Predicting total biogas potential of food waste using the initial output of biogas potential tests as input data to train an artificial neural network

Sarah M. Hunter[a*], Edgar Blanco[b], Adiuan Borrion[a]

[a] Department of Civil, Environmental and Geomatic Engineering, Chadwick Building, University College London, Gower Street, London, UK. WC1E 6BT

[b]Anaero Technology Limited, Cowley Road, Cambridge, UK. CB4 0DL

Abstract

Quantification of biogas potential is important for predicting anaerobic digestion operability and price. This study uses data from 446 biogas potential tests to train and test a multilayer perceptron artificial neural network (ANN) to forecast total biogas production using the evolution at the start of the experiment (3-14 days) as input data. ANN architecture (training algorithm, activation function, hidden nodes, regularisation, and input data) was optimised using response surface methodology. Best conditions (accuracy/computational speed) were obtained using adaptive moment estimation (adam) training algorithm and rectified linear unit (ReLU) activation function. When using three days of biogas production data, the accuracy of the model was reasonable (r^2_{test}=0.881, $r^2_{validation}$=0.879), although this increased significantly for 7 days (r^2_{test}=0.953, $r^2_{validation}$=0.925), or 14 days (r^2_{test}=0.971, $r^2_{validation}$=0.953). The highest accuracy was reported for readily digestible substrates (sugars and carbohydrates) and macronutrient mixtures. The methodology could be used to shorten prediction times of biogas potential tests.

Keywords: biogas yield, multilayer perceptron, response surface methodology, machine learning, activation function, training algorithm

1. Introduction

[*] Corresponding author
Email address: sarah.hunter.18@ucl.ac.uk

Anaerobic digestion (AD) is an effective technology for the treatment of food waste, producing biogas which can be burnt as a source of renewable energy. Optimising the performance and efficiency of AD is challenging, due to the nature of the complex process, varied chemical structure of food waste and the microbial consortiums involved (Enitan et al., 2017; Hunter et al., 2021; Xu et al., 2018). Artificial neural networks (ANNs) have emerged as a powerful tool for modelling and predicting the complex interactions within AD systems (Andrade Cruz et al., 2022; Enitan et al., 2017; Pomeroy et al., 2022). This study investigates the use of ANNs to predict the biogas potential of different macronutrients.

A common problem facing operators of AD is quantifying biogas potential of substrates (Da Silva et al., 2018). A methodology for measuring this is the biogas potential test in which a sample of substrate is placed in a stirred batch reactor containing inoculum at a fixed temperature and monitored until gas production stops (Koch et al., 2020). As well as providing insight into the energetic potential, these tests can also be used to agree substrate prices and predict operability (Strömberg et al., 2015). One challenge associated with this type of test is that it can take days or weeks to complete, during which time large quantities of substrate may need to be stored. Shortening the time taken to estimate biogas potential could reduce delays in supply chains (Da Silva et al., 2018), which would offer significant financial savings.

ANNs have been successfully used to predict various parameters and performance indicators in AD, such as biogas production (Aklilu and Waday, 2021; Mehryar et al., 2017b), methane yield (Almomani, 2020; Nair et al., 2016; Saghouri et al., 2020) and volatile fatty acid concentrations (Casallas-Ojeda et al., 2021; Dibaba et al., 2016). ANNs allow for the modelling of nonlinear relationships and complex interactions among multiple variables and use large data sets to extract patterns that may not be easily identifiable through traditional statistical methods and provide real-time predictions, facilitating process optimization and control. Many authors have indicated high correlations between model results and measured outcomes (r^2>0.95) (Casallas-Ojeda et al., 2021; Khashaba et al., 2022; Saghouri et al., 2020) and have indicated the superiority of such a technique over other statistical approaches (Aklilu and Waday, 2021; Jacob and Banerjee, 2016).

Mougari et. al. (2021) used an ANN to predict cumulative biogas production and methane yield of several organic wastes using volatile and total solid ratio, carbon content, carbon to nitrogen ratio and digestion time. The ANN architecture was optimised using a genetic algorithm to optimise the number of hidden layers, neurons and activation function in each hidden layer. The resulting model achieved excellent agreement between the measured and predicted values ($r^2_{training}$=0.9999, $r^2_{testing}$=0.9998). Although this study clearly demonstrates the ability of ANNs to predict biogas and methane production, it requires timely and expensive analysis of carbon and nitrogen content of food waste. Similarly, the input data was collected from literature studies, therefore is only likely to include experiments with positive outcomes (Mougari et al., 2021).

Another study which evaluated the use of ANNs on predicting biomethane test data was presented by Casallas-Ojeda (2021). In this work the authors evaluate co-digestion of food waste and garden waste using a multilayer perceptron (MLP) trained using Bayesian regularisation using the tangent sigmoid activation function. RMSE and computation load were combined in an objective function and this was optimised using response surface methodology (RSM). Although this study shows excellent model prediction (r^2>0.99), it's application is limited to co-digestion of two particular substrates (food and garden waste) (Casallas-Ojeda et al., 2021).

The aim of the presented study is to evaluate if data collected during the early stages of a biogas potential test can be used to predict the total biogas potential. Unlike the other studies discussed which require some characterisation of the input substrate, using early stages of a biogas potential test could be a cost-efficient way of expediting total biogas production. This has been achieved using statistical (Ponsa et al., 2011; Strömberg et al., 2015) or kinetic modelling approaches (Strömberg et al., 2015), but no studies were identified which use ANNs to reduce biogas estimation time using only batch biogas production data as an input, as opposed to substrate characteristics such as solids or macronutrient content or operating parameters such as organic loading rate or hydraulic retention time. ANNs offer the advantage of being more generally applicable and have frequently been reported to be more accurate for predicting AD using alternative forms of input data (such as substrate characterisation). First some background on ANNs and RSM will be provided, followed by the methodology used. The ANN optimisation results, and associated biogas predictions are discussed, followed by study limitations and potential future work.

1.1 Artificial Neural Networks

An ANN is a machine learning technique which uses an interconnected network of nodes (or neurons) to model and predict real life systems. In an ANN, information is passed through layers of nodes producing a non-linear output by assigning weights to each node and, if a node output exceeds a threshold value, the node is activated and passes the output to the next layer (see Figure 1). The output is calculated using an activation function which adds a bias term to the weighted sum (Nagy, 2018; Wang et al., 2021).

Activation functions generate non-linearity in neural networks (without this the model could be described by matrix multiplication), the most common being the identity (or linear), logistic (or sigmoid), hyperbolic tangent (tanh) and rectified linear unit (ReLU) functions (Nagy, 2018);

$\Phi(v) = v$ Equation 1: Identity/linear activation function

$\Phi(v) = \frac{1}{1+e^{-v}}(v)$ Equation 2: Logistic/sigmoid activation function

$\Phi(v) = \frac{e^{2v}+1}{e^{2v}+1}(v)$ Equation 3: Hyperbolic tangent activation function

$\Phi(v) = max(v, 0)$ Equation 4: ReLU activation functions

An MLP is a feedforward ANN which uses a training algorithm, such as backpropagation (backward propagation of errors) for supervised learning. Backpropagation is an algorithm which uses gradient descent (with respect to the weights of the ANN) to minimise the objective function (Lowe and Lawless, 2021) although many algorithms are available. During the process of training, the weights and biases of the neural network are initialised, and the entire training set is fed forward through the network. The error is calculated, and backpropagation is performed by computing the gradients of the loss with respect to the weights and biases and updated. This whole process, known as an epoch, is then repeated until acceptable ANN performance is achieved (Berlyand and Jabin, 2023).

Python is a popular programming language for the implementation of ANNs. The python sklearn package (version 0.20.3) offers three training algorithms; stochastic gradient descent (SGD), adaptive moment estimation (adam) and the limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) (scikit-learn developers, 2023). SGD is an optimisation method where gradient is estimated from a selection of random data points (Cady, 2017). SGD and adam both use the gradient of the loss function to update the weights and bias', but the adam optimiser can automatically adjust the size of parameter changes, using adaptive estimates (scikit-learn developers, 2023). Specifically, it uses the first and second moments of the gradients where the first is the mean of the gradients and the second is the uncentered variance. This is beneficial as it can be used to provide a different learning rate for each parameter, thereby accounting for more sensitive parameters. It achieves quicker convergence and better generalisation than other optimisers (Sakshii, 2023). L-BFGS approximates the second order partial derivative of a function, using a Hessian matrix and uses the inverse of this to update the weights and biases (scikit-learn developers, 2023).

Optimising ANN structure is a challenging aspect of their application. Chen et. al. (2022) recommends optimising the number of hidden layers, number of neurons in each layer, activation function and training algorithm (Chen et al., 2022) although the balance between good model prediction and overfitting should also be considered. Overfitting occurs when predictions are too closely fitted to the training data and do not generalise well for new data sets. This can become a problem when the number of nodes is increased above the optimum (Chen et al., 2022). Finding the optimum hyperparameter values can be achieved by trial and error (Andrade Cruz et al., 2022; Guclu et al., 2011; Wang et al., 2018), grid search (Al et al., 2019; Fernandes, 2014; Long et al., 2022) or statistical approaches, such as RSM (Antwi et al., 2018; Lujan-Moreno et al., 2018; Nguyen et al., 2022) and global optima determining techniques like genetic algorithms (Jacob and Banerjee, 2016; Saghouri et al., 2020), or similar alternatives (Beltramo et al., 2016; Casallas-Ojeda et al., 2021; Dibaba et al., 2016). Approaches such as trial and error and grid search are computationally demanding and inefficient (Lujan-Moreno et al., 2018) particularly when the number of parameters is large and unlikely to return an optimal result whereas the use of RSM or optimisers requires fewer training runs with better selection of parameters (Antwi et al., 2018; Lujan-Moreno et al., 2018; Mckenzie and Mcdonnell, 2023).

