Sustainable Energy & Fuels

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Sustainable Energy Fuels*, 2023, 7, 5129

A hybrid mechanistic machine learning approach to model industrial network dynamics for sustainable design of emerging carbon capture and utilization technologies†

Abhimanyu Raj Shekhar,[a] Raghav R. Moar‡[b] and Shweta Singh *[acd]

Industrial networks consist of multiple industrial nodes interacting with each other through material exchanges that support the overall production goal of the network. These industrial networks exhibit complex nonlinear dynamics arising due to the multiscale nature of interactions among industries and the inherent dynamics of each industrial node. Furthermore, these overall dynamics have a significant impact on the sustainable design of these networks, along with the resource consumption and emission dynamics of the overall network. However, understanding the overall dynamics of industrial networks is challenging as digital models do not exist for the whole network dynamics, especially for emerging industrial systems, and simulative analyses of the same can be computationally expensive. To overcome this limitation, we propose a hybrid mechanistic machine learning approach based on data-driven system identification to build surrogate dynamic models of industrial nodes, which can be coupled to evaluate the overall industrial network dynamics. Furthermore, we propose utilizing the overall network dynamics to quantify the dynamic carbon footprint and design of industrial networks for a maximum carbon sink. We apply our methodology to evaluate the dynamic carbon footprint of an algal-biodiesel industrial network comprising 5 separate dynamic industrial systems. The redesign of the network with the modified technological parameters informed by overall network dynamics results in an approximately 2% enhanced CO_2 sequestration rate of 29 750.34 kg h^{-1}, with the net CO_2 footprint being accurately calculated as −1485069.47 kg for 50 hours of operation based on the nonlinear model obtained for the network. The dynamic models were also used to analyze the net neutralization time required to completely remove the energy-related CO_2 emissions using this specific algal biodiesel network for a specific region in a particular year, providing insights into the potential of this technology to meet the climate mitigation goals. Hence, the proposed approach establishes a pathway to evaluate industrial network dynamics for any emerging system by relying on mechanistic models and data-driven system identification and informing the sustainable design of future industrial networks.

Received 9th August 2023
Accepted 18th August 2023

DOI: 10.1039/d3se01032e

rsc.li/sustainable-energy

1 Introduction

Industrial decarbonization is a major objective for meeting the climate change goals aimed towards limiting global warming to 1.5 °C set forth by the IPCC.[1] Industrial emissions for the year 2020 accounted for 1426.2 million metric tonnes CO_2 eq. of GHG emissions globally which are 24% of the total US GHG emissions distinguished by economic sectors.[2] In order to meet the goals of industrial decarbonization, several technologies in the US are being proposed for carbon capture & utilization, primarily being operational in 5 industrial sectors, *viz.* chemical production, hydrogen production, fertilizer production, natural gas processing, and power generation.[3] Furthermore, the transitions away from a fossil fuel-based economy towards waste reutilization for the manufacturing of value-added products are being proposed, for instance, conversion of vegetable oil into bioadsorbents for wastewater treatment,[4] thermolysis of waste plastics to liquid fuel,[5] biofuel production from grape marc,[6] upcycled carbon black in the formation of battery anodes,[7] *etc.* These emerging technologies will be embedded in existing industrial networks, which are a group of industries interacting through the exchange of materials to meet a goal of production.

[a]*Agricultural & Biological Engineering, Purdue University, West Lafayette, USA. E-mail: singh294@purdue.edu*

[b]*Microsoft Research, Hyderabad, India*

[c]*Environmental & Ecological Engineering, Purdue University, West Lafayette, USA*

[d]*Davidson School of Chemical Engineering (By Courtesy), Purdue University, West Lafayette, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3se01032e

‡ The author participated in research as an Undergraduate Researcher hosted by the Purdue Undergraduate Research Experience program.

Such industrial networks are complex dynamic systems that interact on multiple scales, and each node in the network represents an individual industrial process governed by specific chemical, biological, or physical mechanisms. Furthermore, there exists inherent nonlinearity within the dynamics of each industrial node, which results in a nonlinear dynamic relationship between different material streams in the overall production network. Hence, the overall dynamics of the industrial network on the macroscale is driven by the dynamics at each node and interactions among these nodes which determine the overall production dynamics of the network, the exact characteristics and their underlying governing equations are generally unknown. However, understanding the overall dynamics of existing and emerging industrial networks is necessary to accurately quantify the resource consumption and emission dynamics from these networks in long term, which is critical to evaluate the sustainability of these systems as a whole rather than at individual industrial nodes.

There are several systematic techniques like Life Cycle Assessment (LCA),[8] Material Flow Assessment (MFA),[9] Agent Based Modeling (ABM),[10] Statistical Process Control,[11–13] and System Dynamics (SD)[14] that are widely being used for sustainability assessment of industrial systems. One principal limitation of LCA and MFA methods is the lack of accounting for the nonlinear dynamics between various system components.[15–17] Capturing such nonlinear relationship is important for accurate assessment of resource consumption and emissions over time as the system response to changes cannot be linearly extrapolated. While the system dynamics method does account for this dynamic behavior, it relies on causal loop diagrams which are generally unknown for emerging industrial networks.[18] Similarly, ABMs rely on accurate information about the effect of participating components on each other, which is also unknown for emerging industrial networks.[19] Hence, to overcome these gaps in modeling the dynamics of industrial networks for sustainability assessment, we propose a hybrid mechanistic machine learning approach inspired by data-driven system identification. Since the existing methods do not capture the overall dynamics of industrial networks and do not focus on the network design for

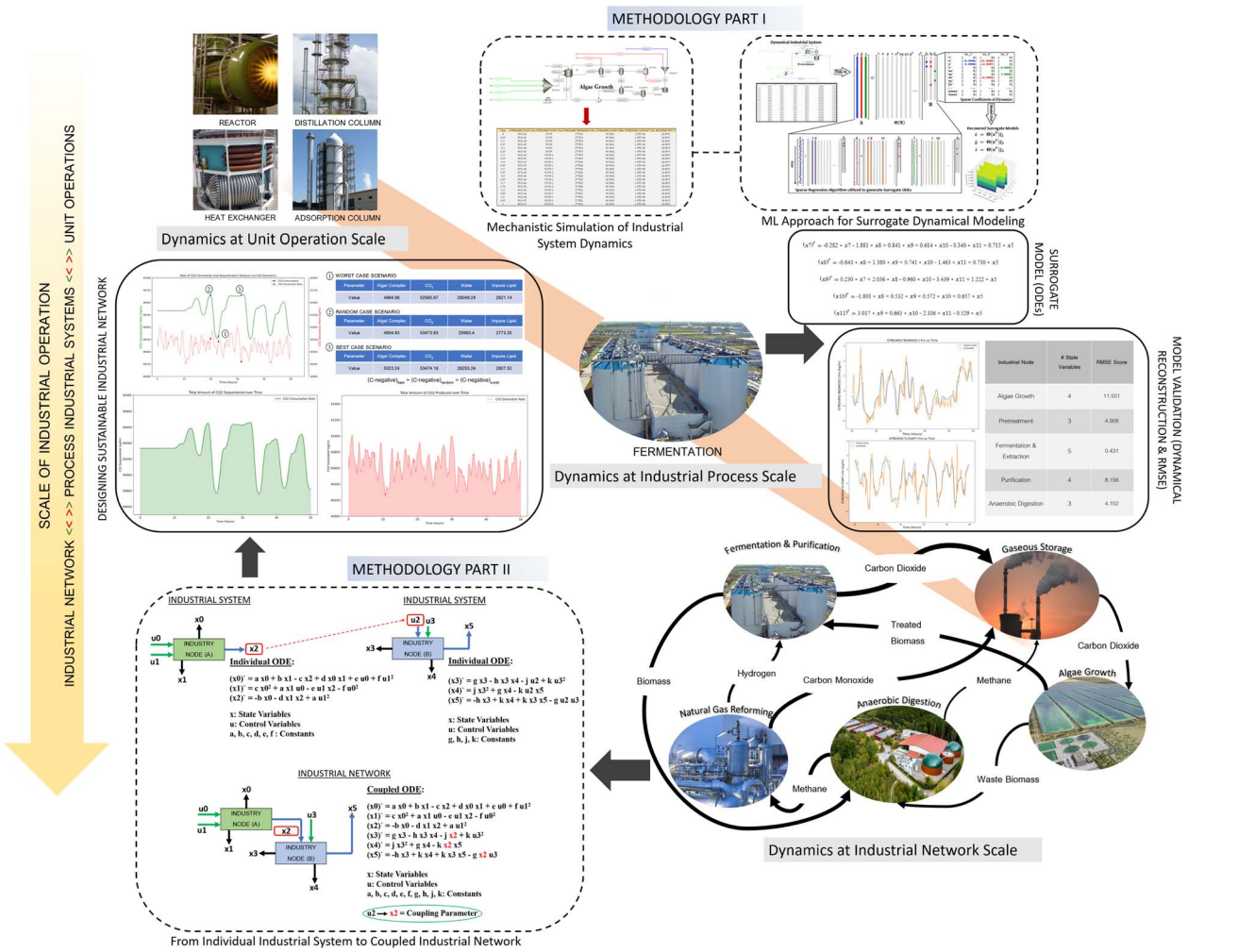


Fig. 1 Industrial production network comprising multiple smaller individual industrial systems operating on different time scales, which further comprises multiple unit operations. The industrial systems are connected together by the material interdependencies existing in a sub-economic supply chain framework.

sustainability, a direct comparative analysis of the proposed approach in this work with existing methods cannot be done.

