SUPPORTING INFORMATION

Early-Stage Environmental Impact Forecasting of Chemicals with Machine Learning and Data Analytics Tools

Harriet Dufie Appiah, Matthew Conway, Jahnvi Patel, Marcella McMahon, Robert Hesketh, and Kirti M. Yenkie*

Department of Chemical Engineering, Rowan University, Glassboro, NJ, USA

Corresponding author email: yenkie@rowan.edu

S.1. Sequential Backward Feature Selection (SBFS)

The initial feature space contained 200 molecular descriptors and 23 thermodynamic properties. SMILES strings for each chemical were obtained using CIRpy (v1.0.2), then passed to the *chemicals* (v1.1.4) and *thermos* (v0.2.26) libraries to retrieve thermodynamic data, while RDKit (v2023.3.3) was used to compute the molecular descriptors. Given the large number of available features, it is important to retain only those that contribute most to model performance hence we used the Sequential Backward Feature Selection with linear regression and 5-fold cross-validated mean squared error to select five features from each feature set.

Mathematically, the SBFS procedure can be written as follows:

Input:

F – set of all features

J – evaluation function (5-fold cross-validated mean squared error)

K – desired number of features ($= 5$)

Initialize:

$t = 0$

$S^{(0)} = F$

While $\left|S^{(t)}\right| > K$:

For each feature $f \in S^{(t)}$:

$\text{score}(f) = J\left(S^{(t)} \backslash \{f\}\right)$

$f^* = \text{argmin}\, \text{score}(f)$

$S^{(t+1)} = S^{(t)} \backslash f^* S^{(t+1)}$

Output:

$S^{(t)}$ with $\left|S^{(t)}\right| = K$

S.2. Machine learning hyperparameter tuning

The performance of the ANN model depends strongly on a set of hyperparameters, such as the number of hidden layers and neurons, activation functions, learning rate etc. In this work, the Adam optimizer was employed to train the ANN, as it uses adaptive learning rates and momentum, which typically lead to faster and more stable convergence compared to standard gradient descent for this type of dataset. The available data w split into training, testing and validation sets, and the tuning of the hyperparameters was conducted using the validation set by selecting the configuration that yielded the lowest validation error.

Table 1 Hyperparameter tuning for ANN models

Hyperparameter	GWP	RUI	HHI	EQI
Input layer activation function	elu	elu	tanh	elu
Input layer #neurons	200	200	140	200
Input layer dropout rate	0.2631	0.2631	0.1834	0.2631
Input layer initializer	GlorotUniform	GlorotUniform	GlorotUniform	GlorotUniform
Number of hidden layers	5	5	4	5
Hidden layer initializer	GlorotUniform, GlorotUniform, GlorotUniform, GlorotUniform, GlorotUniform	GlorotUniform, GlorotUniform, GlorotUniform, GlorotUniform, GlorotUniform	HeNormal, HeNormal, HeNormal, HeNormal	GlorotUniform, GlorotUniform, GlorotUniform, GlorotUniform, GlorotUniform
Hidden layer #neurons	200, 200, 200, 200, 200	200, 200, 200, 200, 200	250, 250, 250, 250	200, 200, 200, 200, 200
Hidden layer activation function	leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051),	leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051),	relu, relu, relu, relu	leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051), leaky_relu(alpha=0.051),
Hidden layer dropout rate	0.3055, 0.3055, 0.3055, 0.3055, 0.3055,	0.3055, 0.3055, 0.3055, 0.3055, 0.3055,	0.3055, 0.3055,	0.3055, 0.3055,
Output layer activation function	relu	relu	relu	relu
Loss function	Mean squared logarithmic error	Mean squared logarithmic error	Mean squared logarithmic error	Mean squared logarithmic error
Learning rate	0.00108	0.00108	0.00127	0.00108

