Systems & Control Transactions

Research Article
Foundations of Computer Aided Process Design (FOCAPD 2024)
Breckenridge, Colorado, USA. July 14-18, 2024
Peer Reviewed Conference Proceeding

PSE PRESS

Machine Learning Methods for the Forecasting of Environmental Impacts in Early-stage Process Design

Emmanuel A. Aboagye[a], Austin L. Lehr[a], Ethan Shumaker[a], Jared Longo[a], John Pazik[a], Robert P. Hesketh[a] and Kirti M. Yenkie[a]*

[a] Rowan University, Department of Chemical Engineering, Glassboro, NJ, USA
* Corresponding Author: yenkie@rowan.edu.

ABSTRACT

Initial design stages are inherently complex and often lack comprehensive information, posing challenges in evaluating sustainability metrics. Machine Learning (ML) emerges as a valuable solution to address these challenges. ML algorithms, particularly effective in predicting environmental impacts of new chemicals with limited data, enable more informed decisions in sustainable design. This study focuses on employing ML for predicting the environmental impacts related to human health, ecosystem quality, climate change, and resource utilization to aid in early-stage environmental impact assessment of chemical processes. The effectiveness of the ML algorithm, eXtreme Gradient Boosting (XGBoost) tested using a dataset of 350 points, divided into training, testing, and validation sets. The study also includes a practical application of the model in a cradle-to-cradle LCA of N-Methylpyrrolidone (NMP), demonstrating its utility in sustainable chemical process design. This approach signifies a significant advancement in the early stages of process design, highlighting the potential of ML in enhancing environmental sustainability in the chemical industry.

Keywords: Machine Learning, Life Cycle Analysis, Process Design, Modelling, Process Synthesis

INTRODUCTION

Amidst growing climate change concerns and heightened environmental awareness, industries are increasingly scrutinized for their environmental impact [1]. This scenario underscores the importance of environmental impact assessment at early-stage process synthesis [2], [3], where operational processes are initially formulated and assessed. Decisions made at this stage have significant implications for the environmental footprint of the entire operation. In this context, Machine Learning (ML) [4], [5] presents a transformative solution. By integrating ML in the early stages, industries can efficiently utilize its potential for rapid, accurate, and comprehensive assessments of sustainability.

ML offers a significant advantage in systems with non-obvious relationships in that, once trained, ML models can be used to predict such environmental metrics, facilitating prompt design modifications and improvements. Its adaptability also allows for easy integration with optimization strategies, aiding industries in developing processes during early-stages that balance environmental and economic factors. Financially, ML is invaluable; addressing sustainability issues early on helps industries avoid costly later-stage modifications, leading to substantial cost savings. In essence, incorporating ML into early-stage process synthesis represents a forward-thinking move for industries striving towards sustainability.

Thermodynamic properties such as enthalpy, entropy, Gibbs free energy [6], [7], among others provide crucial insights into the energy requirements of a process, operational efficiency, and overall feasibility. These attributes significantly influence the energy consumption of the process, impacting essential sustainability metrics such as GWP and total carbon footprint. On the other hand, molecular characteristics [8], [9], including molecular weight, bond energies, and functional groups, among others offer valuable information about the inherent qualities of chemical substances such as reactivity [10], potential toxicity [11], and environmental impacts [12], [13]. Often times, data on both thermodynamic and molecular

properties are available during the initial stages of process design. Therefore, by developing a ML model that incorporates thermodynamic and molecular descriptors as input and sustainability metrics as outputs, it is possible to predict sustainability metrics for both new and existing chemicals that lack established sustainability data. This approach enables a more holistic and informed evaluation of sustainability at the early stages of chemical process synthesis.

Previous research has shown that ML can be utilized to effectively enhance energy efficiency [14] and forecast corporate carbon footprints [15], among others [16], [17]. Building on this, the current study applies ML, specifically, eXtreme Gradient Boosting (XGBoost) [18], [19], to predict endpoint impact metrics for chemicals, particularly new molecules. This approach can lead to the development of safer and more sustainable chemical alternatives and circular process designs. Subsequently, the developed ML model is used together with other methods to predict the entire cradle-to-cradle environmental impact of NMP.

