Journal of Cleaner Production 332 (2022) 130053

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Optimized feed-forward neural networks to address CO_2-equivalent emissions data gaps – Application to emissions prediction for unit processes of fuel life cycle inventories for Canadian provinces

Sayyed Ahmad Khadem [a,b,*], Farid Bensebaa [a], Nathan Pelletier [b]

[a] Energy, Mining, and Environment, National Research Council Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
[b] Irving K. Barber Faculty of Science, The University of British Columbia, 3247, University Way, Kelowna, BC V1V 1V7, Canada

ARTICLE INFO

Handling Editor: Zhen Leng

Keywords:
Life cycle inventory
Life cycle assessment
Greenhouse gases
Machine learning
Genetic algorithm (GA)
Data gaps

ABSTRACT

Life cycle inventory (LCI) data gaps present a crucial challenge in research and development using life cycle assessment (LCA). The present study addresses this challenge by using optimized Artificial Neural Networks (ANNs). ANN's adjustable parameters (i.e. topology and hyperparameters) significantly affect prediction capability. Thus, the optimality of those parameters is of paramount importance. However, performing an optimization on the entire ANN's adjustable parameters is computationally demanding, which may result in an intractable problem. To tackle this challenge, the present study proposes a hybrid approach using heuristics and Genetic Algorithm to systematically design optimal ANNs in a tractable fashion. The proposed approach is then assessed through the prediction of data gaps in the Canadian fuel LCI database, GHGenius (an open-access tool for modeling Canadian fuel pathways). This is achieved using an automation workflow that extracts LCI data from GHGenius at the fuel/province/unit process level. It is found that the resulting optimal ANNs not only are accurate in predicting data to fill CO_2-eq emissions data gaps for unit processes present in the Canadian fuel life cycles but also possess shallower hidden topologies. These salient results are ascribed to the impacts of the input layer on the optimal network. In particular, input layers comprised of categorical and numerical features lead to enhancement of network prediction and/or shallower hidden topology for optimal ANNs. Taken altogether, the present study proposes, implements, and validates a systematic framework to tractably design optimal ANNs for predicting data to fill data gaps in LCI databases. In addition, this study quantifies average contributions of each unit process present in the Canadian fuel life cycles, including fossil-based and renewable pathways, to total CO_2-eq emissions, which is of particular interest in cut-off analysis.

1. Introduction

Life cycle assessment (LCA) is an ISO-standardized method that can be used to determine the amount of greenhouse gases (GHGs) emitted (along with other emissions and related impacts) throughout a product's life cycle (Jolliet et al., 2015). Fuel consumption is ubiquitous in the life cycles of most products, in particular with respect to transportation. According to the International Energy Agency, nearly a quarter of global CO_2 emissions is attributed to the transportation sector (IEA, 2014). Hence, detailed understanding and quantification of factors contributing to fuel life cycle GHG emissions are important and have attracted interest from numerous research organizations (Marais et al., 2019; McKechnie et al., 2015; Sleep et al., 2020) and government agencies (GHGenius, GREET; LCAcommons).

To date, three well-established open-access life cycle-based tools have been developed to assess the environmental impacts of fuels; namely, BioGrace, GREET, and GHGenius. BioGrace is a spreadsheet model used to determine the life cycle GHG emissions associated with biofuels. This model has been maintained and updated by the Institute for Energy and Environmental Research in Germany (Biograce). GREET is a life cycle-based tool developed by Argonne National Laboratory in the United States (GREET). GHGenius is a tool developed to estimate the carbon intensity of fuels for Canadian provinces. It is an excel-based spreadsheet tool developed by (S&T)^2 Consultants for Natural Resources Canada (GHGenius).

Several commercial LCA software tools have also been developed,

* Corresponding author. Energy, Mining, and Environment, National Research Council Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada.
E-mail addresses: sayyed.khadem@ubc.ca, sayyedahmad.khadem@nrc-cnrc.gc.ca (S.A. Khadem).

https://doi.org/10.1016/j.jclepro.2021.130053
Received 17 July 2021; Received in revised form 7 November 2021; Accepted 9 December 2021
Available online 10 December 2021

such as SimaPro. In contrast to the open-access tools described above, commercial tools can be used to build unique new life cycle inventories in order to model specific supply chains. Building such inventories requires time, expertise, and access to third-party life cycle inventory (LCI) databases to characterize supply chain activities in specific sectors. In response to increasing demand for such data resources, several research organizations have already undertaken to build and host databases, hence facilitating quality and timely LCA studies. With respect to data for fuel life cycles, for example, the US Federal LCA commons (LCA-commons) hosts sets of fuel LCI data from the National Renewable Energy Laboratory/USLCI and the University of Washington Biofuels and Bioproducts Laboratory.

Despite such efforts to build exhaustive LCI databases including LCI data pertinent to fuels (Biograce; GHGenius, GREET; LCAcommons), the lack of representative LCI data (i.e. "data gaps") for many fuels/contexts remains a common challenge (Subramanian and Golden, 2016; Turner et al., 2020). Such gaps may refer to entire LCI data sets for given processes, or to gaps with respect to data for particular input/output data within a given data set (Hou et al., 2018; Zhao et al., 2021). Limited interoperability between data sets from different databases also contributes to the ongoing gaps in reported LCA studies (Fritter et al., 2020; Kneifel et al., 2018; Turner et al., 2020). Furthermore, the mapping between input and output flows in LCI data can be nonlinear and complex, making the resolution of LCI data gaps difficult (Zhao et al., 2021). Given the growing importance of having access to quality LCI data along with current challenges in the estimation of missing data, the development of a systematic framework to use known data to accurately fill data gaps merits consideration.

Past studies have shown that supervised machine learning techniques are capable of addressing this challenge in general (Algren et al., 2021; Dawood et al., 2021; Zhao et al., 2021). However, further studies are required to elucidate the full potential of machine learning techniques in this context (Algren et al., 2021). Feedforward Neural Network (FNN) models, which are a data-driven supervised machine learning technique, have been shown to perform well even for complex systems in which the input-output mapping is highly nonlinear (Algren et al., 2021). FNN models perform well with large-scale data sets. Moreover, FNNs are accurate estimators provided that there is a sufficiently large set of known data and the FNN model is properly designed (Hou et al., 2020; Khadem and Rey, 2021; Song et al., 2017). These capabilities suggest the potential for using FNNs to fill LCI data gaps.

Currently, there is no consensus about best practices for FNN design. Some heuristic-based rules of thumb have been proposed for case-specific problems (Al Imran et al., 2018; Ibnu et al., 2019) or particular design aspects such as backpropagation optimizers and activation functions (Goodfellow et al., 2016). However, to the best of the authors' knowledge, few studies (Hou et al., 2020; Song et al., 2017) discuss FNN design for estimation of LCI data. Owing to the fact that there are several parameters in FNN design requiring optimization, prior works made assumptions in order to simplify and reduce the intrinsic computational complexity of the FNN design. One such simplification made in previous studies (Hou et al., 2020) is to consider few pre-specified values for the hidden topology's parameters (e.g. the number of hidden layers and/or the number of neurons per hidden layer), thereby limiting the search domain and, in turn, expediting the optimization of the hidden topology. Another assumption considered previously (Song et al., 2017) is to evaluate and compare all permutations of the number of hidden layers and the number of neurons per hidden layer to find the optimal design. This approach is practical for a small search domain only, as it is time-consuming (Song et al., 2017). This is because, as the search domain increases, the required computing time grows exponentially, thus hindering FNN optimal design (Ibnu et al., 2019). There are hence substantial challenges to optimally designing FNN models to predict LCI data gaps.

The present study focuses on the optimal design of FNN models to estimate CO_2-eq emissions data to fill data gaps for unit processes in fuel life cycles in Canadian provinces using publicly available data. Specifically, the above-mentioned assumptions related to ANN design are lifted, meaning that the number of hidden layers and the number of neurons per hidden layer are simultaneously optimized by an evolutionary GA optimizer in reasonably wide search domains. Furthermore, the impacts of the network's input layer, known as attributes or features, are taken into account; therefore, this study investigates optimality of the entire network topology comprising input layer and hidden topology. Regarding hyperparameters (e.g. activation function, loss function, optimizer algorithm for training, and more), heuristic approaches are employed in order to not involve hyperparameters during optimization of the ANN's parameters, making the optimization more tractable. Furthermore, the optimality of the heuristics used in ANN design is then validated. In addition to the objectives concerning optimal ANN design, unit processes existing in Canadian fuel life cycles are ranked in terms of their contributions to net emissions, which can facilitate cut-off analysis in LCA studies. All data used in this study are extracted from GHGenius through an automation workflow.

The study is organized as follows. In Section 2 (METHODS AND MATERIALS), the methodologies employed in this study are elaborated in detail. Section 3 (RESULTS) and 4 (DISCUSSION) describe and discuss the numerical results. Section 5 (CONCLUSIONS AND RECOMMENDATIONS) summarizes the novelty achieved in this work along with providing recommendations for future studies.

2. Methods and materials

This section presents our approach to extract LCI data from GHGenius (section 2.1), descriptions of the FNN model used (section 2.2), the two scenarios for arranging the attributes (section 2.3), the proposed hybrid strategy to optimally design FNNs for predicting data gaps (section 2.4), and the overall organization of the present study (section 2.5). It should also be noted that even though the present study focuses on Canadian fuel datasets, the methodology described herein could be easily applied to other datasets.

