difference $y_1 - y_2$ describes the difference of the targets. For all comparisons between two process alternatives whose points belong to the first quadrant (green area), the trend can be predicted correctly (18 out of 21 comparisons).

(86%). For the 3 process comparisons for which the trend cannot be predicted correctly, the predicted differences are close to or below 0. The predictive LCA model cannot distinguish these processes with sufficient accuracy because the stoichiometric impacts of the process alternatives differ by only 0.3−0.6 CO_2-equiv. However, the stoichiometric sum of the impacts of the reactants is the most important process feature, to which the predictive LCA model basically adds a component-specific offset. Therefore, the prediction leads to similar GWIs for both process alternatives. In contrast, the target GWIs of the respective process alternatives differ substantially, as the processes have significantly different energy demands. However, these energy demands cannot yet be adequately described with the features used, which leads to poor predictions for energy-intensive processes.

Nevertheless, the majority of the process alternatives is successfully distinguished using the proposed features. Therefore, the predictive LCA framework allows for the screening of process alternatives in early stages of process development.

Process Descriptors Allow Description of the Background System. An advantage of process descriptors is that, in addition to information about the process itself, changes in the background system can also be considered. As an example, the GWI for CO_2-based methanol is considered in Figure 12 for *today* and *future* scenarios. The GWI is compared to the incumbent methanol production from fossil-based natural gas. We chose methanol as an example because we had complete LCAs for different background systems for this case study.

In the *today* scenario, carbon dioxide is considered as a byproduct of the ammonia production. Hence, a part of the process emissions is attributed to this CO_2 following the modeling in the internal database provided by the Gabi software.[66] The hydrogen is provided by steam methane reforming. The electricity is supplied by the current EU grid mix 2020. In the *future* scenario, electricity is supplied by wind power, and as a result, hydrogen with a lower GWI is obtained by electrolysis. The CO_2 is captured from a power plant, obtaining a credit for the avoided emission.

Currently, methanol is produced via fossil-based means by the oxygenation of natural gas. The resulting GWI accounts for 0.38 kg CO_2-equiv. The CO_2-based production of 1 kg methanol, as proposed by Rihko-Struckmann et al.,[67] requires 1.38 kg of CO_2 and 0.197 kg of hydrogen as input.

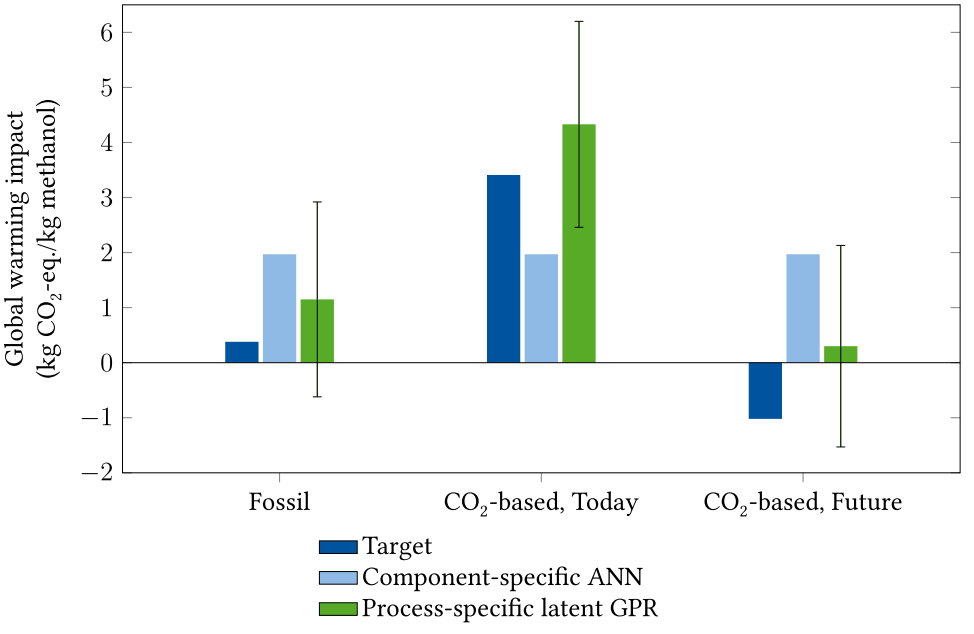


Figure 12. CO_2-based methanol as an example for process specific prediction of the GWI using a state-of-the-art component-specific artificial neural network (ANN, light blue) and a process-specific latent Gaussian Process Regression (GPR, green). The error bars indicate the prediction's standard deviation obtained from the GPR.