Data-driven system identification is an age-old technique for solving inverse problems, that can recover the governing equations for the dynamics of a system based on experimental observations and known physical laws.[20,21] The recovery of the underlying mathematical relationship which governs the systems establishes the governing first principles physical, chemical, and biological informed laws. These laws can then be experimentally validated for smaller systems; however for large scale complex industrial networks, it is not feasible to experimentally validate these inverse equations. Therefore, as the direct application of experiment-based system identification is a challenging as well as nearly impractical procedure to apply to the emerging industrial networks, it becomes important to utilize the multiscale nature of these networks for the evaluation of dynamics. Thus, in our approach, we propose to build surrogate models for individual nodes (industrial system) in the industrial network using machine learning and couple these models to study the overall dynamics of the industrial networks. As it is feasible to experimentally validate the individual node models, it provides confidence in using these surrogate models to evaluate the dynamics of the overall complex industrial network.

The field of data-driven system identification itself has been growing rapidly due to the availability of large-scale data and increased computational power in the last few years. Several recent papers have shown the promise of novel data driven system identification in fields such as fluid mechanics,[22] chemical reaction networks,[23–25] single unit operations such as distillation columns,[26,27] chemical process plants,[28] mechanical systems,[29,30] *etc.* These machine learning algorithms seem promising to build surrogate models for the industrial nodes in large scale industrial networks, which can then be utilized to evaluate the overall dynamics of networks without missing any critical causal relationships that drive the dynamics of overall networks. As mechanistic models are known for several of the industrial systems or can be built when a new industrial technology is being proposed, we propose this hybrid approach that utilizes the strength of mechanistic knowledge to generate data and machine learning to create surrogate dynamic models. We utilize the Sparse Identification of Non-Linear Dynamics (SINDy) algorithm in our work, based on its recent success in model identification in several scenarios such as nonlinear optics,[31] thermal fluids,[32] chemical reaction dynamics,[33] structural modelling,[34] models for partial differential equations,[35,36] and stochastic systems.[37] Recently SINDy has also shown promising results in identifying governing equations for the overall dynamics of single-unit operation and multi-unit operation manufacturing systems.[27,28] We demonstrate our hybrid approach on an emerging industrial network of algal biodiesel production, which is a promising technology for carbon capture & utilization and waste water reutilization for value added production in the economy. Fig. 1 shows an example of multiscale interaction in an industrial network for an algal biodiesel network as modeled in our work. It shows interactions all the way from the unit operations to industrial systems and further to the network. The overall dynamics of this network is then used for sustainability assessment of aspects such as dynamic carbon footprint calculation, identifying optimal values for control parameters to design a "net carbon negative" industrial network and evaluating the time feasibility of meeting net zero carbon goals based on the optimal operational design of industries in the network. We demonstrate these applications using the analyses for the modeled algal bio-diesel network.

The methodology proposed is described in detail in Section 2, results of application and sustainability analysis of the algal bio-diesel industrial network are discussed in Section 3 and conclusions along with future applications or advancements of the method are provided in Section 4.

2 Methodology

We propose a two part methodology shown in Fig. 2 to evaluate the overall dynamics of industrial networks. In the first part of the methodology, we propose a three-step procedure to build a surrogate dynamic model for industrial systems in the production network using a hybrid mechanistic machine learning approach (Section 2.1). The second part of the methodology focuses on coupling the individual node models and evaluating overall network dynamics for scenarios related to sustainability assessment of the system (Section 2.2).

2.1 Hybrid mechanistic machine learning for surrogate dynamic model construction of industrial systems

The hybrid mechanistic machine learning approach is developed based on a data-driven system identification technique that relies on supervised machine learning algorithms. Consequently, there are three steps for developing the dynamic models for the industrial systems in the network. Step 1 involves data generation, where we utilize the computationally designed mechanistic models of each industry to generate reliable time-series data of the state variables based on the excitation variables of the system for each block. Next in step 2, we use data from mechanistic models in step 1 and apply a white-box machine learning algorithm to perform system identification and obtain the governing dynamics model as Ordinary Differential Equations (ODEs) for each industrial system. In the final third step, the obtained ODEs from step 2 are numerically integrated, to test the accuracy of these models for dynamic reconstruction. Thus, the models are validated both qualitatively using visual inspection and quantitatively using appropriate accuracy testing metrics. These evaluations are used for accepting the models with appropriate complexity. Details for implementing these steps and considerations for data generation, model training and validation are discussed next.

2.1.1 Step 1 – Data generation using mechanistic models. The system identification technique relies on the availability of correct data that can capture the mechanisms of the process.[38] For any industrial process, the data can be obtained from the real-time operations and the same can be used. But often such data are proprietary in nature and are not available for the use. Additionally, for emerging technologies that are not yet under

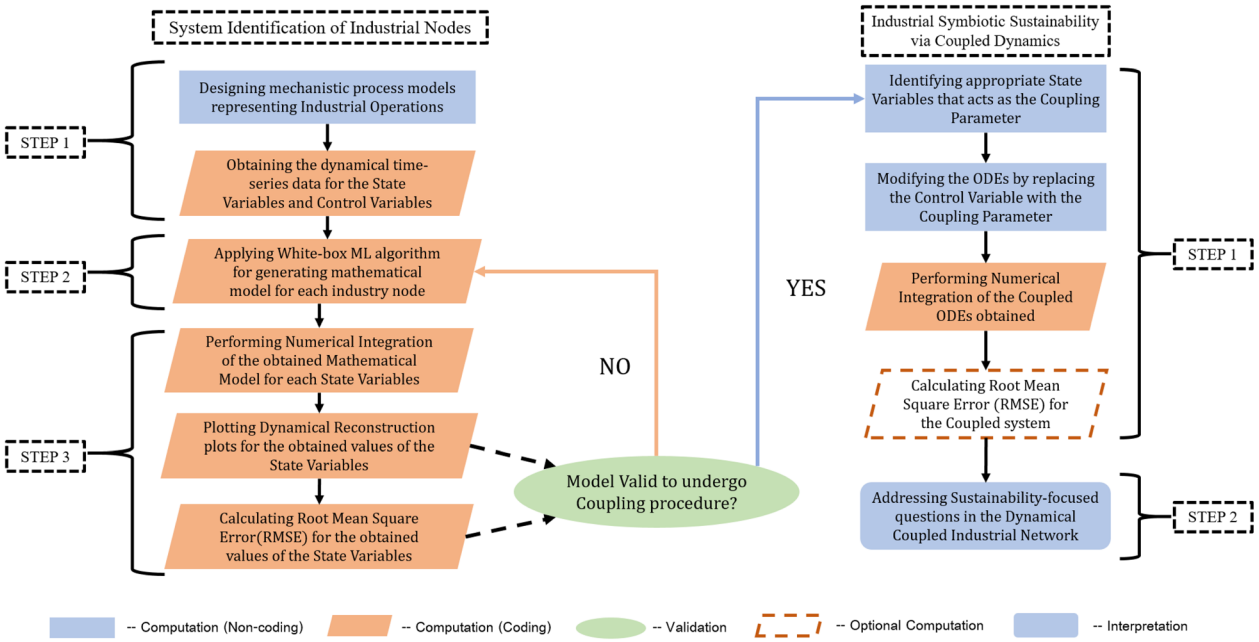


Fig. 2 Methodical schematic of the hybrid mechanistic machine learning approach for industrial network dynamics and sustainability assessment.

commercial operations, such data are not available. Hence, the primary challenge of obtaining data has been proposed to be overcome through the use of mechanistic process simulations. In this work, we rely on Aspen Plus and Aspen Dynamics software, a versatile process simulation tool that can be operated in the static and dynamic modes, respectively. Aspen Plus aids in the design and formulation of a process-scale flow diagram while Aspen Dynamics provides insight into the temporal variation of the state variables (output) upon the excitation of control variables (input). The state variables are defined to be the key parameters of the process system which shows dynamic variation with time and capture the overall state of the system, governed by underlying physical and chemical laws. The control variables are the parameters of a process that can be controlled by the user and the excitation of the same over time results in the dynamic variation pattern of the state variables. For instance, the control variable can be the flow rate or temperature of the input stream, while the state variable can be the flow-rate or temperature of the output stream, the heat duty of the reactor being used in the process as a unit operation, *etc.*

For data generation, first we design a process flow-sheet under "Flow-driven" static conditions using Aspen Plus and thereafter transfer it to the Aspen Dynamics environment. Once the flow-sheet is uploaded in the dynamic environment, the second and most-important step is the qualitative choice of control variables and the state variables that are of interest. Ideally, several system variables are chosen in the state space and based on the non-linear pattern of variation shown by each throughout the dynamic simulation, the state variables are finalized. If a state variable does not show dynamic variation to the excitation used, it is not selected as one of the state variables for model construction. Additionally, among collinear variables only one is selected. In addition to selecting the right state variables from a quantitative *i.e.* system dimensionality, and qualitative *i.e.* model recovery perspective, it is also important to choose appropriate control variables that need to be perturbed for the system. Importantly, such control variables need to be close to the realistic perturbation that occurs during the industrial operations, and have to be meaningful for the later use of models such as in industrial node–node coupling and sustainability analyses.

After the selection of control variables and state variables, excitation input is given to the control variable to generate time series data for state variables in response to the excitation input. In our study, the forcing function for excitation is written in the Aspen Dynamics environment using FORTRAN. Key parameters to set up the dynamic simulation are: (i) type of dynamic jump given to the control variable such as ramp function or sine-ramp function; (ii) total simulation time till which the state variables show dynamics before converging to a stable value; (iii) sampling time for the time stamp to collect data. The sampling time is dependent on the total simulation time and the number of data points required. For instance, if the total simulation time is 100 hours and 1000 data points are desired for training models, the sampling time selected will be 0.1 hours.

