DOI: 10.1111/jiec.13507

METHODS ARTICLE

JOURNAL OF INDUSTRIAL ECOLOGY WILEY

Machine learning for gap-filling in greenhouse gas emissions databases

Luke Cullen[1] | **Andrea Marinoni**[1,2] | **Jonathan Cullen**[1]

[1]Department of Engineering, University of Cambridge, Cambridge, UK

[2]Department of Physics and Technology, UiT the Arctic University of Norway, Tromsø, Norway

Correspondence
Luke Cullen, Department of Engineering, University of Cambridge, Cambridge, UK.
Email: lshc3@cam.ac.uk

Editor Managing Review: Deepak Rajagopal

Funding information
UK Research and Innovation; UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks, Grant/Award Number: EP/S022961/1

Abstract

Greenhouse gas (GHG) emissions datasets are often incomplete due to inconsistent reporting and poor transparency. Filling the gaps in these datasets allows for more accurate targeting of strategies aiming to accelerate the reduction of GHG emissions. This study evaluates the potential of machine learning methods to automate the completion of GHG datasets. We use three datasets of increasing complexity with 18 different gap-filling methods and provide a guide to which methods are useful in which circumstances. If few dataset features are available, or the gap consists only of a missing time step in a record, then simple interpolation is often the most accurate method and complex models should be avoided. However, if more features are available and the gap involves non-reporting emitters, then machine learning methods can be more accurate than simple extrapolation. Furthermore, the secondary output of feature importance from complex models allows for data collection prioritization to accelerate the improvement of datasets. Graph-based methods are particularly scalable due to the ease of updating predictions given new data and incorporating multimodal data sources. This study can serve as a guide to the community upon which to base ever more integrated frameworks for automated detailed GHG emissions estimations, and implementation guidance is available at https://hackmd.io/@luke-scot/ML-for-GHG-database-completion and https://doi.org/10.5281/zenodo.10463104. This article met the requirements for a gold-gold *JIE* data openness badge described at http://jie.click/badges.

KEYWORDS
automation, data completion, data prioritisation, graph representation learning, greenhouse gas emissions, machine learning

1 | INTRODUCTION

Greenhouse gas (GHG) emissions datasets are often incomplete both at an individual facility level, for example, ClimateTRACE (2023), and at a national level, for example, UNFCCC (2023). Most countries, and many companies, are accelerating their emissions reduction strategies in line with the Paris climate agreement and net-zero objectives (Arnold & Toledano, 2021; Christiansen et al., 2023; Erb et al., 2022; Rogelj et al., 2016), but incomplete and inaccurate datasets remain a barrier to understanding and, therefore, to effective policy-making (EPA, 2022a; IPCC, 2021; Marlowe & Clarke, 2022). In regions including the European Union, the United Kingdom, and the United States, large companies are required to use generic emissions intensity factors to convert their facilities' activity data to GHG emissions. These emissions factors are either provided by the relevant government authority such as DEFRA and the EPA (DEFRA, 2009; EPA, 2022b), or obtained from life-cycle assessment (LCA) databases, including EcoInvent (Ecoinvent, 2022). The facility-level data are then aggregated to a national estimate grouped into source types and reported yearly by UNFCCC Annex I parties.

Dataset incompleteness can originate from either the facility-level calculation or the national-level aggregation. At a facility level, the three main causes of dataset incompleteness are: companies excluded from reporting regulations, where data are simply not calculated; lack of transparency, sometimes as a result of industrial secrecy; and non-compliant companies (de Souza Leao et al., 2020; Marlowe & Clarke, 2022). Furthermore, generic emissions factors used in calculations are non-specific to the production methods and supply chain routes used by a facility and can lead to significant uncertainty in emissions reports (Cullen et al., 2024). For companies, large uncertainties can lead to possibly erroneous net-zero calculations, wrong conclusions, and missed emissions. At a national-level, only 43 out of 198 countries are considered Annex I, accounting for approximately 20% of global emissions, while the remaining Non-Annex I countries have inconsistent reporting (UNFCCC, 2023). National reports are further limited by the standardized breakdown into categories used by the UNFCCC, which can result in difficulties for sector-specific mitigation strategies. For national and international governmental organizations, the accumulation of uncertainties in emissions data can lead to significant misunderstanding of progress toward mitigation targets and misinformed policy decisions (Cullen et al., 2024).

The most accurate solution to missing data would be to conduct, or require all emitters to conduct, thorough accounting and aggregation of all emissions for which they are responsible within the scope of the database. For country-level databases such as the UNFCCC, this is supposed to occur, but in reality, reporting is sparse (UNFCCC, 2023). For facility-level databases covering emissions from scopes 1, 2, and 3 of the GHG protocol (World Resources Institute, 2004), this would require full LCAs for all products from all emitters, which is not viable due to the expense and difficulty in gathering necessary data (Jusselme et al., 2018; Potrč Obrecht et al., 2020). A second option, which we will explore in this study, is to improve data coverage and quality by "gap-filling" emissions datasets from the data already available.

To improve data coverage, leveraging multiple databases, also known as data fusion, is increasingly popular in industrial ecology with examples including EXIOBASE, combining multi-regional input–output databases (Stadler et al., 2018), and the IEDC data repository (Pauliuk et al., 2019). These databases can provide complete coverage in some cases at a country-wide or industry-wide scale but lack the granularity of facility or company-specific data. High-resolution, company-specific data are available through some publicly accessible databases, including the EPA's GHGRP interface (EPA, 2022a), satellite-based ClimateTRACE (Climate TRACE, 2021), or spatially zoned but non-company-specific Vulcan (Gurney et al., 2020). These databases can have facility-level granularity but have incomplete coverage of facilities.

Gap-filling can either be mechanistic, by theoretically simulating processes to output predicted environmental impacts, or data-driven, by predicting impacts through proxy data or relationships between processes (Zargar et al., 2022). Data-driven methods are more scalable and applicable when guiding decisions at a company-wide or country-wide scale and will be the subject of this paper. In LCA, regression-based methods have been used to accurately estimate emission intensity factors for non-reporting coal facilities (Steinmann et al., 2014) and to reduce the number of required parameters for calculating environmental impacts (Pascual-González et al., 2015). More complex methods based on utilizing the similarities between facilities with accurate and inaccurate data have emerged to improve unit process impact assessments (Hou et al., 2018; Zhao et al., 2021). Machine learning (ML) has been implemented in optimizing specific parts of LCAs, but has so far been unsuccessful in capturing the complexity and types of inputs required for full GHG emissions estimates (Algren et al., 2021; Donati et al., 2022; Ghoroghi et al., 2022). In material flow analysis (MFA), Bayesian approaches are used to reduce uncertainty by systematically updating model predictions given new data inputs (Dong et al., 2023; Lupton & Allwood, 2018). This is an effective approach in a known framework with constructed priors but is not directly transferable to completing GHG emissions databases. More widely in industrial ecology, ML has been implemented to improve data collection through data mining (Arbabi et al., 2022; Vilaysouk et al., 2022) and to facilitate decision-making, even in complex circular economy scenarios (Alavi et al., 2021). Overall, ML has enabled more accurate impact evaluations than the "taking the average" approach often used in LCA and MFA (Ebrahimi et al., 2022; Larrea-Gallegos & Vázquez-Rowe, 2022).