METHODOLOGY

In this section, we discuss the data gathering process, preprocessing, ML model building, and evaluation metrics.

Data Acquisition

A comprehensive dataset of 350 common solvents, including alcohols, esters, hydrocarbons, and ethers, was compiled for this study. The dataset is divided into two parts: the feature set and the label set. The feature set consists of thermodynamic and molecular descriptor data, encompassing the chemical properties used for model training. The label set, on the other hand, represents the data that the model aims to predict. For each chemical, 15 thermodynamic properties are gathered, such as critical temperature, pressure, volume, heat capacity, boiling point, and standard Gibbs-free energy. The data collection process begins with extracting the SMILES string [20] and chemical formula for each solvent.

These SMILES strings are used to extract the corresponding thermodynamic properties. This is achieved through two Python libraries: “chemicals” [21] and “thermo”, which host extensive databases of pure and calculated chemical properties. The local databank in these libraries includes over 20,000 chemicals and their properties, compiled from sources like the National Institute of Standards and Technology (NIST), Design Institute for Physical Properties (DIPPR), PubChem, CRC Handbook, Perry’s Chemical Engineers’ Handbook, and various scientific papers and publications.

The molecular descriptor properties for each chemical were obtained using RDKit [22] (version 2023.3.3), an open-source Python library renowned in cheminformatics. For this study, RDKit was employed to acquire 200 molecular descriptors for each chemical, covering various properties. These include molecular weight, carbon count, maximum partial charge, functional group, number of heterogeneous atoms, number of radical atoms, and the number of aliphatic rings, among others.

For the label data, SimaPro® [23] (version 9.4.0.2) is used to gather the cradle-to-gate metrics for each chemical. The metrics include human health impact (HHI), ecosystem quality impact (EQI), global warming potential (GWP), and resource utilization impact (RUI). These four endpoint metrics are chosen due to decision-making relevance, ease of communication, and depth of analysis.

Data Preprocessing

The initial step is to address the issue of missing data in the label set. While removing rows with missing data is a typical solution, ML models benefit from larger datasets. Therefore, the k-Nearest Neighbors (kNN) method, a well-established technique in data imputation, was employed for this analysis. Upon completion, the feature dataset was scaled to be in a range of 0 and 5.

Given the large number of features available in the dataset, it was necessary to identify and select those features that contribute most significantly to the model. This process of feature selection not only reduces computational time but also eliminates redundant features, thereby enhancing model accuracy. A balanced approach was adopted, choosing a total of 10 features with an equal number (5) from both the thermodynamic and molecular feature sets. This equal representation ensures that each feature set contributes fairly to the model. The streamlined selection of just 10 features also adds practical value for users, simplifying the prediction process for specific chemicals by requiring only a limited set of properties. Additionally, the model was tailored to each of the four metrics it predicts, with a unique set of features for each metric. This customization ensures that only the most relevant and impactful features are used for predictions in each specific case, optimizing the effectiveness and precision of the model. The Sequential Backward Feature Selection (SBFS) methodology, with linear regression and Mean-Squared-Error (MSE) criterion, was used to achieve this aim.

Model Training and Hyperparameter Tuning

Once the feature set for each label is finalized, the next step is to build the ML model for the prediction. Two models were developed, XGBoost and Artificial Neural Network (ANN), but in this paper, we present the XGBoost model.

XGBoost, an advanced ensemble ML model, is an

efficient implementation of the gradient boosting framework, particularly suitable for optimizing large-scale ML problems. It functions by iteratively building and refining models, each new model correcting the inaccuracies of its predecessors. This refinement is guided by the gradient descent method, which addresses the weaknesses in the existing ensemble by adding new decision trees, continuing until a predetermined error limit is reached or a specified number of trees is included.