Eventually the process flow system is initialized at time $t =$ 0 hours with the chosen forcing function of control variables. The system is then run in the dynamic mode where the dynamic variation time is assigned in a way similar to the time variation that occurs in the actual industrial operation for the corresponding control variables. This is done to maintain the consistency of the computational model with the real world processes. This dynamic simulation generates a time series array of data for the selected state variables and control

variables, which is exported as a CSV file serving to be the input file for the machine learning algorithm.

For the case of material flows occurring within a large complex industrial operation comprising smaller sub-industries, a decoupling scheme can be utilized in order to ease the computational load on dynamic data generation providing an easier approach for the convergence of the solution while using Aspen Plus and Aspen Dynamics. In order to decouple the industry into sub-industrial nodes, the selection of nodes for an industrial network is based on the division of the processes in an industry and the material exchanges occurring between different industries, in which case the nodes represent the individual industries that can be easily modeled. To decouple, the material flows from one "industry" to another "industry" within the complex is identified and each industry is modeled separately. This is realistically valid since there exists different processes comprising several unit operations in an industry and each process is connected by shared material interdependencies flowing *via* pipelines.

2.1.2 Surrogate dynamic model construction using the machine-learning approach. In this step, we use a data-driven system identification technique to create surrogate models governing dynamics for each individual industrial node *via* utilizing the time series dynamic data generated in the previous step. Data-driven system identification approaches include a diverse range of techniques like Symbolic regression,[39] Sparse regression,[40] Gaussian processes,[41] Sure-independence-screening sparsifying-operator regressor (SISSO),[42] and Deep learning.[43] For this particular work, we utilize the Sparse identification of non-linear dynamics (SINDy)[40] approach for recovering the governing mathematical model for an industrial node operation. SINDy is based on sparse linear regression that is highly extensible and requires significantly fewer data in comparison to other techniques, for instance, neural networks. It makes two assumptions about the structure of the model; the first one is that only a few essential terms in the space of possible functions actually govern the dynamics of the chosen system, thus reducing the dimensionality of the governing equations making it parsimonious. The second assumption is that the space of possible functions comes from a predefined set of library functions that a user can either choose or define. Both these assumptions hold for a wide range of complex physical systems following reduced order representation for the dynamics of the system. Furthermore, SINDy models have certain advantages over other methods, like having the characteristics of interpretability, tending to generalize the system over a wide range of control dynamics and also preventing the model from overfitting.

The goal of SINDy is to discover a model for dynamic systems in the form given in eqn (1).

$$\frac{\mathrm{d}(x(t))}{\mathrm{d}t} = f(x(t)) \tag{1}$$

where $x(t) \in R^n$ is time-series data of the state variables and the dynamics encoded by the function f.

For modeling of non-linear dynamic systems with a known external forcing function given as $u(t) \in R^n$, $f(x(t),u(t))$ is a linear combination of non-linear functions of $x(t)$ and $u(t)$. This leads to eqn 2

$$\dot{x}_t = \sum_{i=1}^{k} \xi_i \theta_i(x(t), u(t)) \tag{2}$$

where θs are non-linear functions called the candidate terms of $f(x,u)$ that can comprise a user provided function which can be of the form constant, polynomial, Fourier, or any custom defined function as well, while ξs are the coefficients of the terms identified by the SINDy algorithm. It is crucial to choose the library of candidate functions carefully, and can include any function that might describe the data. Additionally, numerical differentiation methods such as finite differences or smoothed finite differences method are used to calculate the time derivative of state variable dynamic data that are used in SINDy.

The system in eqn (2) can be written in terms of these data matrices

$$\dot{X}_t = \Theta(X(t),u(t))\Xi \tag{3}$$

SINDy uses a sparsity-promoting optimization algorithm to identify the sparse matrix of coefficients Ξ, *i.e.* to select only a few model terms from a library of candidate functions. The optimization algorithm is typically based on a regularization approach that can be L1 regularization such as (LASSO)[44] or Sequentially Thresholded Least Squares (STLSQ),[45] L2 regularization (Ridge Regression),[46] TV regularization such as Sparse Relaxed Regularized Regression (SR3),[47] Elastic Nets,[48] *etc.* The sparsity is controlled by the use of λ, which is the regularization hyperparameter. For the implementation of SINDy, we use PySINDy,[49] a python package that provides various tools to apply the SINDy approach for model discovery. PySINDy is scikit-learn compatible and also includes options for user customization. The data from step 1 are fed to the algorithm and models for each industrial node, and are iteratively trained with specific λ values, which are tested and validated in the next step.

2.1.3 Model validation and accuracy estimation of surrogate dynamic models. Finally, we validate the surrogate models for dynamic reconstruction using visual and quantitative approaches. The ODEs obtained from training models are numerically integrated using an Initial Value Problem (IVP) approach. To perform this numerical integration, we have utilized an inbuilt Python function in the PySINDY library. The arguments for the integration function include (i) the initial values of the state variable initialized at $t = 0$ hours, (ii) the array of values of the control variable expanding up till the total dynamic simulation time, and (iii) the array of total time up to which the numerical integration needs to be performed with an appropriate time step. The selection of the integrator available also needs to be done in the function. One can choose to use explicit integration techniques like RK45, DOP853, RK23, *etc.* or implicit integration techniques like Radau, BDF, LSODA, *etc.* The choice of integrator type is correlated with the system of ODEs being stiff (implicit integrators) or non-stiff (explicit integrators). Importantly, depending on the stiffness of the ODE system, the time step for the total time of integration needs to be adjusted as well, where the stiff system usually works best for

the extremely small time step. As the selection is dependent on the type of ODE obtained, there is no generalized approach; however most ODEs potentially can be solved using the available solvers. In the terminating step, the solution of the system of ODEs for each of the state variables is obtained as an array of values spanning across the overall time of integrative simulation, which is also the total dynamic simulation time.

These integrated values from ODEs are used for graphical visualization of dynamic construction against the simulated data from mechanistic models, thus qualitatively testing the accuracy of ODE models to trace the dynamics of the system. Furthermore, the data from ODE integration are also used for the calculation of quantitative metrics for model accuracy testing. From the data obtained *via* the dynamic simulation of the mechanistic models, 10% of the dataset has been used for the quantitative accuracy calculation. There are several metrics available in the machine learning (ML) literature that can be used such as Mean Absolute Error (MAE), Mean Square Error (MSE), R-Squared Error (R^2), Root mean Square Error (RMSE), *etc.* In our study, we have used RMSE to test the accuracy of the obtained ODE models for each industrial node. RMSE is very similar to the Euclidean distance between two data points, except it takes into consideration all the n data points in the test set and heuristically defines the normalized distance between the vector of numerically integrated values and the vector of dynamically simulated values. RMSE is calculated using eqn (4) and the model with minimum RMSE is accepted as the surrogate model to represent the dynamics of the node.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

where $\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots\ldots, \hat{y}_n$: numerically integrated values of the system identified ODEs and $y_1, y_2, y_3, \ldots\ldots, y_n$: values obtained *via* Aspen Dynamics simulation n: number of data points in the class of the test data set.

2.2 Coupled nonlinear dynamics analysis for sustainability assessment of industrial networks

The second part of the methodology focuses on the application of the nonlinear dynamic models obtained in part 1 for sustainability assessment of coupled industrial networks. To achieve this goal, the first step is to couple the models of each industrial node in the production network *via* identification of appropriate coupling parameters in the form of state variables. The second step is the calculation of selected sustainability metrics (*e.g.* resource consumption, emissions, *etc.*) for long term using integrated dynamics for the coupled system and scenario analysis based on varying dynamics of key forcing variables (*e.g.* water availability).

2.2.1 Coupling dynamic models of industrial nodes for overall industrial network dynamics. The validated ODEs upon the accuracy check obtained from the first part represent the governing equations for the individual industry node. In order to couple models to evaluate the coupled dynamics, we first identify the coupling parameters, which is used to modify the mathematical model (ODEs) obtained for the state variables of each individual node. Such modification is performed by replacing the control variable term of the next node with the coupling variable term that establishes a conjoining link between the previous and the next node. The modification of the ODEs in this manner provides a mathematical model structure of higher dimension, which consists of all the state variables of the coupled industrial network. A binary coupling consisting of two individual systems