The industrial ecology community and the wider carbon accounting community are calling on researchers to leverage data-driven innovations to make best possible use of all data available to inform climate change mitigation strategies (Donati et al., 2022; He et al., 2022; Marlowe & Clarke, 2022). For the specific problem of gap-filling GHG emissions datasets which this paper addresses, explicit calculation of emissions through traditional industrial ecology methods, including LCA, remains the most accurate solution. However, when real-world constraints render full calculations infeasible, data-driven techniques may help reduce large uncertainties by effectively exploiting the large quantity of data that are available. For the

TABLE 1 Levels of gap filling within greenhouse gas (GHG) emissions datasets.

Level	Gap composition	Set relationship	Inference type	Typical use case
1	Time step	$E_{test} \subseteq E_{train}$ $P_{test} \subseteq P_{train}$	Supervised learning	· Inconsistent data reporting from well-regulated facilities.
2	Emitter	$E_{test} \nsubseteq E_{train}$ $P_{test} \subseteq P_{train}$	1D transfer learning	· Non-reporting small facilities. · Poor transparency.
3	Emitter with unknowns	$E_{test} \nsubseteq E_{train}$ $P_{test} \nsubseteq P_{train}$	Multidimensional transfer learning	· UNFCCC gaps, especially Non-Annex-I countries. · Poorly mapped products.

E denotes the set of emitters and *P* denotes the set of properties of those emitters. Inference type refers to the machine learning literature and typical use cases are not exhaustive.

implementation of such methods, industrial ecologists must be aware of which modelling techniques can be used to tackle which problems. The rest of this paper seeks to address this gap as follows. Section 2 will introduce the types of gaps in emissions datasets and a series of increasingly complex models that can be applied to gap-filling problems. Section 3 will assess the performance of each model when applied to gap-filling problems across 3 GHG emission databases. Finally, section 4 will discuss the implications of the results to inform the community of the correct tools to explore for the problem at hand. An interactive guide is available at https://hackmd.io/@luke-scot/ML-for-GHG-database-completion. This guide can serve as a reference for the community and will provide a road map for future development of a multi-input consistent framework for inferring company-specific emissions estimations, inline with urgent needs for companies and governments in approaching net-zero targets.

2 | METHODS

Gap filling is based on using a subset of data to infer additional data where they are not available. This is analogous to inferring output data y_{test} given a set of input data X_{test}, when the model has been trained to infer y_{train} from X_{train}. In this section, we begin by defining the types of gaps in emissions databases, followed by a presentation of the three datasets we will use to quantify gap-filling performance of different methods. Individual classification techniques are then discussed, with extra focus given to graph-based models due to their novelty and tangibility for mapping industrial ecology data. Finally, we consider secondary model outputs that may help with future data collection.

2.1 | Gap definition

In Table 1, we present three types of gaps addressing increasingly challenging problems in GHG emissions datasets. These gaps are equivalent to different training and test set splits for classification. Level 1 considers an emitter with some reporting records but with missing time steps, such as a missing yearly report. Level 2 considers the situation of an emitter with no reports at any time step but with properties that are shared with other emitters in the database that do have available reports. Level 3 considers an emitter with no reports at any time step and with at least one property that is not shared with any other emitter in the database, for example, for a facility in a country with no reporting facilities at all.

2.2 | Datasets

In this study, the three gap-filling problems presented in Table 1 will be considered across the three datasets presented in Table 2. These datasets cover a range of typical properties of GHG emissions databases, including different resolutions and number of features. Where data sources include facilities with multiple co-products, allocation between them has already been conducted by the sources cited.

To apply a wide range of classification techniques and standardize the performance evaluation, this study discretizes emissions estimation into a four-class classification problem. Class boundaries are set according to the quantiles of emissions values in each dataset with labels as stated in Table 3.

2.3 | Classification techniques

To each of the gaps presented in Table 1, considered for each of the datasets presented in Table 2, we will apply a series of classification techniques and evaluate their performance. The key to understanding the abilities of different techniques on different problems is a set of consistent performance metrics. When applying this methodology to new problems, performance on known test data should always be evaluated before applying

TABLE 2 Properties of the three datasets used in this study.

Dataset	Spatial resolution	Temporal resolution	Features	Number of features
UNFCCC (UNFCCC, 2023)	Country	Yearly	Party (Country), Category	2
ClimateTRACE (ClimateTRACE, 2023)	Facility	Monthly	Country, Sector, Type, Capacity, Latitude, Longitude	6
Petrochemical (Cullen et al., 2024)	Facility	Yearly	Country, Product, Company, Route, Technology, Site, Plant#, Complex, Licensor, Start year, End year, Latitude, Longitude	13

See Supporting Information S1 section 1.1 for full descriptions of dataset features.

TABLE 3 Definition of emissions estimation as a four-class classification problem.

Label	Quantile	Class
0	$0 \rightarrow 0.25$	Low emissions
1	$0.25 \rightarrow 0.5$	Medium emissions
2	$0.5 \rightarrow 0.75$	High emissions
3	$0.75 \rightarrow 1$	Very-high emissions

the models to fill gaps. For each database, the data are split into a training set, which is used to build the model and a test set, which the model does not see during training and is used to evaluate the model's performance by comparing the model's prediction to the real test set data. For the level 1 problem, the timesteps for each facility are randomly allocated to training and test sets. For the level 2 problem, facilities are randomly allocated between sets, and for the level 3 problem, combinations of properties are randomly allocated between sets. Performance metrics must avoid bias from over-fitting to a single category, which would imply that the models are simply learning the most likely category overall rather than predicting categories based on individual data points. To this end, we will use average accuracy and *F*1 score as summary metrics defined in Equations (1) and (2), where *C* represents the total number of classes *c*, four in this study; and *TP*, *FP*, and *FN* represent the number of true positives, false positives, and false negatives, respectively. Taking into account the *F*1 score avoids biases from exclusively relying on average accuracy by also considering the number of false positive predictions. Although less intuitive than average accuracy, it quantifies the ratio between correct predictions and misclassification and is applied across all categories at once, known as the "micro" method:

$$\text{Accuracy} = \frac{1}{C}\sum_{c=0}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}; \tag{1}$$

$$F1\ \text{score} = \frac{\text{TP}}{\text{TP} + 0.5(\text{FP} + \text{FN})}. \tag{2}$$