S.3. Data points for the Gate-to-Gate and end-of-life phase model

This phase leverages a power-law regression correlation to develop regression models that estimate the GWP of the different technologies. The regression equation inspired by cost estimation equation is written as:

$$\left(\frac{\mathrm{GWP_{new}}}{\mathrm{GWP_{ref}}}\right) = \left(\frac{\mathrm{F_{new}}}{\mathrm{F_{ref}}}\right)^{\alpha} \left(\frac{\mathrm{E_{new}}}{\mathrm{E_{ref}}}\right)^{\beta}$$

where F denotes the feed throughput and E denotes the energy consumption of the technology, and $\alpha$and β are regression coefficients which tell if the GWP of the technology is dependent on the feed throughput or energy consumption respectively.

Linearized form:

$\ln \mathrm{X1} = \alpha \ln \mathrm{X2} + \beta \mathrm{Y}$

Where;

$$\mathbf{X_1} = \frac{\mathrm{F_{new}}}{\mathrm{F_{ref}}},\ \mathbf{X_2} = \frac{\mathrm{E_{new}}}{\mathrm{E_{ref}}}, \text{ and } \mathbf{Y} = \frac{\mathrm{GWP_{new}}}{\mathrm{GWP_{ref}}}$$

Table 2 and **Table 3** present the literature-derived data points used to construct the distillation and pervaporation graphs shown in this work. In each case, a reference data point was chosen to normalize the full dataset, and the resulting normalized points were plotted and fitted to obtain the regression coefficients $\alpha$and β.

Table 2 Data points for distillation obtained from literature

	Feed Rate (kg/s)	**Energy (kW)**	**GWP (kg CO_2-eq/kg feed)**
Chea et al. [1]	0.2778	1175.3	0.2723
Do Thi and Toth. [2]	0.2778	223.3	0.0706
Caballero-Sanchez et al. [3]	0.0130	20.1	0.1295
Cavanagh et al. [4]	0.1261	407.2	0.2067
Cavanagh et al. [4]	0.4856	324.7	0.0500
Cavanagh et al. [4]	0.3089	1022.9	0.2219

Reference data:

Do Thi et al. [2]	0.2778	223.3	0.0706

Table 3 Normalized dataset used to generate the distillation graph

Data	$\mathbf{X_1}$	$\mathbf{X_2}$	**Y**
1	1	5.2627	3.8590
2	1	1	1
3	0.0466	0.0901	1.8350
4	0.454	1.8232	2.9294
5	1.748	1.4541	0.7082
6	1.112	4.5806	3.1445

Power-law model for the distillation unit:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0706}\right) = \left(\frac{\mathrm{F_{new}}}{0.278}\right)^{-0.83} \left(\frac{\mathrm{E_{new}}}{223.3}\right)^{0.799}$$

The same procedure described above for distillation was applied to pervaporation, using Meng et al. as the reference data.

Table 4 Data points for pervaporation obtained from literature

	Flux (L/m².h)	**Energy (kJ/h)**	**GWP (kg CO_2-eq/kg feed)**
Meng et al.[5]	4.3357	1.22E+07	1.46E-02
Lee et al.[6]	1.3870	2.05E+05	5.500E-03
Norkobilov et al.[7]	0.262	1.54E+06	7.420E-03
Norkobilov et al.[7]	0.2876	6.16E+05	3.160E-03
Do Thi and Toth[2]	1.3000	1.66E+05	1.320E-03

Reference data:

Meng et al.[5]	4.3357	1.22E+07	1.46E-02

Table 5 Normalized dataset used to generate the pervaporation graph

Data	$\mathbf{X_1}$	$\mathbf{X_2}$	**Y**
1	1	1	1
2	0.3199	0.0167	0.3767
3	0.0604	0.1256	0.5082
4	0.0663	0.0504	0.2164
5	0.2998	0.0135	0.0904

Power-law model for the pervaporation unit:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0146}\right) = \left(\frac{\mathrm{F_{new}}}{4.3357}\right)^{0.0154} \left(\frac{\mathrm{E_{new}}}{1.22 * 10^7}\right)^{0.411}$$

These datasets were used to fit a regression model and obtain the scaling coefficients α and β, which were then used to calculate the GWP of the distillation unit in all three scenarios from the case study.

