

# COSMO-susCAMPD: Sustainable solvents from combining computer-aided molecular and process design with predictive life cycle assessment



Lorenz Fleitmann<sup>a,b,1</sup>, Johanna Kleinekorte<sup>b,1</sup>, Kai Leonhard<sup>b</sup>, André Bardow<sup>a,b,c,\*</sup>

<sup>a</sup>Energy and Process Systems Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Tannenstrasse 3, 8092 Zurich, Switzerland

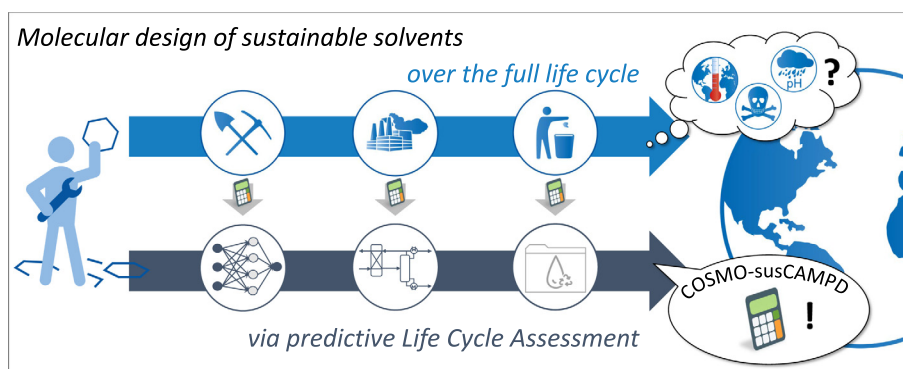
<sup>b</sup>Institute for Technical Thermodynamics, RWTH Aachen University, Schinkelstrasse 8, 52062 Aachen, Germany

<sup>c</sup>Institute for Energy and Climate Research - Energy Systems Engineering (IEK-10), Forschungszentrum Jülich GmbH, Wilhelm-Johnen Strasse, 52425 Jülich, Germany

## HIGHLIGHTS

- Computer-aided design of solvents and processes with an environmental objective.
- Predictive life cycle assessment for each solvent candidate from cradle-to-grave.
- Cradle-to-gate impacts predicted by artificial neural network.
- Gate-to-grave impacts from pinch-based process models.
- Consistent molecular descriptors from quantum-chemistry-based COSMO-RS.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

### Article history:

Received 30 October 2020

Received in revised form 23 March 2021

Accepted 2 June 2021

Available online 17 June 2021

### Keywords:

Predictive thermodynamics  
Artificial neural network  
Pinch-based process models  
Extraction-distillation  
 $\gamma$ -valerolactone

## ABSTRACT

Sustainable solvents are crucial for chemical processes and can be tailored to applications by Computer-Aided Molecular and Process Design (CAMPD). Recent CAMPD methods consider not only economics but also environmental hazards and impacts. However, holistic environmental assessment needs to address the complete life cycle of solvents. Here, we propose a CAMPD framework integrating Life Cycle Assessment (LCA) of solvents from cradle-to-grave: COSMO-susCAMPD. The framework builds on the COSMO-CAMPD method for predictive design of solvents using COSMO-RS and pinch-based process models. Cradle-to-grave LCA is enabled by combining predictive LCA from cradle-to-gate using an artificial neural network with gate-to-grave life cycle inventory data from the process models. The framework is applied to design solvents in a hybrid extraction-distillation process. The results highlight the need for cradle-to-grave LCA as objective function: Heuristics, economics, or cradle-to-gate LCA lead to suboptimal solvent choices. COSMO-susCAMPD thus enables the holistic environmental design of solvents using cradle-to-grave LCA.

© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Growing environmental concerns have recently increased the awareness of sustainability in the design of chemical processes (Bakshi, 2019). The term sustainability is defined by three dimensions: an economic, a social and an environmental dimension

\* Corresponding author.

E-mail address: [abardow@ethz.ch](mailto:abardow@ethz.ch) (A. Bardow).

<sup>1</sup> L.F. and J.K. contributed equally to this work.

(Brown et al., 1987). Both economics and environmental impacts of many chemical processes depend strongly on the employed solvents (Clarke et al., 2018; Jimenez-Gonzalez, 2019; Zhou et al., 2020). Therefore, sustainable processes with economic success and minimum environmental impact require the selection of optimal solvents.

Today, optimal solvents can be tailored to chemical processes by Computer-Aided Molecular and Process Design (CAMPD). CAMPD methods optimise the molecular structure of candidate solvents and the process simultaneously for optimum process performance (Papadopoulos et al., 2018). Thereby, CAMPD methods allow for the targeted exploration of vast molecular design spaces at low cost. All CAMPD methods employ predictive thermodynamic models to link the molecular structures to the process performance. The estimation of the thermodynamic properties from the molecular structures bridges the scales between molecules and processes to predict process performance.

So far, CAMPD methods in literature have mainly focused on economics and technical process performance (Zhou et al., 2020; Gertig et al., 2020). However, the design of environmentally sound chemical processes requires a broader objective in CAMPD: Not only process performance and economics but also environmental impacts need to be optimised. To capture environmental impacts, CAMPD needs to integrate environmental assessment (Zhou et al., 2020; Gertig et al., 2020).

Several CAMPD methods already integrated environmental assessment, e.g. the assessment of environmental impact potentials or hazards. Many of these approaches are based on metrics and guidelines for green solvents (Soh and Eckelman, 2016). In particular, the systematic assessment of indicators for Environmental, Health and Safety hazards (EHS) (Adu et al., 2008) has been applied successfully in CAMPD problems. For example, systematic screening approaches evaluate candidates based on environmental databases and Quantitative Structure-Activity Relationship (QSAR) toolboxes (McBride et al., 2018; Linke et al., 2020; Song et al., 2020). If CAMPD problems are formulated as an integrated mathematical optimisation problem, solution algorithms require an automated, integrated evaluation of EHS criteria. For this purpose, predictive models are frequently employed, e.g. group-contribution models fitted to experimental data (Papadopoulos et al., 2010; Schilling et al., 2017; Ten et al., 2017; Ooi et al., 2018; Jonuzaj et al., 2019; Ten et al., 2020; Ten et al., 2021). All approaches have in common that they evaluate environmental impact potentials from molecular properties of the candidate molecules.

However, environmental assessment has to go beyond the environmental impact potential of the molecules, which is a molecular property, such as the global warming potential (Hellweg et al., 2004). For a holistic assessment, CAMPD needs to consider the environmental impacts of the full life cycle of a molecule, including emissions caused during production, use and disposal (Jimenez-Gonzalez, 2019; Chemmangattuvalappil, 2020). A broadly accepted method for the holistic environmental assessment is Life Cycle Assessment (LCA). LCA is an ISO-normed method (ISO 14040, 2006) considering emissions of all life cycle stages from cradle to grave of a substance. For this purpose, detailed mass and energy balances summarise all energy and mass flows from and to the environment.

As a consequence of the holistic analysis, LCA helps to avoid problem shifting between life cycle stages or environmental impacts. However, a cradle-to-grave LCA generally requires much information on a substance: mass and energy flows of production, use and disposal (Hellweg and Milà i Canals, 2014).

In CAMPD, available data on candidate solvents is usually minimal, in particular on *in silico* designed solvents. For economic objectives, CAMPD methods have already been equipped with pre-

dictive tools to close data gaps: Predictive thermodynamic models estimate thermodynamic properties so that process simulation can be performed for economic assessment. Analogously, CAMPD needs to integrate predictive LCA approaches for environmental assessment. Similarly to the prediction of thermodynamic properties from thermodynamic models, environmental impacts of candidate solvents need to be predicted given their molecular structure (Kleinekorte et al., 2020).

In literature, predictive LCA has been approached by two main routes: (1) The prediction of Life Cycle Inventory (LCI) and (2) the direct prediction of the Life Cycle Impact Assessment (LCIA). The life cycle inventory is the basis for life cycle impact assessment and provides the “bill of materials” of the life cycle. To yield ultimately environmental impacts, the LCI needs to be multiplied by characterisation factors. LCI is frequently predicted from estimates for energy and mass flow from generic flowsheets (Righi et al., 2018; Parvatker and Eckelman, 2020). In contrast, the direct prediction of the LCIA has been investigated by multi-linear regression (Calvo-Serrano et al., 2018; Calvo-Serrano et al., 2019) and Artificial Neural Networks (ANN) (Song et al., 2017).

Recently, predictive LCA has successfully been combined with molecular design for the first time to the best of our knowledge: Papadopoulos et al. formulated an integrated CAMD problem including predictive LCA and predictive EHS scores (Papadopoulos et al., 2020). For the predictive LCA, the authors use the ANN-based FineChem model (Wernet et al., 2009) to estimate the specific impacts of solvent production per kilogram solvent. For the predictive EHS scores, (Papadopoulos et al., 2020) employ group contribution and molecular similarity approaches. By combining these prediction approaches into one integrated multi-objective CAMD problem, desired solvent properties are optimised simultaneously with environmental impact scores, e.g. maximising specific solvent density and minimising specific global warming impact and EHS scores.