The objective of this study is to understand which techniques can be implemented for particular gap-filling problems. To this end, models of increasing complexity will be applied to each level of problem for each dataset. The models used are assembled into five types as follows:

1. Interpolation (Süli & Mayers, 2003)—mean filling, polynomial fit. For level 1 gaps, interpolation is carried out across the time dimension. For level 2 and 3 gaps, the mean emissions value of entities with matching properties is used.
2. Shallow learning models (Pedregosa et al., 2011)—logistic regression, stochastic gradient descent (SGD), support vector classifier (SVC), passive aggressive classifier, naive Bayes, decision tree, k-nearest neighbors, and perceptron. These models take as input all properties of the predicted entity and the time step for which the prediction is to be made. They learn to classify entities by iteratively updating model weights based on current performance.
3. Ensemble models (Sagi & Rokach, 2018)—adaboost and random forest. These models amalgamate the learning capabilities of many shallow learning models to improve the performance.
4. Deep learning models (Goodfellow et al., 2016)—multilayer perceptron, deep fully connected network (DeepNet), long short-term memory network (LSTM), and residual network (ResNet). Functioning in the same manner as shallow learning models, these models are able to learn more complex relationships between input variables and outputs due to a very large number of model weights.

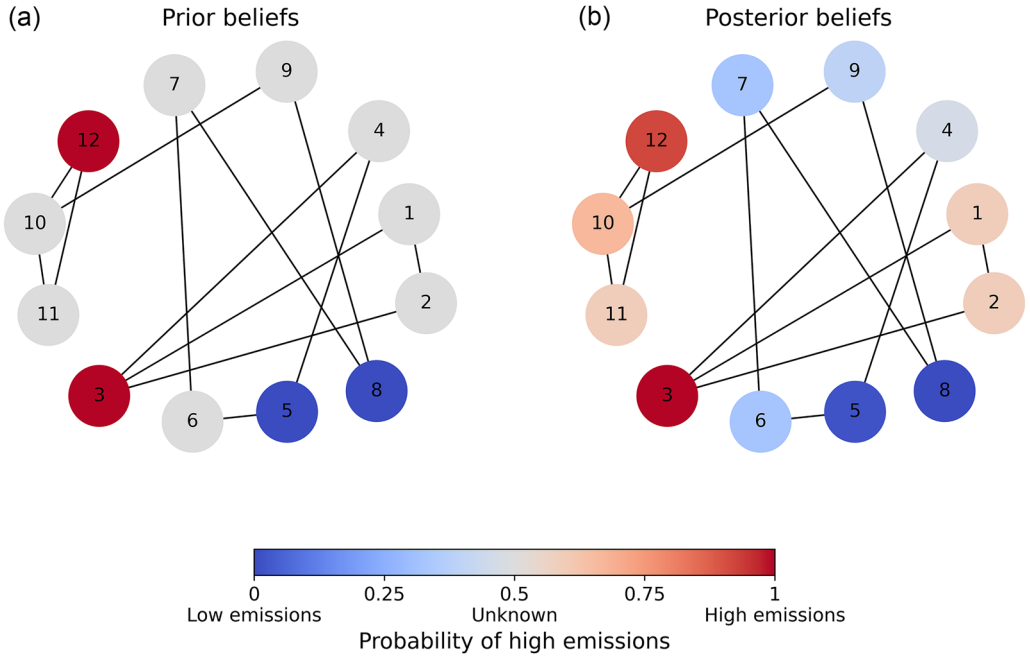

FIGURE 1 Belief propagation in graphs. (a) Nodes are embedded with input data values to form prior beliefs and linked by edges according to embedding similarities. (b) Beliefs are then propagated by message-passing along edges resulting in posterior beliefs for all nodes, including those without prior knowledge.

5. Graph representation learning models (Hamilton, 2020)—graph convolutional network (GCN) and graphSAGE network (SAGE). Above simple deep-learning models, these models explicitly assign a structure to the data by linking entities with the same properties with an edge as explained in the next subsection. For this study, we use two properties to construct edges to remain consistent across all three datasets: Country and Category/Sector/Product for UNFCCC, ClimateTRACE, and Petrochemical, respectively.

Model structure details can be found in Supporting Information S1 section 1.2 and implementation information can be found at https://hackmd.io/@luke-scot/ML-for-GHG-database-completion. Cross-validation is used to optimize hyperparameters and prevent overfitting in all groups except interpolation; details can be found in Supporting Information S1 section 1.3 and Supporting Information S2 section 1.4. For graph representation learning models, input datasets must be manipulated into a graph form. Graph representation can be particularly beneficial to industrial ecologists due to its flexibility in incorporating multiple data types and relationships between emitting entities. Therefore, we will lay out an example graph linking emitting facilities in some more detail to provide intuition to the community.

2.4 | Graph example

To simplify this example and provide intuition, we will consider a classification problem with just two categories: low emissions and high emissions. Figure 1 considers a set of 12 facilities. We know that facilities 5 and 8 are high emitters and facilities 3 and 12 are low emitters. To construct the a priori graph G, shown in Figure 1a, we consider each facility as an individual node $u \in V$ and draw edges $e \in E$ between facilities if they produce identical products or share owners. Equation (3) denotes the formal definition of this graph with a set of nodes V and edges E:

$$G = (V, E). \tag{3}$$

Given the a priori information, we would like to infer the category of the nodes for which we have no direct information. This can be done through belief propagation that is based on message-passing between nodes. Equation (4) from Hamilton (2020) explains message-passing where the weight h_u of each node u at time $k+1$ is calculated according to an update rule applied to the weight h_u at time k and an aggregation of the weights h_v of all nodes v within the neighborhood N of node u:

$$h_u^{(k+1)} = \text{UPDATE}^{(k)}(h_u(k), \text{AGGREGATE}^{(k)}(h_v^{(k)}, \forall v \in N(u)). \tag{4}$$

Applying one step of this algorithm, using the Dirichlet multinomial distribution-based Netconf framework as update and aggregation rules (Eswaran et al., 2017) yields a probability of belonging to the high-emitting category for each previously unknown node, as shown in Figure 1b.

