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

Khadem, Sayyed Ahmad; Bensebaa, Farid; Pelletier, Nathan

Publisher's version / Version de l'éditeur:
https://doi.org/10.1016/j.jclepro.2021.130053
Journal of Cleaner Production, 332, 2022-01

Optimized Feed-Forward Neural Networks to Address CO_2-equivalent Emissions Data Gaps – Application to Emissions Prediction for Unit Processes of Fuel Life Cycles Inventories for Canadian Provinces

Sayyed Ahmad Khadem[†‡], Farid Bensebaa[†], and Nathan Pelletier[‡]*

[†]Energy, Mining, and Environment, National Research Council Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada

[‡] Irving K. Barber Faculty of Science, The University of British Columbia, 3247 University Way, Kelowna, BC V1V 1V7, Canada

*Corresponding author. E-mail addresses: sayyed.khadem@ubc.ca

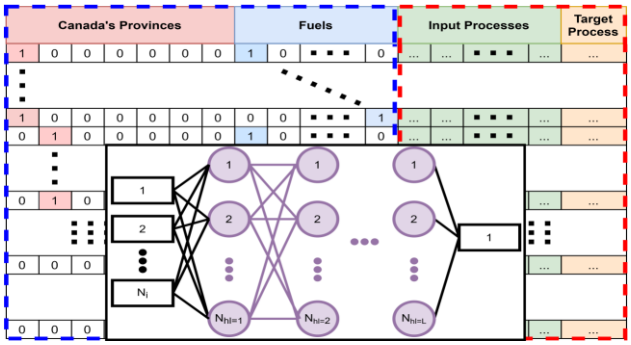
ABSTRACT

Data gaps present a crucial challenge in life cycle assessment studies. In the present study, we address this challenge by using optimized Artificial Neural Networks (ANNs) in the context of fuel life cycle inventory models. We extract emissions data at the fuel/province/unit process level from GHGenius (an open-access tool for modelling Canadian fuel pathways). Unit processes are ranked according to their contributions to total CO_2-eq emissions. Thereafter, we focus on the optimal design of ANNs, predicting CO_2-eq emissions from each unit process. Since optimization is computationally demanding, we propose a tractable hybrid approach using heuristics and Genetic Algorithm (GA). Decision variables are categorized into input layer (attributes), topology of hidden layers (hidden topology), and parameters affecting learning (hyperparameters). Two attributes scenarios are proposed. Hidden topology is optimized through GA for each scenario and the resulting impacts are analyzed to find the optimal scenario. We found that attributes scenarios can significantly affect the optimal network performance and/or the optimal hidden topology. Regarding hyperparameters, we rely on well-known heuristics and validate their optimality. Taken altogether, the hybrid optimization proposed herein is a tractable approach to design optimal ANNs which not only are accurate in addressing data gaps but also possess shallower hidden topologies.

Keywords: Life Cycle Inventory, Life Cycle Assessment, Greenhouse Gases, Machine Learning, Genetic Algorithm, Sustainability, GHGenius

Synopsis. Life cycle inventory data are indispensable for sustainability assessment studies. However, missing data is inevitable. This article proposes an approach to estimate missing data, facilitating sustainability research

Abstract Art



.

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[1]. 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[2]. Hence, detailed understanding and quantification of factors contributing to fuel life cycle GHG emissions are important and have attracted interest from numerous research organizations[3-6] and government agencies[7-10].

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[8]. GREET is a life cycle-based tool developed by Argonne National Laboratory in the United States[9]. 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[10].

Several commercial LCA software tools have also been developed, 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 governments and others 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[7] 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[7-10], the lack of representative LCI data (i.e. "data gaps") for many fuels/contexts remains a common challenge[11, 12]. 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[13, 14]. Limited interoperability between data sets from different databases also contributes to the ongoing gaps in reported LCA studies[11, 15, 16]. Furthermore, the mapping between input and output flows in LCI data is nonlinear and complex, making the resolution of LCI data gaps difficult[13]. 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 (e.g. LCI data estimation) in general[13, 14, 17-22]. However, further studies are required to elucidate the full potential of machine learning techniques in this context[19]. 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[19]. 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[17, 18, 23, 24]. These capabilities suggest the potential for using FNNs to fill LCI data gaps.

Currently, there are no consensus best practices for FNN design. However, heuristic-based rules of thumb have been proposed for case-specific problems[25, 26] or particular design aspects such as backpropagation optimizers and activation functions[27]. In particular, few studies[17, 18] discuss FNN design for estimation of LCI data. In addition, prior works made significant assumptions in order to simplify and reduce the computational complexity of the FNN design. For example, few pre-defined values were taken into account for the number of hidden layers and/or the number of neurons per hidden layer, making the search domain limited. Also, [17] evaluated and compared 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 very time-consuming [17]. This is because, as the search domain is increased, the required computing time grows exponentially, thus hindering FNN optimal design[25]. There are hence substantial challenges to optimally designing FNN models to predict LCI data gaps.

In the present study, we focus 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. The study is organized as follows. Section 2 (METHODS AND MATERIALS) presents our approach to data extraction from GHGenius (section 2.1), descriptions of the FNN model used (section 2.2), the two scenarios for arranging the attributes (section 2.3), and our proposed hybrid strategy to optimally design FNNs to predict CO_2-eq emission data gaps (section 2.4). Section 3 (RESULTS) and 4 (DISCUSSION) describe and discuss the numerical results.

2. METHODS AND MATERIALS

2.1. Life cycle inventory database development for fuels. The life cycle GHG emissions data reported in Version 5.01 of the GHGenius model[10] 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). In some fuel pathways, the actual number of unit processes may be less than eleven. For example, life cycles associated with fossil fuels do not include "fertilizer manufacture" unit processes. To reflect the absence of a unit process, emission levels are set to zero for that process. It should also be pointed out that the first ten unit processes listed above, (i-x), possess either positive or zero values, showing GHG emitted to the environment. However, the last unit process, (xi), manifests the amount of GHG emissions displacement due to co-products; hence, the emission values are either negative or zero.

To estimate the Global Warming Potential of each unit process as CO_2 equivalent (CO_2-eq) emissions, we use characterization factors for each greenhouse gas[1]. 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 we use "unit process" and "contributor" interchangeably 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. We hence consider location and fuel name 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 Figure S1 for details). By using the

automation algorithm depicted in Supplementary Figure 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 of Canada's provinces and 131 fuel pathways. In total, emissions estimates for 7×131 or 917 fuel life cycles are obtained. 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. We refer readers to Supplementary Tables S1 and S2 for the complete lists of locations, fuels, and unit processes considered in this study. Note that for simplicity and without loss of generality, the year was set to 2021.