2.1. Life cycle inventory database development for fuels

The life cycle GHG emissions data reported in Version 5.01 of the GHGenius model is used in this work. GHGenius 5.01 includes 11 unit processes: fuel dispensing (i), fuel distribution and storage (ii), fuel production (iii), feedstock transmission (iv), feedstock recovery (v), feedstock upgrading (vi), land-use changes and cultivation (vii), fertilizer manufacture (viii), gas leaks and flares (ix), CO_2 and H_2S removed from natural gas (NG) (x), and displaced emissions from co-products (xi).

To estimate the Global Warming Potential of each unit process as CO_2 equivalent (CO_2-eq) emissions, characterization factors are used for each greenhouse gas (Jolliet et al., 2015). Hence, eleven CO_2-eq emissions values corresponding to the 11 unit processes are obtained for each fuel life cycle. It should be further noted that "unit process" and "contributor" are interchangeably used as each unit process is a contributor to the total CO_2-eq emission.

The amount of CO_2-eq emissions for a given unit process depends on several factors. GHGenius 5.01 provides location-specific pre-defined values for the parameters to estimate CO_2-eq emissions for each fuel pathway. Hence, location and fuel name are considered as key factors and rely on assumptions applied in GHGenius 5.01 for other contributing factors. CO_2-eq emissions are extracted using bidirectional communications based on the "Component Object Model (COM)" protocol between Python and Excel (see Supplementary Fig. S1 for details). By using the automation algorithm depicted in Supplementary Fig. S1, emissions per functional unit of energy delivered to the end-user (in $KgCO_2$-eq/GJ) were extracted from GHGenius 5.01 for each of 7 Canada's provinces and 131 fuel pathways. In total, emissions are estimated for 7×131 or 917 fuel life cycles. Since each fuel life cycle is modeled

based on eleven unit processes, 11 × 917 or 10,087 discrete CO_2-eq emissions data points were extracted in total. In other words, emission values were extracted per location, fuel, and unit process. Below, $c_{L,\ F,\ U}$ is used to represent emission values where L, F, and U are location, fuel, and unit processes, respectively. Readers are referred to Supplementary Tables S1 and S2 for the complete lists of locations, fuels, and unit processes considered in this study. Furthermore, Supplementary File S1 contains a sample of the extracted data. Note that for simplicity and without loss of generality, the year was set to 2021.

2.2. MISO-FNN

Fig. 1 shows a representative topology of a Multiple-Input Single-Output (MISO) Feedforward Neural Network (FNN) model as used in this study (hereafter MISO-FNN). Each MISO-FNN possesses N_i inputs, which are equal to the number of attributes (see section 2.3 for more details), and one output. In addition, the architecture of hidden layers can generally be described by a vector showing the number of neurons in the hidden layers, which reads

$$\mathbf{H} = [N_{hl=1},\ N_{hl=2},\ \ldots,\ N_{hl=L}] \tag{1}$$

where $N_{hl=i}$ stands for the number of neurons in the ith hidden layer, therefore, the length of vector **H** also shows the number of hidden layers (see Fig. 1). For convenience, we refer to the topology of hidden layers, **H**, as hidden topology. **H*** is the optimal hidden typology.

Data are randomly split into three categories: training (60%), validation (20%), and testing (20%). Note that although the distribution percentages are flexible, the ratios used here correspond to common practice in the literature (for example, see (Hou et al., 2020; Sun et al., 2020)). Given that there are 917 samples for each unit process (see section 2.1), training, validation, and test sets contain 550, 183, and 184 samples respectively for each MISO-FNN. The training set is employed to train the MISO-FNNs through which the model parameters (i.e. weights and biases) are optimized. In the present study, the training of MISO-FNNs is performed in Keras library (Chollet, 2015). Although the validation set is not used during the training phase, validation errors are monitored to avoid overfitting. Specifically, the early stopping heuristic is applied based on cross validation, meaning that the training process is terminated before the maximum epoch is reached if the validation error increases for a certain number of epochs (Fiszelew et al., 2007). Here, the training is automatically stopped if the validation error increases in 25 epochs. In addition, the validation error is used as a fitness function that is required to be minimized using GA. As described in sections 2.4.1 and 3.2, this helps to obtain the optimal or near-optimal hidden topology. The testing set was not incorporated in the network training and the optimization of the hidden topology. Since the test set is not used in network training, nor in finding optimal hidden topology, test sets are called "unseen data". Therefore, the test error is considered as a yardstick to evaluate the generality of the trained network. Note that the Root Mean Square Error defined by $\mathrm{RMSE} = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}}$ is used as a standard measure of error throughout this study.

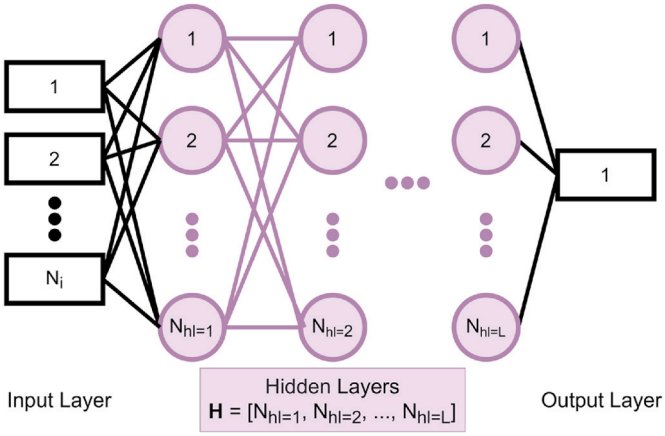

Fig. 1. A generic MISO-FNN topology used in this study.

2.3. Attributes scenarios

It is assumed that data gaps may originate from each unit process. Available data, including emissions from other unit processes, are then used to fill the data gaps. As shown in Fig. 2, two distinct scenarios are identified to arrange the attributes to feed the MISO-FNNs.

- **Scenario I.** As discussed above, there are 11 unit processes for each fuel life cycle. The CO_2-eq emission for a unit process is considered as the neural net output (labeled as "Target Process" in Fig. 2) and the CO_2-eq emissions of the ten remaining unit processes are then fed to the neural net as inputs (labeled as "Input Processes" in Fig. 2). The underlying idea behind Scenario I is that the CO_2-eq emission of a unit process depends on several independent variables, as described in equation (2)

$$y_k = f_k(\mathbf{x}),\ k \in \text{unit processes} \tag{2}$$

  where **x**, y_k, and f_k refer to a vector of independent variables, the amount of CO_2-eq emissions corresponding to the unit process k, and a nonlinear mapping function between independent and dependent variables, respectively. Although the CO_2-eq emission amounts are fed to the network, the network can implicitly benefit from independent variables. Indeed, as shown in equation (2), the CO_2-eq emissions fed to the network are also affected, in theory, by the independent variables. Furthermore, Scenario I presents an advantage as the independent variables are not required to be fed to the network explicitly. Thus, Scenario I does not suffer from uncertainty nor a lack of independent variables.
- **Scenario II.** In each fuel life cycle, location and fuel are readily accessible information. In Scenario II, these two pieces of information, locations and fuels, are augmented with Scenario I. In other words, in Scenario I, network inputs are exclusively based on dependent variables (CO_2-eq emissions). In Scenario II, both dependent variables (CO_2-eq emissions) and independent variables (locations and fuels) are incorporated. Owing to the fact that locations and fuels possess non-numeric values, they must be converted to numerical values so that they become suitable as inputs for the MISO-FNN model. The locations and fuels attributes are essentially nominal variables, meaning that there is no intrinsic order in these categories. The approach through which nominal variables are appropriately converted to numerical ones is via the *one-hot encoder*. Through this conversion, each nominal value is considered as a single attribute and, thereafter, the active and inactive attributes are indicated by 1 and 0, respectively. Fig. 2 depicts a schematic representation of the dataset. Additionally, in the one-hot encoder approach for nominal variables, one value in each categorical attribute can be arbitrarily ignored to avoid redundancy in the input layer. For example, there are 7 provinces in the nominal attribute "location". If Alberta (AB) is assumed to be excluded, then the records with 6 zeros in location attributes represent AB.

2.4. Optimal network exploration

In general, MISO-FNNs are capable of estimating any nonlinear multivariable function provided that the network topology and the hyperparameters are properly tuned. To achieve the best performance, both network topology and hyperparameters should be optimized. The MISO-FNN topology is defined by N_i and **H**, and the network hyperparameters are parameters associated with learning such as learning