The current approach limits the LCA scope to a so-called cradle-to-gate system boundary, considering only emissions caused during the solvent production per kilogram solvent. However, the amount of solvent required by the process varies greatly depending on the solvent performance in the process. Moreover, the process corresponds to the use phase of the solvent life cycle, and the solvent properties directly impact the process performance and the emissions of the use phase. Finally, the emissions from solvent disposal depend on the solvent loss during the use phase. Thus, a cradle-to-gate assessment does not capture the full environmental impacts of the candidate solvents. To avoid problem shifting between life cycle stages, CAMPD needs to consider all solvent-related emissions within a cradle-to-grave system boundary.

In this work, we propose a CAMPD framework with predictive LCA that accounts for the full solvent life cycle, including use and disposal. The CAMPD framework builds on the COSMO-CAMPD method (Scheffczyk et al., 2018) that is extended by environmental assessment using predictive LCA. As a result, we present COSMO-CAMPD with an environmental objective: COSMO-susCAMPD.

COSMO-susCAMPD overcomes the limitations of previous CAMPD approaches by exploiting process data from the process model in COSMO-CAMPD as LCI for solvent use and disposal. To predict the specific cradle-to-gate impacts, molecular descriptors from predictive thermodynamics serve as an input for an ANN. By combining the LCI from process models and cradle-to-gate impacts of solvent production, we achieve a full LCA for every candidate solvent covering the system boundary from cradle-to-grave. As a result, COSMO-susCAMPD provides a framework for integrated computer-aided design of solvents and processes for both maximal process performance and minimal environmental impact with cradle-to-grave system boundary.

This paper is structured as follows: In Section 2, the COSMO-susCAMPD framework is described. We explain not only the details of the framework, but also how we set up the ANN for predictive LCA and discuss its accuracy. Section 3 introduces a case study and describes the application of COSMO-susCAMPD. We present results of the optimisation and discuss consequences from the integrated design. Finally, conclusions are drawn in Section 4.

## 2. COSMO-susCAMPD: A framework for the design of sustainable solvents and processes

The design of optimal solvents for maximal process performance and minimal environmental impacts requires the combination of two methods: (1) A method for integrated molecular and process design (CAMPD) and (2) a method for predictive cradle-to-grave Life Cycle Assessment (LCA).

1. Solution algorithms for CAMPD problems usually combine three steps (Papadopoulos et al., 2018):
  - (a) First, candidate molecules are generated by an algorithm, e.g. by a combination of functional groups or molecular fragments. The algorithm needs to be able to change the molecular structures systematically to explore a given design space, while structural feasibility is ensured for all candidates.
  - (b) Secondly, for each candidate molecules, thermodynamic properties are predicted using predictive thermodynamic models. Thermodynamic properties are required to bridge the scales between the candidate molecules and the process, e.g. by prediction of activity coefficients or vapour pressures.
  - (c) Finally, the candidate molecules are evaluated by an objective. The objective function quantifies the fit of the candidate molecules to the process application, e.g. by a particular thermodynamic property, a process variable or an economic metric.
2. A method for predictive LCA of candidate molecules requires the assessment of all stages of a molecule's life cycle: the production, the use phase and the disposal. In literature, predictive methods for particular life cycle stages have been proposed:
  - (a) The environmental impacts from the production can be estimated by molecular structure models that use molecular descriptors to predict the cradle-to-gate LCIA, e.g. using multi-linear regression (Calvo-Serrano et al., 2018) or ANN (Song et al., 2017).
  - (b) The use phase of a molecule can be modelled by generalised flowsheets estimating the gate-to-gate energy demand of processes (Jiménez-González et al., 2000; Parvatkar and Eckelman, 2020). Afterwards, this LCI is translated into LCIA by multiplying the energy demand with the corresponding characterisation factors.
  - (c) Predictive LCA approaches dealing with the disposal of molecules have not been published so far. However, proxies from LCI databases for generic wastewater treatment or waste incineration can be used (Canals et al., 2011).

Methods for predictive LCA of each life cycle stage are combined with a CAMPD method in the proposed COSMO-susCAMPD framework (Fig. 1) to yield a fully predictive framework with cradle-to-grave environmental assessment. COSMO-susCAMPD is introduced in the following Section 2.1.

### 2.1. Implementation of the COSMO-susCAMPD framework

The COSMO-susCAMPD framework expands COSMO-CAMPD by a predictive LCA method as follows:

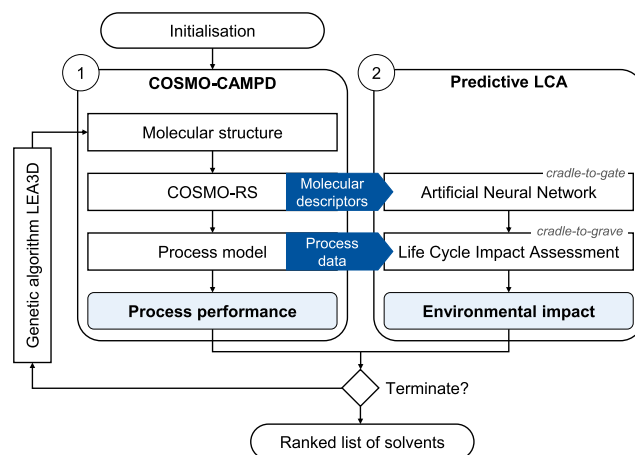
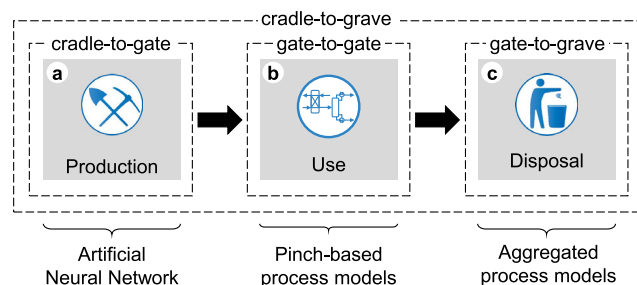


Fig. 1. COSMO-susCAMPD: Fully automated framework to design environmentally beneficial solvents by combining COSMO-CAMPD with predictive LCA.

1. The basis for COSMO-susCAMPD is the COSMO-CAMPD method initially developed to design solvents for optimum process performance and economic objectives (Scheffczyk et al., 2018). The CAMPD algorithm involves three steps in each iteration of the optimisation procedure:
  - (a) Generation of a molecular structure: The generation of candidate solvents is part of the molecular optimisation using the genetic algorithm LEA3D (Douguet et al., 2005). LEA3D builds molecules from 3D-molecular fragments. The fragments are specified in the initialisation of the algorithm via a fragment database. The fragment database is created by the users to reflect their preferences. The algorithm starts by randomly combining fragments for the first generation of molecules. After these molecules have been evaluated by the constraints and the objective function (Steps 1b and 1c), LEA3D alters the population of molecules for each following generation using genetic operations on the candidate molecules, i.e. crossover and mutation. Thereby, LEA3D explores the vast molecular design space towards an objective function. After a predefined number of generations is reached, the molecular optimisation stops. Already during the generation of the molecular structures, LEA3D ensures chemical feasibility of the molecules, e.g., all candidate molecules fulfil the octet rule. Moreover, the 3D-structure of the candidate molecules allows evaluating constraints on the molecular size or functional groups. If such constraints exist, undesirable candidate molecules can already be discarded before the time-consuming computational steps.
  - (b) Prediction of thermodynamic properties: For each candidate solvent of each generation, thermodynamic properties are obtained using the predictive thermodynamic model COSMO-RS (Klamt et al., 2010). COSMO-RS uses surface charge interactions from quantum chemical Density Functional Theory DFT (Kohn and Sham, 1965). The computationally expensive DFT-calculations are performed in parallel for each generation based on the 3D-molecular structure of the candidate solvents and provide the surface charge densities of the molecules serving as input to the COSMO-RS calculations. By applying statistical thermodynamics to the interactions between the surface charges, COSMO-RS is then able to predict many thermodynamic properties of pure components and mixtures, such as activity coefficients, Liquid-Liquid-Equilibria (LLE) or vapour pressures with low computational effort.

The thermodynamic properties are used for the evaluation of constraints, e.g., the existence of LLE or an appropriate boiling point. These constraints on thermodynamic properties can reduce the search space and help to identify feasible solvents. In addition, in COSMO-susCAMPD, the thermodynamic properties serve as an input for the ANN to predict cradle-to-gate impacts.