To determine the effectiveness of the model fit and identify any overfitting or underfitting, Goodfellow et. al. (2016) recommends comparing the error associated with the training and test sets. If the training set error is large, the model is underfit, whereas if the gap between the training error and test error is large, the model is overfit (Goodfellow et al., 2016). In addition, statistical tools such as the Akaike information criterion (AIC), Bayesian information criterion

(BIC) can also be used to assess ANN performance, particularly when evaluating model parsimony (Maier et al., 2010; Wu et al., 2014).

1.2 Response Surface Methodology

RSM is a statistical method which seeks to model the surface of an output, or response (Lujan-Moreno et al., 2018). It systematically evaluates the effect of reaction variables, their ranges, and their combined effects on an output or response variable. In a factorial design, experiments will be conducted at the corner of a design space, by running combinations of the input factors (Freddi and Salmon, 2019; Lujan-Moreno et al., 2018). The output or response can then be expressed as a linear combination of the inputs and interactions between the inputs;

$$y_1 = a_1 + a_2x_1 + a_3x_2 + a_3x_1x_2 \qquad \text{Equation 5}$$

Where y_n represents the outputs, x_n represents the inputs and a_n represents the coefficients. Replicates are performed to determine the error (Freddi and Salmon, 2019). RSM provides a methodology to maximise sample variability and is beneficial when an exhaustive search is impractical, expensive or, as in this case, time consuming (Mckenzie and Mcdonnell, 2023).

To build an RSM, factorial or fractional factorial designs are commonly applied. Central composite designs build on these factorial designs by including axial points, allowing estimation of curvature in the model, and therefore fitting a second order polynomial in the RSM. Distance of the axial points from the centre points (and the number of centre points) is dependent on the specific system being investigated, with circumscribed/distance-centred, inscribed/inner-face centred and face centred all possible alternatives (Freddi and Salmon, 2019).

2. Materials & Methods

2.1 Biogas potential tests

The biogas potential of various substrates was quantified using Anaero Technology Nautilus units (Anaero Technology, n.d.) and standard methodology (Holliger et al., 2016). All biogas potential tests were performed at mesophilic conditions (36˚C). Each reactor in the Nautilus units were connected to a cell in a calibrated Anaero Gas Flowmeter which recorded real time biogas production data. Reactors were preloaded with microorganism rich inoculum, and a known mass of a single substrate was added. A range of substrate to inoculum (S:I) ratios were used (1:2 to 1:6) as this impacts the kinetics and therefore produces a more varied dataset. Substrates were analysed at least in duplicate (for full details refer to Supplementary Data) and continued until biogas production plateaued (20-80 days). Control reactors (inoculum only) were included in each set of tests to evaluate the biogas activity of the inoculum.

2.2 Materials

A total of 112 different substrates were analysed. This includes pure substrates (various carbohydrates, proteins, fats, sugars and amino acids) as well as mixtures (a total of 446 individual experiments). These were sourced from supermarkets and food grade suppliers (Tesco, Bulk Powders ltd., Special Ingredients ltd.). The inoculum was sourced from a mesophilic, food waste, industrial AD facility.

2.3 Analytical Methods

Analysis of total solids (TS) and volatile solids (VS) for all substrates and inoculum (to determine S:I) was done according to the American Public Health Association (APHA) standard methods for water and wastewater (APHA, 1999).

2.4 Artificial Neural Network

Biogas potential raw data was imported into python (version 3.7.3) as a dataframe in pandas (version 0.24.2) and no preprocessing was performed. Biogas production data was collected on an hourly basis, and this was used for training the model, but for ease of understanding it will be referenced in days. Data was scaled (within the range 0-1) using the sklearn (version 0.20.3)

MinMaxScaler package and an MLP ANN (sklearn MLPregressor) was trained using back propagation and various ranges of input data (from 3 to 14 days).

2.5 ANN Optimisation using CCF and RSM

To determine the optimal ANN structure, a full factorial, face centred composite design (CCF) was used to train and test multiple model structures in silico. The use of a CCF design allows interpretation of the interactions between hyperparameters (as these are inherently non-linear). CCF was used (as opposed to a circumscribed design) to avoid negative hyperparameter values, which are not feasible. Table 1 indicates the CCF design used for each combination of training algorithm (adam, SGD, LFBGS) and activation function (identity, logistic, ReLU, tanh), resulting in this design being executed a total 12 times. Centre points were included for each of the 12 CCF designs and performed in triplicate to account for the variation generated by randomised partition of training and validation data.

The ranges for the four input variables were defined; input data size (3-14 days of biogas production), number of nodes in layer 1 (1-9x the size of the input data), number of nodes in layer 2 (3-30x the size of the input data) and strength of the regularisation term ($\alpha=1x10^{-5}$ to 10). The input data size was chosen based on a review of the raw data, as some of the substrates analysed have a lag time of 1-2 days, a minimum of 3 days was deemed appropriate. Similarly, some of the substrates analysed produce the majority of biogas within 14 days and therefore a larger input data set would not offer any value for shortening prediction times. The number of nodes in each layer was selected using the widest possible range from existing heuristics (Chen et al., 2022), although this was verified by comparing model accuracy (r^2 and MSE) and complexity (AIC and BIC) between training and validation data to avoid overfitting. Finally, the regularisation strength was selected using the python default ($\alpha=1x10^{-5}$) and, as was identified during screening simulations, increasing this appeared to have beneficial impact on accuracy, therefore an upper value (α=10) was selected to take advantage of the effect whist limiting the potential risk of overfitting.

After initial data screening and exclusion of the worst performing combinations of training algorithm and activation functions, RSM was fitted and analysed using MODDE® software. This approach was taken as an efficient way to explore a wide parameter space as opposed to using a global optimiser, which would inevitably find the larger input data sets give better prediction or grid search which would be laborious with the variable input data set.

3. Results

Initial screening was used to evaluate the effect of different training algorithms on MSE to assess the impact of the categorical variables and reduce the number of RSM models required to model the continuous variables. By plotting the outputs from the CCF design (varying number of input days, number of nodes per layer, and regularisation parameter alpha as shown in Table 1) it is possible to evaluate the categoric variables.

3.1 Comparison of Training Algorithms

As can be seen from Figure 2a, the MSE was much higher when using the LBFGS training algorithm, than when using the adam or SGD algorithm for all activation functions. In addition, adam and SGD generally show greater alignment between the training and validation data, indicating better model fit. On this basis, only the adam and SGD algorithms were considered for further optimisation.

3.2 Comparison of Activation Functions

As can be seen from Figure 2b, the ReLU activation function generally performed best across the range of hyperparameters and input days, although tanh also performed well at a low number of input days. As expected, the most accurate predictions can be achieved when using the highest number of days (14 days), this applied for all training algorithms except the logistic function. The logistic activation function is prone to the vanishing gradient problem which could be the cause of this poor performance (Salam et al., 2021). Response surfaces were modelled for ReLU and tanh activation functions only.

3.3 Response Surfaces

RSM models were fitted to the CCF data and used to explore the hyperparameters (number of nodes, strength of regularisation and size of the input data set) for combinations of the best performing training algorithms (adam and SGD) and activation functions (ReLU and tanh). As can be seen in Table 2, the RSM model showed good prediction of the ANNs MSE ($r^2 \geq 0.92$, $r^2_{adjusted} \geq 0.84$). Similarly, the model shows good predictive capability ($Q^2 \geq 0.73$) for all but the combination of adam and tanh and high reproducibility in all cases ($R \geq 0.959$).

A contour plot (Figure 3) was produced from the RSM model, showing how the MSE varies with different structures. Although equivalent MSE can be achieved using the adam and SGD optimiser, the ability of the SGD optimiser to generalise can be seen; lower MSE values can be achieved for a much broader range of hyperparameters. This may be because SGD is more locally unstable and can, therefore, converge to the minima at the flat or asymmetric valleys (Zhou et al., 2020). For both optimisers, ReLU can clearly be seen to outperform tanh. This is discussed more in the following sections.

3.4 Effect of Number of Input Days

As expected, increasing the number of days used as input data resulted in more accurate predictions (lower MSE and r^2), for both training functions and activation functions. It has been previously reported that biogas production at day 14 is highly correlated with total biogas production (Ponsa et al., 2011) and these results further support this. From Figure 3, it can be seen that it is also possible to get low MSE (<15,000) using 7 days of input data, when trained with the SGD algorithm and ReLU activation function. Good predictions can also be attained using adam and ReLU and SGD and tanh, although approximately 10 days of data would be required to achieve equivalent performance.