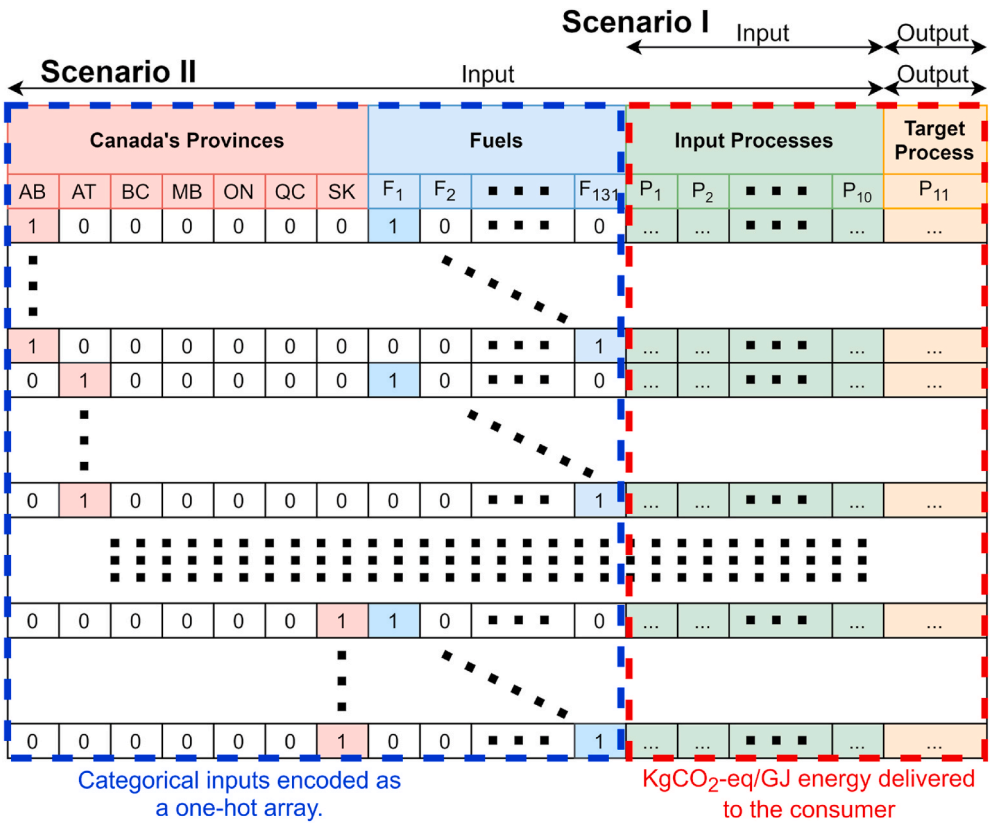

Fig. 2. General data structure showing two possible scenarios for the networks' input layer. See Supplementary Table S1 for abbreviations used for Canadian provinces.

rate, epoch, batch size, activation function types in hidden neurons, optimization methods for obtaining weights and biases (known as model parameters), etc. In practice, simultaneous optimization of both network topology and hyperparameters require significant computational resources (e.g. powerful hardware requirements and long running time) when using data sets of moderate to large size. Besides computation requirements, training a given data set is often challenging. As a consequence, trade-offs are used to address the intractability of finding the optimal MISO-FNN. The trade-off approach is described in detail in sections 2.4.1 and 2.4.2.

2.4.1. *Optimal MISO-FNN topology*

The optimal network topology is achieved by finding the optimal attributes scenario and the optimal hidden topology **H***. For the former, the search domain is small as there are only two scenarios, hence we rely on the grid search approach. This means that we separately evaluate the performances of each scenario discussed in Section 2.3, and thereafter the best scenario is identified. As detailed in Section 3.2, for each

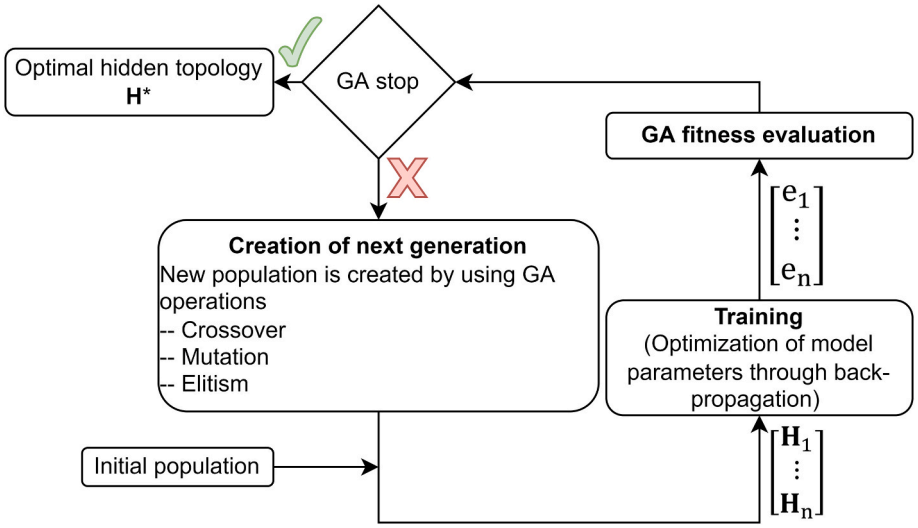

Fig. 3. The workflow for optimization of the hidden topology, **H**. As an initial population, a number of random **H** is generated. Each individual, **H**, is trained and results in a validation error. Then, in GA fitness evaluation, individuals are ranked according to validation errors. If any GA stopping criteria (e.g. the maximum number of generations) is not satisfied, based on the current generation and their fitness (i.e. validation error), the next generation is then created using GA operators.

attributes scenario, the optimal hidden topology is obtained for each unit process (i.e. contributor), and then the attributes impacts are compared. Regarding hidden topology optimization, the search domain is wide, often leading to the failure of approaches such as random walk and grid search due to exponential time complexity (Hou et al., 2020; Ibnu et al., 2019). However, the GA approach has demonstrated successful performance in finding the hidden topology (Hou et al., 2020). In the present work, the GA is thus used to simultaneously find both the number of hidden layers and the number of neurons per hidden layer in wide search domains. This enables finding the optimal hidden topology, **H***, hence overcoming assumptions made in previous studies (Hou et al., 2020; Song et al., 2017). The general schema for finding **H*** using GA is shown in Fig. 3.

Using the GA approach to obtain the optimal hidden topology gives rise to two main challenges, which are explained and addressed in detail as follows.

- **The varying length of H through evolutionary optimization.** Following the GA terminology, for a given generation, **H** and **H*** are the individual and the best individual in the population, respectively. The first challenge is that the length of individuals **H** (i.e. the number of hidden layers) can vary through evolution. We found that this challenge has been well-addressed in (Wirsansky, 2020). On this basis, a constant length for **H** is considered. Specifically, **H** is assumed to have 5 elements, showing the upper limit for the length of **H.** Additionally, zero and negative values are allowed in the GA search space except for the first hidden layer. However, zero or negative values are used as an indicator that the rest of the layers are not added. For instance, $\mathbf{H} = [5 \quad 4 \quad 3 \quad -1 \quad 10]$ represents a three-hidden-layer network with 5, 4, and 3 neurons, respectively; thus, the first presence of zero or negative indicates where the hidden layers are terminated.
- **Uncertainty of fitness function.** From a mathematical standpoint, the training of a neural network is a minimization problem, in which the loss function is high-dimensional and non-convex in general (Li et al., 2017). Even though a robust algorithm for finding the global extrema of non-convex functions has not been found to date, gradient methods are still practical in finding local extrema depending on the initialization of model parameters (Li et al., 2017). Indeed, the model parameters obtained through the training stage and the resulting validation errors (i.e. fitness function) are often sensitive to the initialization of model parameters. In other words, the training of a neural network is often a non-convex optimization; hence, there can likely be several local minima in the loss function with respect to the model parameters and, in consequence, each local minimum could represent a trained network. The less the local minimum is, the more accurate the trained network could be. To date, the available optimizers can succeed in capturing one of these local minima depending on two prime factors: the optimizer algorithm and the initial model parameters. Given that the optimizer algorithm and its adjustment parameters are set properly (see Table 1) in this study, exploration of the local minimum (i.e. near-optimal model parameters, which represents a trained network) can depend on the initialization of model parameters. Supplementary Note S1 provides an example showing the initialization-dependency of the training stage.

Table 1
The optimal hyperparameters used in the present study.

Hyperparameter	Near-optimal values/method
activation function	ReLU
optimizer	adam
loss function	MSE
learning rate	0.001
maximum epoch	750
batch size	550

The above-mentioned numerical uncertainty is unavoidable and may lead to confusion in the GA decision-making approach because GA produces the next generation, i.e. the next guesses of optimal hidden topology (**H***), based on the individuals' fitness (i.e. validation errors) in the current generation; therefore, it is most likely that the uncertainty related to validation errors adversely affects the effectiveness of the next generation. This challenge can also be appreciated from Fig. 3 where the "GA Fitness Evaluation" and "Creation of Next Generation" stages are performed after the training stage. In partial summary, to obtain the optimal hidden topology, **H***, there is the need for an iterative GA optimization as shown in Fig. 3. In each iteration, there can also be uncertainty in the evaluation of the fitness function, which is the validation error. The origin of this uncertainty stems from the fact that training itself requires another iterative optimization, which can depend on the initialization of model parameters due to the non-convex nature of the loss function with respect to the model parameters. Therefore, the uncertainty can adversely impact GA performance.

To address this challenge, in the training stage, for each individual (i.e. hidden topology) proposed by GA, the model parameters are randomly initialized with different sets of values; thereafter, the performance of each resulting network is compared and the most accurate network is selected. This multi-initialization approach significantly improves the probability of finding a more accurate network for any given hidden topology. Finally, each individual (i.e. hidden topology) and the most accurate network are reported to the GA for the creation of the next generation through applying GA operations (e.g. mutation, cross-over, elitism) on the current generation. In the present study, the number of initializations per individual in the training stage was chosen as 15 unless another number is specifically mentioned. In short, the multi-initialization method allows finding the optimal or near-optimal trained networks at expense of computational costs.

Regarding the implementation of GA, we used a package developed by one of the authors (Khadem et al., 2014) with the following assumptions: GA operators are mutation (0.1), crossover (0.8), and migration (0.1); the search domain considered for the number of neurons is [1, 200] for the first layer and [−50, 200] for the rest of the layers; and the population size is 25. Furthermore, the evolutionary process continues for 150 generations. We used an in-house high-performance computer (16 processors with 1.5–2.5 GHz and 256 GB memory) and observed that such optimizations are more processor-intensive than memory-intensive as all 16 available processors were engaged whereas approximately 13 GB of memory was required. In view of the computational power used, the running time of each GA optimization was nearly 1.5 days and 2.5 days for scenarios I and II, respectively.