Figure 1 demonstrates a simple case of inference through belief propagation, where nodes with no prior data are assigned a probability of emissions category given their relationship to other nodes. For example, nodes 6 and 7 are more likely to be low emitters than high emitters, and conversely, nodes 2 and 10 are more likely to be high emitters than low emitters. The limiting factor of this procedure is the a priori knowledge that shared product type and ownership positively correlate with shared emissions category. In real-world scenarios, far more links between facilities exist and can be incorporated into the graph, for example, geographical proximity, production output, suppliers, and age. Also, data can be leveraged from thousands rather than 12 facilities. This increase in scale allows us to overcome the limitation of needing to impose a priori knowledge, by employing learning algorithms that will learn how to propagate beliefs through experience. An added benefit of graph representations versus non-graph neural networks is improved explainability. Explainability not only allows for justification of estimates but also prioritization of future data gathering to which we will now turn our attention.

2.5 | Prioritizing data collection

Ranking the importance of individual input features in determining a model's prediction allows us to establish the most valuable features to collect during future data gathering. Shallow classifier feature importance is determined through the highest weighted factors in logistic regression and most weighted branches in decision trees. Ensemble importance is an aggregation of feature importance outputs of the simple models composing them. For deep models, explainability remains an active area of research, and in this study, we will use gradient-based attribution methods (Nielsen et al., 2022; Sundararajan et al., 2017) to rank the relative importance of features in determining outputs. Graph models benefit from additional explainability relative to deep models, and we will use the GNNExplainer method (Ying et al., 2019) in this study.

Beyond feature importance, graph-based models also allow for data collection prioritization based on node importance. Consider a node (e.g., facility or country) where data are secret and cannot be acquired, then node importance will determine which other nodes, where data may be more easily acquired, are the most valuable targets for supplementary information leading to an estimate of emissions at the secretive node. This could be a valuable tool for industrial ecology given the inaccessibility of many data points.

3 | RESULTS

This study has tested a range of classification methods for gap-filling GHG emissions databases. In this section, we will present the output of a range of models in predicting missing GHG emissions values across the three levels described in Section 2.1: (1) missing time steps, (2) missing emitters, and (3) missing emitters with a previously unseen property. The competence of models under each scenario will be evaluated first according to the average accuracy at a fixed train/test set split, second according to accuracy with different training set availability, and finally according to the range of output statistics models can provide to inform future data collection. Figure 2 displays the average accuracy for each model applied to each dataset and for each level of gap-filling, given a constant 70%–30% train/test set split.

In a simple level 1 gap-filling problem with missing years in the UNFCCC dataset, interpolation is the best solution with 97% average accuracy seen in the top left of Figure 2. In datasets with more input features, shallow learning algorithms based on decision trees and nearest neighbors can perform as well or better than simple interpolation at level 1. Deep-learning models can achieve up to ~70% accuracy but are not well suited to optimally predict values for a single emitter and thereby do not perform as well as some simpler methods.

For level 2 gap-filling with missing emitters, interpolation accuracy drops significantly and is no longer the best option for any dataset. Decision trees and the ensemble random forest method are viable solutions across all datasets with 60%–70% average accuracy. For the UNFCCC dataset with only two input features, deep learning is ineffective, but for datasets with more input features, deep learning and graph representation learning methods perform as well as the best shallow methods.

In the most complex level 3 gap-filling problem, interpolation and shallow models have poor accuracy. For the ClimateTRACE and petrochemical datasets, random forests, deep-learning, and graph-based models all perform well. Once again, a low number of input features is a strongly limiting factor for UNFCCC dataset prediction where deep learning is ineffective. In the UNFCCC case, random forest and decision tree classifiers obtain the best performance with 60% and 61% average accuracies.

Deep-learning and graph-based models maintain a steady average accuracy as gap-filling problems become more complex. It should also be noted that this study used uniform hyperparameter tuning across all models to maintain consistency, but it is likely that performance for individual neural network-based models could be increased with network structure adjustments and further parameter tuning. Considering simple gap-filling problems and datasets with fewer input features, simple models outperform deep learning in output average accuracy. As gap-filling problems become more complex and datasets have more input features from which models can learn, more complex deep-learning models output the best average accuracy scores. Figure 2 assumes the availability of 70% of the data for training (i.e., a gap consisting of 30% of the data), but in real-life gap-filling problems, data availability is variable. Figure 3 shows the evolution of average accuracy as data availability increases for three of the best-performing models from different groups.

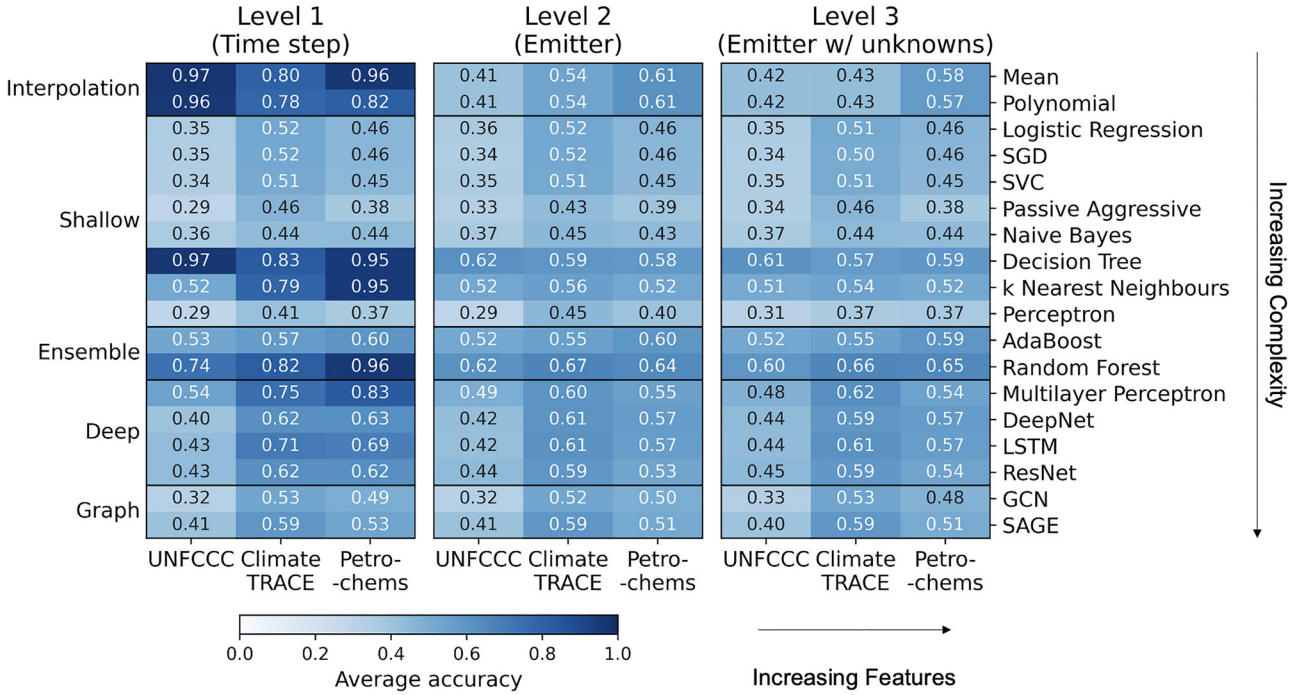

FIGURE 2 Average accuracy scores for emissions classification task across each level of gap filling for the three datasets considered. A train/test split of 70%/30% is used, and trainable models were run for a maximum of 100 epochs. Random guessing in a four-class classification problem would result in 0.25 average accuracy. See Supporting Information S1 section 1.2 for model details and section 2 for supplementary performance metrics including *F*1-score.