3.5 Effect of Hyperparameters

3.5.1 Number of Nodes

Optimising the number of nodes results in improved model performance (see Figure 3). It should be noted, however, that higher number of nodes increases model complexity and increases computational load and can result in overfitting (Andrade Cruz et al., 2022; Maier et al., 2010). Therefore, it is desirable to minimise the number of nodes required to achieve the desired model accuracy. Generally, the models with the broadest range of high accuracy were in the centre of the explored ranges (i.e. hidden nodes in the first and second layers are 5x and 16.5x the number of input nodes), except for the adam/tanh model, which performed best with high nodes in the first and second layer (9x and 30x input nodes).

3.5.2 Effect of Alpha

Increasing the regularisation term (alpha) can improve overfitting by encouraging smaller weights (Wu et al., 2014). Higher alpha values improve the models ability to generalise by reducing the effect that small variations in the inputs have on the output (Tarca et al., 2007). In this study higher alpha values generally improve predictions when using SGD, whereas the adam optimiser performed best with median alpha values.

3.6 Overfitting

For most of the model configurations assessed, the MSE of the training data and validation data are similar indicating good model fit, but the extent of this does vary. As can be seen from Figure 4, choice of training algorithm and input data (number of days) can impact the extent of overfitting. When using the adam algorithm, the difference in the MSE for the training and validation data is smaller when using a lower number of input days. This suggests that increasing the quantity of input data, increases the likelihood for overfitting. The SGD training algorithm, on the other hand, has a similar difference regardless of the number of days.

3.7 Optimised Result

Using the RSM model, an optimised model structure was determined for the best performing models (adam and SGD, both using the ReLU activation function) and this was used to train the ANN. As can be seen from Figure 5, the two training algorithms give very similar performance for optimised conditions (shown in Table 3) although SGD slightly outperforms adam. As seen in the RSM model, increasing the number of days used as input data gives increasing model accuracy (increasing r^2) for both algorithms.

To further evaluate the performance of each model, the computational load was considered. The number of epochs required for training each model is shown in Table 4. The adam training algorithm is generally much faster to converge than SGD (scikit-learn developers, 2023), and this was apparent in this study with SGD requiring 5-7x more epochs than when training with adam. Although accuracy is comparable between the models, adam clearly offers a significant advantage in terms of computational time.

A selection of biogas curves predicted by the trained MLP is shown in Figure 6 and accuracy of the models considered per substrate group are shown in Figure 7. The cases, taken from the test set, show the difference in prediction capability for various substrates. Generally, the ANN is capable of more accurate prediction for mixed composition. This includes mixtures of carbohydrates and proteins, as well as fully synthetic mixtures (also including fats) and real food waste, but many of the pure substrates analysed can also be well predicted. The examples given in Figure 6 show how biogas production from fats and sugars are well modelled. On the other hand, poor predictability of inoculum only (control) systems is also apparent, this could be attributed to the variation in activity of inoculum, and also the magnitude of difference in overall biogas production compared to substrates and this can be seen in Figure 7 where, despite the large error, inoculum actually has a relatively low RSME. Similarly fatty acids were poorly predicted, which is likely due to the smaller size of this subgroup (8 experiments), which is significantly limiting the effectiveness of the ANN.

4. Discussion

The presented results indicate that it is possible to expedite the results of a biogas potential test by using the early experimental output (biogas production) and an ANN to predict the remaining gas production. An optimised model selection can only be made by considering a balance of prediction error, extent of overfitting and model computation time. Of the parameters explored (training algorithms, activation functions, number of nodes, quantity of input data and alpha regularisation) the most robust models, capable of good predictions over a wide range of input data, were trained using SGD and adam in combination with the ReLU activation functions, although the SGD training algorithm also achieved reasonable results with tanh. Overall, a model trained with adam, using the ReLU activation function is recommended for accuracy and computational time. The optimised network structure depends on the total number of days used as input data (3-14 days); the number of nodes in the first and second hidden layers was optimised at 4-8x and 17-20x the total number of nodes in the input layer. The regularisation parameter (α) was found to be optimal in the range 6.2-6.5.

The methodology and results of this study have several implications which expand knowledge in the field. This study validates the use of this type of data (3-14 biogas production) as inputs when training ANNs to predict total biogas production. This differs from other published studies which typically use substrate characterisation data or operating parameters to estimate biogas potential, and this is advantageous as it reduces the need for analytical testing which can be costly or time consuming. Another key aspect of this research is that the broad set of substrates used to generate the training and validation data demonstrates general applicability of ANNs. This expands on studies in the literature which generally focus on models developed for a single substrate or co-digestion feedstocks.

Another novel aspect of the work is the use of a CCF design and RSM to optimise ANN structure, which builds on methodologies developed by other authors (Antwi et al., 2018; Lujan-Moreno et al., 2018; Nguyen et al., 2022). In this case it is advantageous as the variable input data size increases the number of potential ANN structures, and the use of an RSM model allows for a much more visual interpretation of the impact of hyperparameters than using grid search alone.

This study also demonstrates the ability of ANNs to predict biogas production using raw data, with no pre-processing or noise reduction. ANNs generally have excellent capacity to handle noisy data (Khashei and Bijari, 2010), but many similar studies have used data presented in the literature (Mougari et al., 2021), simulated by the ADM1 (Beltramo et al., 2016) or fitted with a kinetic model (Casallas-Ojeda et al., 2021; Khashaba et al., 2022). By using raw data, this methodology could be applied directly to industry, where biogas potential tests are used to evaluate substates and set commodity prices, and large datasets are available. Due to the nature of biogas potential tests, this can be a slow process and using a ANN to give a prediction of biogas potential, within a few days, operators could potentially reduce these delays and allowing substrates to be fed into a digestor without the need to wait for a completed test, reducing the need for storage times and simplifying supply chains. Although this also has limitations as the performance of ANNs relies on the quality and representativeness of the training dataset and the use of raw data likely reduced the model accuracy.

One limitation of this work is the variation in experimental data. Although efforts were made to ensure consistency in protocol, inherent variation exists, such as inoculum microorganism consortium and activity. Similarly, a larger data would inevitably improve the model accuracy, which is particularly important when comparing the accuracy of different substrates, as those with a greater number of individual experiments were typically found to be more accurate.

5. Further Work

Although the presented model shows promise for predicting total gas production from early biogas potential test results, some improvements are recommended for future development. Increasing the size of the data set would improve the model's robustness. Although 446 individual experiments were used to train and test the model, other works have suggested that over 1000 experiments are required to develop a model of this complexity (Alwosheel et al., 2018). Other works have highlighted a similar challenge and propose a collaborative approach among researchers and industry to create a database for model development (Andrade Cruz et al., 2022; Das Ghatak and Ghatak, 2018).

As other authors have pointed out, although the use of design of experiment methodology and RSM is unlikely to have identified a global optimum (Lujan-Moreno et al., 2018) but it does allow for a much more extensive exploration of the parameters considered. This approach was particularly valuable for considering the optimisation over a range of number of input days, providing a better understanding of the interactions between parameters, but combining this with an optimisation technique, such as a grid search or genetic algorithm would allow users to definitively fix the ANN structure for a selected number of input days.

In addition, a very wide range of substrates were evaluated as part of this study. Pure carbohydrates, proteins and fats were evaluated and these each have a distinct digestion kinetic profile. Mixtures of these are much more common in AD substrates. Although this demonstrates the ability of ANNs to model a wide variety of inputs, narrowing the range of these to those relevant to an industrial AD would likely produce a more relevant model with potentially improved accuracy.

Another improvement would be the inclusion of biogas composition as well as total production as an output of the ANN. To achieve this, online gas analysis would be required to provide the accuracy and frequency of measurements required (particularly when using hourly data). Such capability is not currently common standard equipment and therefore was deem infeasible for this study. Other works have reported lower accuracy when predicting methane yield compared to biogas (Holubar et al., 2000; Mehryar et al., 2017a), likely due to limited availability of measurements.

6. Conclusion

An ANN has been shown to accurately predict biogas production of single substrates and mixtures of food waste components. A large data set of biogas potential tests (446 experiments), analysing various substrates (carbohydrates, proteins, and fats) was used to train the ANN. To determine optimal model architecture, RSM was used to evaluate hyperparameters (hidden nodes, regularisation), input data set (3-14 days), training algorithm (adam, SGD and LBFGS) and activation function (identity, logistic, ReLU and tanh). The optimal conditions (lowest MSE and maximised r^2) were achieved using adam and ReLU and SGD and ReLU, although the former required lower computational load.

7. Acknowledgements

The research received support from the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R512143/1.

8. References

Aklilu, E.G., Waday, Y.A., 2021. Optimizing the process parameters to maximize biogas yield from anaerobic co-digestion of alkali-treated corn stover and poultry manure using artificial neural network and response surface methodology. Biomass Convers. Biorefinery 13, 12527–12540. https://doi.org/10.1007/s13399-021-01966-0

Al, R., Behera, C.R., Zubov, A., Gernaey, K. V, Sin, G., 2019. Meta-modeling based efficient global sensitivity analysis for wastewater treatment plants – An application to the BSM2 model. Comput. Chem. Eng. 127, 233–246. https://doi.org/https://doi.org/10.1016/j.compchemeng.2019.05.015

Almomani, F., 2020. Prediction of biogas production from chemically treated co-digested agricultural waste using artificial neural network. Fuel 280, 118573. https://doi.org/10.1016/j.fuel.2020.118573

Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2018. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. J. Choice Model. 28, 167–182. https://doi.org/10.1016/j.jocm.2018.07.002

Anaero Technology, n.d. Automatically Fed Digesters [WWW Document]. URL https://www.anaerotech.com/automatically-fed-digesters (accessed 4.6.22).