2.2. MISO-FNN. Figure 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

$$\mathbf{H} = [N_{hl=1}, N_{hl=2}, \dots, N_{hl=L}] \quad (1)$$

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

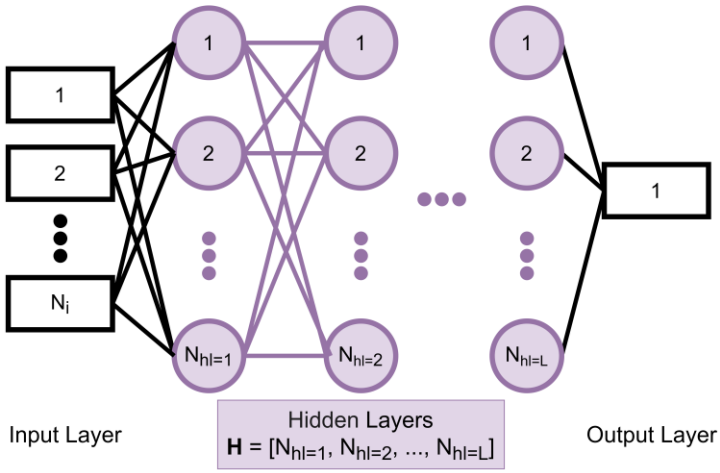


Figure 1. A generic MISO-FNN topology.

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 [18, 28]). 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[29]. Although the validation set is not used during the training phase, validation errors are monitored to avoid overfitting. Specifically, we applied the early stopping heuristic based on crossed validation, meaning that the training process is terminated before the maximum epoch is reached if the validation error increases for a certain number of consecutive epochs[30]. Here, the training is automatically stopped if the validation error increases in three epochs in a row. 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.

2.3. Attributes Scenarios. We assume that data gaps may originate from each unit process. We then use available data, including emissions from other unit processes, to fill the data gaps. As shown in Figure 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 Figure 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 Figure 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)

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

where $\mathbf{x}$, y_k, and f_k respectively refer to a vector of independent variables, the amount of CO_2-eq emission corresponding to the unit process k, and a nonlinear mapping between independent and dependent variables. Although the CO_2-eq emissions are fed to the network, the network can implicitly benefit from independent variables. Indeed, the CO_2-eq emissions fed to the network are also affected in theory by the independent variables, see equation (2). 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 by which the number of attributes fed to the network increases. In other words, unlike Scenario I, in which the network inputs are exclusively based on dependent variables (CO_2-eq emissions), Scenario II obeys a hybrid approach by incorporating both dependent variables (CO_2-eq emissions) and independent variables (locations and fuels). Owing to the fact that locations and fuels possess non-numeric values, the one-hot encoder approach is applied to make these non-numeric values suitable for the MISO-FNNs' input.

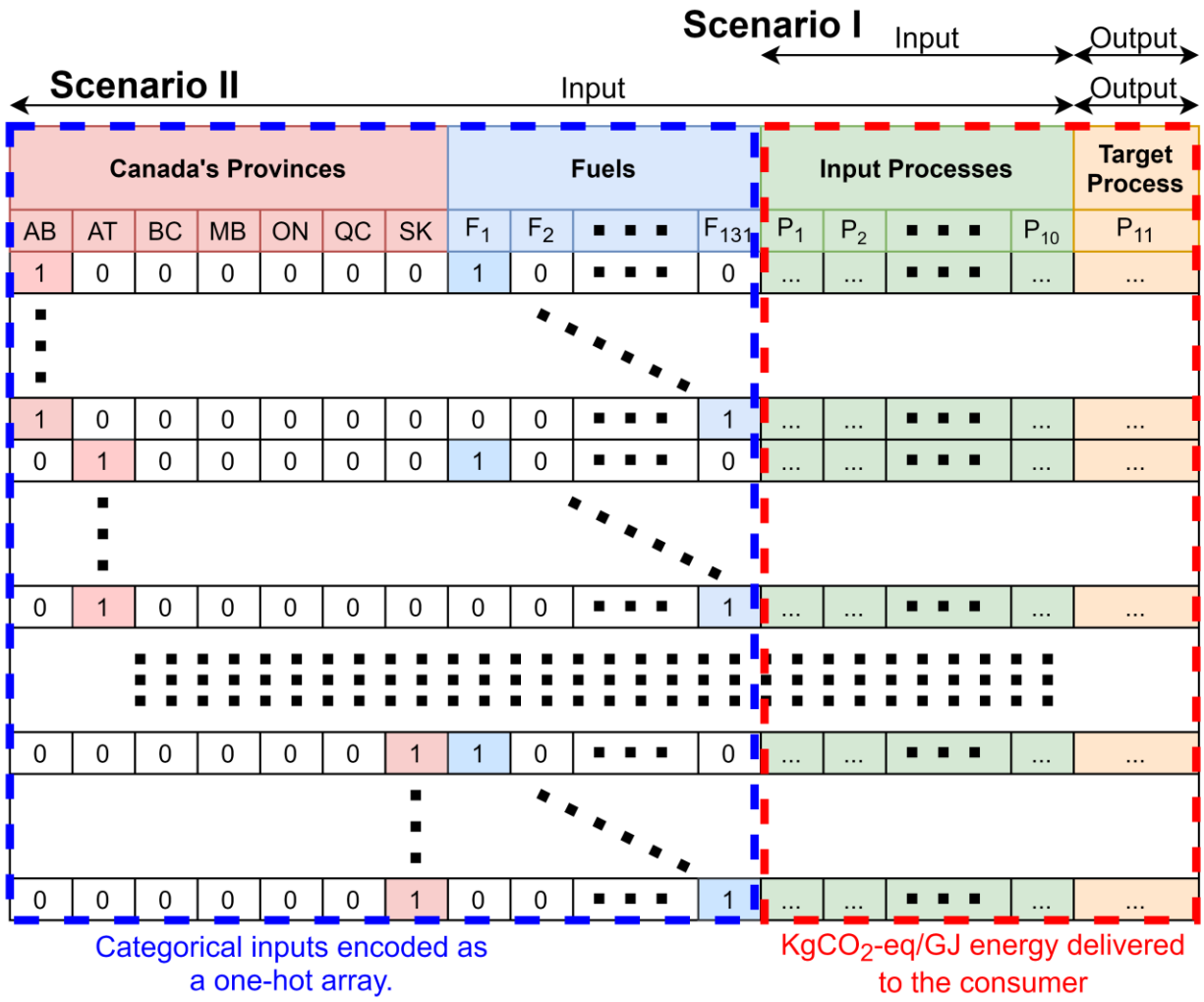


Figure 2. General data structure showing two possible scenarios for the networks' input layer. See Supplementary Table S1 for abbreviations used for Canadian provinces.