2.4.2. Optimal hyperparameters

Heuristic approaches are incorporated to select optimal hyperparameters because there are viable practices by which certain hyperparameter values can be efficiently selected (Goodfellow et al., 2016). As indicated in (Song et al., 2017), we also found that it may be unnecessary to involve each hyperparameter in the optimal FNNs design. This stems from the fact that the default recommendations in certain hyperparameter tunings often lead to good performance. Hence, relying on heuristic approaches for selecting certain hyperparameters allows us to find at least near-optimal hyperparameters without involving a rigorous optimization algorithm.

Although there is no consensus on a single optimal activation function, it has been suggested that the rectified linear activation function, ReLU, can outperform other activation functions such as sigmoid and hyperbolic tangent when using multi-layer feedforward networks. The reason lies in the fact that ReLU is nearly linear and, in consequence, optimization of model parameters can be easier (Brownlee, 2019;

Goodfellow et al., 2016). Similarly, there is no single model-parameter optimizer, however "adam" is considered to be a fairly robust optimizer, in general (Goodfellow et al., 2016). Regarding the loss function in the present regression problem, MSE, which stands for Mean Square Error, is chosen. The underlying reason is that the dataset is generated from a deterministic source, i.e. GHGenius; hence, there should not be outliers in the dataset. However, in the case of outlier presence, other loss functions such as MAE (Mean Absolute Error) or Huber merit consideration. It is observed that the MISO-FNNs used in this study confirmed the utility of these well-known practices. Readers are referred to Supplementary Note S2 for details, showing the optimality of ReLU, adam, and MSE in the present study. Regarding the "learning rate" and the "maximum epoch", we performed optimization tests and were able to obtain near-optimal values. Interestingly, for predicting global warming impact, Song et al. (2017) has also reported the same activation function and learning rate, and a similar maximum epoch as the optimal hyperparameter values. Finally, owing to having sufficient computational power available, the networks are trained using all data in one batch to enable fast network training. The optimal or near-optimal hyperparameters used in the present study are summarized in Table 1.

2.5. Organization

The present study aims to first reveal the contribution of each unit process involved in the Canadian fuel life cycles to the net CO_2-eq emissions, then to optimally design MISO-FNNs to predict CO_2-eq emissions for each unit process in face of data gaps. To these ends, three key steps are taken, as summarized in Fig. 4.

Step (1). As explained in section 2.1, data are collected from GHGenius for CO_2-eq emissions from all of the unit processes of fuel life cycles for each Canadian province.

Step (2). The extracted data are analyzed to determine the contribution of each unit process. For this purpose, the contribution of the unit process P to the total CO_2-eq emissions, $\overline{C}_P$, is quantified by averaging the emissions from the unit process P with respect to locations and fuels in each fuel category. The average contribution of the unit process P thus reads as

$$\overline{C}_P = \frac{\sum_i \sum_j c_{i,j,P}}{\sum_i \sum_j \sum_k c_{i,j,k}} \times 100 \ \ \% \ \ i \in \text{Canada's} \ \ \text{Provinces} j \in \text{Fuels} k \in \text{Unit} \ \ \text{processes} \tag{3}$$

where $c_{i,j,k}$ indicates CO_2-eq emissions corresponding to location i, fuel j, and unit process k. The results are presented in section 3.1.

Step (3). The present study proposes a hybrid approach using both heuristics and GA to tractably optimize MISO-FNNs. As elaborated in section 2.4, the proposed optimization framework divides all decision variables (i.e. parameters required to be optimized) into three categories; (1) Input layer (attributes or features), (2) Topology of hidden layers (hidden topology), and (3) Parameters affecting learning (hyperparameters). To optimally design one MISO-FNN for a unit process, the following procedure is carried out. Hyperparameters are set based on widely used heuristics (see section 2.4.2). Next, for each attributes scenario, the hidden topology is optimized by GA such that validation error is minimized (see section 2.4.1). Finally, the impacts induced by the attributes scenarios on the optimal MISO-FNN (e.g. complexity of the hidden topology and the performance) are compared to find the best optimal attributes scenario (see section 3.2). These workflows are highlighted in green in Fig. 4. As elaborated in detail in Supplementary Note S2, the validity of the chosen hyperparameters is eventually assessed by using networks with optimal topology.

3. Results

This section consists of three sub-sections as follows. Section 3.1 discusses the contributions of each unit process to the total CO_2-eq emissions for fossil-based and renewable fuel pathways. Section 3.2 elaborates on the impacts associated with the attributes scenarios.

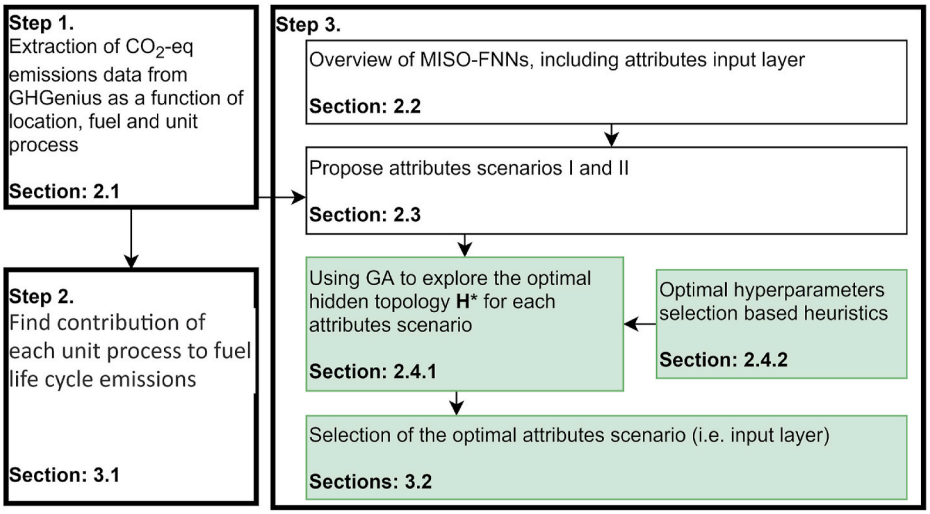

Fig. 4. Three main steps for determination of the contribution of each unit process and for the optimal design of MISO-FNNs to predict missing CO_2-eq emissions for filling data gaps in fuel life cycle inventories. The three steps are delineated by the thick border. Green highlights indicate the determination of optimal decision variables.

Moreover, the optimal design for each of the MISO-FNNs, predicting CO_2-eq emission data gaps for each unit process, are presented. Section 3.3 demonstrates the generality of the optimal MISO-FNNs designed in Section 3.2.

3.1. Contribution of unit processes to fuel life cycle GHGs in Canada

Using the inventory database for fuel life cycles described in Section 2.1, unit processes are ranked according to their contributions to net GHG emissions. For this purpose, we first categorize fuel pathways into two sets; fossil-based and renewable. Of 917 fuel pathways, 273 and 546 pathways fall into the fossil-based and renewable categories, respectively. The database associated with each fuel category was obtained from GHGenius 5.01 under settings reflecting fuel life cycles in Canada's provinces for the year 2021. Fig. 5(a) and Fig. 5(b) show the average contribution of each unit process to the total CO_2-eq emissions in the fossil and renewable fuel life cycles, respectively.

To determine the major contributors for each fuel category, we apply a cumulative cut-off of 95%, which is a reasonable cut-off for LCA. The major contributors are summarized in Table 2.

As can be concluded from Fig. 5 and Table 2, for fossil fuel pathways, Feedstock upgrading (1%), Land-use changes, cultivation (1%), CO_2, H_2S removed from NG (1%), and Fertilizer manufacture (0%) can be reasonably ignored. Additionally, between Feedstock transmission (2%) and Gas leaks and flares (2%), only one is required to be considered in the LCA to meet the 95% cumulative contribution. The individual cut-off applied for fossil fuel life cycles is thus 2%. In a similar vein, Fuel dispensing (2%), Gas leaks and flares (1%), and CO_2, H_2S removed from NG (0%) contribute negligibly to the net emissions of renewable fuel pathways and, in consequence, can be ignored. Therefore, 2% is also the individual cut-off applied for the renewable fuel life cycles.