Andrade Cruz, I., Chuenchart, W., Long, F., Surendra, K.C., Renata Santos Andrade, L., Bilal, M., Liu, H., Tavares Figueiredo, R., Khanal, S.K., Fernando Romanholo Ferreira, L., 2022. Application of machine learning in anaerobic digestion: Perspectives and challenges. Bioresour. Technol. 345, 126433. https://doi.org/10.1016/j.biortech.2021.126433

Antwi, P., Li, Jianzheng, Meng, J., Deng, K., Koblah Quashie, F., Li, Jiuling, Opoku Boadi, P., 2018. Feedforward neural network model estimating pollutant removal process within mesophilic upflow anaerobic sludge blanket bioreactor treating industrial starch processing wastewater. Bioresour. Technol. 257, 102–112. https://doi.org/10.1016/j.biortech.2018.02.071

APHA, 1999. Standard Methods for the Examination of Water and Wastewater. American Public Health Association, Washington, DC.

Beltramo, T., Ranzan, C., Hinrichs, J., Hitzmann, B., 2016. Artificial neural network prediction of the biogas flow rate optimised with an ant colony algorithm. Biosyst. Eng. 143, 68–78. https://doi.org/10.1016/j.biosystemseng.2016.01.006

Berlyand, L., Jabin, P.-E., 2023. Mathematics of Deep Learning. Walter de Gruyter GmbH, Boston, USA.

Cady, F., 2017. The Data Science Handbook, The Data Science Handbook. John Wiley & Sons, Hoboken, NJ, USA. https://doi.org/10.1002/9781119092919

Casallas-Ojeda, M., Soto-Paz, J., Alfonso-Morales, W., Oviedo-Ocaña, E.R., Komilis, D., 2021. Optimization of Operational Parameters during Anaerobic Co-digestion of Food and Garden Waste. Environ. Process. 8, 769–791. https://doi.org/10.1007/s40710-021-00506-2

Chen, W.Y., Chan, Y.J., Lim, J.W., Liew, C.S., Mohamad, M., Ho, C.D., Usman, A., Lisak, G., Hara, H., Tan, W.N., 2022. Artificial Neural Network (ANN) Modelling for Biogas Production in Pre-Commercialized Integrated Anaerobic-Aerobic Bioreactors (IAAB). Water 14, 1410. https://doi.org/10.3390/w14091410

Da Silva, C., Astals, S., Peces, M., Campos, J.L., Guerrero, L., 2018. Biochemical methane potential (BMP) tests: Reducing test time by early parameter estimation. Waste Manag. 71, 19–24. https://doi.org/10.1016/j.wasman.2017.10.009

Das Ghatak, M., Ghatak, A., 2018. Artificial neural network model to predict behavior of biogas production curve from mixed lignocellulosic co-substrates. FUEL 232, 178–189. https://doi.org/10.1016/j.fuel.2018.05.051

Dibaba, O.R., Lahiri, S.K., T'Jonck, S., Dutta, A., 2016. Experimental and Artificial Neural Network Modeling of a Upflow Anaerobic Contactor (UAC) for Biogas Production from Vinasse. Int. J. Chem. React. Eng. 14, 1241–1254. https://doi.org/10.1515/ijcre-2016-0025

Enitan, A.M., Adeyemo, J., Swalaha, F.M., Kumari, S., Bux, F., 2017. Optimization of biogas generation using anaerobic digestion models and computational intelligence approaches. Rev. Chem. Eng. 33, 309–335. https://doi.org/10.1515/revce-2015-0057

Fernandes, L.S.M., 2014. Modeling Anaerobic Digestion with Artificial Neural Networks. Unpubl. Case Study 1–9.

Freddi, A., Salmon, M., 2019. Design of experiment, Springer Tracts in Mechanical Engineering. https://doi.org/10.1007/978-3-319-95342-7_6

Goodfellow, I., Bengio, Y., Courvill, A., 2016. Deep Learning. The MIT Press, Cambridge, Massachusetts.

Guclu, D., Yilmaz, N., Ozkan-Yucel, U.G., 2011. Application of neural network prediction model to full-scale anaerobic sludge digestion. J. Chem. Technol. Biotechnol. 86, 691–698. https://doi.org/10.1002/jctb.2569

Holliger, C., Alves, M., Andrade, D., Angelidaki, I., Astals, S., Baier, U., Bougrier, C., Buffière, P., Carballa, M., De Wilde, V., Ebertseder, F., Fernández, B., Ficara, E., Fotidis, I., Frigon, J.C., De Laclos, H.F., Ghasimi, D.S.M., Hack, G., Hartel, M., Heerenklage, J., Horvath, I.S., Jenicek, P., Koch, K., Krautwald, J., Lizasoain, J., Liu, J., Mosberger, L., Nistor, M., Oechsner, H., Oliveira, J.V., Paterson, M., Pauss, A., Pommier, S., Porqueddu, I., Raposo, F., Ribeiro, T., Pfund, F.R., Strömberg, S., Torrijos, M., Van Eekert, M., Van Lier, J., Wedwitschka, H., Wierinck, I., 2016. Towards a standardization of biomethane potential tests. Water Sci. Technol. 74, 2515–2522. https://doi.org/10.2166/wst.2016.336

Holubar, P., Sani, L., Hager, M., Froschl, W., Radak, Z., Braun, R., 2000. Modelling of anaerobic digestion using self-organizing maps and artificial neural networks. Water Sci. Technol. 41, 149–156. https://doi.org/https://doi.org/10.2166/wst.2000.0259

Hunter, S.M., Blanco, E., Borrion, A., 2021. Expanding the anaerobic digestion map: A review of intermediates in the digestion of food waste. Sci. Total Environ. 767, 144265. https://doi.org/10.1016/j.scitotenv.2020.144265

Jacob, S., Banerjee, R., 2016. Modeling and optimization of anaerobic codigestion of potato waste and aquatic weed by response surface methodology and artificial neural network coupled genetic algorithm. Bioresour. Technol. 214, 386–395. https://doi.org/https://doi.org/10.1016/j.biortech.2016.04.068

Khashaba, N.H., Ettouney, R.S., Abdelaal, M.M., Ashour, F.H., El-Rifai, M.A., 2022. Artificial neural network modeling of biochar enhanced anaerobic sewage sludge digestion. J. Environ. Chem. Eng. 10, 107988. https://doi.org/10.1016/j.jece.2022.107988

Khashei, M., Bijari, M., 2010. An artificial neural network (p, d, q) model for timeseries forecasting. Expert Syst. Appl. 37, 479–489. https://doi.org/10.1016/j.eswa.2009.05.044

Koch, K., Hafner, S.D., Astals, S., Weinrich, S., 2020. Evaluation of common supermarket products as positive controls in biochemical methane potential (BMP) tests. Water 12, 1223. https://doi.org/10.3390/W12051223

Long, F., Fan, J., Xu, W., Liu, H., 2022. Predicting the performance of medium-chain carboxylic acid (MCCA) production using machine learning algorithms and microbial community data. J. Clean. Prod. 377, 134223. https://doi.org/10.1016/j.jclepro.2022.134223

Lowe, A., Lawless, S., 2021. Artificial Intelligence Foundations - Learning from Experience. BCS

The Chartered Institute for IT, Swindon, UK.

Lujan-Moreno, G.A., Howard, P.R., Rojas, O.G., Montgomery, D.C., 2018. Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. Expert Syst. Appl. 109, 195–205. https://doi.org/10.1016/j.eswa.2018.05.024

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ. Model. Softw. 25, 891–909. https://doi.org/10.1016/j.envsoft.2010.02.003

Mckenzie, M.C., Mcdonnell, M.D., 2023. Hperparameter Selection Selection in in Reinforcement Reinforcement Learning Learning Using Using the the " Design of Experiments " Method. Procedia Comput. Sci. 222, 11–24. https://doi.org/10.1016/j.procs.2023.08.140

Mehryar, E., Ding, W., Hemmat, A., Hassan, M., Talha, Z., Kafashan, J., Huang, H., 2017a. Modeling and multiresponse optimization for anaerobic codigestion of oil refinery wastewater and chicken manure by using artificial neural network and the taguchi method. Biomed Res. Int. 2017. https://doi.org/10.1155/2017/2036737

Mehryar, E., Ding, W., Hemmat, A., Talha, Z., Hassan, M., Mamat, T., Hei, K., 2017b. Anaerobic co-digestion of oil refinery wastewater with bagasse; evaluating and modeling by neural network algorithms and mathematical equations. BioResources 12, 7325–7340. https://doi.org/10.15376/biores.12.4.7325-7340

Mougari, N.E., Largeau, J.F., Himrane, N., Hachemi, M., Tazerout, M., 2021. Application of artificial neural network and kinetic modeling for the prediction of biogas and methane production in anaerobic digestion of several organic wastes. Int. J. Green Energy 18, 1584–1596. https://doi.org/10.1080/15435075.2021.1914630

Nagy, Z., 2018. Artificial Intelligence and Machine Learning Fundamentals. Packt Publishing, Birmingham, UK.