S.4. Overall GWP estimation for the Case Study

In the case study presented by Do Thi and Toth [2], the authors examine the recovery of high-purity ethanol and isobutanol from dilute alcohol–water waste streams using hybrid distillation–pervaporation processes. They evaluate the environmental impacts of the different separation configurations, including climate change, alongside their process performance. In our subsequent calculations, we adopt their ethanol–water mixture as the reference case and estimate the GWP of all scenarios using our model, then compare these values with the GWPs reported in their study.

1. **Gate-to-gate GWP calculation for scenario 1 (D+PV):**

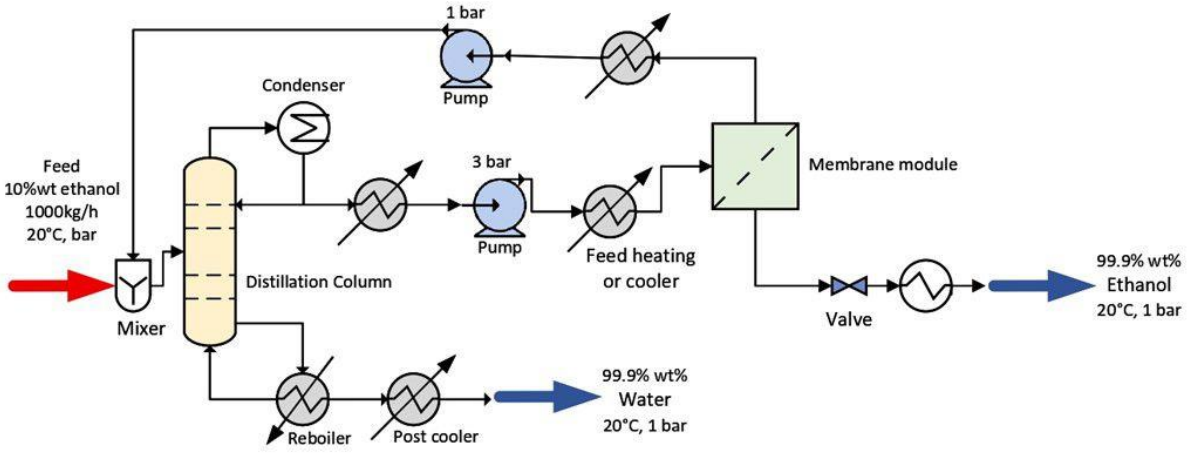


Table 6 Summary of material and energy balance

Technology	Parameter	Value	Unit
Distillation	Feed Mass Flow	1000	Kg/hr.
	Ethanol mass fraction	10	%
	Water mass fraction	90	%
	Reboiler duty	554.64	MJ/hr.
	Condenser duty	-234.58	MJ/hr.
Pervaporation	Retentate heating	35.58	MJ/hr.
	Permeate cooling	-38.54	MJ/hr.
	Total membrane area	200	m^2

GWP for distillation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0706}\right) = \left(\frac{0.278}{0.278}\right)^{-0.83} \left(\frac{219.23}{223.3}\right)^{0.799}$$

$\mathrm{GWP_{new}}$ = 0.0696 kg CO_2-eq

GWP for pervaporation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0146}\right) = \left(\frac{0.0728}{4.3357}\right)^{0.0154} \left(\frac{74.12 * 10^3}{1.22 * 10^7}\right)^{0.411}$$

GWP_{new} = 0.016829 kg CO_2-eq

Table 7 GWP of the various technologies involved in scenario 1

Technology	**GWP (kg CO_2-eq/kg feed)**
Distillation	0.0696
Pervaporation	0.0016829
Total	**0.0728829**