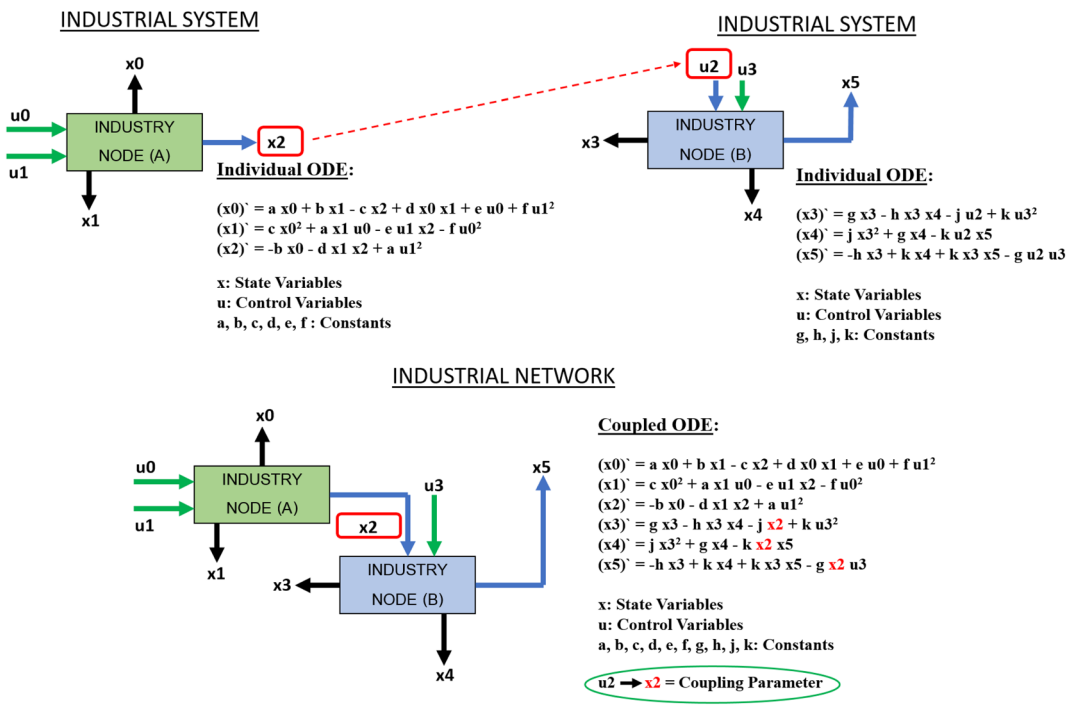


Fig. 3 Variable-transformation coupling of the surrogate models obtained for the individual industrial systems generates a single coupled mathematical model for an industrial network.

has been portrayed in Fig. 3 where the first system consists of 3 state variables and 2 control variables, while the second system also consists of 3 state variables and 2 control variables, but one of the state variables of System A is the control variable of System B, which is the coupling parameter. This can be a direct coupling when industries are co-located and streams are connected or indirect coupling (also called tele-coupling) in the case when industries are at a distance but System A provides the input to System B. Both scenarios can be simulated using the coupled ODE system set up.

With the modified equations to represent overall network dynamics, the coupled ODEs can be integrated for user specified time. The time specified for the numerical integration here should be less than or equal to the time utilized for running the dynamic simulation for the least time-consuming mechanistic process model, since the coupling beyond this time value will not make physical sense for the coupled node operation. For integration, the stiffness of the coupled mathematical model needs to be checked and the type of integrator, whether implicit or explicit, needs to be selected based on the stiffness of the ODEs.

2.2.2 Sustainability assessment of industrial networks *via* coupled dynamics. The overall industrial network dynamics using the coupled ODEs can be utilized to provide insights into sustainability issues including resource utilization, minimization of waste/emissions over time in the whole network, design parameters for each node to attain overall carbon neutrality in the supply chain over time *etc.* In this work, we address three key challenges for the design of sustainable industrial networks utilizing the dynamics of the overall network, as described below.

2.2.2.1 Dynamic carbon footprint of industrial networks. A net dynamic carbon footprint of the whole industrial network can be calculated using the CO_2 sequestration and CO_2 generation profiles for each node in the network. For this, it is proposed to use the coupled surrogate models for resource consumption (when CO_2 is feedstock) and emissions generation (CO_2 as emissions from nodes) integrated over time to calculate the net carbon footprint over time. The benefit of utilizing this coupled dynamics approach is that a more accurate representation of carbon utilization and emissions in the whole network is done by accounting for the non-linear dynamics of each node in the network. This overcomes the existing limitation of carbon footprint calculations in life cycle analysis, which only accounts for static representation of carbon footprint and does not account for nonlinear interactions between different subsystems of the overall network. For the ease of understanding and application, the concept is shown on CO_2 only, which can be extended to other GHGs associated with carbon footprint in future.

2.2.2.2 Identification of optimal control parameter values to create a carbon negative industrial network. From the dynamical graphs for overall CO_2 sequestration and CO_2 emissions, the control variable values at which the overall network attains the state of maximum net negative can be obtained, implying the rate of net CO_2 sequestration becomes highest under the steady state operations. Since these dynamic graphs were obtained by perturbations of the control variables, the time stamp of the net negative system can be used to obtain the parameter values of control variables for all nodes in the network to create an overall carbon negative industrial network. These identified control variables values can then be used to operate the individual nodes at a steady state which will lead to achieve the desired net negative industrial network on the macroscale. We apply this technique to design the algal biodiesel network towards a net-negative industrial network.

2.2.2.3 Quantifying time for sequestering regional energy emissions. With the information on the rate of carbon sequestration in the overall network as discussed above, the total time required to neutralize the energy production related emissions can be calculated. The net neutralization time is defined as eqn (5), where $\tau_{\text{neutralize}}$ is the time needed to neutralize the energy-related CO_2 emissions, $[(C)_{\text{emission}}]_{\text{year}=A}$ is the overall amount of energy-related CO_2 emissions of a particular region extracted from the EIA in a specific year A, and $(C)_{\text{in}}$ and $(C)_{\text{out}}$ are the rates of CO_2 sequestration and generation, respectively, in the industrial network. These values are calculated with the optimal control parameters from the previous step, in kg y^{-1}. The value η signifies the number of similar industrial networks working towards the neutralization of the energy-related CO_2 emission. The assessment of the net neutralization time becomes important to inform technological scale up necessary for meeting the climate mitigation goals. For example, if $\tau_{\text{neutralize}}$ is large such as 100 years, η will need to be increased by higher investment in the specific technology. This can also be used to perform a comparative analysis of time required by different technologies for carbon sequestration.

$$\tau_{\text{neutralize}} = \frac{\left[(C)_{\text{emission}}\right]_{\text{year}=A}}{\eta \times \left[\left(\dot{C}\right)_{\text{in}} - \left(\dot{C}\right)_{\text{out}}\right]} \tag{5}$$

3 Results

The hybridized approach proposed has been demonstrated on the algal biodiesel production network. We consider the conversion of CO_2 and nutrients into biodiesel *via* the utilization of technology involving algal strains, which is a promising carbon capture & utilization technology. Five industrial nodes are considered in this industrial network, where the prime function of each node are algae growth, pretreatment, fermentation & extraction, purification, and anaerobic digestion. Process models for each of these nodes were developed following an overall process given by NREL.[50] The characteristic details of each process are given in ESI Section 2† describing the process flow diagrams of each industrial node in Fig. S2–S6,† along with the initial values of the state space parameters and the characteristics of the forcing function for the control variables. Henceforth, we discuss the results of applying the proposed two part methodology to this industrial production network.

3.1 Surrogate dynamic models for nodes in the algal biodiesel industrial network using the hybrid mechanistic machine learning approach

3.1.1 Step 1: Data generation using mechanistic models. To obtain surrogate dynamic models for the network,

Table 1 Control variables and state variables chosen for each industrial node in the algal biodiesel production network

Industrial node	Time of operation	Control variables (CV)	CV composition	State variables (SV)	SV composition
Algae growth	500 hours	Flowrate stream G120	Algal complex	Flowrate stream BIOMASS	Algal biomass
		Flowrate stream G300	CO_2	Flowrate stream O-EVAP	Gas mixture
				Density stream LOSS	Water loss
				Reactor B4 temperature	Initial algae growth reactor
Pretreatment	400 hours	Flowrate stream BIOMASS	Algal biomass	Flowrate stream TANKPROD	Pretreated slurry
		Flowrate stream 170	Water	Flowrate stream FLASH	Water
				Reactor NH_4TNK temperature	Ammonia mixing reactor
Fermentation & extraction	1000 hours	Flowrate stream TANKPROD	Pretreated slurry	Flowrate stream OILPROD	Algal oil
				Flowrate stream RECTBOT	Water
				Flowrate stream 230	CO_2
				Flowrate stream 500	Extraction mixture
				Flowrate stream O-ETOH	Ethanol
Purification	1000 hours	Flowrate stream OILPROD	Algal oil	Flowrate stream COOLRDB	Renewable diesel blendstock
				Flowrate stream O-NAPTHA	Long carbon chain naptha
				Flowrate stream 510	CO + CO_2 + H_2 + propane
				Flowrate stream 450	Phosphoric acid
Anaerobic	1000 hours	Flowrate stream 510	CO + CO_2 + H_2 + propane	Flowrate stream FLUGASLP	Flue gas
Digestion		Flowrate stream 520	Hexane + lipid impurities + water	Flowrate stream 550	Centrifugation mixture
				Reactor COMBUST temperature	Combustion reactor

mechanistic models (MMs) were designed for each sub-industrial node which is the process flowsheet generated using Aspen Plus. These models were firstly generated at a steady state, and then converted to flow-driven Aspen dynamics models to obtain time series data of each state variable in response to perturbations in control variables (CVs) for each of the industrial nodes, thus facilitating dynamic modeling. The control variables and state variables (SVs) to obtain the time series data for all the 5 industrial nodes are given in Table 1, along with the underlying details for the characteristics of each of the variables such as the composition of the stream. For this work, the CVs have been chosen as the input stream flow rates in each of the nodes that represent material flow streams, therefore providing an ease in the coupling of industrial nodes and drive the overall network dynamics. These CVs were also selected in such a way that the reduced order surrogate models can be used to study the overall material flow dynamics aligned with the sustainability goals such as total carbon sequestration, additional resource consumption and carbon neutrality of the whole production network over long term. Each node was run for different numbers of hours, and the total amount of time for each node is given in Table 1. Based on the dynamic simulation runs, a total of 10 000 data points were generated for each node in the

Table 2 Characteristics of the surrogate model obtained upon the implementation of SINDy, and the performance of the model in terms of the RMSE score metric

Industrial node	State variables	Regularization approach	Regularisation hyperparameter (λ)	Function library	RMSE score	Complexity	Complexity ($\lambda = 0$)
Algae growth	4	STLSQ	8	Polynomial (degree = 3)	11.551	38	336
Pretreatment	3	STLSQ	7	Polynomial (degree = 3)	4.908	45	168
Fermentation & extraction	5	STLSQ	0.2	Polynomial (degree = 3)	0.431	24	140
Purification	4	TrappingSR3	0.1	Fourier (exclude = cosine)	8.156	30	30
Anaerobic digestion	3	STLSQ	3	Polynomial (degree = 2)	4.102	8	63

network, which was fed to the SINDy algorithm. For training the ML model, a total of 9000 data points were used as a training set and 1000 data points were reserved for the surrogate model accuracy testing. The obtained data were standardized before training the model, a standard protocol followed in the domain of ML.