Nair, V. V., Dhar, H., Kumar, S., Thalla, A.K., Mukherjee, S., Wong, J.W.C.C., 2016. Artificial neural network based modeling to evaluate methane yield from biogas in a laboratory-scale anaerobic bioreactor. Bioresour. Technol. 217, 90–99. https://doi.org/10.1016/j.biortech.2016.03.046

Nguyen, V.T., Ta, Q.T.H., Nguyen, P.K.T., 2022. Artificial intelligence-based modeling and optimization of microbial electrolysis cell-assisted anaerobic digestion fed with alkaline-pretreated waste-activated sludge. Biochem. Eng. J. 187, 108670. https://doi.org/10.1016/j.bej.2022.108670

Pomeroy, B., Grilc, M., Likozar, B., 2022. Artificial neural networks for bio-based chemical production or biorefining: A review. Renew. Sustain. Energy Rev. 153. https://doi.org/10.1016/j.rser.2021.111748

Ponsa, S., Gea, T., Sanchez, A., 2011. Short-time estimation of biogas and methanepotentials from municipal solid wastes. J. Chem. Technol. Biotechnol. 86, 1121–1124. https://doi.org/https://doi.org/10.1002/jctb.2615

Saghouri, M., Abdi, R., Ebrahimi-Nik, M., Rohani, A., Maysami, M., 2020. Modeling and optimization of biomethane production from solid-state anaerobic co-digestion of organic fraction municipal solid waste and other co-substrates. Energy Sources, Part A Recover. Util. Environ. Eff. https://doi.org/10.1080/15567036.2020.1767728

Sakshii, 2023. Adam optimizer: A Quick Introduction [WWW Document]. AskPython. URL https://www.askpython.com/python/examples/adam-optimizer (accessed 8.28.23).

Salam, A., Hibaoui, A. El, Saif, A., 2021. A comparison of activation functions in multilayer neural network for predicting the production and consumption of electricity power. Int. J. Electr. Comput. Eng. 11, 163–170. https://doi.org/10.11591/ijece.v11i1.pp163-170

scikit-learn developers, 2023. 1.17. Neural network models (supervised) [WWW Document]. scikit-learn. URL https://scikit-learn.org/stable/modules/neural_networks_supervised.html (accessed 9.3.23).

Strömberg, S., Nistor, M., Liu, J., 2015. Early prediction of Biochemical Methane Potential through statistical and kinetic modelling of initial gas production. Bioresour. Technol. 176, 233–241. https://doi.org/10.1016/j.biortech.2014.11.033

Tarca, A.L., Carey, V.J., Chen, X. wen, Romero, R., Drăghici, S., 2007. Machine learning and its applications to biology. PLoS Comput. Biol. 3, 0953–0963. https://doi.org/10.1371/journal.pcbi.0030116

Wang, H., Czerminski, R., Jamieson, A.C., 2021. Neural Networks and Deep Learning, in: The Machine Age of Customer Insight. Emerald Publishing Limited, Cham, Switzerland, pp. 91–101. https://doi.org/10.1108/978-1-83909-694-520211010

Wang, X., Bai, X., Li, Z., Zhou, X., Cheng, S., Sun, J., Liu, T., 2018. Evaluation of artificial neural network models for online monitoring of alkalinity in anaerobic co-digestion system. Biochem. Eng. J. 140, 85–92. https://doi.org/10.1016/j.bej.2018.09.010

Wu, W., Dandy, G.C., Maier, H.R., 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ. Model. Softw. 54, 108–127. https://doi.org/10.1016/j.envsoft.2013.12.016

Xu, F., Li, Yangyang, Ge, X., Yang, L., Li, Yebo, 2018. Anaerobic digestion of food waste – Challenges and opportunities. Bioresour. Technol. https://doi.org/10.1016/j.biortech.2017.09.020

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S., Weinan, E., 2020. Towards theoretically understanding why SGD generalizes better than ADAM in deep learning. Adv. Neural Inf. Process. Syst. 2020-Decem, 19–21.

Figure 1: A visual representation of the artificial neural network structure used in this study, where nodes in the input, hidden and output layers are represented by I_n, H_n and O_n respectively, showing how biogas potential data is partitioned to form the input and outputs layers.

Figure 2: a) Effect of training algorithm on MSE, based on 12 executions of the 27 experiments in the CCF design (324 simulations) varying hidden layers and number of nodes, number of days input data and activation function and b) Effect of activation function on MSE, based on 8 executions (LBGFS results have been excluded) of the 27 experiments in the CCF design (216 simulations) varying hidden layers and number of nodes, number of days input data and training algorithm.

Figure 3: Contour plot of RSM models showing MSE at varying numbers of nodes in hidden layers, input days and regularisation (alpha) using a) SGD and ReLU b) adam and ReLU c) SGD and tanh & d) adam and tanh.

Figure 4: Saddle plots for MSE of all (light grey), training (dark grey) and validation (grey) data, with varying numbers of nodes in each layer, and optimal alpha for a) SGD and 3 days of input data, b) adam and 3 days of input data, c) SGD and 14 days of input data and d) adam and 14 days of input data.

Figure 5: Measured biogas production and predicted values (using optimal ANN structure) for a) SGD and 3 days of input data, b) adam and 3 days of input data, c) SGD and 7 days of input data, d) adam and 7 days of input data, e) SGD and 14 days of input data and f) adam and 14 days of input data

Figure 6: A sample from the test group of biogas potential raw data and associated predictions by the trained MLP using 3, 7 and 14 days of input data

Figure 7: Percentage error for different substrate groups when using a) 3 days, b) 7 days and c) 14 days of input data

Table 1: Full CCF design used to build the RSM model for the ANN MSE with varying hyperparameters. Design was applied to train and test an ANN using every combination of training algorithm (adam, SGD, LBFGS) and activation function (identity, logistic, ReLU, tanh).

Table 2: Accuracy of RSM model predictions of the ANN MSE across a range of hyperparameters when using different combinations of training algorithms (adam and SGD) and activation functions (ReLU and tanh)

Table 3: Hyperparameters for optimal ANN structure (as defined by RSM) using 3, 7 and 14 days of input data for a) adam and b) SGD

Table 4: Number of epochs required to reach optimal ANN configuration using SGD and adam training algorithms.