2. Gate-to-gate GWP calculation for scenario 2 (D+PV+D):

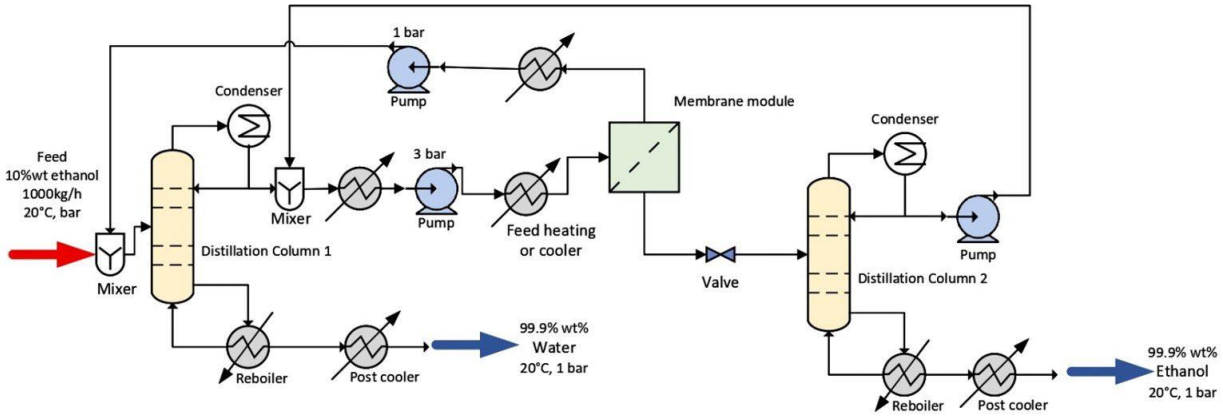


Table 8 Summary of material and energy balance

Technology	**Parameter**	**Value**	**Unit**
1st Distillation	Feed Mass Flow	1000	Kg/hr.
	Ethanol mass fraction	10	%
	Water mass fraction	90	%
	Reboiler duty	713.56	MJ/hr.
	Condenser duty	-396.13	MJ/hr.
Pervaporation	Retentate heating	14.86	MJ/hr.
	Permeate cooling	-17.86	MJ/hr.
	Total membrane area	130	m^2
2nd Distillation	Retentate heating	1080.19	MJ/hr.
	Permeate cooling	-1095.41	MJ/hr.

GWP for 1st Distillation:

$$\left(\frac{GWP_{new}}{0.0706}\right) = \left(\frac{0.278}{0.278}\right)^{-0.83} \left(\frac{308.25}{223.3}\right)^{0.799}$$

GWP_{new} = 0.0913 kg CO_2-eq

GWP for pervaporation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0146}\right) = \left(\frac{0.04915}{4.3357}\right)^{0.0154} \left(\frac{32.72 * 10^3}{1.22 * 10^7}\right)^{0.411}$$

$\mathrm{GWP_{new}}$ = 0.001195 kg CO_2-eq

GWP for 2nd Distillation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0706}\right) = \left(\frac{0.02764}{0.278}\right)^{-0.83} \left(\frac{604.33}{223.3}\right)^{0.799}$$

$\mathrm{GWP_{new}}$ =1.0626 kg CO_2-eq

Table 9 GWP of the various technologies involved in scenario 2

Technology	**GWP (kg CO_2-eq/kg feed)**
Distillation Column 1	0.0913
Pervaporation	0.001195
Distillation Column 2	1.0626
Total	**1.155095**