98 out of 917 fuel pathways cannot be definitively categorized in the

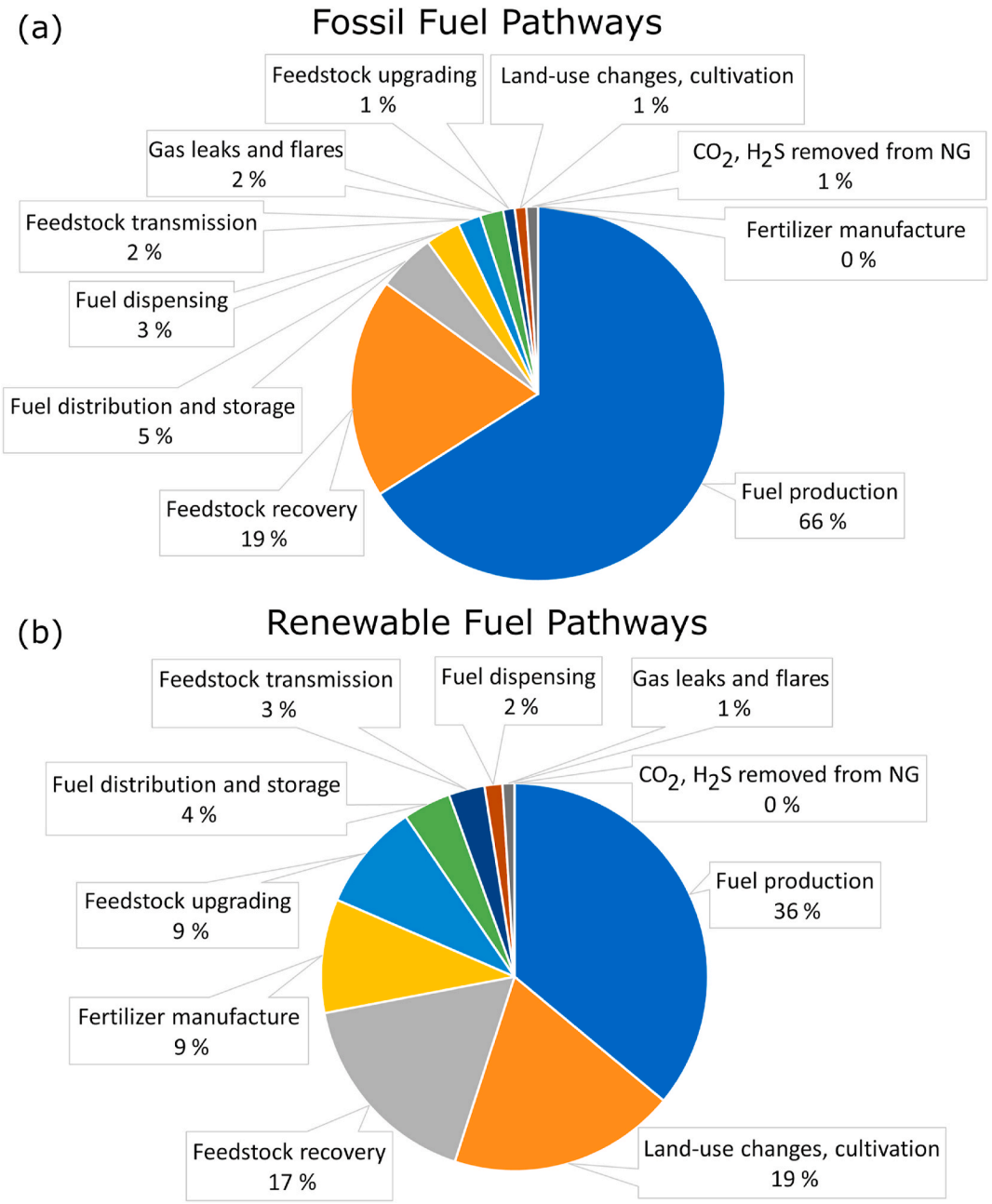

Fig. 5. Average contributions, $\overline{C}_P$, of the unit processes to total CO_2-eq emissions in fuel life cycles for (a) fossil fuel pathways and (b) renewable fuel pathways.

Table 2
The top unit processes, accounting for 95% or larger cumulative contributions to Canadian fuel pathway GHG emissions.

Fossil Fuel Pathways		Renewable Fuel Pathways	
Unit Process	Contribution (%)	Unit Process	Contribution (%)
Fuel production	66	Fuel production	36
Feedstock recovery	19	Land-use changes, cultivation	19
Fuel distribution and storage	5	Feedstock recovery	17
Fuel dispensing	3	Fertilizer manufacture	9
Feedstock transmission OR Gas leaks and flares	2	Feedstock upgrading	9
		Fuel distribution and storage	4
		Feedstock transmission	3
Cumulative Contribution	95	Cumulative Contribution	97

fossil-based or renewable fuel categories because it depends on the source of fuels; for example, electricity and hydrogen pathways. Applying equation (3) to the entire database yields the overall distribution shown in Fig. 6.

3.2. Attributes impacts on topologically optimal networks

To accurately compare the impact of the attributes scenarios described in Section 2.3 on the capacity of the MISO-FFNs to predict the CO_2-eq emissions, it is imperative to eliminate impacts from other factors. Thus, care should be taken with respect to the hyperparameters and hidden topology as they can also affect the network performance. For this reason, identical hyperparameters are used throughout all comparisons. The optimal hyperparameters used in this study are listed in Table 1. Regarding the hidden topology, **H**, there are two reasonable approaches. First, **H** is evolutionarily optimized using GA in order to reveal the best network performance for each attributes scenario. Second, **H** also remains identical for each contributor. The former is the primary objective of the present study and is elaborated in this section, and the latter is also discussed in Supplementary Note S3. To assess the capability of MISO-FNNs, we focus on the design of optimal MISO-FNNs to predict CO_2-eq emissions of all eleven unit processes in fuel life cycles regardless of the fuel type (i.e. fossil-based or renewable). Hence, all data extracted from GHGenius are used through training, validation, and testing of MISO-FNNs.

Fig. 7 demonstrates the impacts of the attributes scenarios (i.e. Scenario I and Scenario II) on the network performance whose hidden topologies are optimized through GA. Details of the GA results validation are provided in Supplementary Note S4. As explained earlier in section 2.4.1, the upper boundaries for the number of neurons per hidden layer and the number of hidden layers are assumed to be 200 and 5, respectively. Fig. 7 confirms the validity of our *a posteriori* approach concerning the upper boundaries because the maximum number of neurons per hidden layer and the maximum number of hidden layers are 100 and 4, respectively, which are less than the upper boundaries.

Based on the data shown in Fig. 7, we found that the attributes scenarios can affect not only the optimal network performance (i.e. training, validation, and testing errors) but also the optimal hidden topology (i.e. **H***). In the rest of this section, the impacts of the attributes scenarios on the optimal networks are discussed in order of their overall contributions to the net emissions (as illustrated in Fig. 6). Note, as a convention, the network performance is shown by a triplet whose elements depict the RMSE for training, validation, and testing sets, respectively.

Fuel production

Fig. 7(a) shows that the network performance of Scenario I and Scenario II are (19.01, 36.50, 51.04) and (14.73, 23.06, 23.53), respectively. Furthermore, the optimal topology of hidden layers obtained by GA for Scenario I and Scenario II are $\mathbf{H}^* = [50, 94, 88, 89]$ and $\mathbf{H}^* = [7, 95, 67]$, respectively. Consequently, in comparison to Scenario I, Scenario II leads to more accurate performance and a shallower optimal hidden topology for the prediction of CO_2-eq emissions for the "Fuel production" unit process. This finding is of particular importance because "Fuel production" is by far the largest contributor among unit processes, accounting for, on average, 58% of overall contributions to the total emissions in a fuel life cycle (Fig. 6). Specifically, the unit process "Fuel production" gives rise to 66% and 36% of average contributions in fossil and renewable fuel life cycle pathways, respectively (Fig. 5).

Feedstock recovery

As shown in Fig. 7(b), in the case of Scenario I, the optimal performance is (1.61, 3.34, 3.5) with $\mathbf{H}^* = [36, 85, 85, 36]$ and, in the case of Scenario II, the optimal performance is (1.37, 2.31, 3.45) with $\mathbf{H}^* = [9]$. For "Feedstock recovery" Scenario II requires a remarkably shallower network compared to Scenario I in order to perform optimally. Moreover, Scenario II outperforms Scenario I to some extent in terms of network performance. "Feedstock recovery" is the second-largest

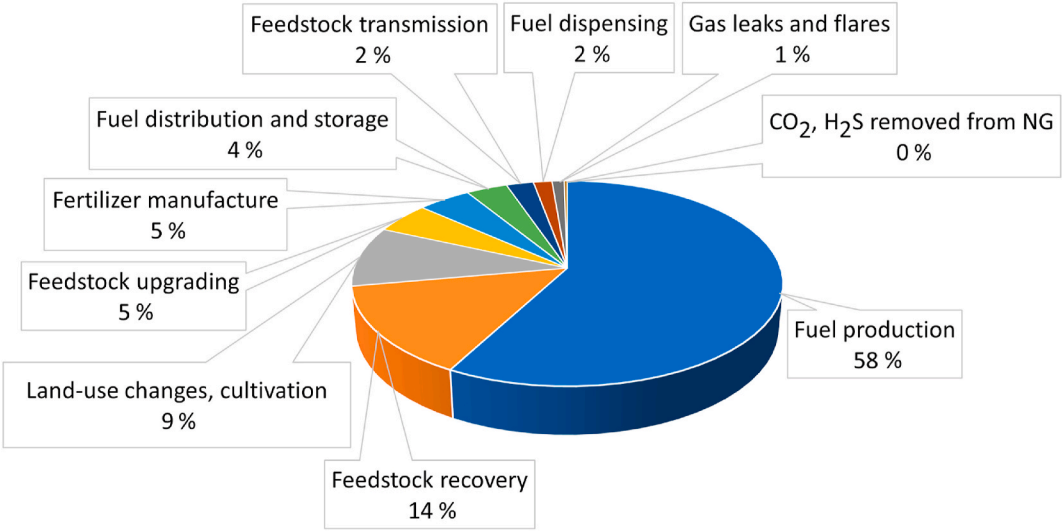

Fig. 6. Overall contributions of the unit processes to total CO_2-eq emissions in fuel life cycles.