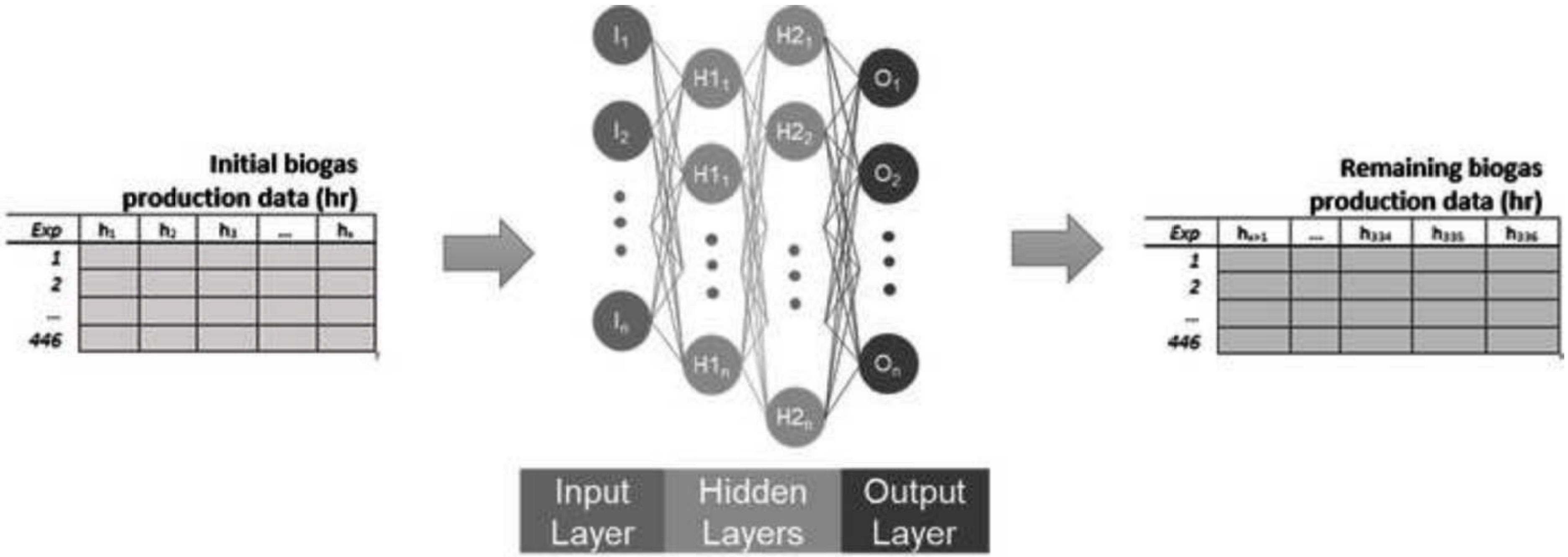
Declaration of interests

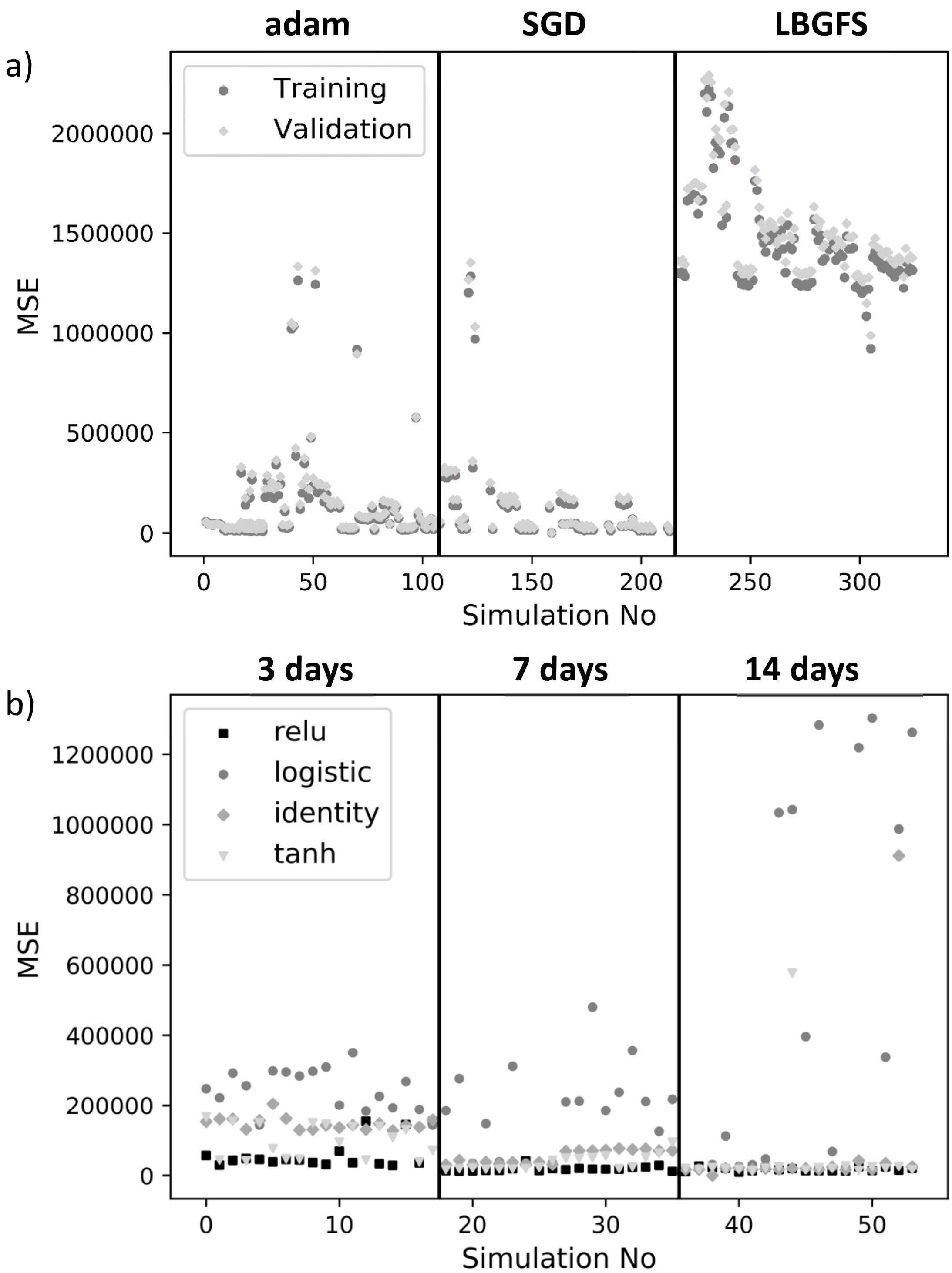
☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

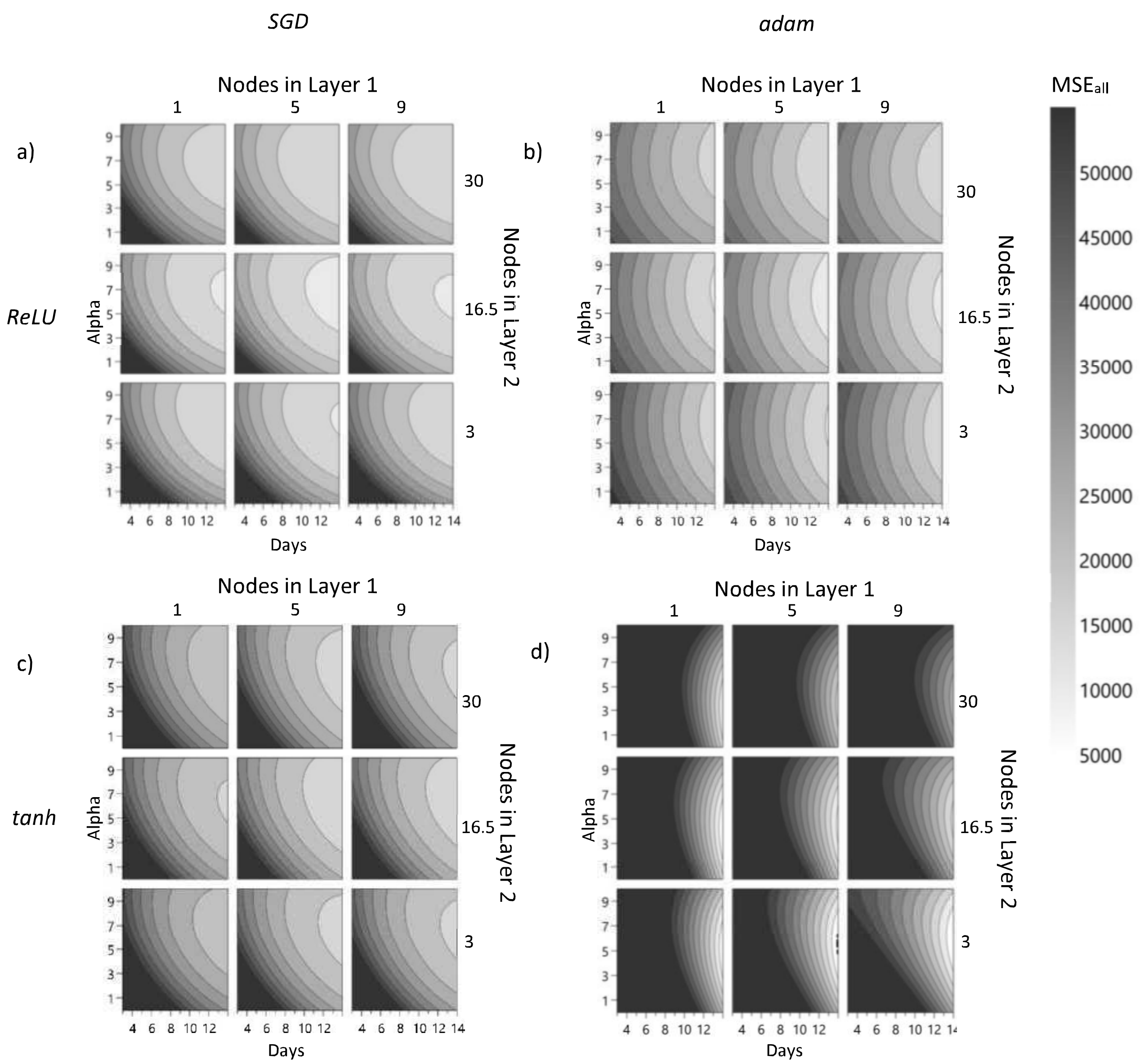
☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Click here to access/download
Electronic Annex
SupplementaryData_ExperimentSummary.xlsx

Figure1


Initial biogas production data (hr)
Exp
1
2
...
446
Input Layer
Hidden Layers
Output Layer
Remaining biogas production data (hr)


a)
adam
SGD
LBGFS
Training
Validation
MSE
2000000
1500000
1000000
500000
0
0
50
100
150
200
250
300
Simulation No
b)
3 days
7 days
14 days
relu
logistic
identity
tanh
MSE
1200000
1000000
800000
600000
400000
200000
0
0
10
20
30
40
50
Simulation No


SGD
adam
a)
b)
c)
d)
ReLU
tanh
Nodes in Layer 1
1
5
9
Nodes in Layer 2
30
16.5
3
Alpha
9
7
5
3
1
Days
4 6 8 10 12 14
MSEall
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000