Figure 3 shows that for level 1 problems, each model reaches within 10% of their optimal accuracy with just 20% of the training data available. This implies that if inferring a yearly category for an emitter, among the four classes available in this study, using data points every 5 years is sufficient to reach within 10% accuracy of using data points every 2 years. The difference in these accuracies may increase with finer resolution classification or regression problems. For level 2 and 3 gap-filling problems with missing emitters, mean-filling performance plateaus with just 10% training data as it is unsuitable for these problems as noted in Figure 2. The accuracy of learning algorithms for levels 2 and 3, with random forest and multilayer perceptron shown here, is within 20% of the optimal accuracy with 20% of training set availability but continues to improve uniformly up to ~ 70% of training set availability, after which the accuracy does not vary more than 5%. The exception to this is the multilayer perceptron performance on the UNFCCC dataset, which is poor due to the low number of input features. From a practical perspective, it is unsurprising that a level 1 problem may achieve higher accuracy with less training data as the variation between emissions at a facility between different years is generally lower than the variation between different facilities.

Across all levels, classification models are able to be used effectively with just 20% of the training set data available. Further data gathering beyond 20% of the dataset is more worthwhile for level 2 and 3 problems with missing emitters than for level 1 problems with missing time steps. In practice, resources are limited and it is worthwhile prioritizing data collection toward the features that will be most effective in improving accuracy. An explanation of model choices through feature importance analysis quantifies which dataset features are the most important for the model's decision-making. Figure 4 shows an example of feature importance, which should then be used to prioritize future data collection.

Figure 4a–c shows that some features, including "Category" for the UNFCCC, "Capacity" for ClimateTRACE, and "Product" for petrochemicals are more valuable for data collection than other features and should be prioritized. Figure 4d presents that if data for emitter 0 are not accessible, the best locations to gain data that would improve our knowledge of emitter 0's emissions are emitters 1, 2, and 4; and information for emitter 5 would be more valuable than information from emitter 3 despite it being less directly linked. Basing additional data gathering on these outputs will accelerate the improvement in the accuracy of classifiers and could save time and effort by discounting the collection of features that are not useful for emissions estimates.

4 | DISCUSSION

Gaps in emissions reporting databases are a problem for mitigating climate change. Manual data completion is not scalable and can require numerous assumptions. Are learning-based classifiers able to provide an automated solution to this problem? This study shows that learning-based

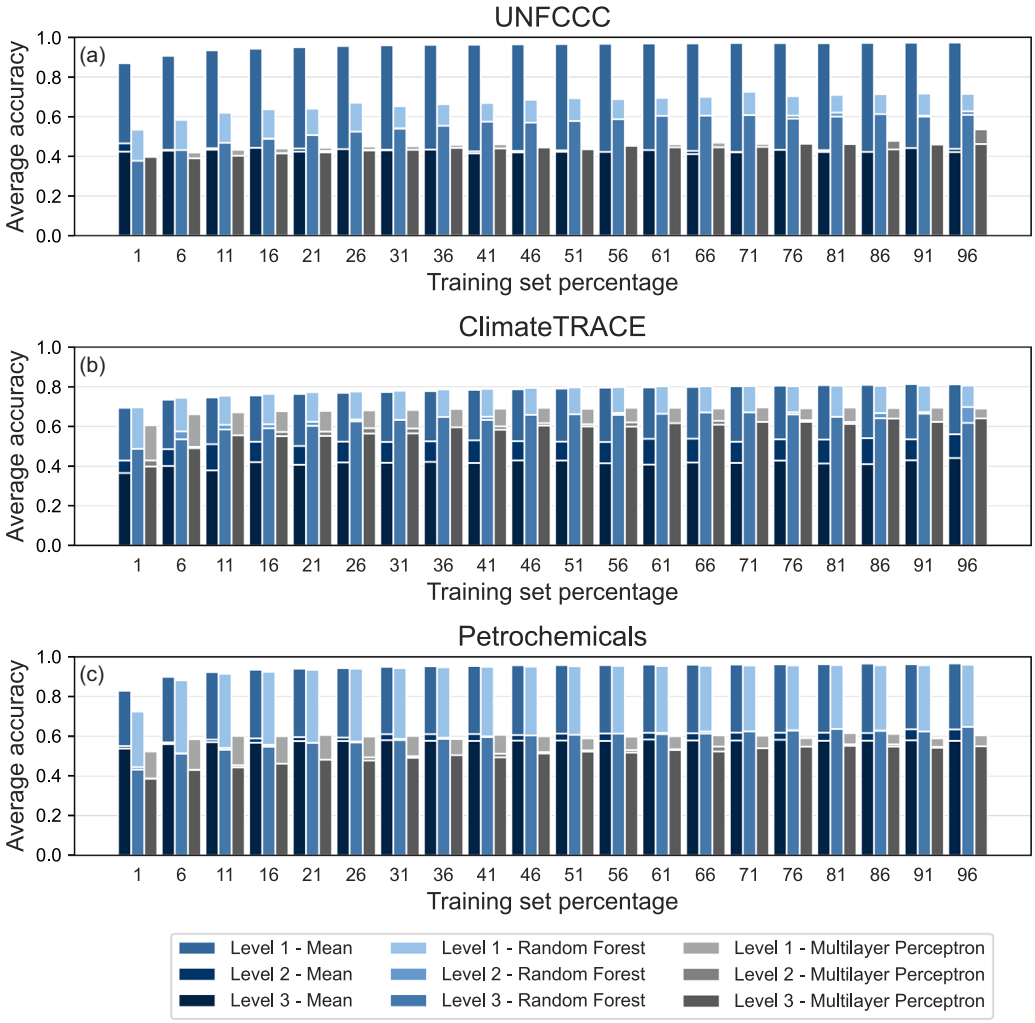

FIGURE 3 Average accuracy of three different models across the three gap-filling levels relative to training set size. Training set size is shown as a percentage of the total dataset for: (a) UNFCCC, (b) ClimateTRACE, and (c) Petrochemicals. Underlying data for Figure 3 are available in tabular form in Supporting Information S2 section 2.2.

classifiers can be used to effectively and scalably complete emissions datasets in some cases. However, different models are appropriate for different types of gap-filling problem and complex classifiers, including neural networks, are not the best solution in many cases. Industrial ecologists and emissions analysts should not rush to use machine learning if their problem is not suitable as an input for learning-based models. Figure 5 presents a decision tree that can be used as a rule of thumb by researchers when choosing appropriate models for emissions database gap-filling problems.