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 reveal 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 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 with moderate to large sizes. 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 $\mathbf{H}^*$. 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 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[18, 25]. However, the GA approach has demonstrated successful performance in finding the hidden topology[18]. We thus use the GA to simultaneously find both the number of hidden layers and the number of neurons per hidden layer in a wide search domain in order to find the optimal hidden topology, $\mathbf{H}^*$, hence overcoming assumptions made in previous studies[17, 18]. The general schema for finding $\mathbf{H}^*$ through GA is shown in Figure 3.

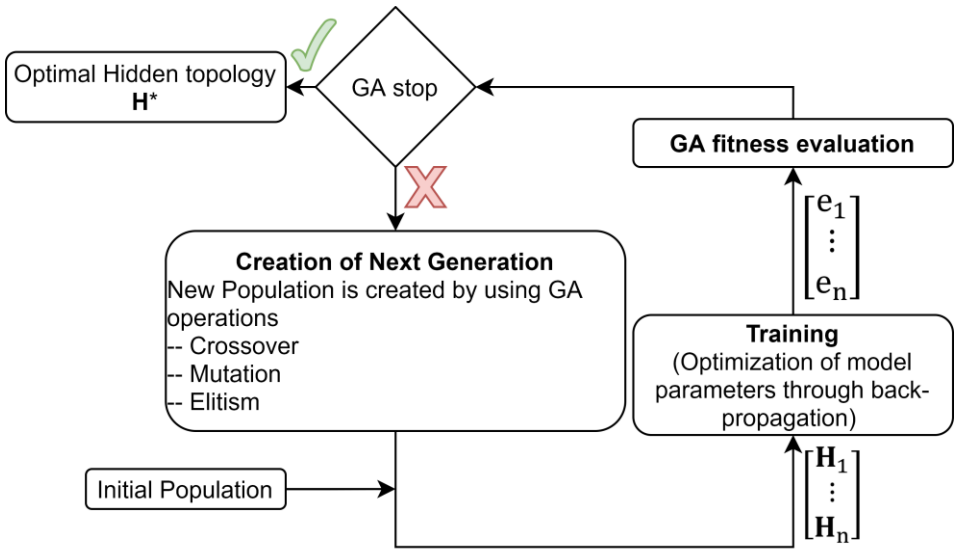


Figure 3. The workflow for optimization of the hidden topology, **H**. As an initial population, a number of random **H** is generated. Each individual, **H**, is trained and results in a validation error. Then, in GA fitness evaluation, individuals are ranked according to validation errors. If any GA 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.

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 $\mathbf{H}^*$ 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 [31]. On this basis, we considered a constant length for **H**. 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[32]. 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[32]. 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 (see Figure 3). This unavoidable numerical uncertainty may lead to confusion of the GA decision-making approach because the next generation is created based on the individuals' fitness (i.e. validation errors) in the current generation; therefore, the uncertainty related to validation errors likely affects the effectiveness of the next generation. This challenge is illustrated in Figure 3 where the "GA fitness evaluation" and "Creation of Next Generation" stages are performed after the training stage. To address this challenge, the model parameters are randomly initialized with different sets of values to significantly improve the probability of finding optimal solutions. Afterward, based on the comparison of the results obtained for different model initializations, the best one is selected and reported for use in GA. The number of initializations per individual in the training stage was chosen as 15. 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[33] 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 2.5 to 1.5 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 12 GB 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. We incorporate a heuristic approach to select optimal hyperparameters because there are viable practices by which certain hyperparameter values can be efficiently selected[27]. As indicated in [17], 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 a heuristic approach 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 feedforward networks. The reason lies in the fact that relu is nearly linear and, in consequence, optimization of model parameters can be easier[27, 34]. Similarly, there is no single model-parameter optimizer, however "adam" is considered to be a fairly robust optimizer, in general[27]. We observed that the MISO-FNNs used in this study

confirmed the utility of these well-known practices. Readers are referred to Supplementary Note S1 for details, showing the optimality of relu and adam 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, [17] 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.

Table 1. The optimal hyperparameters used in the present study.

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

3. RESULTS

We focus on determination of contribution of each unit process and optimal design for MISO-FNNs to predict CO_2-eq emissions from each unit process (also referred to as contributors) in fuel life cycles. To this end, three key steps are taken, as illustrated in Figure 4.

Step (1). As explained in section 2.1, data are collected from GHGenius[10] 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, see section 3.1.

Step (3). The optimal design for FNNs is performed in order to accurately estimate CO_2-eq emissions for the contributors in face of data gaps. Regarding the optimal design of FNNs, we propose a hybrid approach using both heuristics and GA algorithm. We categorize all decision variables (i.e. parameters required to be optimized) into three primary sets (highlighted in green in Figure 4) comprising (1) the input layer, (2) hidden topology, and (3) hyperparameters. Based on the extracted data, two attributes scenarios are proposed for the input layer, see sections 2.2 and 2.3. According to strategies elaborated in sections 2.4.1 and 2.4.2, the hidden topology is optimized with optimal hyperparameters (Table 1) for each attributes scenario. To find the optimal attributes scenario, the impacts of two attributes scenarios are then assessed separately for each contributor, see section 3.2.