In this study, the data was divided into training, validation, and testing sets. This split was not fix but was determined based on the label being predicted. To enhance the performance of the model, key hyperparameters of XGBoost (version 1.7.6) were optimized. This optimization involved selecting and tuning four to six hyperparameters that most significantly affect the model. This process was facilitated by the "hyperopt" [24] (version 0.2.7) library, which employs a Bayesian optimization framework [25]. The hyperparameters adjusted include the maximum depth of a tree, learning rate, number of trees, minimum child weight in a node, subsample fraction for growing trees, and the fraction of features chosen for tree development. The optimal hyperparameters were determined using the validation set, with an objective function designed to minimize the MSE between the actual and predicted values after training on the training set. The test set, crucially, was reserved exclusively for evaluating the generalizability and overall performance of the model.

For model evaluation, two key metrics were used: R-squared (R^2) value and the Root-Mean-Squared-Error (RMSE). The R^2, also known as the coefficient of determination, indicates the proportion of the variance in the dependent variable that the independent variables in the model can explain. On the other hand, the RMSE measures the average magnitude of the errors between the predicted and actual outcomes, providing a direct assessment of the accuracy of the model. This metric gives an absolute measure of the fit of the model, quantifying the average deviation in the predictions.

RESULTS AND DISCUSSION

In this section we discuss the results from the feature selection, the developed XGboost model and an NMP case study.

Feature Selection Result

Table 1 outlines the chosen features for evaluating various sustainability metrics, following the application of SBFS. The selected features underscore the importance of both thermodynamic properties and molecular descriptors in providing a comprehensive assessment of sustainability. Notably, critical temperature and heat capacity emerge as common thermodynamic features across all metrics, highlighting their universal applicability in sustainability evaluations. These properties are fundamental in understanding the energy dynamics and efficiency of chemical processes. Furthermore, the inclusion of XLogP and boiling point in three out of the four metrics signifies their relevance in assessing different environmental impacts. Molecular descriptors, particularly HallKierAlpha, have been selected for their ability to represent the three-dimensionality of molecules, a factor crucial in understanding the environmental behavior and impact of chemicals. HallKierAlpha, selected for three of the four metrics, specifically captures shape representation and molecular branching, aspects essential for evaluating the environmental compatibility of chemical substances.

Table 1: Selected features for each endpoint metric

Metric	Thermodynamic Feature	Molecular Descriptor Feature
HHI	heat of vaporization, heat capacity, XLogP, acentric factor, critical temperature	Chi0, HallKierAlpha, SMR_VSA7, VSA_EState6, NumValenceElectrons
EQI	heat capacity, standard formation enthalpy (gas), boiling Point, critical temperature, critical volume	Chi2v, BertzCT, HallKierAlpha, qed, fr_halogen
GWP	heat capacity, boiling point, XLogP, critical temperature, critical molar volume	BertzCT, ExactMolWt, HallKierAlpha, PEOE_VSA6, NOCount
RUI	heat capacity, boiling point, XLogP, critical pressure, critical temperature	ExactMolWt, MaxAbsPartialCharge, MaxPartialCharge, NumRotatableBonds, SMR_VSA2

Model Result

Figure 1 presents a parity plot of the predictive accuracy of the XGBoost model across the various environmental metrics. This parity plot offers a comprehensive view of the model's performance, highlighting its strengths and areas for further improvement.

Starting with the Human Health Impact (HHI) metric, as depicted in Figure 1 (a), the model demonstrates remarkable accuracy. The test set notably outperforms both the training and validation sets, achieving an R^2 of 0.997. This high score indicates a strong correlation

between the predicted and actual values, signifying the efficacy of the model in predicting HHI metric. Additionally, the RMSE values across the train-validation-test sets are closely aligned, further underscoring the reliability of the predictions. The range of HHI predictions, spanning 0.63 – 12 (x 10^{-6}) DALY/kg_{chem} with a 95% confidence interval, reflects the model's comprehensive coverage of potential human health impacts.

The Ecosystem Quality Impact (EQI) metric, illustrated in Figure 1 (b), presents a slightly different picture. Although the RMSE values remain within acceptable limits, suggesting general reliability, the model exhibits a significant discrepancy in the R^2 value for train-validation sets bur performs wells on the test set. This variance indicates a need for refinement in the model to achieve a more dependable R^2 value for EQI predictions. The predictions for EQI range from 0.022 – 3.0 PDF.m^2.yr/ kg_{chem}.