The UNFCCC dataset has only two features for each model to learn from. This is a strongly limiting factor for using complex models that are less accurate on the UNFCCC dataset than the ClimateTRACE and petrochemicals datasets with 6 and 13 features, respectively. The UNFCCC database could be associated with additional features supplementing those of "Category" and "Party," and subsequently input to a classification model that performs well with higher numbers of features. This presents an opportunity to accelerate the expansion of reliable UNFCCC database coverage and target the tightening of UNFCCC regulation to those parties and features that are the most informative for the overall dataset. This will be the subject of future work.

Data collection is the basis of all emissions reporting and complete data collection will continue to be more accurate than inference from incomplete data. However, data collection relies on the cooperation of the parties providing the data and a consistent effort to aggregate data. First, feature importance outputs from classifiers may allow for a targeted approach that could reduce the burden on organizations such as ClimateTRACE and other enterprises seeking carbon footprint transparency, as they could maintain data completeness without explicitly collecting all data points. Some data are more difficult to acquire than other data, and an evaluation of data accessibility would need to be added to feature importance outputs to target data collection in practice. Second, the ability to infer reliable data across unobserved facilities or parties may be a useful filter in detecting possible misreporting. If the predicted value is significantly different to the value reported, this could suggest some verification is required.

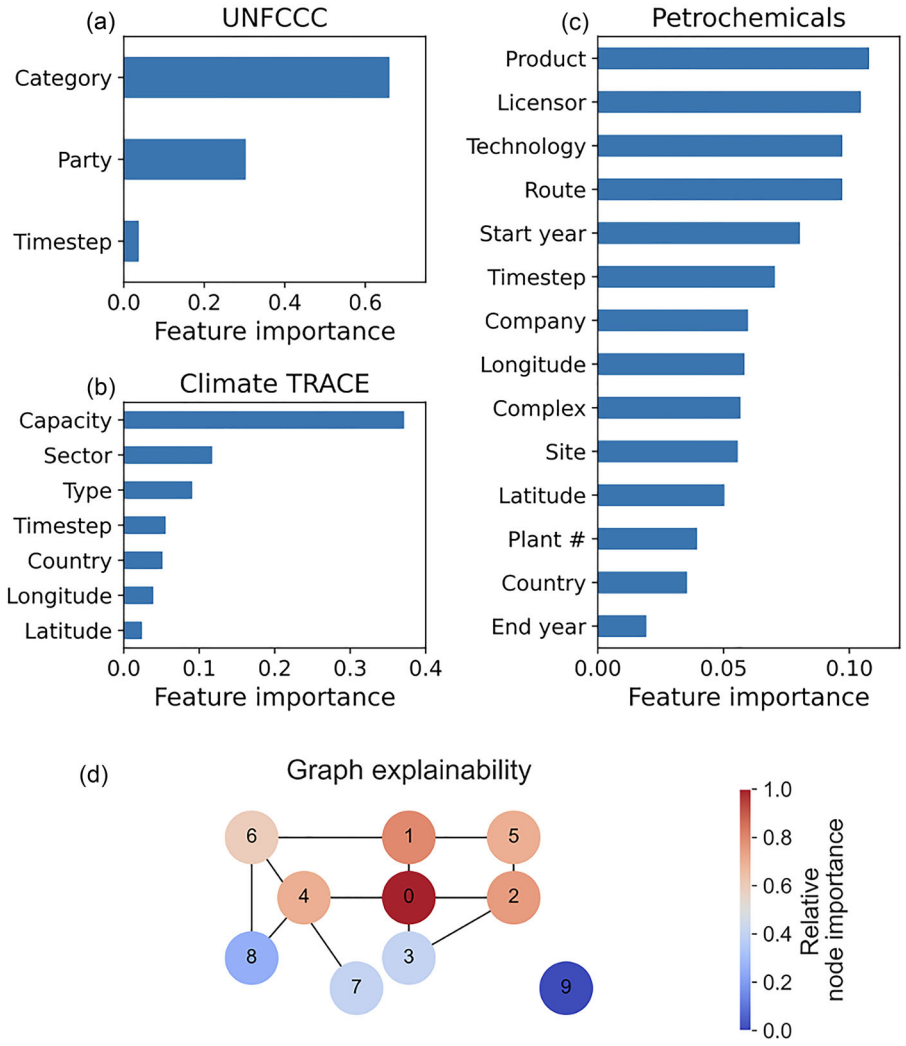

FIGURE 4 Model explainability representations. Relative feature importance for the decision tree classifier in the level 2 gap-filling problem is shown for each dataset: (a) UNFCCC, (b) ClimateTRACE, and (c) petrochemicals. The sum of all importance values equals 1 for each dataset. Plot (d) shows the output of a node importance analysis in predicting the class of node 0. Relative importance is calculated via the weight attributed to each node in updating the value of node 0 during the final step of the iterative graph learning model. Underlying data for Figure 3 are available in tabular form in Supporting Information S2 section 2.3.