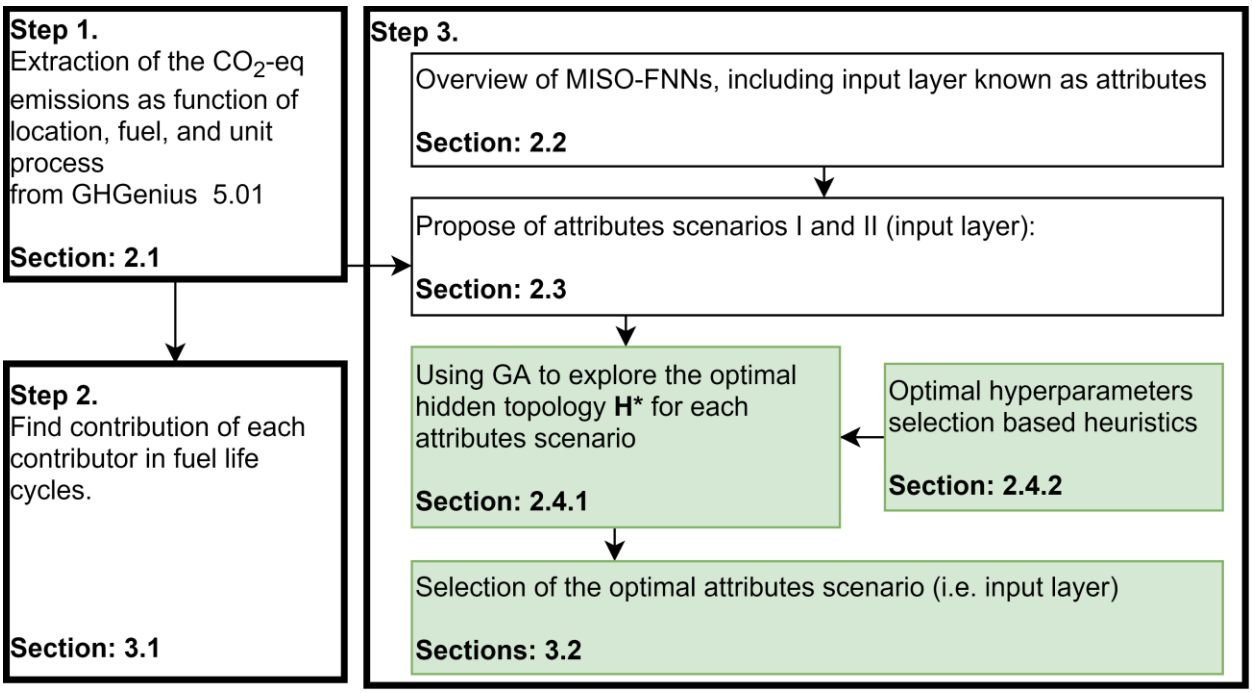


Figure 4. Three main steps for determination of the contribution of each unit process and for the optimal design of FNNs to predict missing CO_2-eq emissions for filling data gaps in fuel life cycle inventories. The three steps are delineated by the thick border. Green highlights signify determination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

3.1. Contribution of unit processes to fuel life cycles in Canada. Using the inventory database for fuel life cycles described in Section 2.1, we ranked unit processes according to their contributions to net GHG emissions. For this purpose, we first categorize fuel pathways into two prime sets; fossil-based and renewable. Of 917 fuel pathways, 273 and 546 pathways fall into the fossil-based and renewable categories, respectively. Thereafter, we quantify the average contribution of the unit process P to the total CO_2-eq emissions, $\bar{C}_P$, 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

$$\bar{C}_P = \frac{\sum_i \sum_j c_{i,j,P}}{\sum_i \sum_j \sum_k c_{i,j,k}} \times 100\%$$

$$i \in \text{Canada's Provinces} \qquad (3)$$

$$j \in \text{Fuels}$$

$$k \in \text{Unit processes}$$

where $c_{i,j,k}$ indicates CO_2-eq emissions corresponding to location i, fuel j, and unit process k. It is worth mentioning that 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[10]. Figure 5(a) and Figure 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.