- (c) **Process model evaluation:** For each candidate solvent of each generation that fulfils property constraints, a process flowsheet is evaluated. The process is modelled using pinch-based process models for each unit operation (Bausa et al., 1998; Redepenning et al., 2017). Pinch-based process models are reduced-order models that provide an accurate and efficient calculation of process units assuming minimum thermodynamic driving force. By this assumption, computationally demanding tray-by-tray calculations can be omitted, but no simplifications on thermodynamic modelling are required. In literature, it has been shown that the pinch-based process models agree well with results from rigorous tray-by-tray models for operation near the thermodynamic minimum (Scheffczyk et al., 2018; Redepenning et al., 2017). As a result, the pinch-based process models yield a maximum achievable process performance for each solvent considering full equilibrium thermodynamics. Due to the computational efficiency of the process evaluation, we can optimise the process flowsheet for each solvent. In COSMO-susCAMPD, the process model not only evaluates process performance but also provides process data as LCI of the use phase for gate-to-gate LCIA.
2. To enable an environmental objective in COSMO-susCAMPD, we add a predictive LCA for every candidate solvent to COSMO-CAMPD. For this purpose, we divide the life cycle of the candidate solvents into three stages (Fig. 2): (a) solvent production (cradle-to-gate), (b) solvent use in the process (gate-to-gate) and (c) solvent disposal (gate-to-grave).
- (a) **Solvent production:** We estimate environmental impacts from solvent production (cradle-to-gate system boundary) using an ANN. As shown by Wernet et al. (2008), ANNs outperform other regression methods such as multi-linear regression in LCA applications. The ANN uses molecular and thermodynamic solvent properties as input as already proposed in the literature (Song et al., 2017; Calvo-Serrano et al., 2018; Papadopoulos et al., 2020). In particular, thermodynamic properties of the candidate solvents from COSMO-RS are included. Properties calculated from COSMO-RS have already been proven to be suitable molecular descriptors by Calvo-Serrano et al. (2019). In COSMO-susCAMPD, molecular descriptors from COSMO-RS provide the additional advantage that a consistent set of descriptors is used for both the LCA and the techno-economic assessment. More details on the training and set-up of ANN are given in Section 2.2.
- (b) **Solvent use:** Impacts related to the solvent use in the process (gate-to-gate system boundary) are calculated from the life cycle inventories provided by the process model. Process evaluation solves the mass and energy balances providing all required LCI information for LCIA. In particular, the minimum amount of solvent used in the process is determined accurately by the pinch-based process models. Knowledge of the amount of solvent used allows for a comparison of candidate solvents in terms of process-specific objectives rather than a specific comparison per kilogram of solvent. The LCIA is performed by multiplying the LCI with specific environmental impacts from LCA databases or the ANN prediction. For example, the process heat



**Fig. 2.** Life cycle stages of a solvent and frequently used system boundaries in environmental assessment. In COSMO-susCAMPD, cradle-to-grave system boundaries are enabled by combining an artificial neural network with pinch-based and aggregated process models.

demand is converted into emissions using the specific impact for natural gas combustion per megajoule heat obtained from the GaBi database (Thinkstep, 2012). Additional emissions, such as fugitive emissions, are not considered.

- (c) **Solvent disposal:** For the disposal of solvents (gate-to-grave system boundary), aggregated process models are known in the literature. Here, we model solvent disposal by LCIA for wastewater treatment based on the mass of wastewater including solvent contamination. The literature model yields a specific impact per kilogram of wastewater (Ruiz, 2019). The gate-to-grave LCIA is completed by multiplying the specific impact with the flow rate of wastewater. Both the flow rate and the contamination of wastewater with the solvent result from the process model evaluation.

By combining COSMO-CAMPD and predictive LCA as described, COSMO-susCAMPD yields a fully automated and predictive framework for solvent design. As an objective for the design, process performance, environmental impacts from cradle-to-grave as well as combined objective functions are possible. Alternatively, the predictive LCA can serve as a constraint.

## 2.2. Training and accuracy of the artificial neural network

The Artificial Neural Network (ANN) is used as a regression model, which is trained on known environmental impacts of solvents from databases or literature. After training, the ANN is capable of predicting environmental impacts for candidate solvents similar to the solvents from the training data. Here, we use consistent cradle-to-gate LCA data from the GaBi Database (Thinkstep, 2012) on 73 solvents for training purpose. While the data set is small, it is important to use consistent, high-quality data and to avoid data based on generic heuristics. Thus, the present data set is the largest high-quality data set available to the authors. To facilitate the set-up of the ANN, we use an automated framework in four steps (Fig. 3) as already outlined by Kleinekorte et al. (2019).

1. First, suitable features for the ANN are selected from various molecular descriptors using linear stepwise regression as a feature selection method (Draper and Smith, 1998; Lindsey and Sheather, 2010). For all molecules in the training data, various molecular descriptors are calculated as prospective features, e.g. information on the molecular structure, such as the number of carbon or oxygen atoms, or thermodynamic properties from COSMO-RS, such as the normal boiling point or the standard enthalpy of formation. The molecular descriptors which show the highest correlation with the environmental impacts are selected as features (see Supporting Information for details).



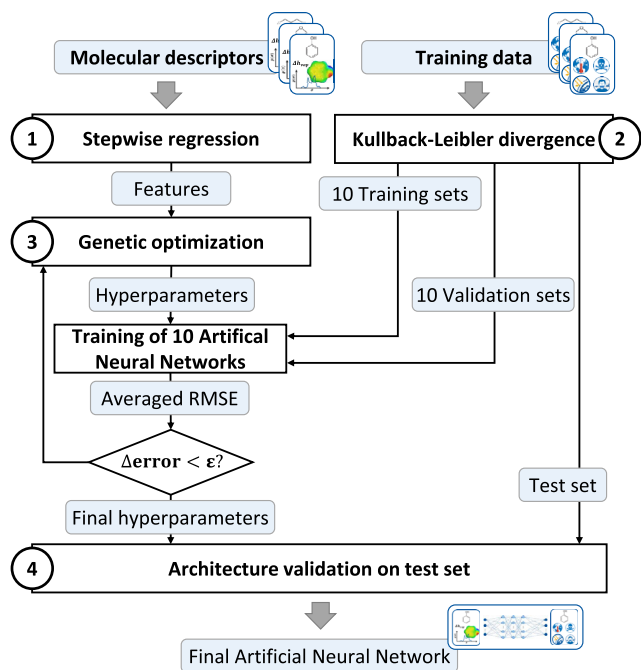


Fig. 3. Flow diagram for the automated set-up of the artificial neural network.

- Secondly, the training data is split into three sets to allow for training, validation and testing of the ANN (Goodfellow et al., 2016). At first, a test set is separated from the training data for final accuracy evaluation of the ANN. The test set contains approximately 10% of the training data and is not used within the training or validation of the ANN to obtain a final accuracy value on unseen data. Extreme points at the edge of the data set are not selected for the test set due to the low extrapolation capability of the ANN beyond the training set. Afterwards, the remaining training data is split ten times into a training and validation set to increase the generalizability of the final architecture. Validation sets are specified to include approximately 10% of the training data as well.

All sets are chosen so that the statistical distribution of the test, training and validation sets are similar (Goodfellow et al., 2016). Therefore, we first randomly generate various test, training and validation sets. For each random set, we calculate the Kullback-Leibler divergence based on the features as a measure of statistical distribution for data sets (Kullback and Leibler, 1951). A low Kullback-Leibler divergence indicates similar and uniform statistical distribution between the data sets, which is a requirement for the training and application of ANN. Therefore, the test set with the lowest Kullback-Leibler divergence is chosen for final accuracy evaluation. For training and validation sets, the ten splits with the lowest Kullback-Leibler divergence are chosen for the training of the ANN.

- Thirdly, the hyperparameters of the ANN, e.g. the number of layers or the number of neurons per layer, are selected. Setting the hyperparameters is not trivial and has a considerable influence on the accuracy of the ANN. Therefore, we use a Genetic Algorithm (GA) (The MathWorks, 2018) to find optimal hyperparameters. The objective of the GA is the minimisation of the average Root Mean Squared Error (RMSE) of the ANN predictions on the validation sets:

$$\min \sum_{i=1}^{10} \frac{\text{RMSE}_i^{\text{val}}}{n} \quad (1)$$

For each instance of the GA, 10 ANNs are trained with the same hyperparameters using the 10 training sets. Afterwards, each ANN is used to predict the corresponding validation set, and the RMSE of the prediction is calculated. By averaging the RMSE over the 10 sets, we flatten extreme prediction errors due to the small set sizes and enable bootstrapping and accuracy evaluation (Carney et al., 1999).

To avoid local optima due to the statistical optimisation, we perform 100 runs of the GA from random starting points by varying the initial hyperparameters.

- Finally, we train the ANN with the optimised architecture on the combined training and validation set to perform an accuracy evaluation by predicting the test set. The test set has neither been used for training of the ANN nor the optimisation of the hyperparameters. Therefore, the ANN predicts the unseen test set with similar accuracy as the molecules designed within the COSMO-susCAMPD framework.

After applying the described set-up, we obtain one trained ANN with optimal hyperparameters and an estimation of its accuracy for one impact category. The ANN can directly be integrated for the prediction of cradle-to-gate impacts for candidate solvents.

In the following, we investigate the accuracy of the ANN predictions before we move to the application of COSMO-susCAMPD. Using the described framework, we set up one ANN for each of the 17 midpoint impact categories from the ReCiPe method (Goedkoop et al., 2008). In the main text, we focus only on the two LCA impact categories, for which the most reliable LCIA methods are available: Climate Change (CC) and Ozone Depletion (OD) (European Commission-Joint Research Centre, 2011). Details on all 17 impact categories can be found in the Supporting Information.

We measure the accuracy of predictions with the coefficient of determination ( $R^2$ ) and the normalised RMSE (nRMSE):

$$R^2 = \frac{\left[ \sum (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) \right]^2}{\sum (\hat{y}_i - \bar{\hat{y}})^2 \sum (y_i - \bar{y})^2} \quad (2)$$

$$\text{nRMSE} = \frac{\sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n}}}{y_{\max} - y_{\min}} \quad (3)$$

The coefficient of determination  $R^2$  indicates the trend-capturing correlation between the ANN predictions and the database values (Alexander et al., 2015). The nRMSE indicates how widely the predictions deviate on average from the database values (Otto et al., 2018). We report the normalised RMSE, which is normalised by the range of the database values so that all impact categories are comparable.