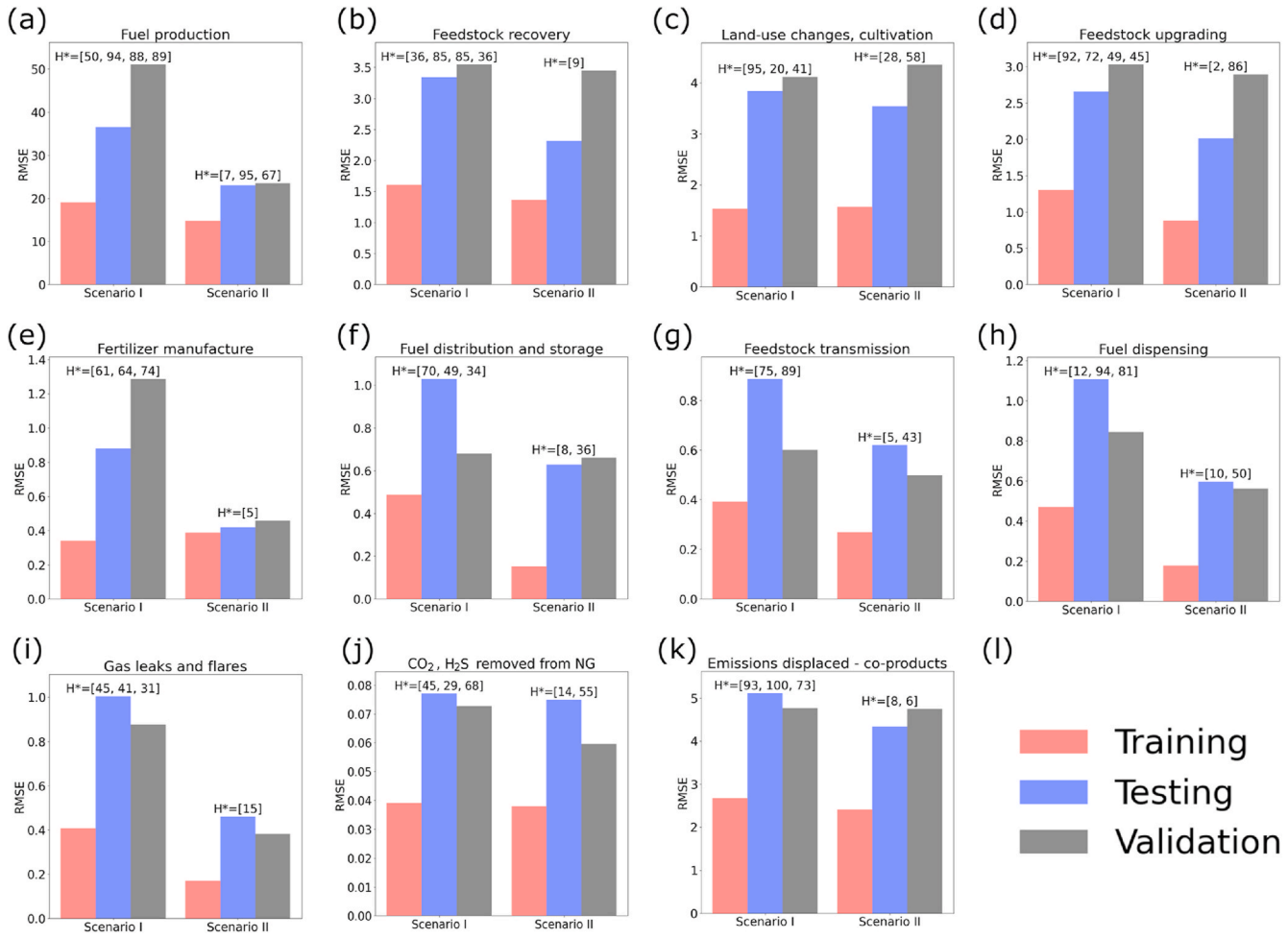

Fig. 7. (a–k) Attributes impacts on the performance of MISO-FNNs after optimization of hidden topologies and hyperparameters. (l) shows the legend for all panels.

contributor to CO_2-eq emissions in fuel life cycles, with a 14% overall contribution (Fig. 6). The contribution of this unit process is significant in both fossil and renewable fuel pathways, with 19% and 17% average contributions, respectively (Fig. 5).

Land-use changes, cultivation

As can be seen in Fig. 7(c), the network performances are (1.53, 3.84, 4.12) and (1.57, 3.53, 4.34) for Scenario I and Scenario II, respectively. Moreover, the optimal hidden topologies are $\mathbf{H}^* = [95, 20, 41]$ and $\mathbf{H}^* = [28, 58]$ for Scenario I and Scenario II, respectively. As a result, Scenario II also shows superiority for "Land-use changes, cultivation" because Scenario II makes the optimal hidden topology shallower compared to Scenario I. Nonetheless, Scenarios I and II result in roughly similar network performances. The overall contribution of "Land-use changes, cultivation" is, on average, 9% in the fuel life cycles (Fig. 6), and is the third-largest contributor to CO_2-eq emissions in the fuel life cycles. As shown in Fig. 5, "Land-use changes, cultivation" primarily contributes to renewable fuel pathways (19%) compared to fossil fuel pathways (1%).

Feedstock upgrading, and fertilizer manufacture

Fig. 6 show that "Feedstock upgrading" and "Fertilizer manufacture" contributions are 5% overall and are the fourth-largest contributor to the net emissions in fuel life cycles. These unit processes contribute primarily to renewable fuel pathways rather than to fossil fuel pathways because, as shown in Fig. 5, "Feedstock upgrading" and "Fertilizer manufacture" contributions in fossil fuel pathways are 1% and 0% respectively, which are negligible. In contrast, each of these unit processes contributes 9% in renewable fuel pathways, which are significant. In terms of the optimal attributes scenario, Scenario II is superior to Scenario I for both unit processes (see Fig. 7(d and e)). For "Feedstock upgrading", the network performances of Scenarios I and II are (1.30, 2.65, 3.03) and (0.88, 2.01, 2.89), respectively. Moreover, the resulting optimal hidden topologies are $\mathbf{H}^* = [92, 72, 49, 45]$ and $\mathbf{H}^* = [2, 86]$, respectively. Therefore, Scenario II yields a slightly more accurate network performance and shallower hidden topology. For "Fertilizer manufacture", Scenario II performs more accurately with noticeably simpler hidden topology. The network performances and optimal hidden topologies of Scenario I and II are, respectively, (0.34, 0.88, 1.29) and (0.39, 0.42, 0.46), $\mathbf{H}^* = [61, 64, 74]$ and $\mathbf{H}^* = [5]$.

Fuel distribution and storage

This unit process contributes roughly equally to fossil and renewable fuel life cycles, resulting in 5% and 4% average contributions, respectively (see Fig. 5), and a 4% overall contribution (see Fig. 6). The results of GA optimization demonstrate that, for this unit process, Scenario II with the shallower hidden topology, $\mathbf{H}^* = [8, 36]$, leads to more accurate performance, (0.15, 0.63, 0.66). Scenario I results in $\mathbf{H}^* = [70, 49, 34]$ and (0.49, 1.03, 0.68) (see Fig. 7(f)).

Feedstock transmission, and fuel dispensing

As displayed in Figs. 5 and 6, these two unit processes have 2–3% contribution in the fuel pathways, whether fossil-based or renewable. For "Feedstock transmission", both attributes scenarios lead to a two-layer hidden topology while Scenario II slightly outperforms Scenario I in terms of prediction capabilities. The network performances and optimal hidden topologies for Scenario I and Scenario II are (0.39, 0.89, 0.60), $\mathbf{H}^* = [75, 89]$ and (0.27, 0.62, 0.50), $\mathbf{H}^* = [5, 43]$, respectively (see Fig. 7(g)). For "Fuel dispensing", Scenario II performs more accurately with a shallower hidden topology. As illustrated in Fig. 7(h), for

Scenarios I and II, the network performances and the optimal hidden topologies are (0.47, 1.11, 0.84), **H*** = [12, 94, 81], and (0.18, 0.60, 0.56), **H*** = [10, 50].

Gas leaks and flares, and CO_2, H_2S removed from NG

As depicted in Figs. 5 and 6, these two unit processes have equal or less than 2% contribution in the fuel pathways, whether fossil-based or renewable. Fig. 7(i) shows that Scenario II causes a shallower optimal hidden topology, predicting the emissions from the "Gas leaks and flares" unit process more accurately in comparison to Scenario I. Scenario I and II leads to (0.41, 1.00, 0.88), **H*** = [45, 41, 31], and (0.17, 0.46, 0.38), **H*** = [15]. For "CO_2, H_2S removed from NG", Scenarios I and II lead to nearly similar network performance, but Scenario II requires a shallower optimal hidden topology compared to Scenario I (see Fig. 7(j)). The network performances and the optimal hidden topologies are (0.039, 0.077, 0.073), **H*** = [45, 29, 68] and (0.038, 0.075, 0.059), **H*** = [14, 55] for Scenarios I and II, respectively.

Emissions displaced - co-products

This unit process reflects the system expansion, and thus emission values are often zero or negative. For this reason, this unit process is not included in the unit processes' contributions as represented in Figs. 5 and 6. Regarding network design for this unit process, as can be seen in Fig. 7(k), Scenario II causes the shallower hidden topology **H*** = [8, 6] to perform more accurately (2.41, 4.33, 4.75) compared to Scenario I in which the network performance and optimal hidden topology are (2.67, 5.11, 4.76) and **H*** = [93, 100, 73], respectively.