3. Gate-to-gate GWP calculation for scenario 2 (D+PV+D+HI):

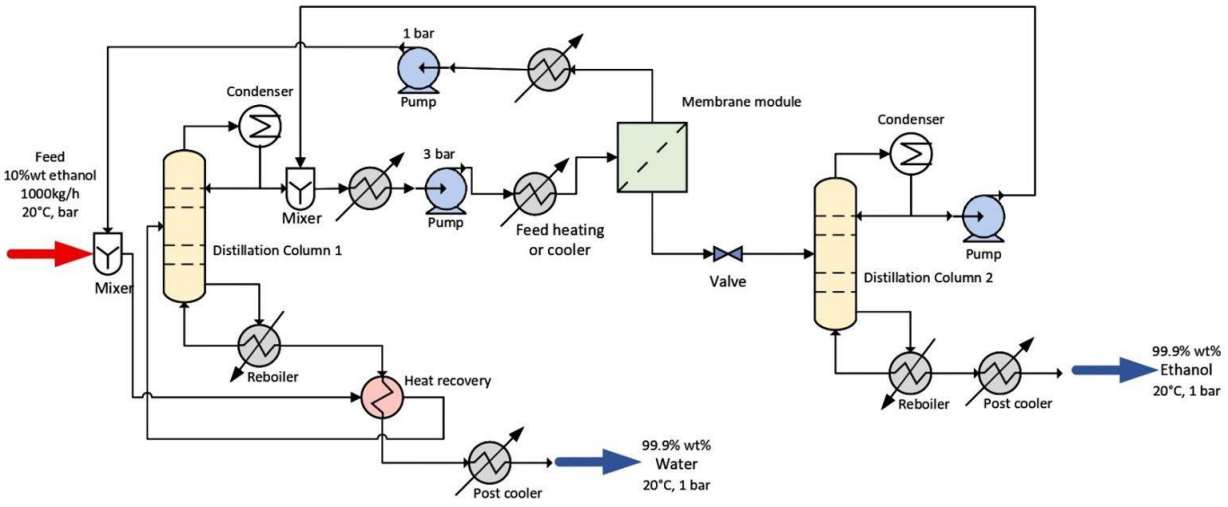


Table 10 Summary of material and energy balance

Technology	Parameter	Value	Unit
1st Distillation	Feed Mass Flow	1000	Kg/hr.
	Ethanol mass fraction	10	%
	Water mass fraction	90	%
	Reboiler duty	414.49	MJ/hr.
	Condenser duty	-396.87	MJ/hr.
Pervaporation	Retentate heating	15.02	MJ/hr.
	Permeate cooling	-18.09	MJ/hr.
	Total membrane area	130	m^2
2nd Distillation	Retentate heating	1078.65	MJ/hr.
	Permeate cooling	-1093.86	MJ/hr.

GWP for 1st Distillation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0706}\right) = \left(\frac{0.278}{0.278}\right)^{-0.83} \left(\frac{225.38}{223.3}\right)^{0.799}$$

$\mathrm{GWP_{new}}$ = 0.0711 kg CO_2-eq

GWP for Pervaporation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0146}\right) = \left(\frac{0.049397}{4.3357}\right)^{0.0154} \left(\frac{33.11 * 10^3}{1.22 * 10^7}\right)^{0.411}$$

$\mathrm{GWP_{new}}$ = 0.00120 kg CO_2-eq

GWP for 2nd Distillation:

$$\left(\frac{\mathrm{GWP_{new}}}{0.0706}\right) = \left(\frac{0.027772}{0.278}\right)^{-0.83} \left(\frac{603.48}{223.3}\right)^{0.799}$$

$\mathrm{GWP_{new}}$ =1.0572 kg CO_2-eq

Table 11 GWP of the various technologies involved in scenario 3

Technology	GWP (kg CO_2-eq/kg feed)
Distillation Column 1	0.0711
Pervaporation	0.00120
Distillation Column 2	1.0572
Total	**1.1295**

Once the GWPs for the production phase have been obtained from the ANN model and those for the use phase from the regression models described above, the overall life-cycle GWP of the process is calculated as:

$$GWP_{overall} = GWP_{production} + GWP_{use-phase} + GWP_{EoL}$$

Table 12 GWP for all three scenarios of ethanol-water mixture case study

Scenario	**GWP (ANN) (kg CO_2-eq/kg feed)**	**GWP (Regression) (kg CO_2-eq/kg feed)**	**Overall GWP (kg CO_2-eq/kg feed)**
D+PV	3.102	0.0728829	3.17
D+PV+D	3.102	1.155095	4.30
D+PV+D+HI	3.102	1.1295	4.23

Having estimated the GWP for our case study, we next summarize how Do Thi and Toth calculated their climate change values in order to enable a consistent comparison.