Currently, the availability of LCA data on solvents is limited for the training of an ANN. Our training data contains only 73 solvents, which is a comparably small number for machine learning approaches (Alwosheel et al., 2018). Therefore, the current accuracy of the ANN predictions is limited as well (c.f. Table 1). On average, the ANN achieves an already acceptable nRMSE of 10%, but the average  $R^2$  is low with a value of only 0.43. The low  $R^2$  can be explained from the small training data set: If very few data points are used, the  $R^2$  value is highly sensitive. Due to the small set sizes, inaccurate predictions for a few solvents decrease the  $R^2$  already significantly despite otherwise acceptable predictions. Therefore, it is important to focus not only on the  $R^2$  but also consider the (n)RMSE. For Ozone Depletion, for example, the nRMSE of the validation and of the test set match very well, indicating acceptable predictions despite large differences in the  $R^2$ . In particular, the

**Table 1**

Prediction accuracy of the artificial neural network for the impact category Climate Change and Ozone Depletion, as well as an average of all 17 regarded impact categories in terms of coefficient of determination ( $R^2$ ) and normalised root mean squared error (nRMSE).

Data set	Climate Change		Ozone Depletion		Average of all impact categories	
	$R^2$	nRMSE	$R^2$	nRMSE	$R^2$	nRMSE
Training set	0.44	17%	0.81	8%	0.57	12%
Validation set	0.56	14%	0.76	16%	0.56	14%
Test set	0.51	9%	0.08	15%	0.43	10%

predictions deviate significantly from the database values for areas of sparse training data. In these areas, a high variance between the 10 ANN predictions can be observed as well, indicating high sensitivity on the training set due to limited data. For example, for the impact on Climate Change (CC), the predictions vary between 5–8 kg CO<sub>2</sub>-eq. kg<sup>-1</sup> chem. for a few solvents (see Fig. 4A). Therefore, some solvents with extreme impacts on CC at the edges of the training data are currently predicted inaccurately and need future improvement. Similarly, the ANN yields a few physically not meaningful results, i.e. negative values for some impact categories, which are removed when applying the ANN in COSMO-susCAMPD.

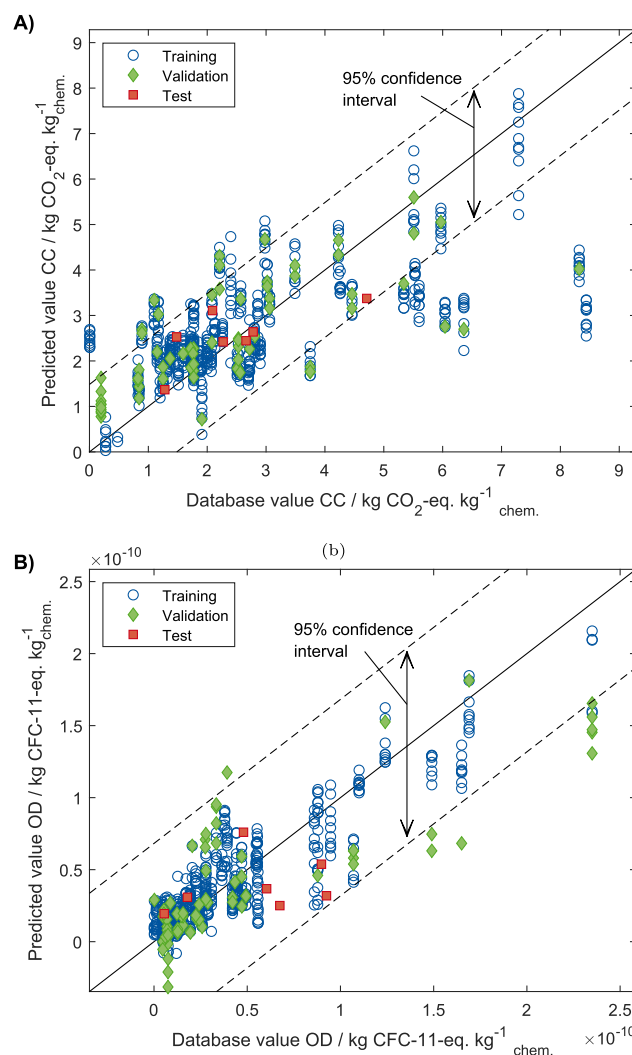
However, an already acceptable accuracy of prediction is achieved for the majority of solvents and, in particular, in ranges with sufficient data (Fig. 4). Generally, the predictions meet the database values with acceptable confidence except for a few strong outliers in sparse regions. The accuracy is comparable to the state-of-the-art in literature: E.g. the estimation of CC had a coefficient of determination  $R^2$  of 0.41 in work by Wernet et al. (2009), or a coefficient of determination  $R^2$  of 0.48 in work by Song et al. (2017). Future improvement of accuracy is expected with more data available. For the design of solvents, ultimately, the uncertainty of the final cradle-to-grave environmental impact is most relevant. Therefore, we investigate how the uncertainties caused by the ANN predictions propagate to the cradle-to-grave impact in Section 3.2.

### 3. Case study and results: Design of benign solvents for hybrid extraction-distillation of $\gamma$ -valerolactone

To demonstrate the application of COSMO-susCAMPD, we investigate the hybrid extraction-distillation of  $\gamma$ -valerolactone (GVL), as proposed by Murat Sen et al. (2012). Recently, GVL has attracted attention as a bio-derived platform chemical, a green solvent or a renewable fuel (Zhang, 2016). A promising pathway to GVL is the production from lignocellulosic biomass and purification from aqueous solution using hybrid extraction-distillation. As an extraction solvent, n-butyl acetate has been suggested in the literature (Murat Sen et al., 2012). Therefore, n-butyl acetate serves as a benchmark for the solvent design with COSMO-susCAMPD.

#### 3.1. Problem specification

We consider the process of GVL purification consisting of an extraction column, a distillation column and a decanter (Fig. 5). A mixture of GVL and water is fed to the extraction column, where the solvent extracts the GVL entirely into the extract stream. The resulting extract is split in the distillation column into pure GVL at the bottom and a water-solvent stream at the top of the distillation column. The water-solvent stream is recycled to the extraction column. If the water-solvent stream splits into two liquid phases, the aqueous phase is separated from the organic phase in a decanter, and only the organic phase is fed back into the extraction col-



**Fig. 4.** Accuracy of the prediction of the ANN for the LCA impact categories Climate Change (CC) and Ozone Depletion (OD). The confidence interval is calculated from the standard deviation of the predictions on the test set.

umn. Both the raffinate and the aqueous phase from the decanter, if present, are sent to wastewater treatment.

Candidate solvents are considered for property prediction and process evaluation if they are expected to be stable within the extraction process based on their functional groups and if they are smaller than 13 non-hydrogen atoms. The process specifications further constrain suitable candidate solvents based on their physical properties: Suitable candidate solvents must have a liquid-liquid-equilibrium with water. Furthermore, the candidate solvents must not exceed the boiling point for GVL to allow for separation of GVL in the bottom of the distillation column. For simple

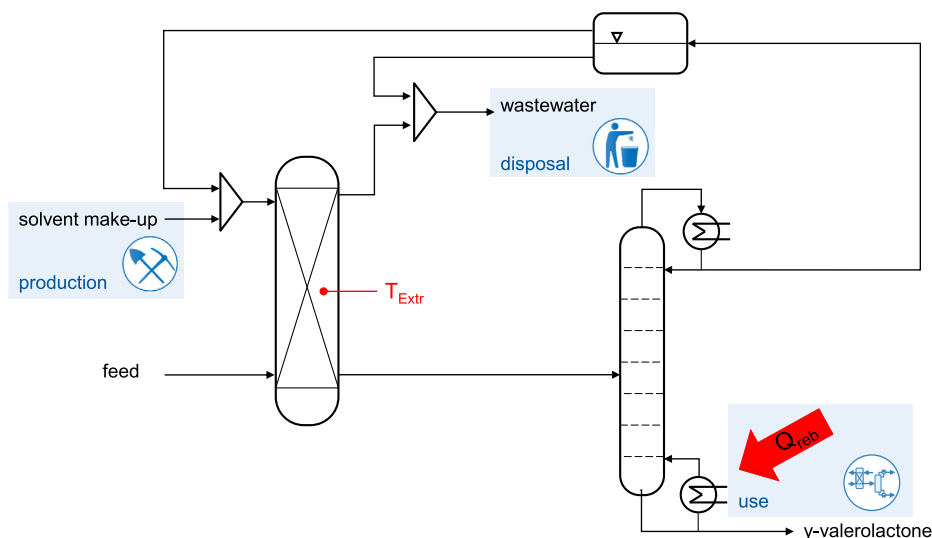


Fig. 5. Flowsheet of the extraction-distillation process for  $\gamma$ -valerolactone purification.

distillation, candidate solvents also must not form an azeotrope with GVL. The constraints on the molecular properties are evaluated for each candidate solvent in each generation of the genetic algorithm with the thermodynamic properties predicted by COSMO-RS (Step 1b of the COSMO-susCAMPD framework). Candidate solvents that do not fulfil these requirements are discarded and not considered as suitable candidate solvent for subsequent process optimisation and environmental assessment.