In the case of the Global Warming Potential (GWP) metric, the model shows good predictive performance, though it is not without its challenges. The GWP model displays a tendency to generalize to a good degree on the validation set but performs less efficient on the testing set based on the R^2 however, the RMSE for both validation and training set are similar in magnitude and order. The predicted values for GWP, ranging from 0.81 to 9.0 $kgCO_2$-eq/kg_{chem} within a 95% confidence interval, demonstrate capability of the model in this domain.

Lastly, the Resource Utilization Impact (RUI) metric, which can be interpreted as the Cumulative Energy Demand (CED) for chemical production, performs less in terms of prediction on the test set. The predictions for RUI, ranging from 4.5 to 15 (x10^1) MJ-primary/kg_{chem} with a 95% confidence interval, indicate a high degree of accuracy and reliability, showcasing the model's strengths in this area. In this study, the 95% confidence intervals were derived through a bootstrapping technique. This resampling approach enables us to assess the variability in the predictions made by our models. Specifically, for each model (HHI, EQI, GWP, RUI), we generated 1,000 bootstrap samples. This was achieved by randomly selecting 50% of the dataset with replacement in each iteration. Subsequently, we utilized the models to make predictions for each bootstrap sample. The construction of the 95% confidence intervals involved determining the 2.5th and 97.5th percentiles from the distribution of these bootstrap predictions. It is important to note that these intervals reflect the variability associated with the model predictions themselves, rather than the uncertainty in the hyperparameters of the XGBoost model or the variability inherent in the dataset. Table 2 also shows the comparison of the 95% confidence intervals for the actual data and the developed XGBoost model indicating a good agreement with the original data.

Table 2: Comparison of confidence interval for actual data and XGBoost model

Metric	Actual Data	XGBoost Model
HHI (x10^{-6})	0.55 – 12	0.63 – 12
EQI	0.022 – 3.1	0.022 – 3.0
GWP	0.68 – 9.8	0.81 – 9.0
RUI (x10)	3.5 – 17	4.5 – 15

Case Study: Cradle-to-cradle Life Cycle Assessment (LCA) of N-Methylpyrrolidone (NMP)

(NMP) is a polar aprotic solvent, notable for its high boiling point, and is widely used in the chemical industry, particularly in the production of polymers. Its role in polymer manufacturing is significant, but it also raises environmental concerns. The issue with NMP lies in its non-consumptive use in synthesis and processing, leading to its release as waste, a common occurrence in the fine and specialty chemical industries. The environmental and health risks associated with the disposal of NMP are well-recognized. However, the lack of suitable and safer alternatives to NMP and similar dipolar aprotic solvents has resulted in its continued widespread use in specialty chemical applications. Given these circumstances, the importance of solvent recovery after its usage becomes paramount. Recovering NMP not only mitigates the environmental and health risks but also addresses waste management concerns in the chemical industry, emphasizing the need for sustainable and responsible solvent usage practices.

Process Description

We consider a specific case where fresh solvents (n-methyl-2-pyrolidone (NMP)) and reagents (trifluoroacetic acid (TFA), hydroxyethyl methacrylate (HEMA), hydrochloric acid (HCl)) are initially sent to a reactor together with monomers (oxydianiline (ODA), pyromellitic dianhydride (PMDA)) for making a resin (polyimide (PI)) precursor. The role of the NMP is to dissolve the ODA while the reagents are added to the reaction medium to improve the photosensitivity of the PI produced. Once the reaction is complete, the product stream flows to the washing stage where ultrapure water is used to wash the produced resin resulting in three main streams: 1) the resin precursor stream which is sent to a filter press for further processing, 2) a hazardous waste stream containing NMP, and 3) a wastewater stream. We look at three main stages in the impact assessment: 1) the cradle-to-gate (production phase) impact which entails the feed stream containing the solvents, reagents and monomers, 2) gate-to-gate (usage phase) impact which entails the energy and water usage from the reaction and washing stages, and 3) gate-to-cradle (end-of-life phase) which entails the recovery of NMP from the hazardous waste

stream. Here, the XGboost model is used to predict the cradle-to-gate impact metrics, which represent the production phase. Additionally, we use the predicted values in the gate-to-cradle phase which represent the end-of-life phase of the waste solvent. Table 3 highlights the specifications for the case study.