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

Furthermore, the energy demand accounts for 4.82 MJ electricity. Overall, the production of 1 kg CO_2-based methanol results in 3.41 kg CO_2-equiv in the today scenario and decreases substantially to −1.02 kg CO_2-equiv in the future scenario (Figure 12). The negative value is due to the scope from cradle-to-gate and the credit for avoided emissions.

Since the molecular descriptors could not represent the changing background system, the predicted impacts of the component-specific ANN predict an impact of 1.97 kg CO_2-equiv, which equals the average impact for methanol in the training set. In contrast, the process-specific latent GPR uses the stoichiometric sum of the reactants' impacts as well as the estimated energy demand of the process as a feature and thus obtains information about the scenario under consideration, e.g., the changing GWI of the hydrogen and CO_2 supply. As a result, the predicted impacts reflect the trends in the GWI correctly: Changing the today scenario from a fossil-based method to a CO_2-based production of methanol is predicted to increase the impact by 3.2 kg CO_2-equiv (from 1.2 to 4.3 kg CO_2-equiv), which is in close agreement with the 3 kg CO_2-equiv difference between the target impacts. Furthermore, the substantial decrease in the GWI for CO_2-based methanol when changing the background scenario is predicted correctly as a reduction potential of approximately 4 kg CO_2-equiv (from 4.33 to 0.3 kg CO_2-equiv).

The presented predictive LCA framework further offers the advantage that the prediction's uncertainty can be quantified. The prediction's standard deviation ranges from 1.77 to 1.87 kg CO_2-equiv for the fossil-based and CO_2-based production methods of methanol, respectively. Considering the error bars reveals that the predictive LCA framework meets the target impacts within one standard deviation. Furthermore, the trend between the CO_2-based production processes can be reflected correctly for varying background scenarios despite the high uncertainty. As a result, the presented predictive LCA framework can serve as decision-supporting tool at TRL 2.

DISCUSSION AND OUTLOOK

A fully automatized predictive LCA framework for the cradle-to-grave environmental impacts is presented that is based on newly developed features, i.e., latent representation of the main product and process features. The GWI is studied as an exemplary impact category. However, the proposed framework can be applied accordingly to any other impact category.

The highest prediction performance is achieved using a combination of latent representation and process features as input for the GPR-based framework. However, the advantage of the GPR over the ANN is canceled out when the prediction performance is averaged over 30 randomly generated test sets. Instead, the choice of features affects the prediction performance substantially. In conclusion, the model choice is less important compared to the choice of suitable features.

A feature importance analysis identifies the stoichiometric sum of the reactants' impact as the most important process feature, which can be explained with the high correlation between this feature and the feedstock-related contributions to the overall impact. The predictive LCA model then adds a component-specific offset based on molecular features. This offset is most closely correlated with the seventh dimension of the latent representation, encoding the aromaticity of the component.

Other process features such as the estimated energy demand contribute only marginally to the prediction, although the impacts of most considered chemical processes are mainly caused by their energy demand. Therefore, process energy should be an important feature. In this work, the used pinch-based process models provide minimum energy demands and thus largely underestimate the process energies resulting in a low correlation with the target GWI. Subsequent work should improve process models to estimate the energy demands more accurately, e.g., by using more rigorous models for reaction and separation including recycling, side reactions, waste treatment, and potentially heat integration. Furthermore, process modeling also needs to consider solids as well as inorganic chemicals.

Side reactions leading to valuable byproducts and process waste also affect the LCA impacts of the main product. In this work, economic allocation is used to solve multifunctionality. However, the effect of allocation methods on the predicted impact needs further investigation. Subsequent work could even consider allocation as a user input.

Our results further show that the developed process features allow for distinguishing between process alternatives and considering changing background systems. In the comparison of process alternatives, where two each lead to the same product, the trend of the impacts could be correctly identified in 86% of the cases. Similarly, the trend for CO_2-based methanol was predicted correctly assuming varying CO_2 and H_2 sources. Therefore, the presented predictive LCA framework can serve as decision-supporting tool at TRL 2. Notably, with increasing TRL, process design might change significantly from the assumed generic flowsheet. The generic process descriptors employed can also not be expected to catch the effect of highly innovative solutions (e.g., process intensification). A more detailed LCA has to refine the estimated impact and reduce uncertainty as process development progresses. One advantage of the presented framework is that these uncertainties can be quantified. Nevertheless, the uncertainties of the predicted impacts are still high. More and better data is urgently needed to improve the prediction of LCA results.