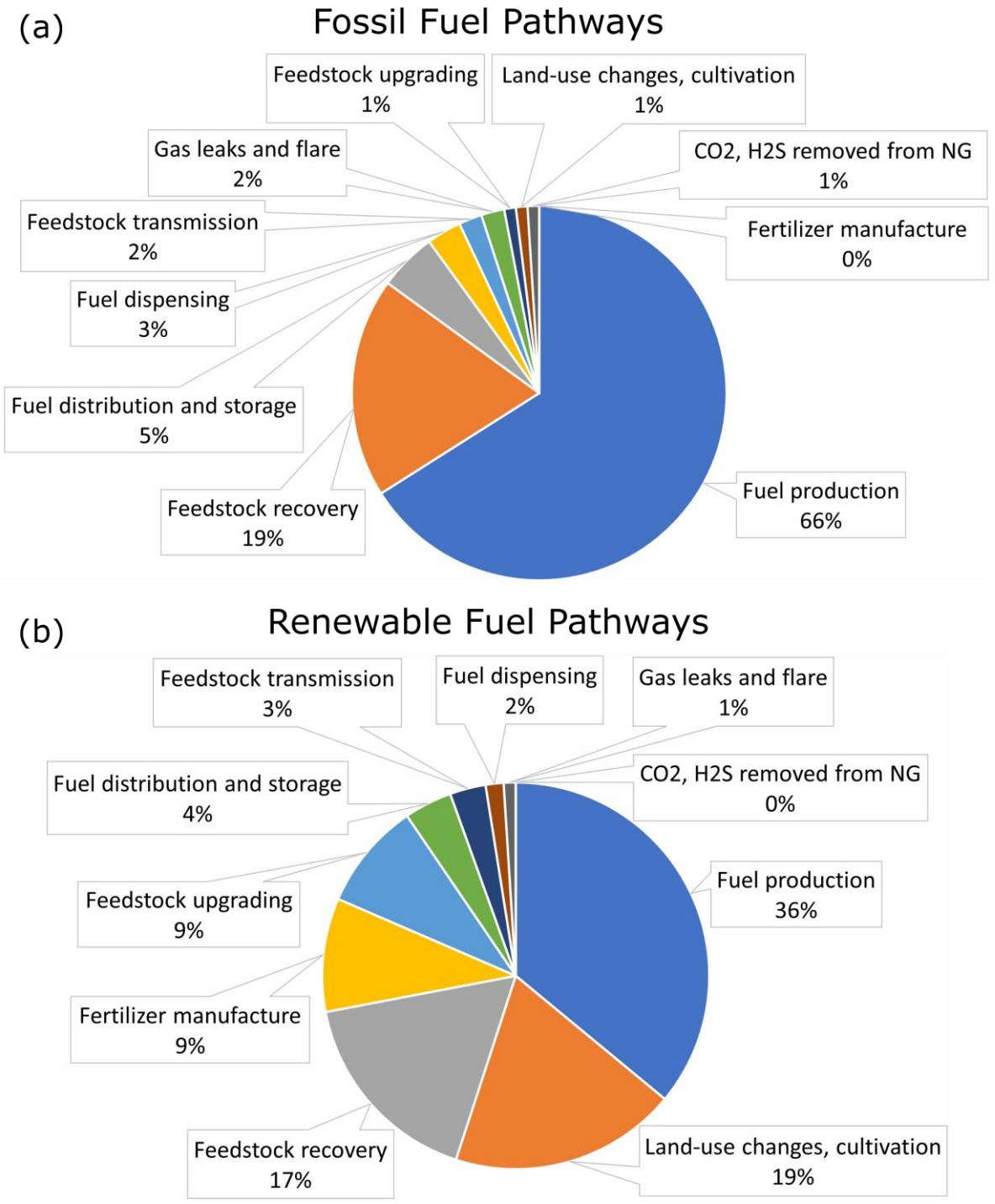


Figure 5. Average contributions, $\bar{C}_P$, of the unit processes to total CO_2-eq emissions in fuel life cycles for (a) fossil fuel pathways and (b) renewable fuel pathways.

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. Doing so, the major contributors are summarized in Table 2.

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	9.5
Feedstock transmission OR Gas leaks and flares	2	Feedstock upgrading	9
		Fuel distribution and storage	4
		Feedstock transmission	3
Cumulative Contribution	95	Cumulative Contribution	97.5

As can be concluded from Figure 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 fossil-based or renewable 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 Figure 6.

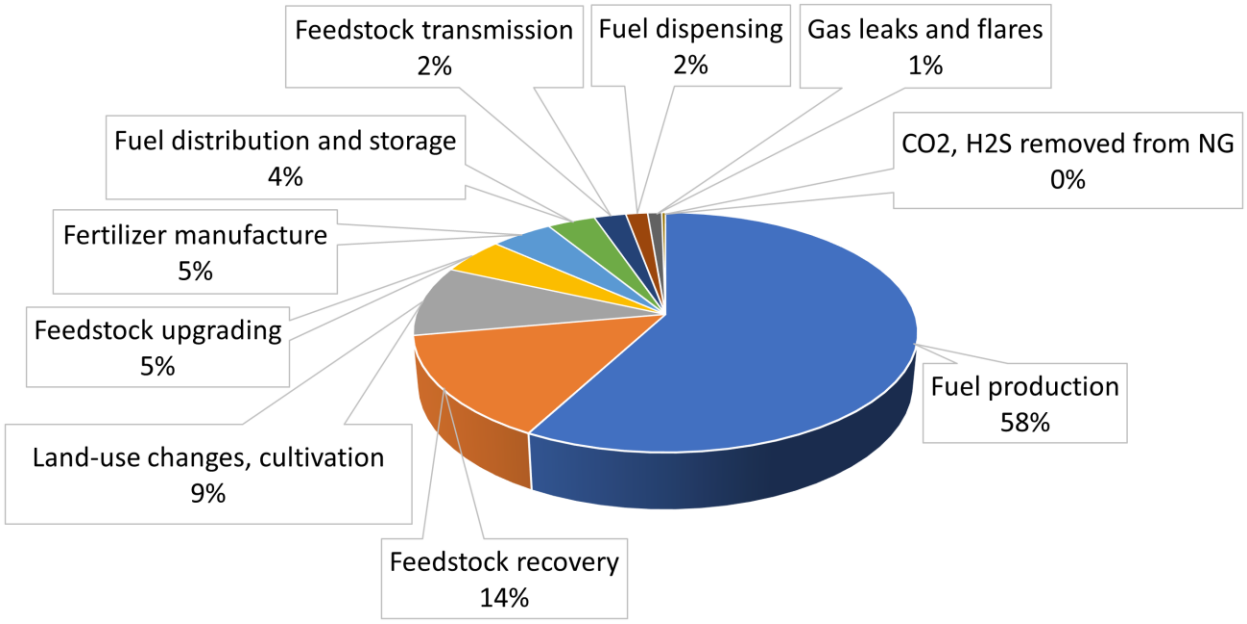


Figure 6. Overall contributions of the unit processes to total CO_2-eq emissions in fuel life cycles.

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 about 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 S2. 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, we incorporate all data extracted from GHGenius through training, validation, and testing of MISO-FNNs.

Figure 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. Supplementary Figure S2 illustrates the corresponding fitness evolution for the unit processes under study. 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. Figure 7 confirms the validity of our *a posteriori* approach concerning the upper boundaries because the maximum number of hidden layers and the maximum number of neurons per hidden layer are 4 and 100, respectively, which are less than the upper boundaries.

Based on the data shown in Figure 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. $\mathbf{H}^*$). 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 Figure 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.