For each suitable candidate solvent, we first optimise the process settings, i.e., the temperature of the extraction, to obtain the minimum energy demand for distillation (Step 1c). For the environmental assessment (Step 2), we consider three types of emissions for the process: The emissions from solvent production due to solvent make-up, the emissions from solvent use due to energy consumption in the distillation reboiler, and the emissions from solvent disposal in wastewater treatment. For each candidate solvent, we predict the cradle-to-gate impacts of all 17 LCA impact categories using the ANNs (Step 2a) and conduct the full cradle-to-grave LCA exploiting the LCI from the process evaluation for use phase (Step 2b) and solvent disposal (Step 2c). All emissions are calculated for the functional unit of 1 kmol of GVL produced in this process.

In total, four optimisation runs of the genetic algorithm LEA3D are performed to find an optimal solvent for the GVL purification. For molecular design, all functional groups are included that were in the training set of the ANN, e.g. alkane-, benzene-, amine-, sulfone- keto- or hydroxyl-fragments (see Supporting Information). Thus, all molecules that are designed should be predictable by the ANN without forcing the ANN to extrapolate. For all optimisation runs, the objective is to minimise the cradle-to-grave impact on Climate Change ( $CC_{\text{cradle-to-grave}}$ ) by summing the impacts on Climate Change of the three life cycle stages of this process: solvent production ( $CC_{\text{Production}}$ ), solvent use in the process ( $CC_{\text{Process}}$ ) and solvent disposal ( $CC_{\text{Disposal}}$ ):

$$\min CC_{\text{cradle-to-grave}} = CC_{\text{Production}} + CC_{\text{Process}} + CC_{\text{Disposal}} \quad (4)$$

The impact on Climate Change of the process ( $CC_{\text{Process}}$ ) is linearly proportional to the energy demand. The energy demand in the distillation column captures the operating cost of the process. Thus, economically attractive solvents have a low impact from the use phase. Therefore, the optimisation of the cradle-to-grave impact on Climate Change yields solvents with a balanced contri-

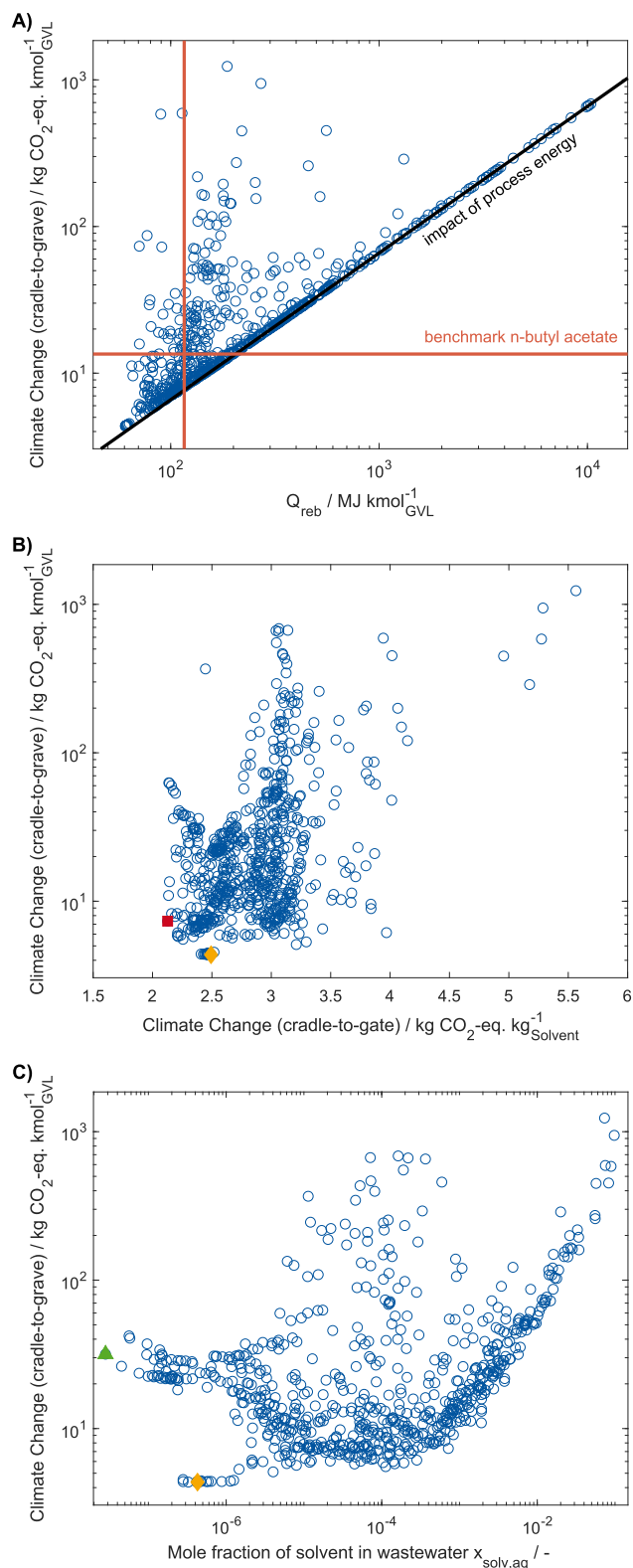
bution from all life cycle phases and low operational cost. If desired, multi-objective optimisation could be employed to explicitly optimise cost and impact on Climate Change.

### 3.2. Results and discussion

In total, the optimisation generates more than 1600 unique solvents, which are evaluated in the 4 design runs in about 5 days (121 h) on an Intel Xeon CPU E5-1660 v3 @ 3.00 GHz using parallel computation on 8 cores. From all candidate solvents, 703 solvents fulfil the property constraints and are suitable for the process. Therefore, we obtain a ranking of 703 solvents according to their cradle-to-grave impact on Climate Change as a result (Fig. 6A).

The solvent with the highest reduction in the impact on Climate Change is 2,3,3,5-tetramethyl-hexane with a cradle-to-grave impact on Climate Change of about 4.4 kg CO<sub>2</sub>-eq. kmol<sup>-1</sup><sub>GVL</sub>. In comparison to the benchmark n-butyl acetate (Murat Sen et al., 2012), 2,3,3,5-tetramethylhexane reduces the impact on Climate Change by about 68%. More generally, 291 of the 703 candidate solvents have a lower impact on Climate Change than the benchmark and 169 solvents outperform the benchmark both in terms of Climate Change and process energy demand  $Q_{\text{reb}}$ . COSMO-susCAMPD thus designs successfully many suitable alternatives. For the top 15 candidates, we find very similar solvents: The top 15 solvents are all alkanes and alkenes, most of them highly branched and therefore not yet commercially available. The highest-ranking commercially available bulk chemical is n-octane on rank 8. N-octane reduces the impact on Climate Change by about 67.5% compared to the benchmark solvent, which is very close to the impact reduction of the optimal solvent.

To challenge the use of the cradle-to-grave impact as an objective function, we compare the cradle-to-grave impact on Climate Change with the gate-to-gate impact from process energy demand during solvent use. The impact on Climate Change from process energy depends linearly on the process energy demand  $Q_{\text{reb}}$  (black line in Fig. 6A) and thus represents the result of an economic optimisation for minimum process energy demand as typically used in CAMPD. Intuitively, one might expect that energy demand in the use phase captures the cradle-to-grave impact on Climate Change already well. However, in this case study, the cradle-to-grave impact of 187 candidate solvents deviates by more than 50% from the impact of process energy (Fig. 6A). The deviation from the impact caused by the process energy is due to the production



**Fig. 6.** (A) Predicted cradle-to-grave impacts on Climate Change (CC) of all solvents designed versus corresponding process energy demand  $Q_{reb}$ . Each blue circle represents one candidate solvent. The black line is the impact resulting from the process energy demand; the red lines stand for the impact on Climate Change and process energy demand of the benchmark solvent n-butyl acetate. (B) Cradle-to-grave impact on Climate Change per kmol GVL versus specific impact from solvent production (cradle-to-grate system boundary) per kilogram solvent. The red square indicates the solvent with the lowest cradle-to-grate impact on Climate Change; the yellow diamond indicates the solvent with the lowest cradle-to-grate impact on Climate Change. (C) Cradle-to-grate impact on Climate Change versus mole fraction of solvent in the wastewater stream. The green triangle indicates the solvent with the lowest solvent loss to wastewater; the yellow diamond indicates the solvent with the lowest cradle-to-grate impact on Climate Change.

and disposal of the candidate solvents. This deviation highlights the importance of the cradle-to-grave system boundary. Still, for this case study, the top 15 solvents with the lowest impact on Climate Change equal the top 15 solvents with the lowest process energy demand  $Q_{reb}$ . For these solvents, the process requires 60.6–62.4 MJ kmol<sup>-1</sup> GVL energy for distillation, corresponding to a reduction of about 46–48% compared to the benchmark n-butyl acetate.