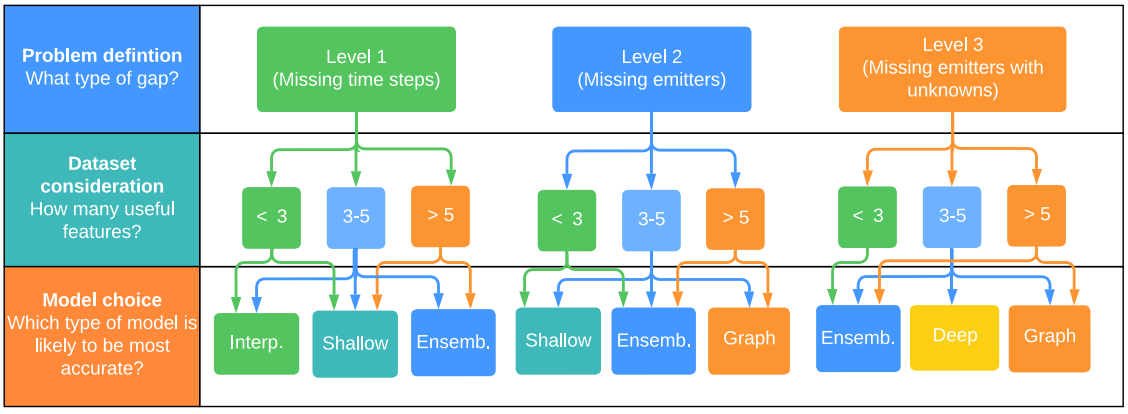

FIGURE 5 Decision tree for model selection during emissions database gap-filling problems. "Interp." signifies interpolation, and "Ensemb." signifies ensemble. The term "useful features" signifies features that will have significant feature importance during classification. Although feature importance cannot be determined a priori, some features are more clearly related to emissions than others. In this study following Figure 4, the number of useful inputs for each dataset are: 2 for UNFCCC, ~3 for ClimateTRACE, and ~6 for petrochemicals. For further guidance on model selection and implementation, visit https://hackmd.io/@luke-scot/ML-for-GHG-database-completion.

The ability to predict emissions categories without imposing any physical knowledge of the relationship between input features and output emissions permits easy incorporation of available data sources and avoids many of the biases involved in manual emissions estimates. In effect, any data source can be added to a learning-based model, and if the data source provides new information, the model predictions will improve, but even if the data are of no value, it will not affect model performance but simply be disregarded during prediction. In industrial ecology, this could open avenues for the incorporation of data from MFA, LCA, and input–output studies into a single framework for process emissions estimates. Graph-based frameworks may be particularly helpful for incorporating MFA and LCA data, which are easily translated to the graphical form. For example, in an LCA, processes can be allocated to nodes, and edges can represent the input and output of materials for each process forming a graph that can be coupled with other data or used in the same fashion as Figure 1 to estimate process emissions. Further afield, a multimodal framework could be a basis for the merging of "top-down" satellite measurements with "bottom-up" emissions reports, as sought out in the remote sensing community (ESA, 2021).

Poor performance in scenarios with insufficient input features or high levels of complexity shows the limitations of the techniques discussed in this paper and should be considered when applied to a new problem. Inference models can only infer from data that are previously available; therefore, predictions for exceptional cases, such as facilities with new technologies, will be inaccurate. Furthermore, results are based on a four-class classification, which is too imprecise for many applications, and uncertainty is not quantified. To address these limitations, our future research will aim to quantify uncertainty across multimodal inputs and output an uncertainty distribution of values instead of a discrete classification. Another limitation to complex network building is training time. Initial training time for all models in this study was less than 1 h on a standard issue 16GB RAM computer. However, different models incorporate new data in different ways, and standard neural networks require some level of re-training for each new data point, which may become impractical and expensive with a regularly updated system. Graph frameworks are able to locally incorporate new data without the need for re-training applied to a whole network and may therefore be a more viable solution for incrementally improving models based on incorporating any additional data across large multimodal networks.

In conclusion, machine learning methods can be used to automate the completion of GHG emissions databases when enough input features are available. Feature importance outputs can guide the targeting of future data collection and should be considered by those engaged in improving GHG emissions accountability. Graph-based frameworks could be particularly adept at handling multimodal and incrementally increasing data, which could be used at the intersection of the remote sensing and industrial ecology communities in future studies.

ACKNOWLEDGMENTS

This work was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks [grant number EP/S022961/1].

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The raw data from the UNFCCC and ClimateTRACE (ClimateTRACE, 2023; UNFCCC, 2023) that support the findings of this study are openly available from https://di.unfccc.int/flex_annex1 and https://climatetrace.org/downloads respectively. The processed versions used during this study are available via https://doi.org/10.5281/zenodo.8279939. Restrictions apply to the availability of the raw petrochemicals dataset that was used under licence from ICIS for this study (ICIS, 2021). It is available for purchase from https://www.icis.com/explore/services/supply-and-demand-database/. Aggregated global emissions from the dataset are available at https://9e4z.short.gy/c-thru-petrochemical-emissions , and an abstracted version of the processed data used for this study is available via https://doi.org/10.5281/zenodo.8279939.

All code used for this study is publicly available at https://github.com/luke-scot/ml-ghg-databases and https://doi.org/10.5281/zenodo.10463104.

ORCID

Luke Cullen https://orcid.org/0000-0002-6179-2430
Jonathan Cullen https://orcid.org/0000-0003-4347-5025

REFERENCES

Alavi, B., Tavana, M., & Mina, H. (2021). A dynamic decision support system for sustainable supplier selection in circular economy. *Sustainable Production and Consumption*, *27*, 905–920.

Algren, M., Fisher, W., & Landis, A. E. (2021). Machine learning in life cycle assessment. In *Data Science Applied to Sustainability Analysis* (pp. 167–190). Elsevier.

Arbabi, H., Lanau, M., Li, X., Meyers, G., Dai, M., Mayfield, M., & Densley Tingley, D. (2022). A scalable data collection, characterization, and accounting framework for urban material stocks. *Journal of Industrial Ecology*, *26*(1), 58–71.

Arnold, J., & Toledano, P. (2021, December 1). *Corporate net-zero pledges: The bad and the ugly*. Columbia Center on Sustainable Investment Staff Publications.

Christiansen, K. L., Hajdu, F., Mollaoglu, E. P., Andrews, A., Carton, W., & Fischer, K. (2023). "Our burgers eat carbon": Investigating the discourses of corporate net-zero commitments. *Environmental Science & Policy*, *142*, 79–88.

Climate TRACE. (2021). *Bringing radical transparency to global emissions*. https://www.climatetrace.org/

ClimateTRACE. (2023). *Manufacturing emissions data*. https://climatetrace.org/downloads

Cullen, L., Meng, F., Lupton, R., & Cullen, J. (2024). Reducing uncertainties in greenhouse gas emissions from chemical production. *Nature Chemical Engineering*, 1, 311-322.

de Souza Leao, E. B., do Nascimento, L. F. M., de Andrade, J. C. S., & de Oliveira, J. A. P. (2020). Carbon accounting approaches and reporting gaps in urban emissions: An analysis of the greenhouse gas inventories and climate action plans in Brazilian cities. *Journal of Cleaner Production*, *245*, 118930.

DEFRA. (2009). Guidance on how to measure and report your greenhouse gas emissions. *Department for Environment, Food and Rural Affairs*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/69282/pb13309-ghg-guidance-0909011.pdf

Donati, F., Dente, S. M., Li, C., Vilaysouk, X., Froemelt, A., Nishant, R., Liu, G., Tukker, A., & Hashimoto, S. (2022). The future of artificial intelligence in the context of industrial ecology. *Journal of Industrial Ecology*, *26*(4), 1175–1181.