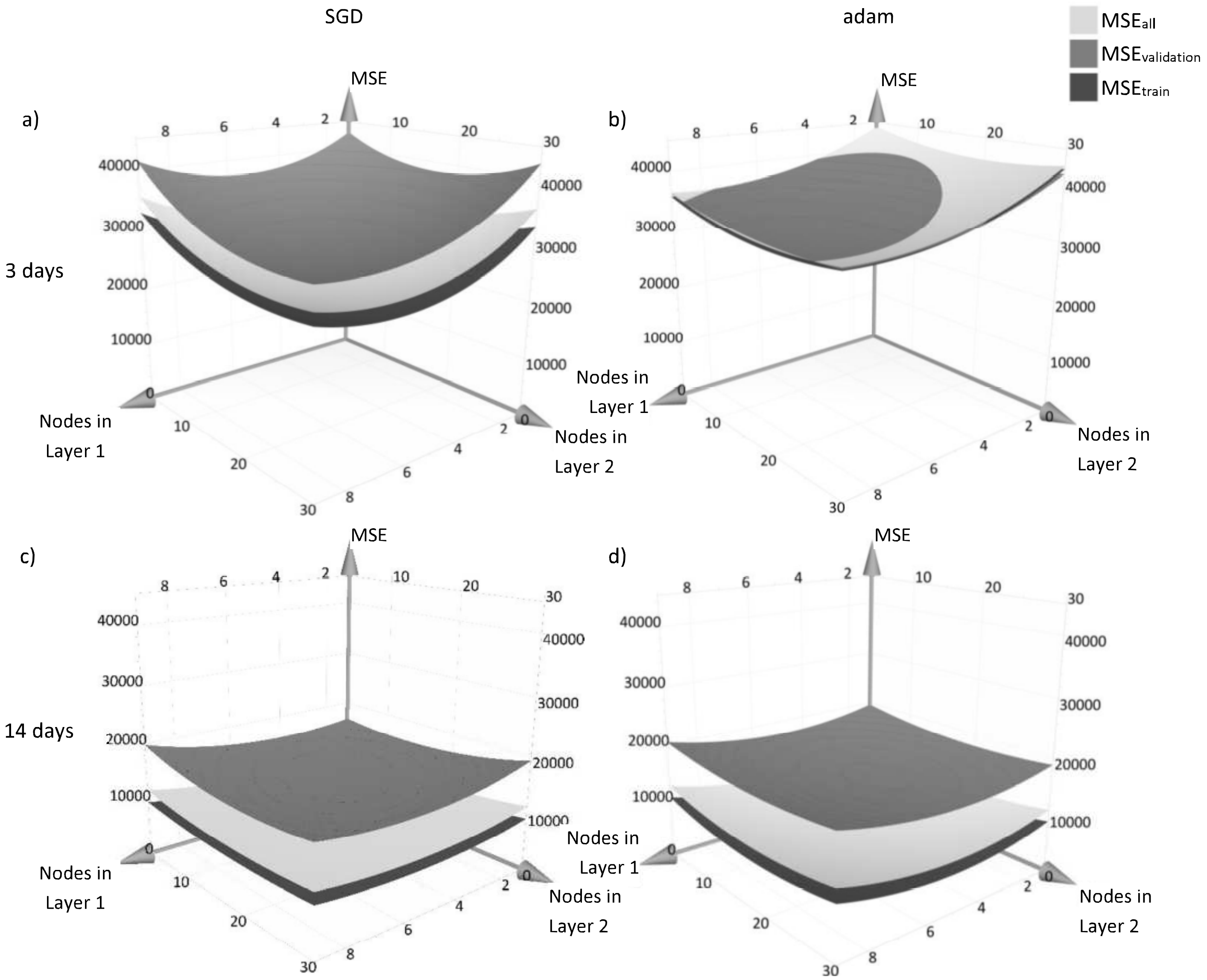

SGD
adam
MSEall
MSEvalidation
MSEtrain
a)
b)
c)
d)
3 days
14 days
MSE
Nodes in Layer 1
Nodes in Layer 2
40000
30000
20000
10000

Figure5

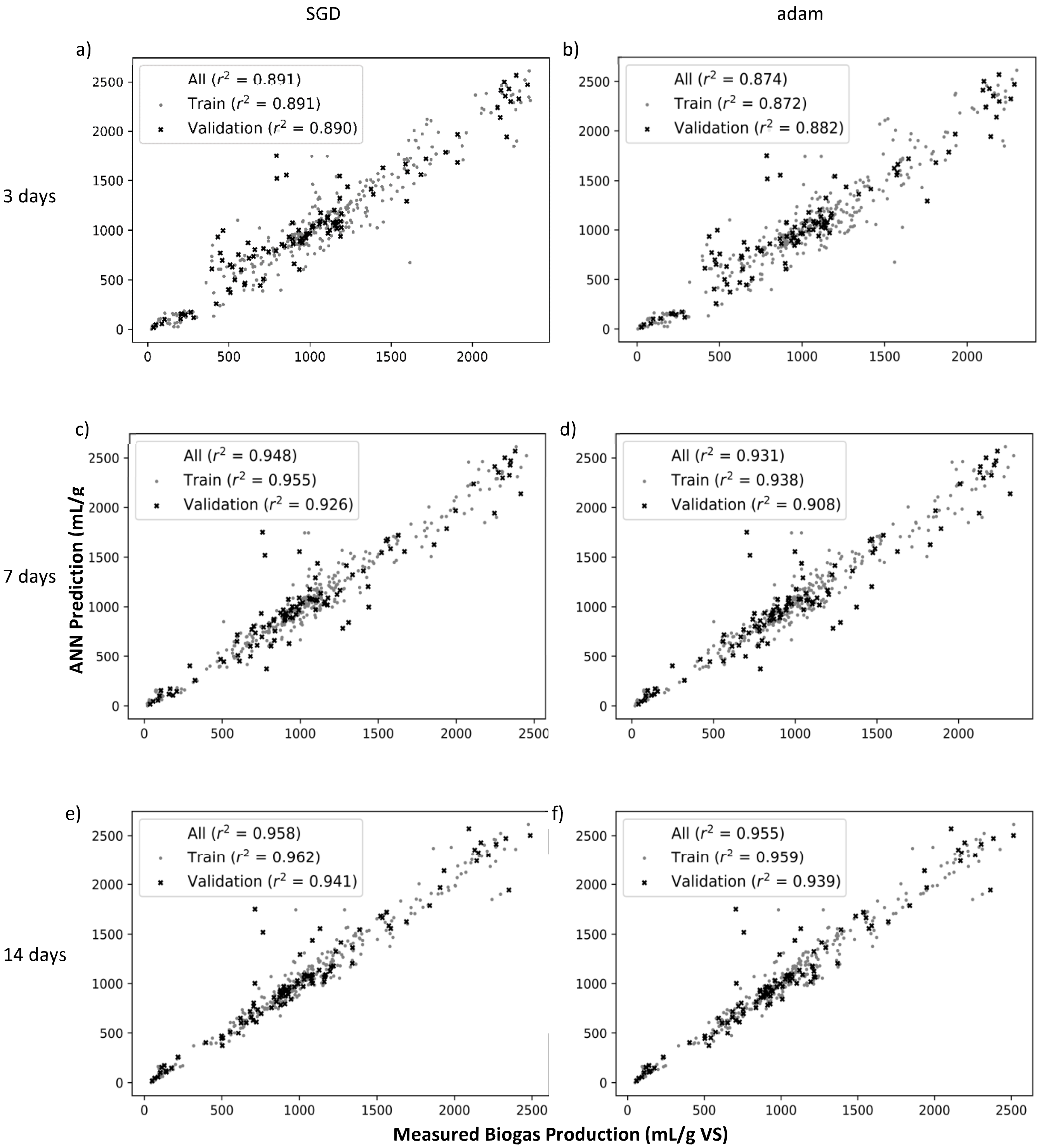
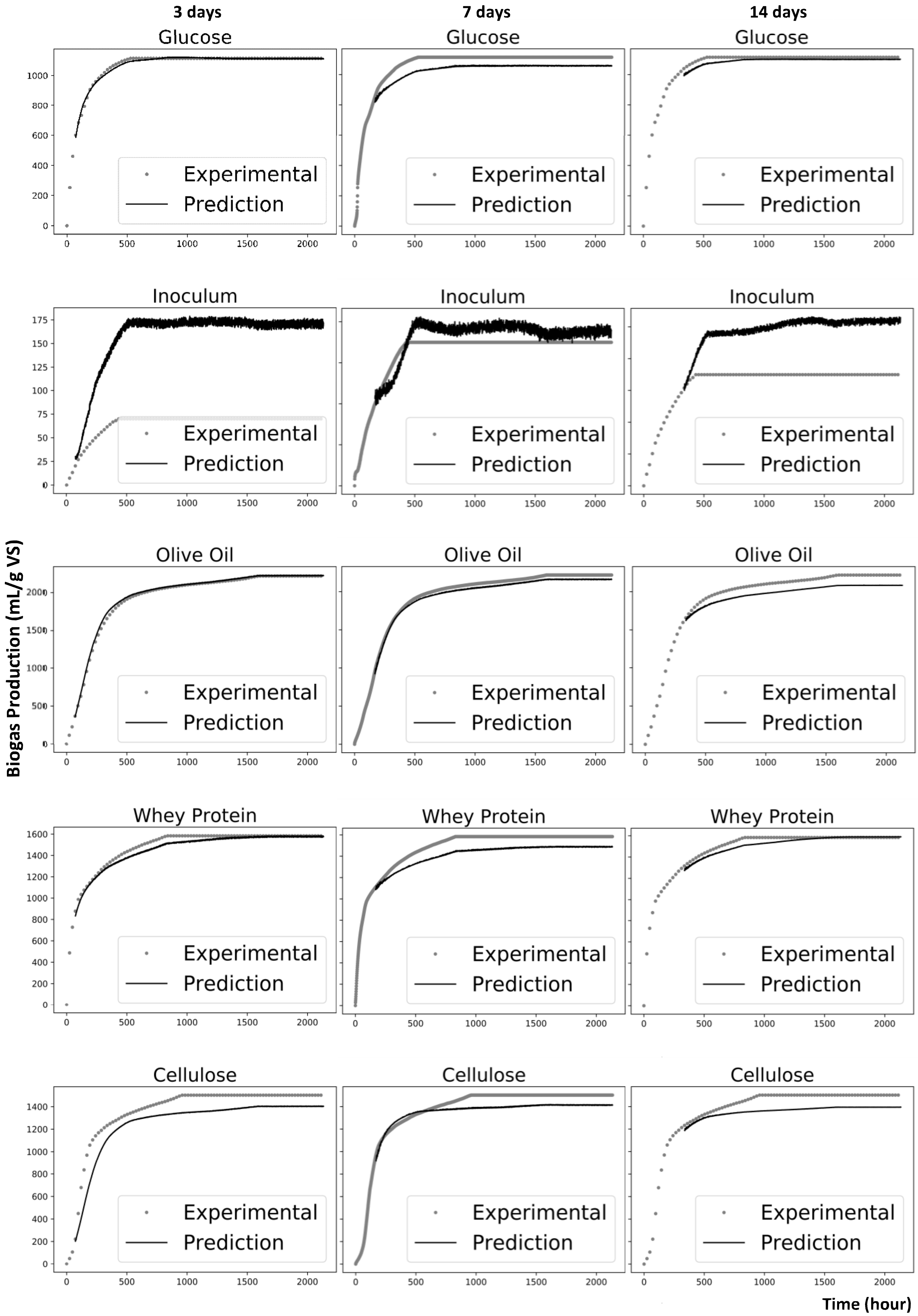
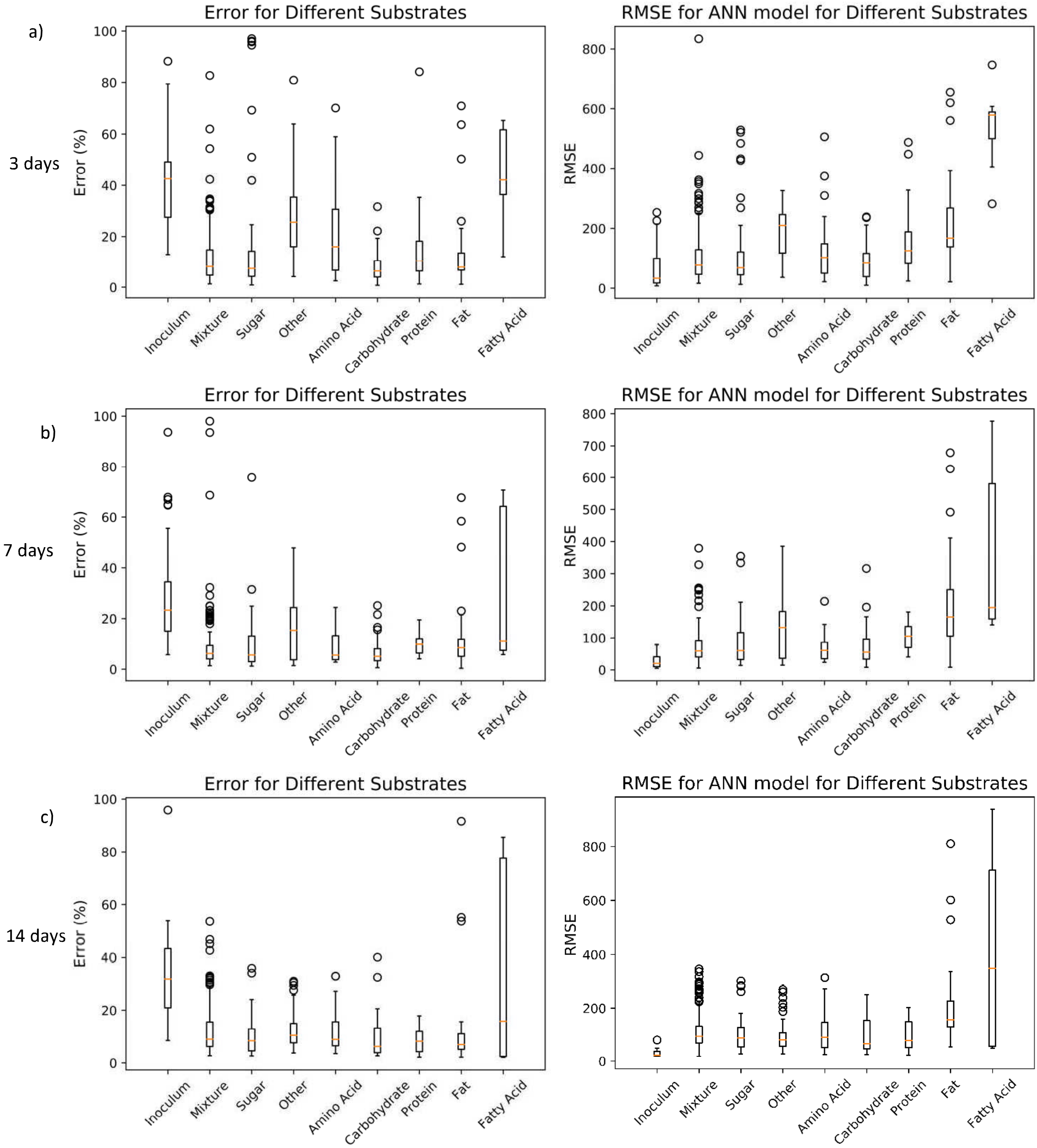