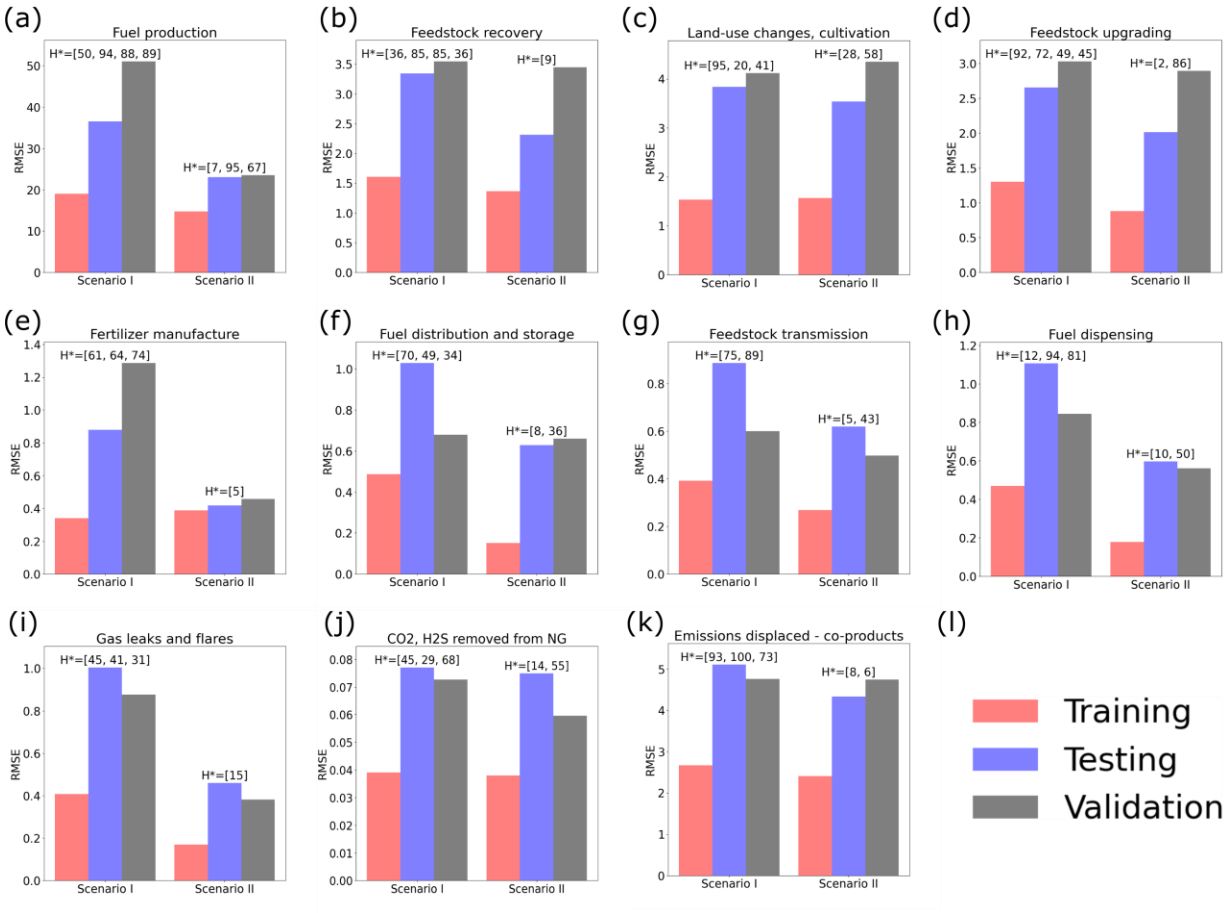


Figure 7. (a-k) Attributes impacts on the performance of MISO-FNNs whose hidden topologies and hyperparameters are optimal. (l) shows the legend for all panels.

Fuel production. Figure 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 [50, 94, 88, 89] and [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 (Figure 6).

Specifically, the unit process "Fuel production" gives rise to 66% and 36% average contributions in fossil and renewable fuel life pathways (Figure 5).

Feedstock recovery. As shown in Figure 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 contributor to CO_2-eq emissions in fuel lifecycles, with a 14% overall contribution (Figure 6). The contribution of this unit process is significant in both fossil and renewable fuel pathway, with 19% and 17% average contributions, respectively (Figure 5).

Land-use changes, cultivation. As can be seen in Figure 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 (Figure 6), and is the third-largest contributor to CO_2-eq emissions in fuel life cycles. As shown in Figure 5, "Land-use changes, cultivation" primarily contributes to renewable fuel pathways (19%) compared to fossil fuel pathways (1%).

Feedstock upgrading, and Fertilizer manufacture. Figure 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 predominantly contribute to renewable

fuel pathways compared to fossil fuel pathways because, as shown in Figure 5, "Feedstock upgrading" and "Fertilizer manufacture" contributions in fossil fuel pathways are 1% and 0% respectively, which are negligible. In contrast, these unit processes each contribute 9% in renewable fuel pathways, which are considerable. In terms of the optimal attributes scenario, Scenario II is superior to Scenario I for both unit processes (see Figure 7(d, 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 is performed 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 approximately contributes equally to fossil and renewable fuel life cycles, resulting in 5% and 4% average contributions in fossil and renewable fuel pathways, respectively (see Figure 5), and a 4% overall contribution (see Figure 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 Figure 7(f)).

Feedstock transmission, and Fuel dispensing. As displayed in Figure 5 and Figure 6, these two unit processes have 2-3% contribution in fuel pathways, whether fossil-based or renewable. For "Feedstock transmission", both attributes scenarios lead to a two-layer hidden topology while Scenario II still 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 Figure 7(g)). For "Fuel dispensing", Scenario II performs more accurately with a shallower hidden topology. As illustrated in Figure 7(h), for Scenarios I and II, the network performances and the optimal hidden topologies are (0.47, 1.11, 0.84), $\mathbf{H}^*$=[12, 94, 81], and (0.18, 0.60, 0.56), $\mathbf{H}^*$=[10, 50].

Gas leaks and flares, and CO_2, H_2S removed from NG. As depicted in Figure 5 and Figure 6, these two unit processes have equal or less than 2% contribution in fuel pathways, whether fossil-based or renewable. Figure 7(i) shows that Scenario II causes a shallower optimal hidden topology, predicting the emission 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), $\mathbf{H}^*$=[45, 41, 31], and (0.17, 0.46, 0.38), $\mathbf{H}^*$=[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 Figure 7(j). The network performances and the optimal hidden topologies are (0.039, 0.077, 0.073), $\mathbf{H}^*$=[45, 29, 68] and (0.038, 0.075, 0.059), $\mathbf{H}^*$=[14, 55] for Scenarios I and II, respectively.

Emissions displaced - co-products. This unit process reflects the system expansion, and thus emission values are zero or negative. For this reason, this unit process is not included in unit processes' contributions as represented in Figure 5 and Figure 6. Regarding network design for this unit process, as can be seen in Figure 7(k), Scenario II causes that the shallower hidden topology $\mathbf{H}^*$=[8, 6] performs 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 $\mathbf{H}^*$=[93, 100, 73], respectively.

In partial conclusion, for all contributing unit processes, regardless of the amount of their contribution 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 hidden layers (see Figure 7).

Figure 8(e-k) illustrates the performance of the optimal networks in which both the attributes scenario and hidden topology are optimal. Moreover, Figure 8(e-k) confirms the excellence of the optimal network in terms of generalization since the performance of test sets, which are unseen data, is highly acceptable. 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.