Equations (1) – (3) gives the environmental impacts of each phase of the Life Cycle Assessment (LCA).

$$LCA_{i,production} = \sum_j^n LCA_{i,j,production} \quad (1)$$

$$LCA_{i,use-phase} = \sum_j^m LCA_{i,k,use-phase} + LCA_{i,water,use-phase} \quad (2)$$

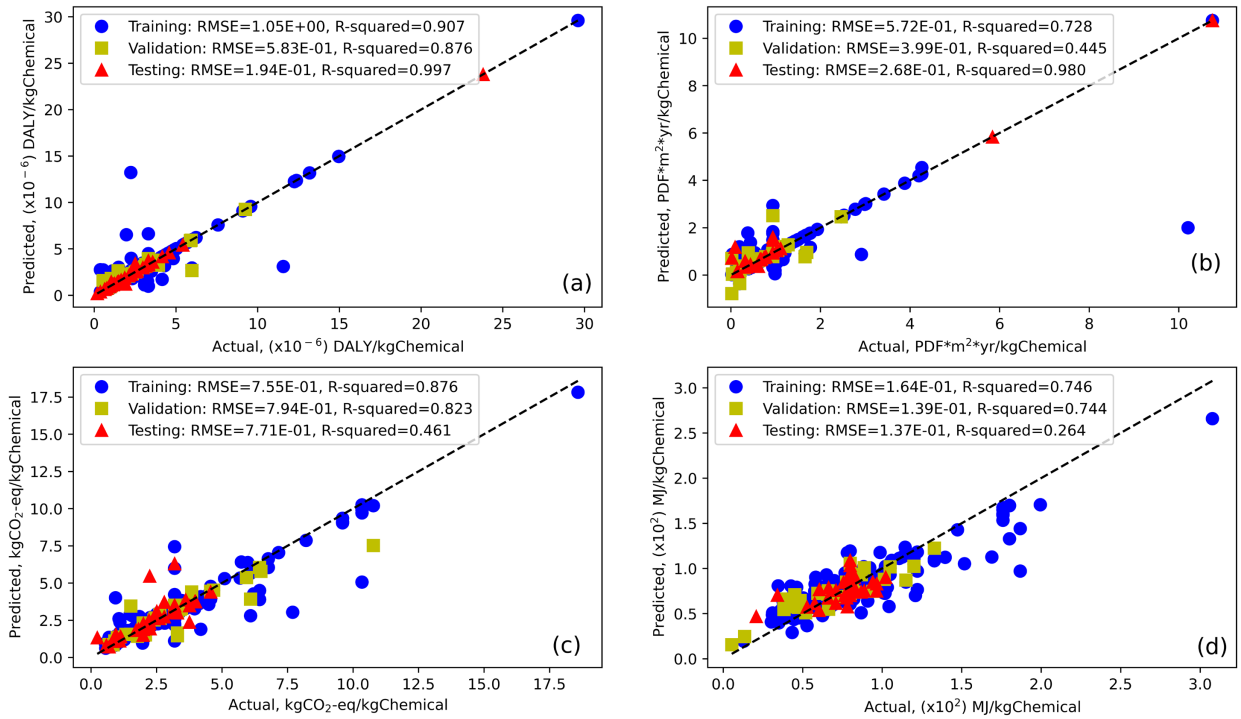

Figure 1: Parity plot for each metric from the XGBoost model. (a) HHI, (b) EQI, (c) GWP, (d) RUI

Table 3: NMP case study specification

Component	Inlet mass flowrate to Reactor (kg/y)	Ultrapure water for washing (kg/y)	Hazardous waste solvent stream composition (%wt)
NMP	183416	-	17
ODA	24055	-	-
PMDA	26202	-	-
HEMA	5448	-	0.5
TFA	5448	-	0.5
HCl	5448	-	0.5
H_2O	-	4114148	81.5

Table 4: Impact metric prediction for each chemical for the production phase from XGboost model