In conclusion, the presented predictive LCA framework can now be used as a decision-supporting tool in early process development. The developed framework is open for future performance improvements by integrating more expressive process features and increasing the training data set size.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssuschemeng.2c07682.

Processes included in the consistent LCA training data set, details on the encoder−decoder neural networks, architecture of the encoder−decoder neural networks, automated flowsheeting according to the Douglas hierarchy, statistics on the data sets to assess the data balance, performance metrics for regression tasks, parity plots of the randomly generated LCA test set, and selected features for the process-specific GPR-based predictive LCA frameworks (PDF)

AUTHOR INFORMATION

Corresponding Author

André Bardow − *Energy & Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, 8092 Zurich, Switzerland; Institute of Energy and*

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319

Climate Research - Energy Systems Engineering (IEK-10), Forschungszentrum Jülich GmbH, Jülich 52428, Germany; orcid.org/0000-0002-3831-0691; Email: abardow@ethz.ch

Authors

Johanna Kleinekorte − *Institute for Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany;* orcid.org/0000-0002-1895-1539

Jonas Kleppich − *Institute for Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany*

Lorenz Fleitmann − *Energy & Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, 8092 Zurich, Switzerland;* orcid.org/0000-0002-6350-5116

Verena Beckert − *Institute for Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany*

Luise Blodau − *Institute for Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssuschemeng.2c07682

Author Contributions

J. Kleinekorte: conceptualization, methodology, investigation, data curation, project administration, visualization, writing - original draft. J. Kleppich: investigation, data curation, software, visualization, writing - review and editing. L.F.: investigation, validation, software, writing - review and editing. V.B.: investigation, data curation, software, visualization, writing - review and editing. L.B.: data curation, formal analysis, software, writing - review and editing. A.B.: conceptualization, supervision, funding acquisition, writing - review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the German Federal Ministry of Education and Research (BMBF) for funding within the project consortium Carbon2Chem under contract 03EK3042C. Simulations were performed with computing resources granted by RWTH Aachen University under projects thes0917 and rwth0745.

■ REFERENCES

(1) Subramaniam, B.; Helling, R. K.; Bode, C. J. Quantitative Sustainability Analysis: A Powerful Tool to Develop Resource-Efficient Catalytic Technologies. *ACS Sustainable Chem. Eng.* **2016**, *4*, 5859−5865.

(2) *ISO 14040, Environmental management - Life cycle assessment - Principles and framework*; International Organization for Standardization, 2006.

(3) Chebaeva, N.; Lettner, M.; Wenger, J.; Schöggl, J.-P.; Hesser, F.; Holzer, D.; Stern, T. Dealing with the Eco-Design Paradox in Research and Development Projects: The concept of Sustainability Assessment Levels. *Journal of Cleaner Production* **2021**, *281*, 125232.

(4) Moni, S. M.; Mahmud, R.; High, K.; Carbajales-Dale, M. Life Cycle Assessment of Emerging Technologies: A Review. *Journal of Industrial Ecology* **2020**, *24*, 52−63.

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

(6) Calvo-Serrano, R.; González-Miquel, M.; Papadokonstantakis, 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.

(7) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbühler, K. Molecular-structure-based Models of Chemical Inventories Using Neural Networks. *Environ. Sci. Technol.* **2008**, *42*, 6717−6722.

(8) Wernet, G.; Papadokonstantakis, S.; Hellweg, S.; Hungerbühler, K. Bridging Data Gaps in Environmental Assessments: Modeling Impacts of Fine and Basic Chemical Production. *Green Chem.* **2009**, *11*, 1826−1831.

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

(10) Kleinekorte, J.; Kröger, L.; Leonhard, K.; Bardow, A. A Neural Network-Based Framework to Predict Process-Specific Environmental Impacts. *Comput.-Aided Chem. Eng.* **2019**, *46*, 1447−1452.