3.1.2 Step 2: Surrogate dynamic model construction for nodes in the algal biodiesel industrial network. Table 2 shows the characteristics of final surrogate models identified for each node using the SINDy algorithm. The final complexity of the models is signified by the number of terms present in the surrogate model with the sparsification approach of SINDy. This implies that the number of terms in the Ordinary Differential Equations (ODEs) for each of the state variables is counted and the net sum calculated is presented as the complexity of the model. Additionally, the column for the model complexity with the regularization parameter $\lambda = 0$ shows the number of terms in the model had it been the case of nonlinear regression and there was no sparsification involved. As can be seen for each node, the nonlinear regularization approach provides a less complex model *i.e.* a reduced order model that captures the overall dynamics of the nodes. This leads to ease of system interpretability and reducing computational load to simulate the dynamics of industrial node coupling. Hence, the approach of creating reduced order surrogate models can provide a powerful tool for studying the overall network dynamics of coupled industrial production networks as compared to mechanistic model simulations over long term due to lower computational needs.

3.1.2.1 Optimization. The surrogate models were obtained using the optimization approach of sequentially thresholded least square (STSLQ)[45] for the algae growth, pretreatment, fermentation and extraction, and anaerobic digestion nodes in the network, and sparse relaxed regularized regression (SR3) approach[47] for the purification node in the SINDy algorithm. These optimizers are computationally more efficient than LASSO and converge in a smaller number of iterations towards a sparse solution, providing bounded ODEs with good accuracy.

3.1.2.2 Regularization for sparsification of models. For obtaining the accurately performing bounded ODEs based on the regularization approach, the selection of appropriate regularization hyperparameter λ becomes important. Fig. 5 shows the selection of λ for the pretreated slurry state variable in the pretreatment industrial node. It can be seen in this figure that model complexity is reduced as the regularization parameter λ increases. However, model accuracy measured by R^2 shows that it remains reasonably close till $\lambda = 7$. Hence, models of lower complexity at $\lambda = 7$ were selected as surrogate models to represent the dynamics of this block. The final values of the regularization hyperparameter λ for all blocks were selected using the same graphical approach and are given in Table 2. This approach is not bound by the choice of the accuracy metric since any other accuracy metric will provide a similar kind of behavior of finding the knee where the accuracy will have a sudden drop at the hyperparameter value. Fig. 5 represents the relationship of model accuracy *vs.* regularization parameter for the pretreatment industrial node in the algae biodiesel network. Additionally, the figure shows the reduction in the complexity of the model as the regularization parameter increases. In the figure, the first knee point for the accuracy plot occurs at $\lambda = 7$, which is then selected as the regularization parameter for the model. The selection of such a non-zero regularization parameter has led to a reduction of model complexity from 168 ($\lambda = 0$) to 45 ($\lambda = 7$).

Fig. 5 also contains the final learned model for the state variable of stream flowrate of pretreatment slurry of the pretreatment node at $\lambda = 0$ and $\lambda = 7$ for comparison. The model at $\lambda = 0$ is the same as the model at $\lambda = 7$ after the first iteration (of the optimisation). Before the second iteration, the coefficients with absolute values less than 7, for instance the terms like $x6$, $x4x6$, $x5^2$ and others, are set to zero. These terms, therefore, don't appear in the final model at $\lambda = 7$. Terms like $x0u2^2$ and $x4^2u2$ with high absolute values at $\lambda = 0$ don't appear in the equation at $\lambda = 7$ because their coefficients in the subsequent iterations have become less than 7, and therefore were set to zero.

3.1.2.3 Model characteristics. For most of the nodes in the algal biodiesel production network, the accurately performing surrogate models had a second or third order polynomial function capturing the nonlinearity of different SVs while the node demonstrating the purification process also shows Fourier terms which possibly capture the periodic behavior of the state variables in the system. A total of 20 ODEs representing the 20 SVs were obtained, which describe the dynamics of each block in the reduced order form driven by the specific set of CVs as given in Table 1 and the model characteristics shown in Table 2.

An example, the surrogate model obtained for the dynamics of the algae growth industrial node is shown in Fig. 4 along with the process flow diagram for this node. This model captures the overall dynamics of the algae growth node as ODEs for the SVs of the biomass flow rate, evaporator loss rate, separator loss stream as density stream, and reactor temperature in response to perturbations to inputs of the flowrate of the algal complex and CO_2. From Fig. 4, it is observed that the surrogate model for the algae growth node has a complexity of 38 (defined as the total number of terms). In contrast, if the regularization approach was not taken and the model is identified by nonlinear regression, the complexity would have been 336, thereby making it a challenging task to analyse and physically interpret the model. Thus, applying the regularization technique reduced the number of terms in the model by almost 10 fold. The identified ODEs for the rest of the industrial nodes that capture the dynamics are given in detail in ESI Section 3.†

3.1.3 Step 3: Model validation and accuracy estimation of surrogate models for the algal biodiesel industrial network. The ODEs obtained as surrogate models for state variables of each node are IVPs which were solved using the LSODA solver of Python. Due to the multiple time scale of variations of different state variables *i.e.* some variables varying at a more rapid rate than other variables in the system, these ODEs were stiff, thus leading to an anomaly during integration. In order to solve the stiffness challenge, the integration for the system of ODEs is therefore performed with an extremely small time step of the order 10^{-5}.[51] For solving the ODEs obtained using the LSODA

$$(x0)' = -40.860\,x1 + -40.041\,u0 + 14.894\,x0^2\,x1 + 31.045\,x0\,x1^2 + 16.129\,x1^3 - 49.031\,x1\,x2\,x3 + 51.280\,x1\,x3\,u1 - 50.611\,x2\,x3\,u0 + 52.873\,x3\,u0\,u1$$

$$(x1)' = 43.519\,x1 + 43.507\,u0 - 47.548\,x0^3 - 73.043\,x0^2\,x1 + 40.590\,x0^2\,u0 - 83.561\,x0\,x1^2 - 32.295\,x1^3 + 62.400\,x1\,x2\,x3 - 65.163\,x1\,x3\,u1 + 65.308\,x2\,x3\,u0 - 68.109\,x3\,u0\,u1 - 14.830\,u0^3$$

$$(x2)' = -27.168\,x2 + 26.829\,u1 + 82.052\,x0^2\,x3 - 90.347\,x0\,x1\,x2 + 78.252\,x0\,x1\,x3 + 61.264\,x0\,x1\,u1 - 76.784\,x0\,x3\,u0 - 26.863\,x0\,u0\,u1 - 69.684\,x1^2\,x2 + 79.140\,x1^2\,u1 - 75.466\,x1\,x3\,u0 + 47.520\,x1\,u0\,u1 + 35.143\,u0^2\,u1$$

$$(x3)' = -55.844\,x1^2\,x2 + 23.995\,x1^2\,u1 - 70.347\,x1\,x2\,u0 - 40.419\,u0^2\,u1$$

$x0 = Stream\ Flowrate\ of\ Algal\ Biomass\ (BIOMASS)$

$x1 = Density\ of\ LOSS\ Stream\ majorly\ comprised\ of\ Fatty\ Acids,\ Glucan,\ and\ Water\ (LOSS)$

$x2 = Stream\ Flowrate\ of\ CarbonDioxide\ and\ Water\ Mixture\ coming\ out\ of\ Evaporator\ (O-EVAP)$