Table 13 Inventory data for the production of 1 kg of product in SimaPro

	D+PV	**D+PV+D**	**D+PV+D+HI**
Feed Streams			
Ethanol (kg)	0.10	0.10	0.10
Water(kg)	0.9	0.9	0.91
Utilities			
Heating duty	5.99	18.34	15.41
Cooling Water	10.96	10.18	1.13
Vacuum energy	0.30	1.21	1.21
Products			
Ethanol	1.00	1.00	1.00
Water by-product	9.07	9.06	9.11
Cooling water(kg)	10.96	10.18	1.13

Using the inventory data reported in **Table 14,** and applying the same Environmental Footprint (EF) (adapted) method with characterization, as described in the reference paper, we attempted to reproduce their impact results. Although we did not obtain same absolute values, likely because the exact background inventories for energy and material flows were not fully specified, we observed similar trends across the three scenarios, particularly for the climate change impact category.

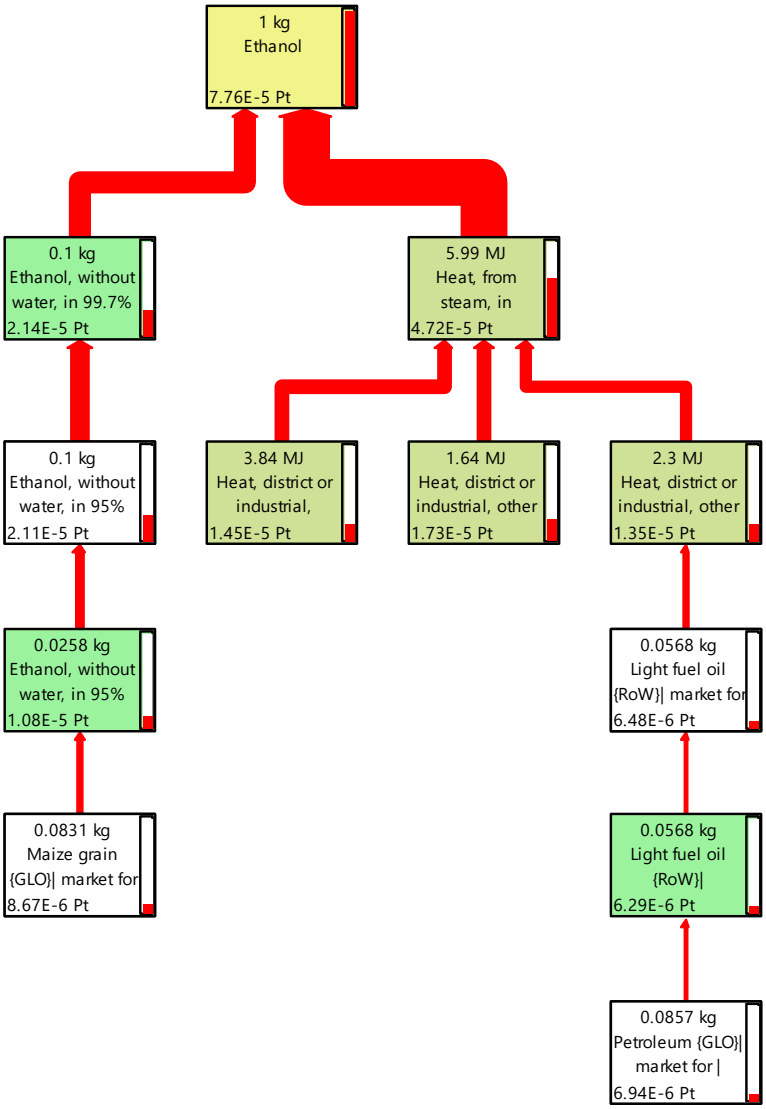
Table 15 Results of the Impact Assessment Using the Adapted EF Method