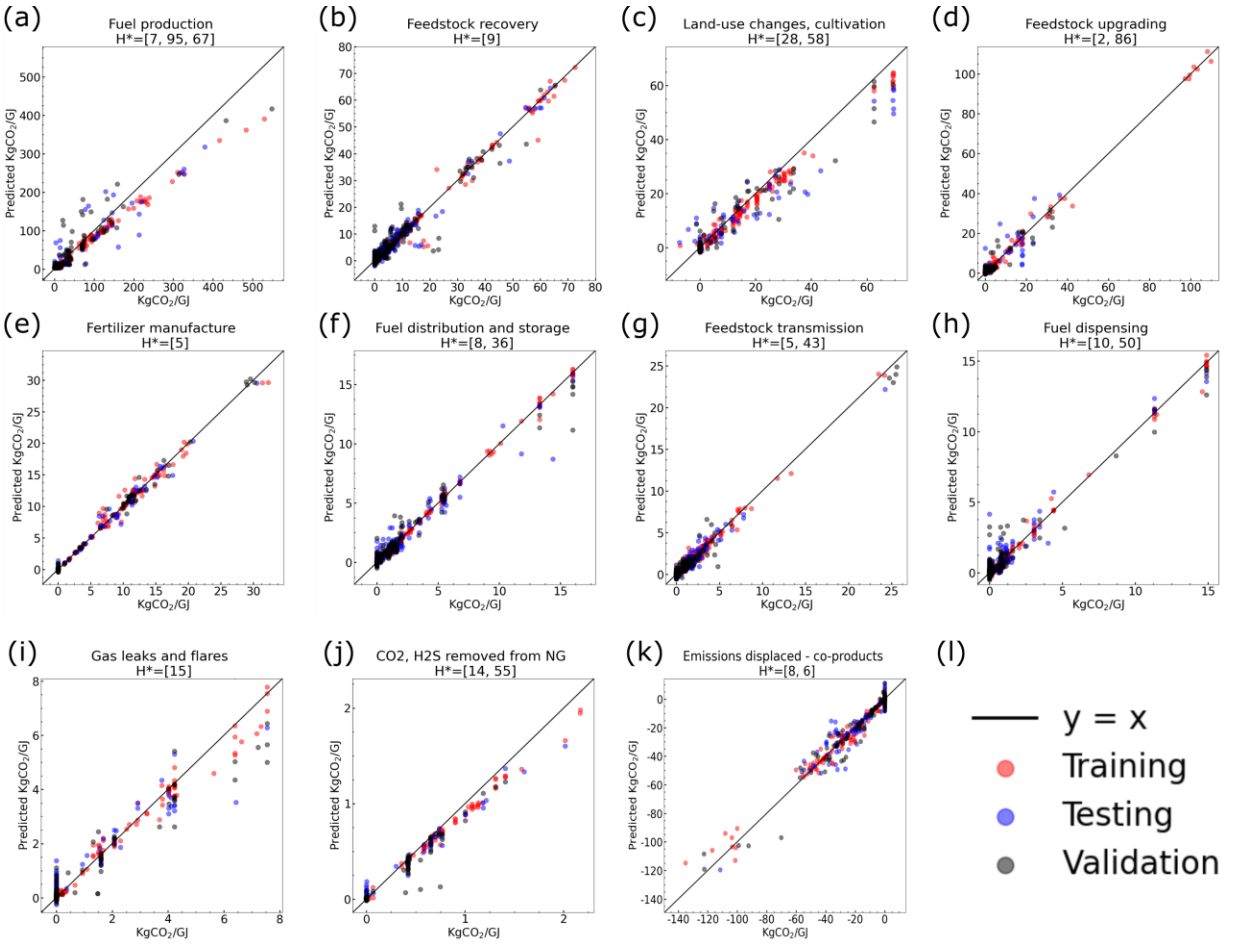


Figure 8. (a-k) The scatter plots depicting the optimal performance, as obtained through nets whose hidden topologies and hyperparameters are optimal and fed by the optimal attributes scenario (i.e. Scenario II). (l) shows the legend for all panels.

4. DISCUSSION

When using non-parametric algorithms such as MISO-FNNs, the number of required samples in the database (i.e. CO_2-eq emissions) often grows exponentially with the dimension of input (i.e. the number of attributes), provided the estimation errors are kept relatively unchanged. This is known as the "curse of dimensionality"[35-38]. Consequently, for a given dataset with a fixed number of samples, learning may worsen if the number of attributes increases. As shown in Figure 2, the number of attributes in Scenario I and II are 148 and 10, 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 Figure 7(a, e-i)), slightly improves the network performance (see Figure 7(b, d, k)), or does not significantly impact the network performance (see Figure 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 in fuel life cycles in Canada (see Figure 5 and Figure 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 Figure 7(b, e, i)). We thus conclude 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[32] that deeper networks (i.e. increase in the number of hidden layers) result in amplifying the non-convexity of 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.

All in all, the hybrid optimization framework proposed in this study is a tractable approach by which optimal MISO-FNNs can be systematically designed. The optimal MISO-FNNs can then be employed to estimate data gaps existing in LCI datasets, thereby addressing a common challenge in the LCA community. 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 of other unit processes) as network inputs can significantly simplify the optimal hidden topology and/or improve the performance of the optimal network. Hence, future studies should explore the impacts of different attributes scenarios, leading to more accurate data gaps estimators. In particular, further research is required to reveal the role of other categorical inputs (e.g. fuel type) as we found that large impacts are induced by categorical inputs. Lastly, the primary objective of the present study is to estimate the CO_2-eq emission of one unit process in face of data gaps. 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 in order to accurately estimate data for data gaps in multiple unit processes.

ACKNOWLEDGMENTS

SAK would like to thank Don O'Connor for his explanations about GHGenius[10]. The authors gratefully acknowledge support in this research from the Materials for Clean Fuels (MCF) Challenge Program of the National Research Council Canada.