$x3 = Temperature\ of\ the\ Algae\ Culture\ Reactor\ (B4)$

$u0 = Stream\ Flowrate\ of\ Algal\ Complex\ Strain\ (G120)$

$u1 = Stream\ Flowrate\ of\ Carbon\ Dioxide\ (G300)$

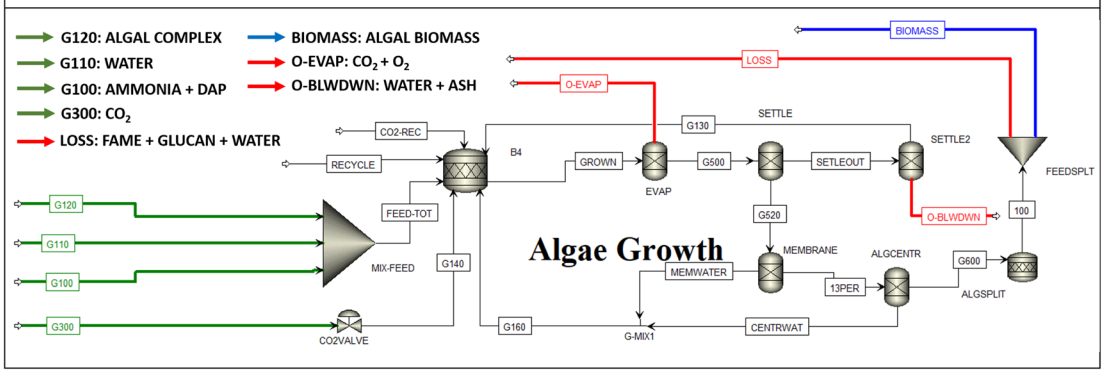


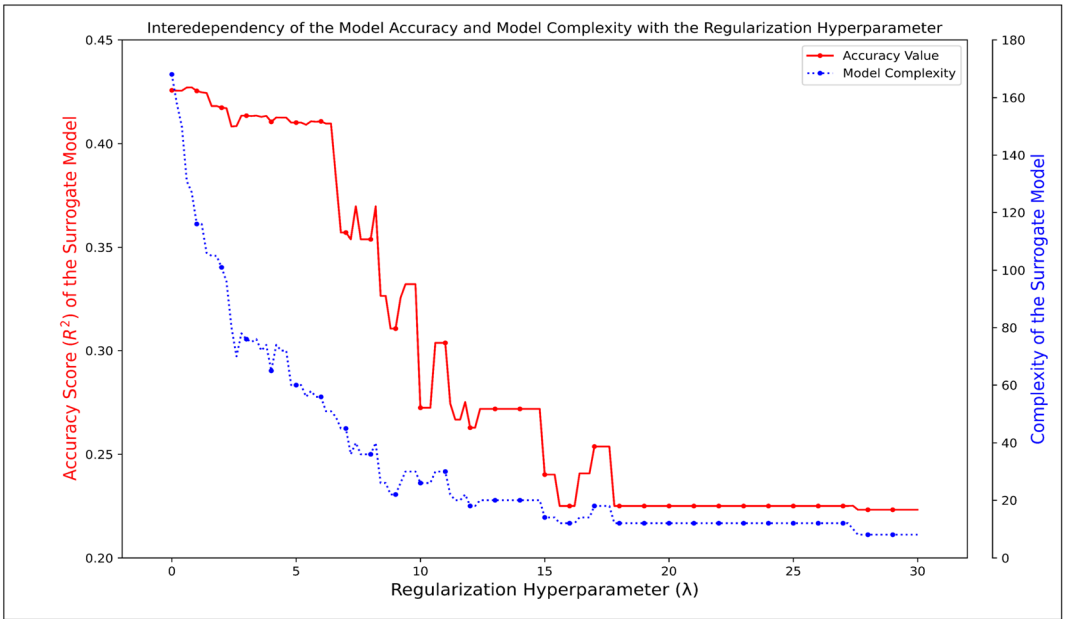
Fig. 4 Surrogate mathematical model for the algae growth node in the algal biodiesel production network.

solver with an extremely small time step, the dataset of control variables has been expanded using the cubic interpolation technique between two subsequent data points in order to generate the optimal number of data points. Since the time step of 0.00005 hours has been used in the integration of ODEs for most of the nodes, the interpolation has been performed to generate t/0.00005 data values, where the 't' signifies the total time of dynamical simulation using Aspen Dynamics.

3.1.3.1 Qualitative testing. The integrated values were used for qualitative accuracy testing by constructing the dynamic reconstruction plots for state variables and comparing against the originally obtained data from mechanistic simulations. The reconstruction plots for each node have been shown in Figures S13–S17 in ESI Section 4,† where the reconstruction has been plotted for the entire dynamic simulation time taken by the individual nodes. An example is shown in Fig. 6, where the reconstruction of state variables for each state variable in the algae growth node has been plotted for a smaller time frame. The graph in blue signifies the originally obtained data from the Aspen Dynamics simulation, while the graph in orange portrays the array of numerically integrated values obtained from the surrogate mathematical models. From the visual representation of the reconstructed plots, it is evident that the surrogate mathematical model performs well for the entire set of the dynamic data that has been obtained for most of the non-temperature based state variables. The poor reconstruction of temperature-based state variable is due to the absence of energy balance in the node. This implies that the mechanistic models that has been designed portrays an optimal material balance but fails to account for the energy balance. The design can be further improved *via* utilization of heat exchangers at the precise location in the node structure flow diagram which will regulate the temperature parameter, thus enabling SINDy to effectively learn from the new training data set for the temperature state variable and allowing a better reconstruction plot for this state variable.

3.1.3.2 Quantitative testing. Integrated values were also used to calculate RMSE for quantitative accuracy testing using eqn (4). RMSE was calculated against the test data set of 1000 data points reserved from the originally obtained data. The model performance is classified to be better when the value of the RMSE score is low. The RMSE score for each node has been tabulated in Table 2.

After the visual and quantitative validation, the ODEs were accepted as surrogate models for further investigation of network dynamics. It is to be noted that most of the white-box ML algorithms which aid in the system identification and interpretation often offer lower accuracy than other black-box predictive algorithms; however they provide ODEs that can be coupled to study overall network dynamics. Additionally, the surrogate models obtained upon the use of the SINDy algorithm also accumulates errors while undergoing the numerical



For the State Variable defining the Stream Flowrate of the Pretreated Slurry

$$When\ Regularization\ Hyperparameter = 7:$$

$$(x5)' = -695.686\,x4 + 6.849\,x5 - 7.979\,x0 + 701.529\,u2 + 7464.239\,x4^3 - 11159.272\,x4^2\,u2 - 161.117\,x4\,x5^2 + 55.215\,x4\,x5\,x0 - 1.048\,x4\,x6^2 - 51.080\,x4\,x6\,x0 + 53.960\,x4\,x0^2 + 161.822\,x5^2\,u2 - 55.673\,x5\,x0\,u2 + 0.328\,x6^2\,u2 + 51.151\,x6\,x0\,u2 - 53.266\,x0^2\,u2 + 3694.917\,u2^3$$

$$When\ Regularization\ Hyperparameter = 0:$$

$$(x5)' = -0.020 - 834.957\,x4 + 7.095\,x5 + 1.332\,x6 - 9.326\,x0 + 841.936\,u2 - 468.638\,x4^2 - 7.717\,x4\,x5 - 2.784\,x4\,x6 + 21.481\,x4\,x0 + 949.467\,x4\,u2 + 0.531\,x5^2 - 1.221\,x5\,x6 + 0.065\,x5\,x0 + 7.980\,x5\,u2 - 0.084\,x6^2 + 0.534\,x6\,x0 + 1.664\,x6\,u2 - 0.077\,x0^2 - 21.162\,x0\,u2 - 480.924\,u2^2 + 660870.145\,x4^3 + 47621.522\,x4^2\,x5 - 6861.791\,x4^2\,x6 - 18179.530\,x4^2\,x0 - 1944287.604\,x4^2\,u2 + 1013.626\,x4\,x5^2 - 490.333\,x4\,x5\,x6 - 654.040\,x4\,x5\,x0 - 93594.198\,x4\,x5\,u2 + 33.824\,x4\,x6^2 + 261.217\,x4\,x6\,x0 + 13417.730\,x4\,x6\,u2 + 80.779\,x4\,x0^2 + 35894.663\,x4\,x0\,u2 + 1906685.416\,x4\,u2^2 - 63.469\,x5^3 - 25.106\,x5^2\,x6 + 136.634\,x5^2\,x0 - 1173.900\,x5^2\,u2 - 0.850\,x5\,x6^2 + 33.151\,x5\,x6\,x0 + 449.513\,x5\,x6\,u2 - 96.076\,x5\,x0^2 + 887.004\,x5\,x0\,u2 + 45837.169\,x5\,u2^2 + 0.356\,x6^3 + 0.786\,x6^2\,x0 - 34.730\,x6^2\,u2 - 10.986\,x6\,x0^2 - 233.534\,x6\,x0\,u2 - 6573.011\,x6\,u2^2 + 22.261\,x0^3 - 163.026\,x0^2\,u2 - 17615.298\,x0\,u2^2 - 623306.372\,u2^3$$

$x4 = STREAMS(\text{" FLASH "}).\ Fm = Water\ Coming\ Out\ of\ the\ Flash\ Unit\ Operation$

$x5 = STREAMS(\text{" TANKPROD "}).\ Fm = Pretreated\ Slurry$

$x6 = BLOCKS(\text{" NH4TNK "}).\ T = Reactor\ Temperature\ for\ Ammonium\ Sulfate\ Production$

$x0 = (Coupling\ Parameter) = STREAMS(\text{" BIOMASS "}).\ Fm = Biomass\ Slurry\ Coming\ From\ Algae\ Growth\ Node$

$u2 = (Control\ Variable) = STREAMS(\text{" 170 "}).\ Fm = Water$

Fig. 5 Correlation of the accuracy and complexity with the regularization hyperparameter. When the hyperparameter exceeds a value of 7, there is a stark drop in the accuracy. The equation below for the pretreated slurry state variable distinguishes between the complexity for hyperparameter values of 7 and 0 respectively.

integration procedure; therefore the low accuracy is also accounted due to the numerical integration operation.

3.2 Coupled nonlinear dynamics analysis for sustainability assessment of the algal biodiesel industrial network

The validated surrogate models as ODEs were next utilized for dynamic sustainability assessment of the overall network. To achieve this goal, we first couple the individual node models for obtaining overall dynamics of network 3.2.1 and then evaluate the dynamic carbon footprint of the overall network along with potential time for sequestering energy related emissions using the algal biodiesel network 3.2.2.