Moreover, the importance of the cradle-to-grave system boundary is shown by comparison to the ranking by cradle-to-grate LCA: A cradle-to-grate LCA based on the specific impacts of solvent production yields a very different ranking (Fig. 6B). The solvent with the lowest cradle-to-grate impact on Climate Change per kilogram solvent is divinyl ether with about 2.1 kg CO<sub>2</sub>-eq. kg<sup>-1</sup> solvent. However, divinyl ether has a cradle-to-grate impact on Climate Change of about 7.3 kg CO<sub>2</sub>-eq. kmol<sup>-1</sup> GVL ranking only 75th in cradle-to-grate impact. 2,3,3,5-tetramethyl-hexane, the solvent with lowest cradle-to-grate impact, ranks only 139th with a higher cradle-to-grate impact on Climate Change of about 2.5 kg CO<sub>2</sub>-eq. kg<sup>-1</sup> solvent. Therefore, concentrating only on the specific cradle-to-grate LCA of the solvent production proves to be a misleading objective. Specific assessment of molecular properties is not sufficient. Instead, the amount of solvent used in the process needs to be considered for solvent selection with an environmental objective. In particular, the specific cradle-to-grate impacts is quite similar for all solvents (x-axis of Fig. 6B) in this case study. In contrast, the cradle-to-grate impact spans multiple orders of magnitude (y-axis of Fig. 6B) yielding a more selective objective.

The differences in the ranking between cradle-to-grate and cradle-to-grave LCA can be explained by the neglect of the solvent use phase: For solvents with a high cradle-to-grate impact, a high amount of solvent is lost in the wastewater stream (Table 2). A high solvent loss to wastewater causes a high make-up demand to run the process in steady-state. Therefore, a high amount of solvent needs to be produced for make-up, causing high absolute impacts from solvent production regarding the functional unit of 1 kmol GVL. Conversely, a low impact of solvent production is only achieved with a small make-up demand of solvent, in particular as the specific cradle-to-grate impacts are within the same order of magnitude for all candidate solvents.

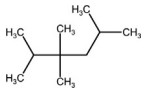
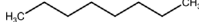
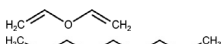
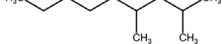
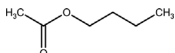
If the solvent loss is small, the use phase impact due to process energy dominates the cradle-to-grate LCA. Furthermore, a low solvent loss reduces uncertainty propagation of the ANN predictions. As a result, the uncertainties of the cradle-to-grate impact decrease (Table 2). Therefore, accurate LCI of the use phase and thus, accurate process modelling and precise property data are crucial. As an indicator for the accuracy, we compare the predicted solubilities of solvent in water from COSMO-RS with experimental data from the literature. The solubilities of the solvents in water are crucial for the LCA because they determine the solvent loss and make-up and consequently, the environmental impact of the solvent production. For the benchmark n-butyl acetate, solubilities of 6.7–8.3 g/l at 25 °C have been determined experimentally (Yalkowsky and He, 2016) compared to 6.1 g/l from our COSMO-RS predictions. For n-octane, COSMO-RS predicts a solubility at 25 °C of 3 mg/l compared to 0.4–0.9 mg/l experimentally (Yalkowsky and He, 2016). Considering the broad range of solubilities over multiple orders of magnitude, the experimental measures are both in good agreement with the COSMO-RS prediction; thus we conclude that COSMO-RS can be used for property prediction to generate accurate LCI for this process, even for the challenging hydrocarbon-water interactions (Klamt, 2003).

Still, the solvent loss alone is also not sufficient as an objective for molecular design (Fig. 6C). The solvent with the lowest solvent loss, 2,4-dimethyl-nonane, ranks only 600th in process energy



**Table 2**

Comparison of the candidate solvents with the lowest cradle-to-grave and the lowest cradle-to-gate impact on Climate Change, as well as the solvent with the lowest solvent loss and the benchmark solvent. The comparison includes the absolute values for the cradle-to-grave impact on Climate Change  $CC_{\text{cradle-to-grave}}$ , the ranking in cradle-to-grave and cradle-to-gate impact on Climate Change (Rank CC cradle-to-grave and Rank CC cradle-to-gate), the ranking in process energy demand  $Q_{\text{reb}}$  (Rank  $Q_{\text{reb}}$ ), as well as the molar fraction of solvent in wastewater (Predicted  $x_{\text{solv,aq}}$ ) predicted by COSMO-RS. The absolute values for  $CC_{\text{cradle-to-grave}}$  also include the 95% confidence interval from the uncertainty propagation of the ANN.

Solvent	Molecular structure	$CC_{\text{cradle-to-grave}} / \text{kg CO}_2 - \text{eq. kmol}^{-1}_{\text{GVL}}$	Rank CC		Rank $Q_{\text{reb}}$	Predicted $x_{\text{solv,aq}}$
			cradle-to-grave	cradle-to-gate		
Lowest Climate Change (cradle-to-grave)		$4.36 \pm 0.0017$	1	139	1	$4.3 \times 10^{-7}$
- Commercially avail.		$4.40 \pm 0.0015$	8	97	9	$4.7 \times 10^{-7}$
Lowest Climate Change (cradle-to-gate)		$7.34 \pm 1.4$	75	1	27	$7.2 \times 10^{-4}$
Lowest solvent loss		$31.7 \pm 0.00012$	524	281	600	$2.8 \times 10^{-8}$
Benchmark (n-butyl acetate)		$13.5 \pm 3.1$	292	247	198	$9.4 \times 10^{-4}$

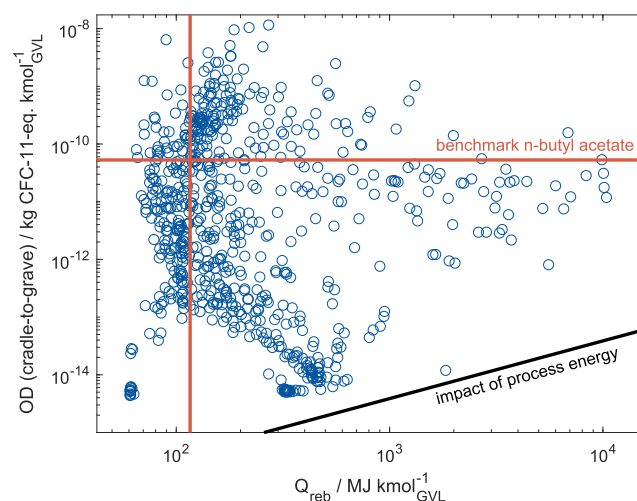
demand  $Q_{\text{reb}}$  and 524th in cradle-to-grave impact on Climate Change. The advantageous low solvent loss does not guarantee a low cradle-to-grave impact on CC, as low solvent loss and low energy demand for separation do not correlate. The high energy demand in distillation outweighs the favourable low solvent loss and make-up. Therefore, top solvents balance solvent loss as well as specific production impact and process energy demand. To include all these relevant factors, the cradle-to-grave LCA is required as objective function.

Besides the impact on Climate Change, we evaluated the other 16 ReCiPe midpoint impact categories (Goedkoop et al., 2008) as well for every candidate solvent in COSMO-susCAMPD. Generally, solvents ranked well in the impact on Climate Change and process energy demand show also a balanced performance in most of the other impact categories. E.g., the top solvent in cradle-to-grave impact on Climate Change is also among the top 10 solvents in 15 of the other 16 impact categories. As for the impact on Climate Change, the low solvent loss in the process combined with low energy demand in separation yields a low cradle-to-grave LCA impact.

For this case study, only the impact category Ozone Depletion (OD) differs from the trend of all other impact categories (Fig. 7). For Ozone Depletion, we find a strong trade-off between the cradle-to-grave impact on Ozone Depletion and the process energy demand for most solvents. As a result, the solvent ranking differs substantially for Ozone Depletion. E.g., 5 of the top 10 solvents in Ozone Depletion occupy ranks 400 and higher in Climate Change or ranks 500 and higher in process energy demand. The change in ranking for Ozone Depletion is due to the fact that the impacts due to solvent production and solvent loss dominate the impact on Ozone Depletion. This outcome is reasonable since process energy is supplied as heat from natural gas combustion with no substantial impact on Ozone Depletion. Thus, the presented method also captures the variable weighting of and trade-offs between the life cycles stages depending on the impact category considered.

#### 4. Conclusion

In this work, we present a framework for the design of solvents and processes with an environmental objective: COSMO-susCAMPD. The COSMO-susCAMPD framework extends state-of-the-art methods for Computer-Aided Molecular and Process Design



**Fig. 7.** Predicted cradle-to-grave impacts on Ozone Depletion (OD) of all solvents designed depending on corresponding process energy demand  $Q_{\text{reb}}$ . Each blue circle represents one candidate solvent. The black line is the impact resulting from the process energy demand; the red lines stand for the impact on Ozone Depletion and process energy demand of the benchmark solvent n-butyl acetate.

(CAMPD) by integrating predictive Life Cycle Assessment (LCA) with a cradle-to-grave system boundary. Cradle-to-grave LCA is achieved by the combination of (1) an Artificial Neural Network (ANN) predicting cradle-to-gate impacts with (2) process optimisation using pinch-based process models providing life cycle inventory for solvent use and disposal. Both the ANN and the process models use molecular and thermodynamic properties calculated from the predictive thermodynamic model COSMO-RS. Therefore, the assessment of environmental impact and process performance is based on one consistent set of descriptors. For simultaneous molecular and process design, the predictive LCA and the process optimisation are combined with the genetic algorithm LEA3D, which optimises 3D-molecular structures based on the results from LCA and process optimisation.