(11) Karka, P.; Papadokonstantakis, S.; Kokossis, A. Environmental Impact Assessment of Biomass Process Chains at Early Design Stages Using Decision Trees. *International Journal of Life Cycle Assessment* **2019**, *24*, 1675−1700.

(12) Baxevanidis, P.; Papadokonstantakis, 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*, No. e17544.

(13) Forrester, A. I. J.; Sóbester, A.; Keane, A. J. *Engineering Design via Surrogate Modelling: A Practical Guide*; Wiley: Chichester, 2008.

(14) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316−1322.

(15) Calvo-Serrano, R.; González-Miquel, M.; Papadokonstantakis, S.; Guillén Gosálbez, G. Cradle-to-gate environmental impact prediction from chemical attributes using mixed-integer programming. *Comput.-Aided Chem. Eng.* **2017**, *40*, 1999−2004.

(16) Golbraikh, A.; Tropsha, A. Beware of q^2! *Journal of Molecular Graphics and Modelling* **2002**, *20*, 269−276.

(17) Karka, P.; Papadokonstantakis, 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.

(18) 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*, 350−367.

(19) Kleinekorte, J.; Fleitmann, L.; Bachmann, M.; Kätelhön, A.; Barbosa-Póvoa, A.; von der Assen, N.; Bardow, A. Life Cycle Assessment for the Design of Chemical Processes, Products, and Supply Chains. *Annu. Rev. Chem. Biomol. Eng.* **2020**, *11*, 203−233.

(20) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31−36.

(21) Buchner, G. A.; Stepputat, K. J.; Zimmermann, A. W.; Schomäcker, R. Specifying Technology Readiness Levels for the Chemical Industry. *Ind. Eng. Chem. Res.* **2019**, *58*, 6957−6969.

(22) Hansen, H. K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. Vapor-liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension. *Ind. Eng. Chem. Res.* **1991**, *30*, 2352−2355.

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.;

Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2016.

(24) Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: A Novel View to Physiological Solvation and Partition Questions. *Journal of Computer-Aided Molecular Design* **2001**, *15*, 355−365.

(25) Minaee, S.; Boykov, Y. Y.; Porikli, F.; Plaza, A. J.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3059968.

(26) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chemical science* **2019**, *10*, 1692−1701.

(27) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, 2016; p 53.

(28) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS central science* **2018**, *4*, 268−276.

(29) Schweidtmann, A. M.; Rittig, J. G.; König, A.; Grohe, M.; Mitsos, A.; Dahmen, M. Graph Neural Networks for Prediction of Fuel Ignition Quality. *Energy Fuels* **2020**, *34*, 11395−11407.

(30) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proc. 34th Int. Conf. Mach. Learning* **2017**, 1263−1272.

(31) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 595−608.

(32) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *arxiv.org (Machine Learning)*, 2019, arXiv:1910.10685, DOI: 10.48550/arXiv.1910.10685.

(33) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific data* **2014**, *1*, 140022.

(34) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(35) Sterling, T.; Irwin, J. J. ZINC 15−Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(36) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design - A Review of the State of the Art. *Molecular Systems Design & Engineering* **2019**, *4*, 828−849.

(37) Marrero, J.; Gani, R. Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilib.* **2001**, *183*−*184*, 183−208.

(38) Patel, A. D.; Meesters, K.; den Uil, H.; de Jong, E.; Blok, K.; Patel, M. K. Sustainability Assessment of Novel Chemical Processes at Early Stage: Application to Biobased Processes. *Energy Environ. Sci.* **2012**, *5*, 8430−8444.

(39) Jung, J.; von der Assen, N.; Bardow, A. Sensitivity coefficient-based uncertainty analysis for multi-functionality in LCA. *International Journal of Life Cycle Assessment* **2014**, *19*, 661−676.

(40) Heijungs, R.; Guinée, J. B. Allocation and 'what-if scenarios in life cycle assessment of waste management systems. *Waste management* **2007**, *27*, 997−1005.

(41) Heijungs, R.; Frischknecht, R. A special view on the nature of the allocation problem. *international journal of life cycle assessment* **1998**, *3*, 321−332.

(42) Douglas, J. M. *Conceptual Design of Chemical Processes*; McGraw-Hill: New York, 1988; pp 99−315.

(43) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS central science* **2018**, *4*, 1465−1476.