Component	HHI ($x10^{-6}$,DALY/kg_{Chem})	EQI (PDF.m^2.yr/kg_{Chem})	GWP ($kgCO_2eq/kg_{Chem}$)	RUI (x10,MJ/kg_{Chem})
NMP	7.6	1.9	2.7	11.3
ODA	1.7	0.6	2.5	14.1
PMDA	3.9	0.9	2.0	16.5
HEMA	1.8	1.0	2.5	7.3
TFA	1.2	0.9	2.7	5.4
HCl	0.7	0.2	0.7	0.2

$$LCA_{i,EoL} = (1 - R_{rec,NMP})LCA_{i,NMP,EoL} + \sum_{j-1}^{n} LCA_{i,k,EoL} + \sum_{j}^{m} LCA_{i,k,EoL} \quad (3)$$

Here, $LCA_{i,j,production}$ is the environment impact metric i for the production of chemical j, and n is the total number of chemicals. $LCA_{i,k,use-phase}$ is the environmental metric i for the energy demand of technology k, m is the total number of technologies in the process, $LCA_{i,water,use-phase}$ is the impact metric for the total amount of water used in the process. $R_{rec,NMP}$ is the amount of NMP recovered for reuse, $LCA_{i,NMP,EoL}$ is the environmental impact metric for NMP, $LCA_{i,j,EoL}$ is the environmental impact of the remaining chemicals not being recovered, and is the environmental impact due to the energy demand of the technologies (pervaporation in this case) for the solvent recovery process. The total cradle-to-cradle impact assessment per kg of NMP is given by Equation (4).

$$LCA_{i,cradle-to-cradle} = LCA_{i,production} + LCA_{i,use-phase} + LCA_{i,EoL} \quad (4)$$

Here, i is the environmental impact indicator (HHI, EQI, GWP, RUI), $LCA_{i,production}$ is the impact metric for the production phase of the chemicals (cradle-to-gate), $LCA_{i,use-phase}$ is the impact metric for the use-phase of the chemicals – in this case the energy demand and water usage in the reaction and washing stages (gate-to-gate) and $LCA_{i,EoL}$ is the impact metric for the EoL phase (gate-to-cradle) for the chemicals in the hazardous waste stream. NMP is the functional unit for the analysis hence the impact metric analysis is per kg of NMP basis. Figure 2 shows the LCA scope and the associated stages of the cradle-to-cradle assessment.

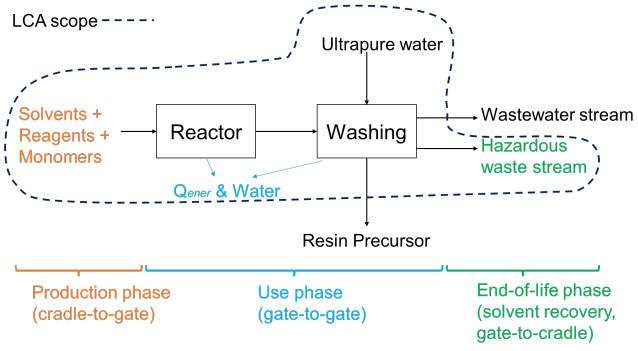

Figure 2: LCA scope and the various aspects of the cradle-to-cradle assessment

Case Study Results & Discussion

Table 4 shows the predictions for each chemical for the production phase assessment from the developed XGBoost model. To give a perspective of how the predictions compare with actual data from the SimaPro® software, the HHI, EQI, GWP, and RUI are 7.6 x 10^{-6}, 1.93, 7.7, and 168.9, respectively for NMP. It should be noted that for water, we used SimaPro® values for the analysis since we did not consider it to be a chemical.

Figure 3 shows the comprehensive impact of the NMP lifecycle.