As an application for COSMO-susCAMPD, we investigate the purification of the bio-based platform chemical  $\gamma$ -valerolactone from aqueous solution by hybrid extraction-distillation. We optimise the process for minimum environmental impact by exploiting the degrees of freedom from molecular and process design. As a

result, we identify promising candidate solvents from a vast design space outperforming the literature benchmark *n*-butyl acetate by reducing the impact on Climate Change by about 68%. The candidate solvents identified exhibit both a high process performance, i.e. a low process energy demand, as well as a low cradle-to-grave environmental impact in various ReCiPe midpoint impact categories.

The results show that a cradle-to-grave assessment is necessary for the design of environmentally beneficial solvents. Simplified objectives, such as cradle-to-gate LCA or economic evaluation alone, lead to suboptimal solutions. Only the cradle-to-grave LCA balances conflicting molecular properties for an optimum result.

The COSMO-susCAMPD framework now provides a method for CAMPD based on process evaluation and environmental assessment using LCA. The results of COSMO-susCAMPD serve as an input for further validation by refined process simulations, life cycle assessment and experiments. In the future, the method can be extended by further criteria, e.g. by inertness of the solvents or environmental assessment of acute exposure and handling, such as the evaluation of Environmental, Health and Safety scores (EHS-criteria). The process modelling can be extended towards processes with other unit operations or product design considering a use phase different from a chemical process. The LCA could also be refined to include other emissions, such as fugitive emissions of the process (Smith et al., 2017). Further work is required to extend the LCA data for training the ANN. Currently, training data for the ANN is rare, leading to limited accuracy of the ANN predictions. An improvement in the prediction quality of the ANN is expected if more consistent LCA data on solvents is available. Importantly, any additional training data needs to be obtained from process data by consistent allocation and with consistent background data. However, for the given case study in this work, the prediction of accurate process data outweighs the influence of inaccuracies of the ANN.

In conclusion, the presented framework COSMO-susCAMPD extends the environmental assessment of state-of-the-art molecular design by predictive cradle-to-grave life cycle assessment to enable the computer-aided design of sustainable solvents and processes.

## CRediT authorship contribution statement

**Lorenz Fleitmann:** Conceptualisation, Methodology, Validation, Investigation, Data curation, Visualisation, Writing - original draft, Writing - review & editing. **Johanna Kleinekorte:** Conceptualisation, Methodology, Validation, Investigation, Data curation, Visualisation, Writing - original draft, Writing - review & editing. **Kai Leonhard:** Methodology, Validation, Writing - review & editing, Supervision. **André Bardow:** Conceptualisation, Methodology, Writing - review & editing, Supervision, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

L.F. gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy and Cluster of Excellence 2186 "The Fuel Science Center" ID: 390919832. J.K.'s work was supported by the German Federal Ministry of Education and

Research (BMBF) within the project consortium "Carbon2Chem" under Contract 03EK3042C.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.ces.2021.116863>.

## References

- Adu, I.K., Sugiyama, H., Fischer, U., Hungerbühler, K., 2008. Comparison of methods for assessing environmental, health and safety (EHS) hazards in early phases of chemical process design. *Process Saf. Environ. Prot.* 86 (2), 77–93. <https://doi.org/10.1016/j.psep.2007.10.005>.
- Alexander, D.L.J., Tropsha, A., Winkler, D.A., 2015. Beware of R<sup>2</sup>: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model.* 55 (7), 1316–1322.
- Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2018. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J. Choice Model.* 28, 167–182.
- Bakshi, B.R., 2019. Toward Sustainable Chemical Engineering: The Role of Process Systems Engineering. *Ann. Rev. Chem. Biomol. Eng.* 10, 265–288. <https://doi.org/10.1146/annurev-chembioeng-060718-030332>.
- Bausa, J., Watzdorf, R.v., Marquardt, W., 1998. Shortcut methods for nonideal multicomponent distillation: I. Simple columns. *AIChE J.* 44 (10), 2181–2198. <https://doi.org/10.1002/aic.690441008>.
- Brown, B.J., Hanson, M.E., Liverman, D.M., Merideth, R.W., 1987. Global sustainability: Toward definition. *Environ. Manage.* 11 (6), 713–719. <https://doi.org/10.1007/BF01867238>.
- Calvo-Serrano, R., González-Miquel, M., Papadokonstantakis, S., Guillén-Gosálbez, G., 2018. Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming. *Comput. Chem. Eng.* 108, 179–193. <https://doi.org/10.1016/j.compchemeng.2017.09.010>.
- Calvo-Serrano, R., González-Miquel, M., Guillén-Gosálbez, G., 2019. Integrating COSMO-Based  $\sigma$ -Profiles with Molecular and Thermodynamic Attributes to Predict the Life Cycle Environmental Impact of Chemicals. *ACS Sustain. Chem. Eng.* 7 (3), 3575–3583. <https://doi.org/10.1021/acssuschemeng.8b06032>.
- Canals, L.M.I., Azapagic, A., Doka, G., Jefferies, D., King, H., Mutel, C., Nemecek, T., Roches, A., Sim, S., Stichnothe, H., 2011. Approaches for addressing Life Cycle Assessment data gaps for bio-based products. *J. Ind. Ecol.* 15 (5), 707–725.
- Carney, J.G., Cunningham, P., Bhagwan, U., 1999. Confidence and prediction intervals for neural network ensembles, IJCNN'99. In: *International Joint Conference on Neural Networks. Proceedings 99CH36339*, pp. 1215–1218.
- Chemmagattuvalappil, N.G., 2020. Development of solvent design methodologies using computer-aided molecular design tools. *Curr. Opin. Chem. Eng.* 27, 51–59. <https://doi.org/10.1016/j.cocoe.2019.11.005>.
- Clarke, C.J., Tu, W.-C., Levers, O., Bröhl, A., Hallett, J.P., 2018. Green and Sustainable Solvents in Chemical Processes. *Chem. Rev.* 118 (2), 747–800. <https://doi.org/10.1021/acs.chemrev.7b00571>.
- Douguet, D., Munier-Lehmann, H., Labesse, G., Pochet, S., 2005. LEA3D: A computer-aided ligand design for structure-based drug design. *J. Med. Chem.* 48 (7), 2457–2468. <https://doi.org/10.1021/jm0492296>.
- Draper, N.R., Smith, H., 1998. *Applied regression analysis*. John Wiley & Sons.
- European Commission-Joint Research Centre, 2011. *International reference life cycle data system (ILCD) handbook: Recommendations for Life Cycle Impact Assessment in the European context, first edition*. EUR. Scientific and technical research series, Publications Office, Luxembourg, vol. 24571.
- Gertig, C., Leonhard, K., Bardow, A., 2020. Computer-aided molecular and processes design based on quantum chemistry: Current status and future prospects. *Curr. Opin. Chem. Eng.* 27, 89–97. <https://doi.org/10.1016/j.cocoe.2019.11.007>.
- Goedkoop, M., Heijungs, R., Huijbregts, M., Schryver, Struijs, J., van Zelm, R., 2009. ReCiPe 2008, A life cycle impact assessment method which comprises harmonised category indicators at the midpoint and the endpoint level 1, 1–126.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep learning*. MIT Press.
- Hellweg, S., Milà i Canals, L., 2014. Emerging approaches, challenges and opportunities in life cycle assessment. *Science* 344(6188), 1109–1113. doi: 10.1126/science.1248361.
- Hellweg, S., Fischer, U., Scheringer, M., Hungerbühler, K., 2004. Environmental assessment of chemicals: Methods and application to a case study of organic solvents. *Green Chem.* 6 (8), 418–427. <https://doi.org/10.1039/b402807b>.
- ISO 14040, 2006. *Life Cycle Assessment: Principles and framework*, Environmental management.
- Jimenez-Gonzalez, C., 2019. Life cycle considerations of solvents, *Current Opinion in Green and Sustainable Chemistry* 18, 66–71. <https://doi.org/10.1016/j.cogsc.2019.02.004>.
- Jiménez-González, C., Kim, S., Overcash, M.R., 2000. Methodology for developing gate-to-gate Life Cycle Inventory information. *Int. J. Life Cycle Assess.* 5 (3), 153–159.
- Jonuzaj, S., Cui, J., Adjiman, C.S., 2019. Computer-aided design of optimal environmentally benign solvent-based adhesive products. *Comput. Chem. Eng.* 130, 106518. <https://doi.org/10.1016/j.compchemeng.2019.106518>.