Dong, J., Liao, J., Huan, X., & Cooper, D. (2023). Expert elicitation and data noise learning for material flow analysis using bayesian inference. *Journal of Industrial Ecology*, *27*, 1105–1122.

Ebrahimi, B., Rosado, L., & Wallbaum, H. (2022). Machine learning-based stocks and flows modeling of road infrastructure. *Journal of Industrial Ecology*, *26*(1), 44–57.

Ecoinvent. (2022). *Ecoinvent life cycle inventory database*. https://ecoinvent.org/the-ecoinvent-database/

EPA. (2022a). *Flight: Facility level information on greenhouse gases tool*. Environmental Protection Agency. https://ghgdata.epa.gov/ghgp/main.do

EPA. (2022b). *Rulemaking notices for GHG reporting. Greenhouse Gas Reporting Program (GHGRP)*. https://www.epa.gov/ghgreporting/rulemaking-notices-ghg-reporting

Erb, T., Perciasepe, B., Radulovic, V., & Niland, M. (2022). Corporate climate commitments: The trend towards net zero. In *Handbook of climate change mitigation and adaptation* (pp. 2985–3018). Springer.

ESA. (2021). Sentinel-1 SAR user guide. European Space Agency. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar

Eswaran, D., Günnemann, S., & Faloutsos, C. (2017). The power of certainty: A dirichlet-multinomial model for belief propagation. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 144–152). SIAM.

Ghoroghi, A., Rezgui, Y., Petri, I., & Beach, T. (2022). Advances in application of machine learning to life cycle assessment: A literature review. *The International Journal of Life Cycle Assessment*, *27*, 433–456.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gurney, K. R., Liang, J., Patarasuk, R., Song, Y., Huang, J., & Roest, G. (2020). The vulcan version 3.0 high-resolution fossil fuel CO_2 emissions for the United States. *Journal of Geophysical Research: Atmospheres*, *125*(19), e2020JD032974.

Hamilton, W. L. (2020). Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, *14*(3), 1–159.

He, R., Luo, L., Shamsuddin, A., & Tang, Q. (2022). Corporate carbon accounting: A literature review of carbon accounting research from the kyoto protocol to the paris agreement. *Accounting & Finance*, *62*(1), 261–298.

Hou, P., Cai, J., Qu, S., & Xu, M. (2018). Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environmental Science & Technology*, *52*(9), 5259–5267.

ICIS. (2021). *ICIS supply and demand data service*. https://www.icis.com/explore/services/analytics/supply-demand-data/

IPCC. (2021). *Climate change 2021: The physical science basis*. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, Eds.). Cambridge University Press.

Jusselme, T., Rey, E., & Andersen, M. (2018). An integrative approach for embodied energy: Towards an LCA-based data-driven design method. *Renewable and Sustainable Energy Reviews*, *88*, 123–132.

Larrea-Gallegos, G., & Vázquez-Rowe, I. (2022). Exploring machine learning techniques to predict deforestation to enhance the decision-making of road construction projects. *Journal of Industrial Ecology*, *26*(1), 225–239.

Lupton, R. C., & Allwood, J. M. (2018). Incremental material flow analysis with bayesian inference. *Journal of Industrial Ecology*, *22*(6), 1352–1364.

Marlowe, J., & Clarke, A. (2022). Carbon accounting: A systematic literature review and directions for future research. *Green Finance*, *4*(1), 71–87.

Nielsen, I. E., Dera, D., Rasool, G., Ramachandran, R. P., & Bouaynaya, N. C. (2022). Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, *39*(4), 73–84.

Pascual-González, J., Pozo, C., Guillén-Gosálbez, G., & Jiménez-Esteller, L. (2015). Combined use of MILP and multi-linear regression to simplify LCA studies. *Computers & Chemical Engineering*, *82*, 34–43.

Pauliuk, S., Heeren, N., Hasan, M. M., & Müller, D. B. (2019). A general data model for socioeconomic metabolism and its implementation in an industrial ecology data commons prototype. *Journal of Industrial Ecology*, *23*(5), 1016–1027.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Potrč Obrecht, T., Röck, M., Hoxha, E., & Passer, A. (2020). Bim and lca integration: A systematic literature review. *Sustainability*, *12*(14), 5534.

Rogelj, J., Den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., Schaeffer, R., Sha, F., Riahi, K., & Meinshausen, M. (2016). Paris agreement climate proposals need a boost to keep warming well below 2○ C. *Nature*, *534*(7609), 631–639.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Stadler, K., Wood, R., Bulavskaya, T., Södersten, C.-J., Simas, M., Schmidt, S., Usubiaga, A., Acosta-Fernández, J., Kuenen, J., & Bruckner, M. (2018). Exiobase 3: Developing a time series of detailed environmentally extended multi-regional input-output tables. *Journal of Industrial Ecology*, *22*(3), 502–515.

Steinmann, Z. J., Venkatesh, A., Hauck, M., Schipper, A. M., Karuppiah, R., Laurenzi, I. J., & Huijbregts, M. A. (2014). How to address data gaps in life cycle inventories: A case study on estimating co2 emissions from coal-fired electricity plants on a global scale. *Environmental Science & Technology*, *48*(9), 5282–5289.

Süli, E., & Mayers, D. F. (2003). *An introduction to numerical analysis*. Cambridge University Press.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, (pp. 3319–3328). PMLR.

UNFCCC. (2023). *Greenhouse gas inventory data—Flexible queries annex I parties*. https://di.unfccc.int/flex_annex1

Vilaysouk, X., Saypadith, S., & Hashimoto, S. (2022). Semisupervised machine learning classification framework for material intensity parameters of residential buildings. *Journal of Industrial Ecology*, *26*(1), 72–87.

World Resources Institute. (2004). *A corporate accounting and reporting standard. The greenhouse gas protocol*. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, *32*, 9240–9251.

Zargar, S., Yao, Y., & Tu, Q. (2022). A review of inventory modeling methods for missing data in life cycle assessment. *Journal of Industrial Ecology*, *26*(5), 1676–1689.

Zhao, B., Shuai, C., Hou, P., Qu, S., & Xu, M. (2021). Estimation of unit process data for life cycle assessment using a decision tree-based approach. *Environmental Science & Technology*, *55*(12), 8439–8446.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cullen, L., Marinoni, A., & Cullen, J. (2024). Machine learning for gap-filling in greenhouse gas emissions databases. *Journal of Industrial Ecology*, *28*, 636–647. https://doi.org/10.1111/jiec.13507