Impact category	Unit	D+PV	D+PV+D	D+PV+D+HI
Climate change	kg CO2 eq	9.31E-01	2.42E+00	2.09E+00
Ozone depletion	kg CFC11 eq	9.92E-08	2.54E-07	2.23E-07
Ionising radiation, HH	kBq U-235 eq	2.96E-02	6.87E-02	6.15E-02
Photochemical ozone formation, HH	kg NMVOC eq	2.05E-03	4.36E-03	4.04E-03
Respiratory inorganics	disease inc.	4.94E-08	1.09E-07	9.75E-08
Non-cancer human health effects	CTUh	9.94E-08	1.91E-07	1.73E-07
Cancer human health effects	CTUh	4.48E-09	7.55E-09	7.26E-09
Acidification terrestrial and freshwater	mol H+ eq	4.50E-03	1.01E-02	8.94E-03
Eutrophication freshwater	kg P eq	1.67E-04	3.58E-04	3.16E-04
Eutrophication marine	kg N eq	1.05E-03	1.72E-03	1.64E-03
Eutrophication terrestrial	mol N eq	1.00E-02	1.71E-02	1.63E-02
Ecotoxicity freshwater	CTUe	7.76E-01	1.12E+00	1.10E+00
Land use	Pt	2.15E+01	2.23E+01	2.27E+01
Water scarcity	m3 depriv.	1.14E+00	1.16E+00	7.17E-01
Resource use, energy carriers	MJ	1.12E+01	2.97E+01	2.57E+01
Resource use, mineral and metals	kg Sb eq	4.93E-07	5.12E-07	5.77E-07
Climate change - fossil	kg CO2 eq	9.05E-01	2.39E+00	2.07E+00
Climate change - biogenic	kg CO2 eq	2.24E-03	2.54E-03	2.48E-03
Climate change - land use and transform.	kg CO2 eq	2.29E-02	2.31E-02	2.30E-02

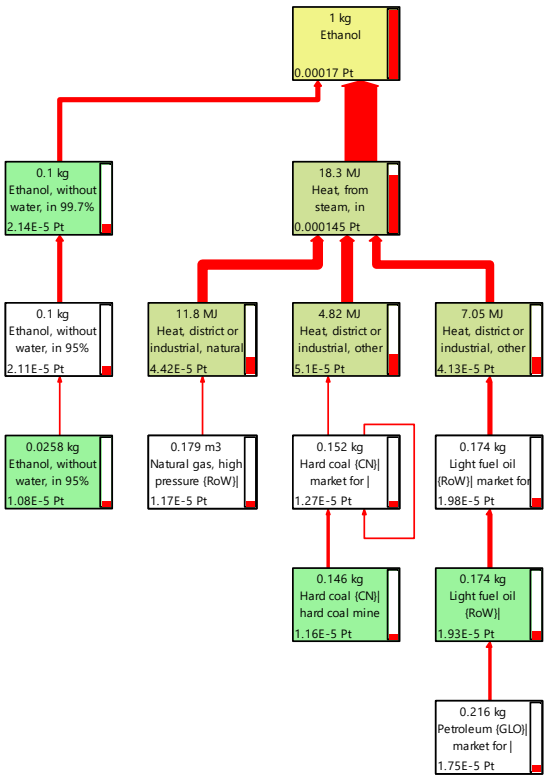
S.5. Contribution Tree Diagrams with Node Cut-Offs

Also presented below are contribution tree diagrams that highlight the processes making the largest contributions to the overall impact in each scenario. A node cut-off is applied so that only processes exceeding a specified share of the total impact are displayed.