SGD
adam
a)
b)
c)
d)
e)
f)
3 days
7 days
14 days
All (r^2 = 0.891)
Train (r^2 = 0.891)
Validation (r^2 = 0.890)
All (r^2 = 0.874)
Train (r^2 = 0.872)
Validation (r^2 = 0.882)
All (r^2 = 0.948)
Train (r^2 = 0.955)
Validation (r^2 = 0.926)
All (r^2 = 0.931)
Train (r^2 = 0.938)
Validation (r^2 = 0.908)
All (r^2 = 0.958)
Train (r^2 = 0.962)
Validation (r^2 = 0.941)
All (r^2 = 0.955)
Train (r^2 = 0.959)
Validation (r^2 = 0.939)
ANN Prediction (mL/g
Measured Biogas Production (mL/g VS)

Figure6


3 days
7 days
14 days
Glucose
Inoculum
Olive Oil
Whey Protein
Cellulose
Experimental
Prediction
Biogas Production (mL/g VS)
Time (hour)
0
500
1000
1500
2000

Fgure7


a)
3 days
b)
7 days
c)
14 days
Error for Different Substrates
RMSE for ANN model for Different Substrates
Error (%)
RMSE
Inoculum
Mixture
Sugar
Other
Amino Acid
Carbohydrate
Protein
Fat
Fatty Acid

Simulation No	Alpha	Nodes in Layer 1 (x input nodes)	Nodes in Layer 2 (x input nodes)	Days
1	0.0001	1	3	3
2	10	1	3	3
3	0.0001	9	3	3
4	10	9	3	3
5	0.0001	1	30	3
6	10	1	30	3
7	0.0001	9	30	3
8	10	9	30	3
9	0.0001	1	3	14
10	10	1	3	14
11	0.0001	9	3	14
12	10	9	3	14
13	0.0001	1	30	14
14	10	1	30	14
15	0.0001	9	30	14
16	10	9	30	14
17	0.0001	5	16.5	5.5
18	10	5	16.5	5.5
19	5.00005	1	16.5	5.5
20	5.00005	9	16.5	5.5
21	5.00005	5	3	5.5
22	5.00005	5	30	5.5
23	5.00005	5	16.5	3
24	5.00005	5	16.5	14
25	5.00005	5	16.5	5.5
26	5.00005	5	16.5	5.5
27	5.00005	5	16.5	5.5

Table2

	adam						*SGD*					
	ReLU			tanh			ReLU			tanh		
	MSE_{all}	MSE_{val}	MSE_{test}	MSEa	MSEv	MSEt	MSEa	MSEv	MSEt	MSEa	MSEv	MSEt
r^2	0.9720	0.9541	0.9734	0.9388	0.9298	0.9417	0.9720	0.9625	0.970	0.9815	0.9694	0.9900
$r^2_{adjusted}$	0.9396	0.9005	0.9425	0.8610	0.8404	0.8677	0.9394	0.9188	0.9341	0.9598	0.9532	0.9846
Q^2	0.8038	0.7388	0.8076	0.5889	0.5188	0.6222	0.8176	0.7812	0.8130	0.8470	0.9522	0.9634
R	0.9820	0.9590	0.9849	0.9991	0.9977	0.9989	0.9974	0.9992	0.9912	0.9726	0.9853	0.9662

	Training algorithm	adam	SGD
	Activation function	ReLU	ReLU
3 days	Nodes in layer 1	7.85	5.99
	Nodes in layer 2	20.48	20.26
	alpha	6.24	8.38
7 days	Nodes in layer 1	6.74	5.61
	Nodes in layer 2	19.39	19.10
	alpha	6.30	7.89
14 days	Nodes in layer 1	4.70	6.65
	Nodes in layer 2	17.70	15.58
	alpha	6.50	6.65

Days	Epochs	
	adam	SGD
3	365	2775
7	621	3673
14	805	4527