In partial conclusion, regardless of the relative contribution of each unit process to the net emissions, Scenario II is superior to Scenario I in terms of more accurate network performance (training, validation, and testing errors) and/or less structural complexity associated with the optimal hidden layers (see Fig. 7). The difference between these attributes scenarios lies in the presence of additional categorical variables (i.e. location and fuel) in Scenario II, as shown in Fig. 2. Consequently, the enhancement of network performance and/or shallower depth of the hidden topology for optimal MISO-FNNs can be ascribed to using the combination of categorical and continuous variables in the network input layer.

Fig. 8(a–k) illustrates the performance of the optimal networks in which both the attributes scenario and hidden topology are optimal. Moreover, Fig. 8(a–k) confirms the excellence of the optimal network in terms of generalization since the performance of test sets, which are unseen data, is acceptable. Another standard approach to measure the generality of trained ANNs is the "learning curve", which is the plot of training and validation errors versus epoch. Supplementary Note S5 discusses the learning curves of each ANN, supporting the generality of the optimal ANNs designed in this section. It should be noted that, with a different distribution of datasets, the optimal networks (i.e. Scenario II and **H***) result in similar performances, confirming their capability in the accurate prediction of CO_2-eq emissions.

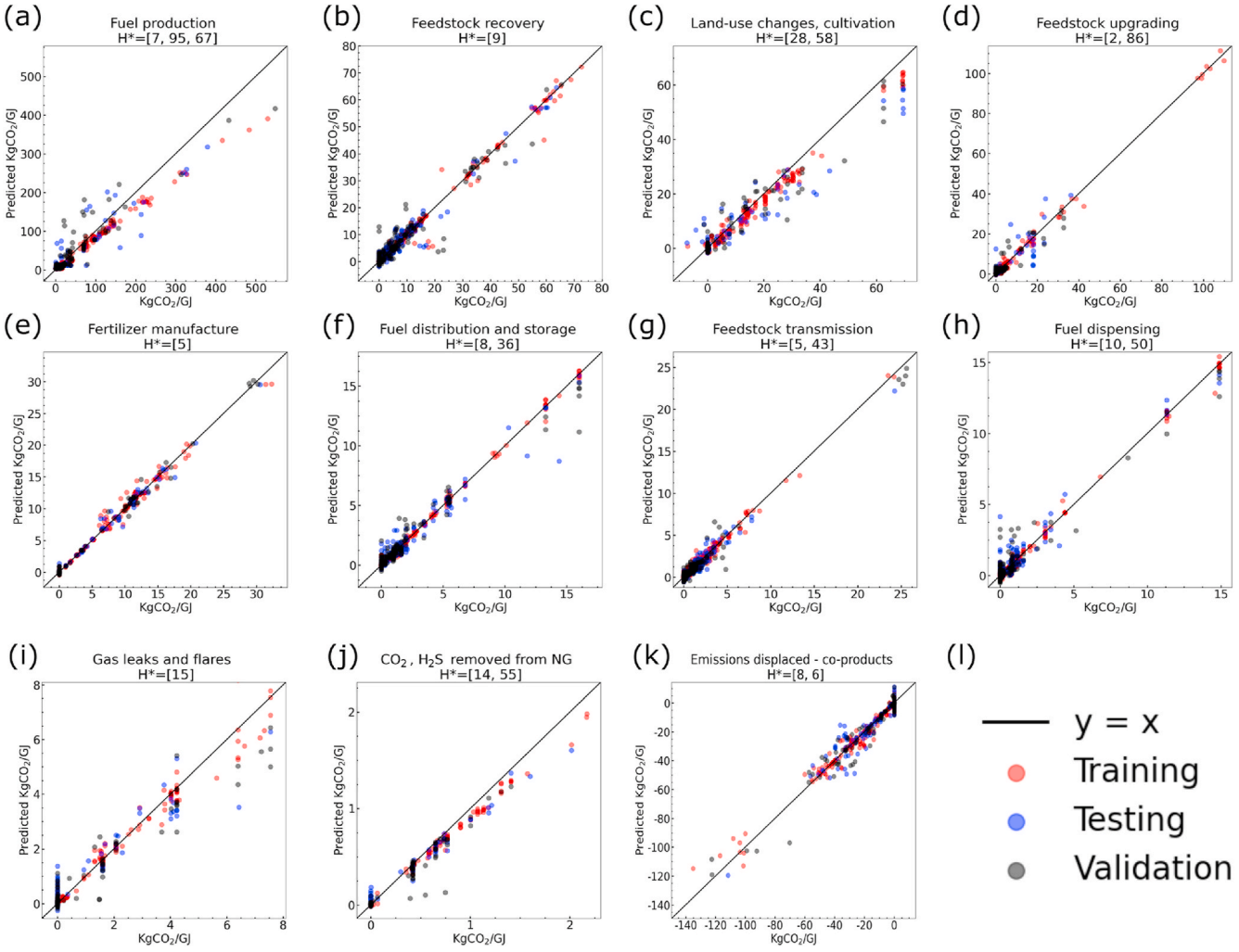

Fig. 8. (a–k) The scatter plots depicting the optimal performance, obtained through MISO-FNNs whose hidden topologies and hyperparameters are optimal and fed by the optimal attributes scenario (i.e. Scenario II). (l) shows the legend for all panels.

3.3. k-fold cross-validation

This is a practical approach, revealing the overall generality of predictive models. In this approach, the dataset is randomly divided into k groups. Thereafter, through each loop over all k groups, the model is trained by k-1 groups and tested by the 1 remaining group. Finally, the errors associated with k models are averaged and the averaged errors are presented as the model's errors (Goodfellow et al., 2016). There is no consensus on the k value in the k-fold cross-validation method; however, 5 or 10 are widely used. Table 3 summarizes the average performances of optimal MISO-FNNs, computed for $k = 10$.

As can be seen in Table 3, the errors of unseen sets are reasonably close to the errors of the corresponding training sets. The conclusion can be drawn that the optimal MISO-FNNs found in section 3.2 are capable of accurately predicting unseen data (i.e. data gaps) for all unit processes.

4. Discussion

The number of required samples in the database often grows exponentially with the dimension of inputs (i.e. the number of attributes), provided the estimation errors are kept relatively unchanged. This is known as the "curse of dimensionality" (Bengio et al., 2005; Silverman, 1986; Verleysen et al., 2003). Consequently, for a given dataset with a fixed number of samples, learning may worsen if the number of attributes increases. As shown in Fig. 2, the number of attributes in Scenario I and II are 10 and 148, respectively; hence, the attributes in Scenario II significantly outnumber those in Scenario I. This might increase the risk of facing the "curse of dimensionality". What might reinforce this risk is the fact that location and fuel could already be introduced to the networks implicitly through other CO_2-eq emissions since the CO_2-eq emissions fed to the networks are dependent on location and fuel. Nonetheless, Scenario II does not give rise to any adverse impact on network performance; instead, surprisingly, Scenario II enhances the accuracy of prediction. Assessment of the attributes scenarios under optimal hidden topologies revealed that Scenario II noticeably enhances the network performance (see Fig. 7(a, e-i)), slightly improves the network performance (see Fig. 7(b, d, k)), or does not significantly impact the network performance (see Fig. 7(c, j)). The improvement achieved through Scenario II in predicting emissions of the "Fuel production" unit process is of particular interest since this unit process is the dominant contributor to emissions from fuel life cycles in Canada (see Figs. 5 and 6).

Since the network performance is generally improved by increasing the number of attributes, two conclusions can be drawn. First, the given number of samples in the training set is sufficiently large so that the network does not suffer from the "curse of dimensionality". Second, the prediction of the CO_2-eq emissions for the unit processes under study is complicated from a mapping standpoint because the presence of more attributes (i.e. further information as networks inputs) is required to enhance the accuracy of network prediction.

Scenario II requires a shallower hidden topology compared to Scenario I to perform optimally, especially for the prediction of emissions from the "Feedstock recovery", "Fertilizer manufacture", and "Gas leaks and flares" unit processes (see Fig. 7(b, e, i)). It is thus concluded that augmentation of readily available features (i.e. fuel and location) as one-hot data with the numerical data (i.e. known CO_2-eq emissions) can lead to simplifying the optimal hidden topology. This finding is of importance owing to recently published results about the connection of network depth with loss function non-convexity. It has been shown (Li et al., 2017) that deeper networks (i.e. increase in the number of hidden layers) result in amplifying the non-convexity of the loss function and, in consequence, the trainability and the generality of networks become more difficult. As a result, shallower optimal hidden topology is another salient impact induced by Scenario II, leading to enhancement of trainability and generality. It should, however, be noted that deeper networks are more powerful in prediction; thus, they can be used for extremely large, complex datasets. In the case that relatively simpler datasets are studied, the network should have a sufficiently shallow depth to avoid the overfitting issue, but the network depth should not be unreasonably shallow so as to avoid the underfitting issue.

5. Conclusions and recommendations

Data gaps in LCI databases impede progress in LCA studies. In general, the present study has addressed this fundamental challenge by optimally designing ANNs to fill data gaps. Specifically, the present study has demonstrated three main outcomes, as follows.

(1) The average contributions of each unit process in the Canadian fuel life cycle pathways (including fossil-based and renewable fuels) to net CO_2-eq emissions are revealed, see Figs. 5 and 6, and Table 2, thereby identifying major and minor contributors to net CO_2-eq emissions, which are beneficial in cut-off analysis. It should be noted that the entire LCI data set was extracted from GHGenius through the automation workflow developed in this study, as shown in Supplementary Fig. S1.