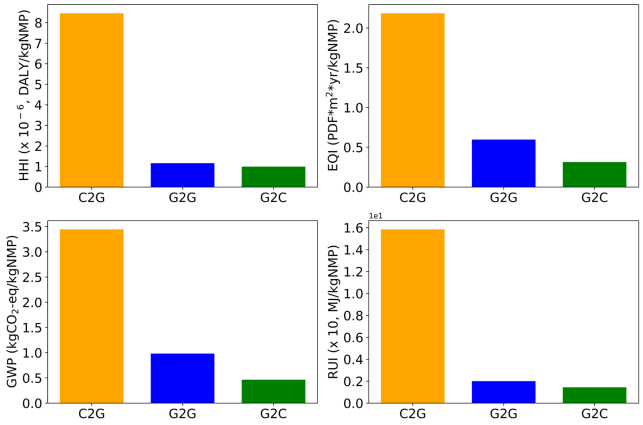

Figure 3: Endpoint impacts of NMP across its lifecycle. C2G:Cradle-to-Gate; G2G:Gate-to-Gate; G2C:Gate-to-Cradle

In the HHI, the results indicate a markedly higher impact in the C2G phase, illustrating the significant health-related implications during the initial stages of NMP production. This could be attributed to the extraction and processing of raw materials, underscoring the need for stringent health and safety measures during these processes. The EQI shows similar trend with the highest impact also observed in the C2G phase. However, the G2G is higher in this case signifying a higher impact contribution from the use-phase. For GWP, the difference between the C2G and G2G is 71.5%, whiles that between the G2G and G2C is 52.7%. Similar trend is observed with the RUI metric.

CONCLUSION

In this work, an XGBoost model is developed to predict four endpoint impact metrics of chemicals based on thermodynamic properties and molecular descriptors. The developed model was subsequently used in a case study where a cradle-to-cradle life cycle assessment is performed with NMP as the functional unit. The model is used to predict the production phase of the various chemicals used, and subsequently used in solvent recovery which is the considered route for the end-of-life phase. The use-phase is analysed using the utilities from the reaction and washing processes. The model results indicate that the human health impact has the best accuracy. While the remaining three metrics had significant improvements on the validation set, the models could be improved to enhance the predictions on the test set. Regarding the life cycle assessment, it is observed that the

cradle-to-gate stage has the significant impact on the lifecycle followed by gate-to-gate, and finally, gate-to-cradle for all four metrics. Additionally, this case study shows that ML model predictions can be used to substitute unknows data for cradle-to-gate, and even gate-to-cradle life cycle assessment. Furthermore, the developed model can be incorporated during the early-design stage to provide initial estimates of impact metrics for better decision-making.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the United State's Environmental Protection Agency Pollution Prevention Grant Program funded by the Bipartisan Infrastructure Law (Grant# 4U – 96236522).