3.2.1 Coupling dynamic models of nodes for overall algal biodiesel network dynamics. In order to couple the surrogate models of individual nodes to obtain overall network dynamics, we first identify the coupling variables to link the dynamics of each node following the approach shown in Fig. 3. The variables mapping to exchange of materials between nodes are used as the coupling variables. For example, the coupling parameter between the Algae Growth Block and the Pretreatment Block in the network is the "BIOMASS" streamflow. Other coupling variables for all the other nodes are shown in Fig. 7, where the green arrows signify the control variable for node dynamics, the black arrows signify the state variables, and the blue arrows

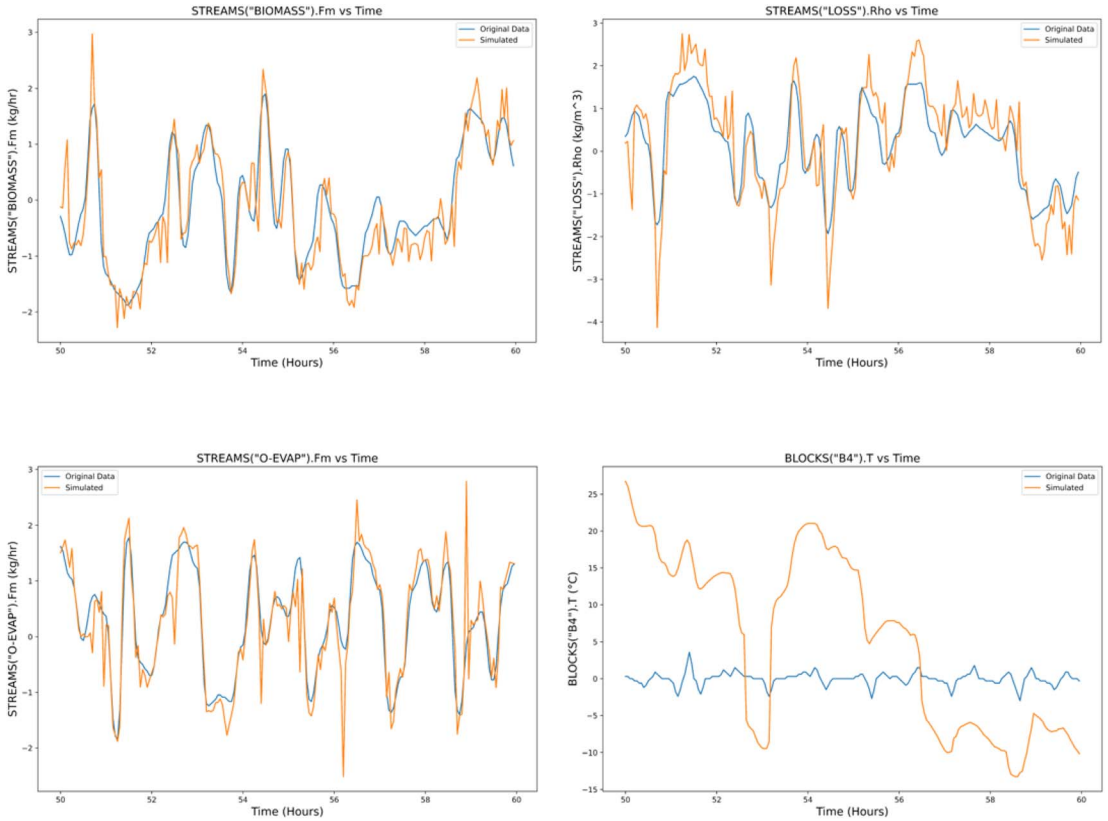


Fig. 6 Dynamic Reconstruction visualization plotted for a time length of 10 hours of the 4 state variables obtained for the algae growth industrial node in the algal biodiesel production network.

signify the state variables which are also the coupling parameter. Once all the coupling variables are identified, the surrogate ODEs for each block are modified by renaming the SVs to link the ODEs of two nodes which creates a coupled ODE model for overall network dynamics. This modification leads to the variable name alteration of the control variable of the subsequent node with the name of the SV of the direct previous node. The modified ODEs that represent a coupled ODE model for overall algal biodiesel network dynamics is given in ESI Section 3.† This coupled ODE model is next numerically integrated for different initialization values of the SVs found at $t = 0$, for a specific set of the CVs and perturbation signals to CVs.

3.2.2 Sustainability assessment of the algal biodiesel industrial network *via* coupled dynamics. The coupled ODE model representing overall network dynamics is solved using the LSODA solver from the *solve_ivp* class in Python. The numerical integration operation was run for an overall compute time of 50 hours for a time step of 0.00005 hours, generating the dynamic projection of the network for 50 hours. In this algal biodiesel network, each node has its own dynamics of CO_2 emissions and sequestration. Particularly, the algae growth node sequesters a high amount of CO_2 while also generating a small amount of CO_2, whereas the fermentation & extraction node and the anaerobic Digestion & CHP node are the major sources of CO_2 emission. With the use of solved values obtained *via* the numerical integration of coupled surrogate ODE models, optimal decision making can be enforced to reduce the carbon footprint of the entire network by accounting for the dynamics, eventually strategizing towards carbon neutrality of the whole industrial network, as we show in the following analyses.

3.2.2.1 Dynamic carbon footprint of the algal biodiesel industrial network. To calculate the dynamic carbon footprint for the network, the streams (SVs) representing CO_2 sequestration and emissions were calculated over time using numerical integration of coupled ODEs for 50 hours. The CO_2 sequestration streams are G300 and CO_2-Rec in the algae growth node, where G300 is also a CV whereas CO_2-Rec is a standalone input state variable (showing resource consumption). The CO_2 emission streams are O-EVAP in the algae growth node, stream 230 from the Fermentation & Extraction node and stream FLUGASLP from the anaerobic digestion & CHP node. The simulation of coupled ODEs for the whole network enables calculation of the net CO_2 footprint in the network accounting for all internal non-linear dynamics of each node and node interactions.

With the given perturbation, the total CO_2 sequestered over 50 hours is 4 516 196.43 kg in the whole network, calculated by numerical integration and adding the streams G300 and CO_2-REC. Similarly, the array of values obtained after numerical integration of coupled ODEs for the streams O-EVAP, 230, and FLUGASLP was combined for the run of 50 hours, and again integrated using the Simpson method to obtain the total

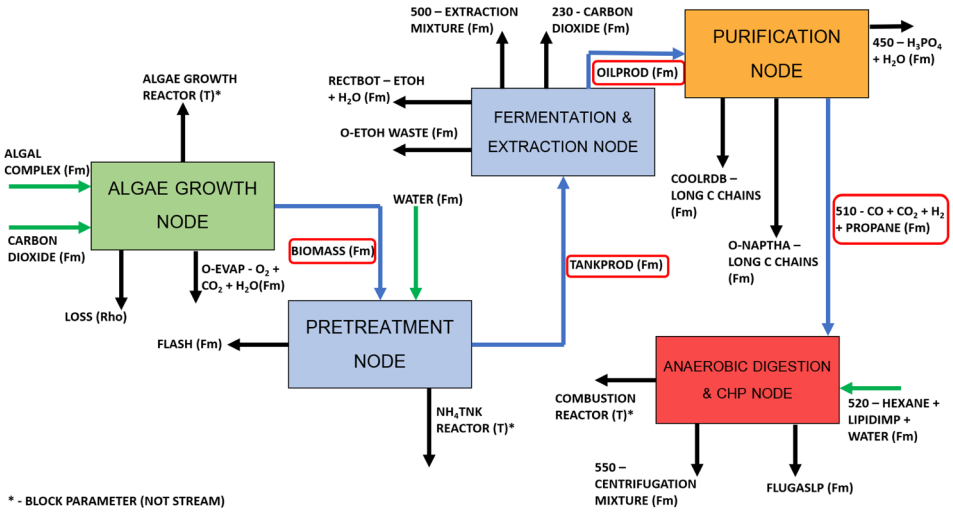


Fig. 7 Block flow visualization of the algal biodiesel production network functioning in the coupled manner with the material exchange shown among the 5 individual industrial systems.

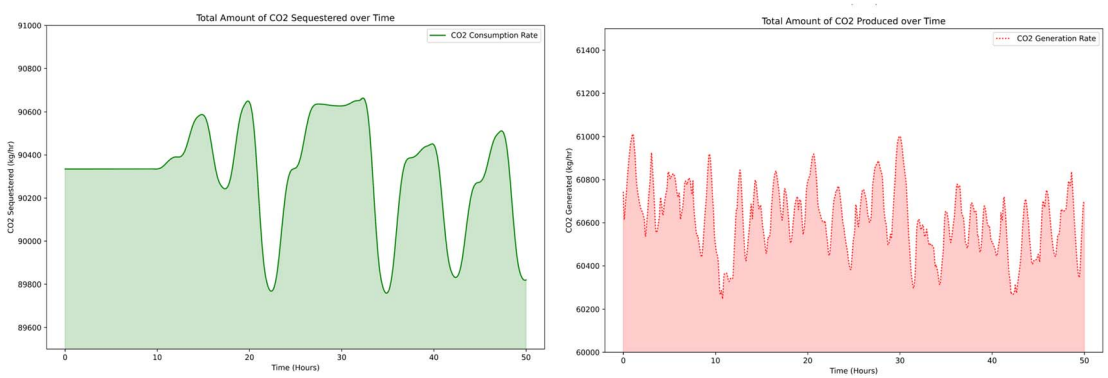
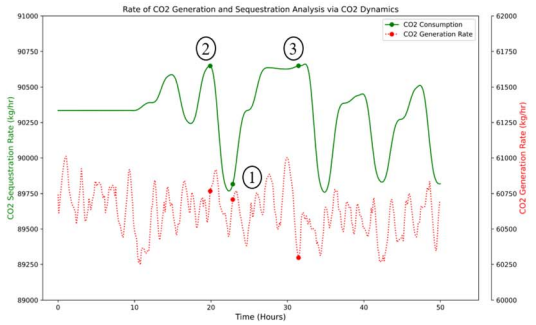


Fig. 8 Net amount of CO_2 sequestered evaluated upon calculating the area under the curve of the net CO_2 sequestration rate (green), and net amount of CO_2 produced evaluated upon calculating the area under the curve of the net CO_2 production rate (red).

amount 3 031 126.96 of CO_2 generated in the network. The net carbon footprint is calculated by subtracting the amount of CO_2 sequestered with the amount of CO_2 generated, thus giving us a net carbon footprint of −1 485 069.47 kg of CO_2, thus establishing that the network in the entirety of its operation has consumed 1 485 069.47 kg of CO_2 for the 50 hours of processing. Visually, temporal CO_2 sequestration and emissions are shown in Fig. 8 and the area under the curve gives the total sequestration and emissions over the time scale.