(2) This study proposes, implements, and validates a hybrid optimization framework based on GA and heuristics by which optimal MISO-FNNs can be systematically designed in a mathematically tractable fashion. The development, implementation, and validation of the proposed methodology are elaborated in sections 2.4, 2.5, 3.2, and 3.3 in detail. We found out that the proposed approach can efficiently lead to optimal MISO-FNNs by which CO_2-eq emission data gaps are accurately estimated. Although the proposed approach was validated for the specific dataset, i.e. GHGenius, it is expected that it can be readily applied for filling data gaps within any LCI database owing to the fact that the proposed framework was generically developed.

(3) This study also shows that the network's input layer deserves significant consideration due to its impacts on the optimal network design. In particular, we found that, for each unit process, augmentation of categorical data (i.e. location and fuel) with numerical data (i.e. known CO_2-eq emissions from other

Table 3
Summary of average performances of optimal MISO-FNNs for 10-fold cross-validation.

Unit Process	Average RMSE of Train Sets	Average RMSE of Test Sets	Unit Process	Average RMSE of Train Sets	Average RMSE of Test Sets
Fuel production	16.57	21.89	Feedstock transmission	0.20	0.47
Feedstock recovery	1.31	2.34	Fuel dispensing	0.18	0.36
Land-use changes, cultivation	1.27	3.31	Gas leaks and flares	0.20	0.37
Feedstock upgrading	1.22	2.41	CO_2, H_2S removed from NG	0.02	0.03
Fertilizer manufacture	0.39	0.50	Emissions displaced - co-products	1.5	3.7
Fuel distribution and storage	0.12	0.46			

unit processes) as the network's input layer can make the optimal hidden topology shallower and/or improve the performance of the optimal network (see section 3.2).

Given the importance of the network's input layer, it is recommended that future studies explore the impacts of different attributes scenarios, leading to more accurate data gap filling estimators. In particular, further research is required to reveal the role of other categorical attributes (e.g. fuel type) augmented with numerical attributes (e.g. CO_2-eq emissions) as we found that large impacts are induced by the combination of categorical and numerical attributes in the network's input layer.

Lastly, as the primary objective of the present study is to estimate the CO_2-eq emissions of one unit process in face of data gaps, the proposed framework is limited to the MISO-FNNs regressor. However, simultaneous data gaps in multiple unit processes can also be expected in practice. For this reason, the optimal design of Multiple-Input Multiple-Output Feedforward Neural Networks (MIMO-FNNs) will be necessary to accurately estimate data for data gaps in multiple unit processes.

CRediT authorship contribution statement

Sayyed Ahmad Khadem: Data curation, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Farid Bensebaa:** Supervision, Formal analysis, Writing – review & editing. **Nathan Pelletier:** Supervision, Formal analysis, Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

SAK would like to thank Don O'Connor for his explanations about GHGenius. The authors gratefully acknowledge support in this research from the Materials for Clean Fuels (MCF) Challenge, Sustainable Protein Production (SPP), and the Advanced Clean Energy (ACE) Programs of the National Research Council Canada.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclepro.2021.130053.

ACRONYMS LIST

LCA	Life cycle assessment
LCI	Life cycle Inventory
GHG	greenhouse gas
CO_2-eq	CO_2-equivalent
FNN	Feedforward Neural Network
MISO-FNN	Multiple-Input Single-Output Feedforward Neural Network
MIMO-FNN	Multiple-Input Multiple-Output Feedforward Neural Network
GA	Genetic Algorithm
COM	Component Object Model

References

Al Imran, A., Amin, M.N., Johora, F.T., 2018. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE, pp. 1–6.

Algren, M., Fisher, W., Landis, A.E., 2021. Machine Learning in Life Cycle Assessment, Data Science Applied to Sustainability Analysis. Elsevier, pp. 167–190.

Bengio, Y., Delalleau, O., Le Roux, N., 2005. The curse of dimensionality for local kernel machines. Tech. Rep. 1258, 12.

Biograce, E.U.. Biograce. https://www.biograce.net/content/ghgcalculationtools/recognisedtool/. (Accessed 24 September 2021).

Brownlee, J., 2019. A gentle introduction to the rectified linear unit (ReLU). https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. (Accessed 24 September 2021).

Chollet, F., 2015. Keras. https://github.com/keras-team/keras. (Accessed 24 September 2021). https://keras.io.

Dawood, T., Elwakil, E., Novoa, H.M., Delgado, J.F.G., 2021. Toward urban sustainability and clean potable water: prediction of water quality via artificial neural networks. J. Clean. Prod. 291, 125266.

Fiszelew, A., Britos, P., Ochoa, A., Merlino, H., Fernández, E., García-Martínez, R., 2007. Finding optimal neural network architecture using genetic algorithms. Adv. Comput. Sci. Eng. Res. Comput. Sci. 27, 15–24.

Fritter, M., Lawrence, R., Marcolin, B., Pelletier, N., 2020. A survey of Life Cycle Inventory database implementations and architectures, and recommendations for new database initiatives. Int. J. Life Cycle Assess. 25 (8), 1522–1531.

GHGenius. GHGenius. https://www.ghgenius.ca/. (Accessed 24 September 2021).

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, Massachusetts, United States.

GREET, A.N.L.. GREET model. https://greet.es.anl.gov/. (Accessed 24 September 2021).

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

Hou, P., Zhao, B., Jolliet, O., Zhu, J., Wang, P., Xu, M., 2020. Rapid prediction of chemical ecotoxicity through genetic algorithm optimized neural network models. ACS Sustain. Chem. Eng. 8 (32), 12168–12176.

Ibnu, C.R.M., Santoso, J., Surendro, K., 2019. Determining the neural network topology: a review. In: Proceedings of the 2019 8th International Conference on Software and Computer Applications, pp. 357–362.

IEA, I., 2014. CO2 Emissions from Fuel Combustion Highlights. International Energy Agency Paris.

Jolliet, O., Saade-Sbeih, M., Shaked, S., Jolliet, A., Crettaz, P., 2015. Environmental Life Cycle Assessment. CRC Press.

Khadem, S.A., Jahromi, I.R., Zolghadr, A., Ayatollahi, S., 2014. Pressure and temperature functionality of paraffin-carbon dioxide interfacial tension using genetic programming and dimension analysis (GPDA) method. J. Nat. Gas Sci. Eng. 20, 407–413.

Khadem, S.A., Rey, A.D., 2021. Nucleation and growth of cholesteric collagen tactoids: a time-series statistical analysis based on integration of direct numerical simulation (DNS) and long short-term memory recurrent neural network (LSTM-RNN). J. Colloid Interface Sci. 582, 859–873.

Kneifel, J., Kneifel, J., O'Rear, E., Lavappa, P., Greig, A.L., Suh, S., 2018. Building Industry Reporting and Design for Sustainability (BIRDS) Low-Energy Residential Incremental Energy Efficiency Improvements Database Technical Manual: Update. US Department of Commerce, National Institute of Standards and Technology.

LCAcommons. US Federal LCA commons. https://www.lcacommons.gov/. (Accessed 24 September 2021).

Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T., 2017. Visualizing the Loss Landscape of Neural Nets arXiv preprint arXiv:1712.09913.

Marais, E.A., Silvern, R.F., Vodonos, A., Dupin, E., Bockarie, A.S., Mickley, L.J., Schwartz, J., 2019. Air quality and health impact of future fossil fuel use for electricity generation and transport in Africa. Environ. Sci. Technol. 53 (22), 13524–13534.

McKechnie, J., Pourbafrani, M., Saville, B.A., MacLean, H.L., 2015. Environmental and financial implications of ethanol as a bioethylene feedstock versus as a transportation fuel. Environ. Res. Lett. 10 (12), 124018.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. CHAPMAN & HALL/CRC press, Boca Raton London New York Washington, D.C.

Sleep, S., Guo, J., Laurenzi, I.J., Bergerson, J.A., MacLean, H.L., 2020. Quantifying variability in well-to-wheel greenhouse gas emission intensities of transportation fuels derived from Canadian oil sands mining operations. J. Clean. Prod. 258, 120639.

Song, R., Keller, A.A., Suh, S., 2017. Rapid life-cycle impact screening using artificial neural networks. Environ. Sci. Technol. 51 (18), 10777–10785.

Subramanian, V., Golden, J.S., 2016. Patching life cycle inventory (LCI) data gaps through expert elicitation: case study of laundry detergents. J. Clean. Prod. 115, 354–361.

Sun, X., Zhang, X., Muir, D.C., Zeng, E.Y., 2020. Identification of potential PBT/POP-like chemicals by a deep learning approach based on 2D structural features. Environ. Sci. Technol. 54 (13), 8221–8231.

Turner, I., Smart, A., Adams, E., Pelletier, N., 2020. Building an ILCD/EcoSPOLD2–compliant data-reporting template with application to Canadian agri-food LCI data. Int. J. Life Cycle Assess. 1–16.

Verleysen, M., Francois, D., Simon, G., Wertz, V., 2003. On the effects of dimensionality on data analysis with neural networks. In: International Work-Conference on Artificial Neural Networks. Springer, pp. 105–112.

Wirsansky, E., 2020. Hands-On Genetic Algorithms with Python: Applying Genetic Algorithms to Solve Real-World Deep Learning and Artificial Intelligence Problems. Packt Publishing Ltd, Birmingham, United Kingdom.

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. Environ. Sci. Technol. 55, 8439–8446.