REFERENCES

1. H. Cabezas and U. Diwekar, *Sustainability: Multi-Disciplinary Perspectives*, 1st ed. Bentham Science Publishers, 2012.
2. A. Argoti, A. Orjuela, and P. C. Narváez, "Challenges and opportunities in assessing sustainability during chemical process design," *Current Opinion in Chemical Engineering*, vol. 26, pp. 96–103, Dec. 2019, doi: 10.1016/j.coche.2019.09.003.
3. P. Karka, S. Papadokonstantakis, and A. Kokossis, "Environmental impact assessment of biomass process chains at early design stages using decision trees," *Int J Life Cycle Assess*, p. 26, 2019.
4. P. Karka, S. Papadokonstantakis, and A. Kokossis, "Digitizing sustainable process development: From ex-post to ex-ante LCA using machine-learning to evaluate bio-based process technologies ahead of detailed design," *Chemical Engineering Science*, vol. 250, p. 117339, Mar. 2022, doi: 10.1016/j.ces.2021.117339.
5. P. Karka, S. Papadokonstantakis, and A. Kokossis, "Predictive LCA - a systems approach to integrate LCA decisions ahead of design," in *Computer Aided Chemical Engineering*, vol. 46, Elsevier, 2019, pp. 97–102. doi: 10.1016/B978-0-12-818634-3.50017-5.
6. K. D. Dahm and D. P. Visco, "Fundamentals of Chemical Engineering Thermodynamics," 2015.
7. S. I. Sandler, *Chemical, biochemical and engineering thermodynamics*, Fifth edition. Hoboken, NJ: Wiley, 2017.
8. R. Parthasarathi and A. Dhawan, "Chapter 5 - In Silico Approaches for Predictive Toxicology," in *In Vitro Toxicology*, A. Dhawan and S. Kwon, Eds., Academic Press, 2018, pp. 91–109. doi: 10.1016/B978-0-12-804667-8.00005-5.
9. S. Hongmao, "Chapter 6 - Quantitative Structure–Property Relationships Models for Lipophilicity and Aqueous Solubility," in *A Practical Guide to Rational Drug Design*, S. Hongmao, Ed., Woodhead Publishing, 2016, pp. 193–223. doi: 10.1016/B978-0-08-100098-4.00006-5.
10. T. Stuyver, F. De Proft, P. Geerlings, and S. Shaik, "How Do Local Reactivity Descriptors Shape the Potential Energy Surface Associated with Chemical Reactions? The Valence Bond Delocalization Perspective," *J. Am. Chem. Soc.*, vol. 142, no. 22, pp. 10102–10113, Jun. 2020, doi: 10.1021/jacs.0c02390.
11. R. Srivastava, "Theoretical Studies on the Molecular Properties, Toxicity, and Biological Efficacy of 21 New Chemical Entities," *ACS Omega*, vol. 6, no. 38, pp. 24891–24901, Sep. 2021, doi: 10.1021/acsomega.1c03736.
12. E. A. Aboagye *et al.*, "Systematic Design of Solvent Recovery Pathways: Integrating Economics and Environmental Metrics," *ACS Sustainable Chem. Eng.*, vol. 10, no. 33, pp. 10879–10887, Aug. 2022, doi: 10.1021/acssuschemeng.2c02497.
13. O. Jolliet *et al.*, "IMPACT 2002+: A new life cycle impact assessment methodology," *Int J LCA*, vol. 8, no. 6, pp. 324–330, Nov. 2003, doi: 10.1007/BF02978505.
14. D. A. C. Narciso and F. G. Martins, "Application of machine learning tools for energy efficiency in industry: A review," *Energy Reports*, vol. 6, pp. 1181–1199, Nov. 2020, doi: 10.1016/j.egyr.2020.04.035.
15. Q. Nguyen, I. Diaz-Rainey, and D. Kuruppuarachchi, "Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach," *Energy Economics*, vol. 95, p. 105129, Mar. 2021, doi: 10.1016/j.eneco.2021.105129.
16. S. Boobier, D. R. J. Hose, A. J. Blacker, and B. N. Nguyen, "Machine learning with physicochemical relationships: solubility prediction in organic solvents and water," *Nat Commun*, vol. 11, no. 1, p. 5753, Dec. 2020, doi: 10.1038/s41467-020-19594-z.
17. A. Carranza-Abaid, H. F. Svendsen, and J. P. Jakobsen, "Surrogate modelling of VLE: Integrating machine learning with thermodynamic constraints," *Chemical Engineering Science: X*, vol. 8, p. 100080, Nov. 2020, doi: 10.1016/j.cesx.2020.100080.
18. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
19. M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, and

R. Liu, “XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System,” *IEEE Access*, vol. 7, pp. 13149–13158, 2019, doi: 10.1109/ACCESS.2019.2893448.
20. “CIRpy — CIRpy 1.0.2 documentation.” Accessed: Aug. 25, 2023. [Online]. Available: https://cirpy.readthedocs.io/en/latest/
21. “chemicals: Chemical properties component of Chemical Engineering Design Library (ChEDL) — Chemicals 1.1.4 documentation.” Accessed: Aug. 25, 2023. [Online]. Available: https://chemicals.readthedocs.io/
22. “RDKit.” Accessed: Aug. 25, 2023. [Online]. Available: https://www.rdkit.org/
23. M. Goedkoop, M. Oele, J. Leijting, T. Ponsioen, and E. Meijer, “Introduction to LCA with SimaPro.” PRe Sustainability, 2016.
24. J. Bergstra, D. Yamins, and D. Cox, “Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms,” presented at the Python in Science Conference, Austin, Texas, 2013, pp. 13–19. doi: 10.25080/Majora-8b375195-003.
25. P. I. Frazier, “Bayesian Optimization,” in *Recent Advances in Optimization and Modeling of Contemporary Problems*, in INFORMS TutORials in Operations Research. , INFORMS, 2018, pp. 255–278. doi: 10.1287/educ.2018.0188.