- Klamt, A., 2003. Prediction of the mutual solubilities of hydrocarbons and water with COSMO-RS. *Fluid Phase Equilib.* 206 (1–2), 223–235. [https://doi.org/10.1016/S0378-3812\(02\)00322-9](https://doi.org/10.1016/S0378-3812(02)00322-9).
- Klamt, A., Eckert, F., Arlt, W., 2010. COSMO-RS: An alternative to simulation for calculating thermodynamic properties of liquid mixtures. *Ann. Rev. Chem. Biomol. Eng.* 1, 101–122. <https://doi.org/10.1146/annurev-chembioeng-073009-100903>.
- Kleinekorte, J., Kröger, L., Leonhard, K., Bardow, A., 2019. A neural network-based framework to predict process-specific environmental impacts. In: Kiss, A.A., Zondervan, E., Lakerveld, R., Özkan, L. (Eds.), 29th European Symposium on Computer Aided Process Engineering, Vol. 47 of Computer-Aided Chemical Engineering. Elsevier, Amsterdam, pp. 1447–1452. <https://doi.org/10.1016/B978-0-128-18634-3.50242-3>.
- Kleinekorte, J., Fleitmann, L., Bachmann, M., Kätelhöhn, A., Barbosa-Póvoa, A., von der Assen, N., Bardow, A., 2020. Life Cycle Assessment for the Design of Chemical Processes, Products, and Supply Chains. *Ann. Rev. Chem. Biomol. Eng.* <https://doi.org/10.1146/annurev-chembioeng-011520-075844>.
- Kohn, W., Sham, L.J., 1965. Self-Consistent equations including exchange and correlation effects. *Phys. Rev.* 140 (4A), A1133–A1138. <https://doi.org/10.1103/PhysRev.140.A1133>.
- Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *Ann. Math. Stat.* 22 (1), 79–86.
- Lindsey, C., Sheather, S., 2010. Variable selection in linear regression. *Stata J.* 10 (4), 650–669.
- Linke, S., McBride, K., Sundmacher, K., 2020. Systematic green solvent selection for the hydroformylation of long chained alkenes. *ACS Sustain. Chem. Eng.* <https://doi.org/10.1021/acssuschemeng.0c02611>.
- McBride, K., Linke, S., Xu, S., Sundmacher, K., 2018. Computer Aided Design of Green Thermomorphic Solvent Systems for Homogeneous Catalyst Recovery. 13th International Symposium on Process Systems Engineering (PSE 2018). Computer Aided Chemical Engineering, vol. 44. Elsevier, pp. 1783–1788. <https://doi.org/10.1016/B978-0-444-64241-7.50292-5>.
- Murat Sen, S., Henao, C.A., Braden, D.J., Dumesic, J.A., Maravelias, C.T., 2012. Catalytic conversion of lignocellulosic biomass to fuels: Process development and techno-economic evaluation. *Chem. Eng. Sci.* 67 (1), 57–67. <https://doi.org/10.1016/j.ces.2011.07.022>.
- Ooi, J., Ng, D.K., Chemmangattuvalappil, N.G., 2018. Optimal molecular design towards an environmental friendly solvent recovery process. *Comput. Chem. Eng.* 117, 391–409. <https://doi.org/10.1016/j.compchemeng.2018.06.008>.
- Otto, S.A., Kadin, M., Casini, M., Torres, M.A., Blenckner, T., 2018. A quantitative framework for selecting and validating food web indicators. *Ecol. Ind.* 84, 619–631.
- Papadopoulos, A.I., Stijepovic, M., Linke, P., 2010. On the systematic design and selection of optimal working fluids for Organic Rankine Cycles. *Appl. Therm. Eng.* 30 (6–7), 760–769. <https://doi.org/10.1016/j.applthermaleng.2009.12.006>.
- Papadopoulos, A.I., Tsivintzelis, I., Linke, P., Seferlis, P., 2018. Computer-Aided Molecular Design: Fundamentals, Methods, and Applications. In: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. Elsevier. <https://doi.org/10.1016/B9780-12-409547-2.14342-2>.
- Papadopoulos, A.I., Shavaliyeva, G., Papadokonstantakis, S., Seferlis, P., Perdomo, F.A., Galindo, A., Jackson, G., Adjiman, C.S., 2020. An approach for simultaneous computer-aided molecular design with holistic sustainability assessment: Application to phase-change CO<sub>2</sub> capture solvents. *Comput. Chem. Eng.* 135, 106769. <https://doi.org/10.1016/j.compchemeng.2020.106769>.
- Parvatkar, A.G., Eckelman, M.J., 2020. Simulation-Based Estimates of Life Cycle Inventory Gate-to-Gate Process Energy Use for 151 Organic Chemical Syntheses. *ACS Sustain. Chem. Eng.* 8 (23), 8519–8536. <https://doi.org/10.1021/acssuschemeng.0c00439>.
- Redepenning, C., Recker, S., Marquardt, W., 2017. Pinch-based shortcut method for the conceptual design of isothermal extraction columns. *AIChE J.* 63 (4), 1236–1245. <https://doi.org/10.1002/aic.15523>.
- Righi, S., Baioli, F., Dal Pozzo, A., Tugnoli, A., 2018. Integrating Life Cycle Inventory and process design techniques for the early estimate of energy and material consumption data. *Energies* 11 (4), 970.
- Scheffczyk, J., Schäfer, P., Fleitmann, L., Thien, J., Redepenning, C., Leonhard, K., Marquardt, W., Bardow, A., 2018. COSMO-CAMPD: A framework for integrated design of molecules and processes based on COSMO-RS. *Mol. Syst. Des. Eng.* 3 (4), 645–657. <https://doi.org/10.1039/c7me00125h>.
- Schilling, J., Tillmanns, D., Lampe, M., Hopp, M., Gross, J., Bardow, A., 2017. From molecules to dollars: Integrating molecular design into thermo-economic process design using consistent thermodynamic modeling. *Mol. Syst. Des. Eng.* 2 (3), 301–320. <https://doi.org/10.1039/c7me00026j>.
- Smith, R.L., Ruiz-Mercado, G.J., Meyer, D.E., Gonzalez, M.A., Abraham, J.P., Barrett, W. M., Randall, P.M., 2017. Coupling Computer-Aided Process Simulation and Estimations of Emissions and Land Use for Rapid Life Cycle Inventory Modeling. *ACS Sustain. Chem. Eng.* 5 (5), 3786–3794. <https://doi.org/10.1021/acssuschemeng.6b02724>.
- Soh, L., Eckelman, M.J., 2016. Green Solvents in Biomass Processing. *ACS Sustain. Chem. Eng.* 4 (11), 5821–5837. <https://doi.org/10.1021/acssuschemeng.6b01635>.
- Song, R., Keller, A.A., Suh, S., 2017. Rapid life-cycle impact screening using artificial neural networks. *Environ. Sci. Technol.* 51 (18), 10777–10785.
- Song, Z., Hu, X., Wu, H., Mei, M., Linke, S., Zhou, T., Qi, Z., Sundmacher, K., 2020. Systematic Screening of Deep Eutectic Solvents as Sustainable Separation Media Exemplified by the CO<sub>2</sub> Capture Process. *ACS Sustain. Chem. Eng.* 8 (23), 8741–8751. <https://doi.org/10.1021/acssuschemeng.0c02490>.
- Ten, J.Y., Hassim, M.H., Ng, D.K.S., Chemmangattuvalappil, N.G., 2017. A molecular design methodology by the simultaneous optimisation of performance, safety and health aspects. *Chem. Eng. Sci.* 159, 140–153. <https://doi.org/10.1016/j.ces.2016.03.026>.
- Ten, J.Y., Hassim, M.H., Chemmangattuvalappil, N.G., 2020. Integration of safety and health aspects in a simultaneous process and molecular design framework. *Chem. Eng. Res. Des.* 153, 849–864. <https://doi.org/10.1016/j.cherd.2019.11.018>.
- Ten, J.Y., Liew, Z.H., Oh, X.Y., Hassim, M.H., Chemmangattuvalappil, N., 2021. Computer-Aided Molecular Design of Optimal Sustainable Solvent for Liquid-Liquid Extraction. *Process Integr. Optimiz. Sustain.* <https://doi.org/10.1007/s41660-021-00166-7>.
- Ruiz, E.M., 2019. Treatment of spent solvent mixture, hazardous waste incineration: Europe without Switzerland (ecoinvent database version 3.6).
- I. The MathWorks, 2018. MATLAB: Global Optimization Toolbox (Release 2018b).
- Thinkstep, 2012. GaBi: LCA software and LCI database. <http://www.gabi-software.com/databases/gabi-databases/>.
- Wernet, G., Hellweg, S., Fischer, U., Papadokonstantakis, S., Hungerbühler, K., 2008. Molecular-structure-based models of chemical inventories using neural networks. *Environ. Sci. Technol.* 42 (17), 6717–6722. <https://doi.org/10.1021/es7022362>.
- Wernet, G., Papadokonstantakis, S., Hellweg, S., Hungerbühler, K., 2009. Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. *Green Chem.* 11 (11), 1826. <https://doi.org/10.1039/B905558D>.
- Yalkowsky, S., He, Y., 2016. Handbook of Aqueous Solubility Data, first ed. CRC Press, Boca Raton. doi: 10.1201/9780203490396.
- Zhang, Z., 2016. Synthesis of  $\gamma$ -Valerolactone from Carbohydrates and its Applications. *ChemSusChem* 9 (2), 156–171. <https://doi.org/10.1002/cssc.201501089>.
- Zhou, T., McBride, K., Linke, S., Song, Z., Sundmacher, K., 2020. Computer-aided solvent selection and design for efficient chemical processes. *Curr. Opin. Chem. Eng.* 27, 35–44. <https://doi.org/10.1016/j.coche.2019.10.007>.