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

1. Jolliet, O.; Saade-Sbeih, M.; Shaked, S.; Jolliet, A.; Crettaz, P., *Environmental Life Cycle Assessment*. CRC Press: 2015.
2. IEA, I., CO2 emissions from fuel combustion highlights. International Energy Agency Paris: 2014.
3. Yang, C.-J.; Leveen, L.; King, K., Ethane as a Cleaner Transportation Fuel. *Environmental Science & Technology* **2015,** *49* (6), 3263-3264.
4. Marais, E. A.; Silvern, R. F.; Vodonos, A.; Dupin, E.; Bockarie, A. S.; Mickley, L. J.; Schwartz, J., Air Quality and Health Impact of Future Fossil Fuel Use for Electricity Generation and Transport in Africa. *Environmental Science & Technology* **2019,** *53* (22), 13524-13534.
5. Sleep, S.; Guo, J.; Laurenzi, I. J.; Bergerson, J. A.; MacLean, H. L., Quantifying variability in well-to-wheel greenhouse gas emission intensities of transportation fuels derived from Canadian oil sands mining operations. *Journal of Cleaner Production* **2020,** *258*, 120639.
6. McKechnie, J.; Pourbafrani, M.; Saville, B. A.; MacLean, H. L., Environmental and financial implications of ethanol as a bioethylene feedstock versus as a transportation fuel. *Environmental Research Letters* **2015,** *10* (12), 124018.
7. US Federal LCA commons, https://www.lcacommons.gov/.
8. EU. Biograce, https://www.biograce.net/content/ghgcalculationtools/recognisedtool/.
9. ANL. GREET model, https://greet.es.anl.gov/.
10. GHGenius, https://www.ghgenius.ca/.
11. Turner, I.; Smart, A.; Adams, E.; Pelletier, N., Building an ILCD/EcoSPOLD2–compliant data-reporting template with application to Canadian agri-food LCI data. *The International Journal of Life Cycle Assessment* **2020**, 1-16.

12. Subramanian, V.; Golden, J. S., Patching life cycle inventory (LCI) data gaps through expert elicitation: case study of laundry detergents. *Journal of Cleaner Production* **2016,** *115*, 354-361.
13. Zhao, B.; Shuai, C.; Hou, P.; Qu, S.; Xu, M., Estimation of Unit Process Data for Life Cycle Assessment Using a Decision Tree-Based Approach. *Environmental Science & Technology* **2021**.
14. Hou, P.; Cai, J.; Qu, S.; Xu, M., Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environmental science & technology* **2018,** *52* (9), 5259-5267.
15. Fritter, M.; Lawrence, R.; Marcolin, B.; Pelletier, N., A survey of Life Cycle Inventory database implementations and architectures, and recommendations for new database initiatives. *The International Journal of Life Cycle Assessment* **2020,** *25* (8), 1522-1531.
16. Kneifel, J.; Kneifel, J.; O'Rear, E.; Lavappa, P.; Greig, A. L.; Suh, S., *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: 2018.
17. Song, R.; Keller, A. A.; Suh, S., Rapid life-cycle impact screening using artificial neural networks. *Environmental science & technology* **2017,** *51* (18), 10777-10785.
18. Hou, P.; Zhao, B.; Jolliet, O.; Zhu, J.; Wang, P.; Xu, M., Rapid Prediction of Chemical Ecotoxicity Through Genetic Algorithm Optimized Neural Network Models. *ACS Sustainable Chemistry & Engineering* **2020,** *8* (32), 12168-12176.
19. Algren, M.; Fisher, W.; Landis, A. E., Machine learning in life cycle assessment. In *Data Science Applied to Sustainability Analysis*, Elsevier: 2021; pp 167-190.
20. Liao, M.; Yao, Y., Applications of artificial intelligence-based modeling for bioenergy systems: A review. *GCB Bioenergy* **2021,** *13* (5), 774-802.
21. Liao, M.; Kelley, S.; Yao, Y., Generating Energy and Greenhouse Gas Inventory Data of Activated Carbon Production Using Machine Learning and Kinetic Based Process Simulation. *ACS Sustainable Chemistry & Engineering* **2020,** *8* (2), 1252-1261.
22. Liao, M.; Kelley, S. S.; Yao, Y., Artificial neural network based modeling for the prediction of yield and surface area of activated carbon from biomass. *Biofuels, Bioproducts and Biorefining* **2019,** *13* (4), 1015-1027.
23. Ma, S.; Zhou, C.; Chi, C.; Liu, Y.; Yang, G., Estimating physical composition of municipal solid waste in China by applying artificial neural network method. *Environmental Science & Technology* **2020,** *54* (15), 9609-9617.
24. Khadem, S. A.; Rey, A. D., 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). *Journal of Colloid and Interface Science* **2021,** *582*, 859-873.
25. Ibnu, C. R. M.; Santoso, J.; Surendro, K. In *Determining the neural network topology: A review*, Proceedings of the 2019 8th International Conference on Software and Computer Applications, 2019; pp 357-362.
26. Al Imran, A.; Amin, M. N.; Johora, F. T. In *Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning*, 2018 International Conference on Innovation in Engineering and Technology (ICIET), IEEE: 2018; pp 1-6.
27. Goodfellow, I.; Bengio, Y.; Courville, A., *Deep Learning*. MIT Press: 2016.

28. Sun, X.; Zhang, X.; Muir, D. C.; Zeng, E. Y., Identification of Potential PBT/POP-Like Chemicals by a Deep Learning Approach Based on 2D Structural Features. *Environmental Science & Technology* **2020,** *54* (13), 8221-8231.
29. Chollet, F., Keras, https://github.com/keras-team/keras, https://keras.io. **2015**.
30. Fiszelew, A.; Britos, P.; Ochoa, A.; Merlino, H.; Fernández, E.; García-Martínez, R., Finding optimal neural network architecture using genetic algorithms. *Advances in computer science and engineering research in computing science* **2007,** *27*, 15-24.
31. Wirsansky, E., *Hands-On Genetic Algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*. Packt Publishing: 2020.
32. Li, H.; Xu, Z.; Taylor, G.; Studer, C.; Goldstein, T., Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913* **2017**.
33. Khadem, S. A.; Jahromi, I. R.; Zolghadr, A.; Ayatollahi, S., Pressure and temperature functionality of paraffin-carbon dioxide interfacial tension using genetic programming and dimension analysis (GPDA) method. *Journal of Natural Gas Science and Engineering* **2014,** *20*, 407-413.
34. Brownlee, J., A Gentle Introduction to the Rectified Linear Unit (ReLU), https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. **2019**.
35. Silverman, B. W., *Density estimation for statistics and data analysis*. CRC press: 1986; Vol. 26.
36. Verleysen, M.; Francois, D.; Simon, G.; Wertz, V. In *On the effects of dimensionality on data analysis with neural networks*, International Work-Conference on Artificial Neural Networks, Springer: 2003; pp 105-112.
37. Bengio, Y.; Delalleau, O.; Le Roux, N., The curse of dimensionality for local kernel machines. *Techn. Rep* **2005,** *1258*, 12.
38. Priddy, K. L.; Keller, P. E., *Artificial Neural Networks: An Introduction*. SPIE Press: 2005.