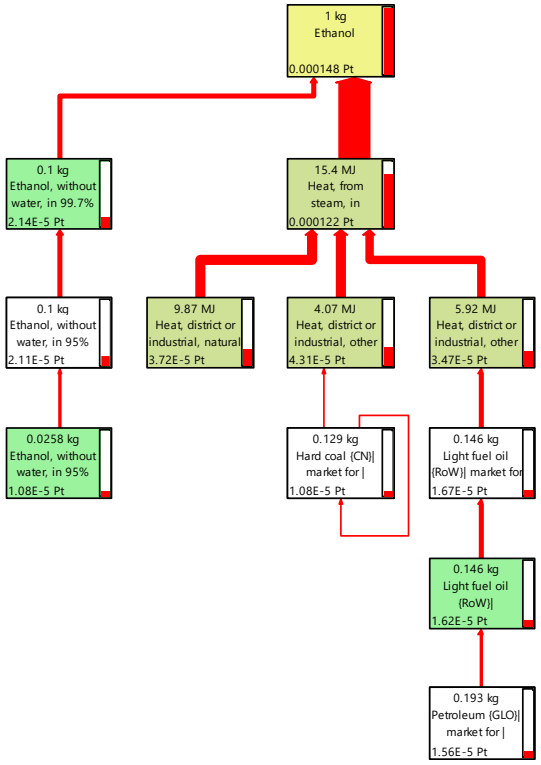
D+PV: 8.11% node cut-off



D+PV+D: 5.78% node cut-off



D+PV+D+HI: 6.73% node cut-off



References

[1] J. D. Chea, A. L. Lehr, J. P. Stengel, M. J. Savelski, C. S. Slater, and K. M. Yenkie, "Evaluation of Solvent Recovery Options for Economic Feasibility through a Superstructure-Based Optimization Framework," *Ind Eng Chem Res*, 2020.

[2] H. T. Do Thi and A. J. Toth, "Environmental evaluation and comparison of hybrid separation methods based on distillation and pervaporation for dehydration of binary alcohol mixtures with life cycle, PESTLE, and multi-criteria decision analyses," *Sep. Purif. Technol.*, vol. 348, p. 127684, Nov. 2024, doi: 10.1016/j.seppur.2024.127684.

[3] L. Caballero-Sanchez, A. A. Vargas-Tah, P. E. Lázaro-Mixteco, and A. J. Castro-Montoya, "Recovery of 1,4-butanediol from aqueous solutions through aqueous two-phase systems with K2CO3," *Chem. Eng. Res. Des.*, vol. 201, pp. 150–156, Jan. 2024, doi: 10.1016/j.cherd.2023.11.015.

[4] E. J. Cavanagh, M. J. Savelski, and C. S. Slater, "Optimization of environmental impact reduction and economic feasibility of solvent waste recovery using a new software tool," *Chem. Eng. Res. Des.*, vol. 92, no. 10, pp. 1942–1954, Oct. 2014, doi: 10.1016/j.cherd.2014.02.022.

[5] D. Meng *et al.*, "Energy, economic and environmental evaluations for the separation of ethyl acetate/ethanol/water mixture via distillation and pervaporation unit," *Process Saf. Environ. Prot.*, vol. 140, pp. 14–25, Aug. 2020, doi: 10.1016/j.psep.2020.04.039.

[6] H.-Y. Lee, S.-Y. Li, and C.-L. Chen, "Evolutional Design and Control of the Equilibrium-Limited Ethyl Acetate Process via Reactive Distillation−Pervaporation Hybrid Configuration," *Ind. Eng. Chem. Res.*, vol. 55, no. 32, pp. 8802–8817, Aug. 2016, doi: 10.1021/acs.iecr.6b01358.

[7] A. Norkobilov, D. Gorri, and I. Ortiz, "Process flowsheet analysis of pervaporation-based hybrid processes in the production of ethyl tert-butyl ether," *J. Chem. Technol. Biotechnol.*, vol. 92, no. 6, pp. 1167–1177, June 2017, doi: 10.1002/jctb.5186.