(44) Scheffczyk, J. D. Integrated Computer-Aided Design of Molecules and Processes Using COSMO-RS. Dissertation, RWTH Aachen University, 2018. DOI: 10.18154/RWTH-2018-225002.

(45) Bausa, J.; Watzdorf, R. v.; Marquardt, W. Shortcut Methods for Nonideal Multicomponent Distillation: I. Simple Columns. *AIChE J.* **1998**, *44*, 2181−2198.

(46) Redepenning, C.; Recker, S.; Marquardt, W. Pinch-based Shortcut Method for the Conceptual Design of Isothermal Extraction Columns. *AIChE J.* **2017**, *63*, 1236−1245.

(47) Perry, R. H.; Green, D. W.; Maloney, J. O. *Perry's Chemical Engineers' Handbook*; McGraw-Hill: New York, 1984; pp 4.34−4.36.

(48) Patel, A. D.; Meesters, K.; den Uil, H.; de Jong, E.; Worrell, E.; Patel, M. K. Early-stage comparative sustainability assessment of new bio-based processes. *ChemSusChem* **2013**, *6*, 1724−1736.

(49) Artz, J.; Müller, T. E.; Thenert, K.; Kleinekorte, J.; Meys, R.; Sternberg, A.; Bardow, A.; Leitner, W. Sustainable Conversion of Carbon Dioxide: An Integrated Review of Catalysis and Life Cycle Assessment. *Chem. Rev.* **2018**, *118*, 434−504.

(50) Sheldon, R. A. Metrics of Green Chemistry and Sustainability: Past, Present, and Future. *ACS Sustainable Chem. Eng.* **2018**, *6*, 32−48.

(51) Huijbregts, M.; Steinmann, Z.; Elshout, P.; Stam, G.; Verones, F. *ReCiPe 2016: A Harmonized Life Cycle Impact Assessment Method at Midpoint and Endpoint Level: Report I: Characterization*; RIVM report 2016-0104; National Institute for Public Health and the Environment, 2016.

(52) Corrales, D.; Corrales, J.; Ledezma, A. How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. *Symmetry* **2018**, *10*, 99.

(53) Parvatker, A. G.; Eckelman, M. J. Simulation-Based Estimates of Life Cycle Inventory Gate-to-Gate Process Energy Use for 151 Organic Chemical Syntheses. *ACS Sustainable Chem. Eng.* **2020**, *8*, 8519−8536.

(54) IHS Markit. Process Economics Program (PEP) Yearbook, 2018; https://ihsmarkit.com/index.html (accessed on April 9, 2021).

(55) Lloyd, S. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory* **1982**, *28*, 129−137.

(56) Kodinariya, T. M.; Makwana, P. R. Review on Determining Number of Cluster in K-Means Clustering. *Int. J. Adv. Res. Comput. Sci. Manage. Stud.* **2013**, *6*, 90−95.

(57) Bao, F. *best_kmenas(X)*, 2021. https://www.mathworks.com/matlabcentral/fileexchange/49489-best_kmeans-x (accessed on May 19, 2021).

(58) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Annals of Mathematical Statistics* **1951**, *22*, 79−86.

(59) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learning Res.* **2011**, *12*, 2825−2830.

(60) Ferri, F. J.; Pudil, P.; Hatef, M.; Kittler, J. Comparative Study of Techniques for Large-Scale Feature Selection. *Machine Intelligence and Pattern Recognition* **1994**, *16*, 403−413.

(61) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, **2006**; Vol. 3; pp 13−19.

(62) Duvenaud, D. Automatic Model Construction with Gaussian Processes. Ph.D. thesis, University of Cambridge, 2014.

(63) Abadi, M. Tensorflow: A System for Large-Scale Machine Learning In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, GA, November 2−4, 2016; USENIX, 2016; pp 265−283.

(64) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.* **2019**, 2623−2631.

(65) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(66) *GaBi Life Cycle Assessment Database*; Sphera Solutions GmbH, 2021.
(67) Rihko-Struckmann, L. K.; Peschel, A.; Hanke-Rauschenbach, R.; Sundmacher, K. Assessment of Methanol Synthesis Utilizing Exhaust CO_2 for Chemical Storage of Electrical Energy. *Ind. Eng. Chem. Res.* **2010**, *49*, 11073−11078.

https://doi.org/10.1021/acssuschemeng.2c07682
ACS Sustainable Chem. Eng. 2023, 11, 9303−9319