3.2.2.2 Identification of the optimal control parameter value to create a carbon negative algal biodiesel industrial network. Next, we utilize the dynamic profile of CO_2 emissions and sequestration to identify the values of control variables that lead to a maximum difference between the rate of CO_2 sequestration and the rate of CO_2 emissions. For this production network where the coupled ODEs have been solved for 50 hours, the time stamp of the maximum difference occurs at 31.45 hours, as observed from Fig. 9. This indicates that if the entire network is operated with the control parameter obtained at $t = 31.45$ hours from the data set of control variables, the network will achieve the best case scenario of carbon negativity, always sequestering a specific amount of CO_2. This provides the optimal value of the control variable to operate the network in a steady state for maximum net carbon sequestration. The control variables for the entire network are the four stream flow rates *viz.* G120 and G300 in the algae growth node, 170 in the pretreatment node, and 520 in the anaerobic digestion node (refer to Fig. 7 for stream composition). An important step is the validation of the hypothesis about the network representing the best case scenario for carbon negativity, which was performed by running the Aspen steady state simulation on the mechanistic models with the values of control variables found at 31.45 hours while the stand alone stable input values were the same, since they do not change.

Furthermore, in order to establish that the parameter values at $t = 31.45$ hours lead to the best case carbon negative ecosystem, the hypothesis was tested by extracting the values of the control parameters at a time stamp of 22.85 hours, which is the worst case carbon negative scenario, and at a random time stamp of 19.9 hours which is neither the best case C-negative



① WORST CASE SCENARIO

Parameter	Algal Complex	CO_2	Water	Impure Lipid
Value	4964.98	52585.67	26049.24	2821.14

② RANDOM CASE SCENARIO

Parameter	Algal Complex	CO_2	Water	Impure Lipid
Value	4954.83	53473.83	25993.4	2773.25

③ BEST CASE SCENARIO

Parameter	Algal Complex	CO_2	Water	Impure Lipid
Value	5023.24	53474.19	26255.24	2807.52

$(C\text{-negative})_{best} > (C\text{-negative})_{random} > (C\text{-negative})_{worst}$

Fig. 9 Control parameters obtained for the three scenarios of the carbon negative network *via* the implementation of the proposed hybrid mechanistic machine learning approach.

nor the worst case C-negative. Again, the "worst case carbon negative" here implies that the network stays carbon negative, but here the difference between the rate of CO_2 sequestration and the rate of CO_2 generation is minimum. Therefore, the parameters found at these three time stamps were used to run the Aspen steady state simulation and the hypothesis was validated with the sequestration rate value obtained for the best case C-negative scenario to be the highest while the value found out in the worst case C-negative scenario was the lowest. From these simulations at the steady state, the carbon negative rate (rate at which CO_2 is being sequestered) for the best case was found out to be 29 750.34 kg h^{-1}, while the carbon negative rate for the worst case and random case scenario was evaluated to be 28 956.14 kg h^{-1} and 29 687.11 kg h^{-1} respectively.

The evaluated values of the sequestration rate upon validation imply that the optimal control parameter for the best case carbon negative scenario for the network exists at t = 31.45 hours for a 50 hour coupled operation. The four control variable

Table 3 Total time required (in years) to sequester the regional energy-related CO_2 emissions that are generated for the year 1970 up till the year 2050 using three different capacities of the algal biodiesel network utilized in this research

Region/year	1970	1975	1980	1985	1990	1995	2000	2005	2010	2015	2020	2025	2030	2035	2040	2045	2050
***n* = 50**	**RDB throughput = 6.273 MMT per year**																
Pacific	36.4	39.5	43.5	42.3	48.4	48.7	53.3	54.2	49.3	48.5	43	46.3	46.6	45.5	46.1	47	48.6
Mountain	16.9	22.1	26.8	29.1	34	36	42.2	45.3	44	42.4	37.4	35.1	34.4	34	34.9	35.6	36.5
West Midwest	29.1	32	34.6	34.8	37.7	41.6	45.5	47.6	47.6	44.7	39.8	38.6	38.8	38.3	37.6	38	38.5
East Midwest	90.5	91.6	88.9	81.2	85.1	89.2	97.5	98.6	90.1	82.3	68.5	70.4	67.2	66.4	67.5	68.4	69.6
New England	18.3	16.6	14.8	14.8	16	15.6	17.1	17.8	15	14.2	11.9	12.5	11.8	11.6	11.5	11.4	11.5
Middle Atlantic	67.3	59.9	58.6	51.5	54.3	55.2	57.5	58.2	51	47.7	39.3	45.3	43	41.8	41.6	42.6	43.1
West South Central	55.5	62.1	77	75.7	84.1	88.7	98.7	94.6	92.5	93.4	88.3	88.7	90.3	91.1	90.6	92.1	95.4
East South Central	28.5	30.8	33.6	33.3	35.8	41	44.8	46	42.8	38.3	32.4	32.7	29.6	28	28	28.7	28.9
South Atlantic	54.6	58.3	66.4	67.4	73.8	79.6	91.4	97	89.1	80.3	68.4	70	69.9	67.7	68.1	68.6	70.2
***n* = 75**	**RDB throughput = 9.41 MMT per year**																
Pacific	24.3	26.3	29	28.2	32.2	32.4	35.5	36.1	32.9	32.3	28.6	30.9	31	30.4	30.8	31.3	32.4
Mountain	11.3	14.7	17.9	19.4	22.7	24	28.2	30.2	29.4	28.2	24.9	23.4	22.9	22.7	23.2	23.7	24.3
West Midwest	19.4	21.3	23	23.2	25.1	27.7	30.3	31.7	31.7	29.8	26.5	25.8	25.9	25.5	25.1	25.3	25.7
East Midwest	60.3	61.1	59.3	54.2	56.7	59.5	65	65.7	60.1	54.9	45.7	46.9	44.8	44.3	45	45.6	46.4
New England	12.2	11.1	9.8	9.9	10.7	10.4	11.4	11.9	10	9.5	7.9	8.3	7.8	7.7	7.7	7.6	7.7
Middle Atlantic	44.9	39.9	39	34.3	36.2	36.8	38.3	38.8	34	31.8	26.2	30.2	28.7	27.9	27.7	28.4	28.7
West South Central	37	41.4	51.3	50.5	56	59.1	65.8	63.1	61.7	62.2	58.9	59.1	60.2	60.7	60.4	61.4	63.6
East South Central	19	20.5	22.4	22.2	23.9	27.3	29.8	30.6	28.5	25.6	21.6	21.8	19.7	18.7	18.6	19.1	19.3
South Atlantic	36.4	38.9	44.3	44.9	49.2	53	60.9	64.7	59.4	53.5	45.6	46.7	46.6	45.1	45.4	45.7	46.8
***n* = 100**	**RDB throughput = 12.546 MMT per year**																
Pacific	18.2	19.7	21.8	21.2	24.2	24.3	26.6	27.1	24.7	24.2	21.5	23.1	23.3	22.8	23.1	23.5	24.3
Mountain	8.5	11	13.4	14.6	17	18	21.1	22.7	22	21.2	18.7	17.6	17.2	17	17.4	17.8	18.3
West Midwest	14.6	16	17.3	17.4	18.8	20.8	22.8	23.8	23.8	22.4	19.9	19.3	19.4	19.1	18.8	19	19.3
East Midwest	45.3	45.8	44.4	40.6	42.6	44.6	48.7	49.3	45.1	41.2	34.3	35.2	33.6	33.2	33.7	34.2	34.8
New England	9.1	8.3	7.4	7.4	8	7.8	8.6	8.9	7.5	7.1	5.9	6.2	5.9	5.8	5.7	5.7	5.8
Middle Atlantic	33.7	30	29.3	25.7	27.2	27.6	28.7	29.1	25.5	23.9	19.6	22.6	21.5	20.9	20.8	21.3	21.6
West South Central	27.8	31.1	38.5	37.9	42	44.3	49.3	47.3	46.3	46.7	44.2	44.4	45.2	45.6	45.3	46.1	47.7
East South Central	14.3	15.4	16.8	16.6	17.9	20.5	22.4	23	21.4	19.2	16.2	16.4	14.8	14	14	14.3	14.5
South Atlantic	27.3	29.2	33.2	33.7	36.9	39.8	45.7	48.5	44.5	40.1	34.2	35	34.9	33.8	34	34.3	35.1

values extracted for each of the three cases are shown in Fig. 9. Thus, utilizing the overall dynamics of the network, we can identify the optimal value of control variables to operate this network at the steady state for creating a net carbon negative industrial network. This steady state rate of carbon sequestration in the network is next used to calculate the time required for mitigating the energy related emissions in a region.

3.2.2.3 Time to sequester regional energy related CO_2 emissions using the algal biodiesel network. We utilize the best case carbon sequestration scenario of 29 750.34 kg h^{-1} based on the optimal control variables at $t = 31.45$ hours to operate the industrial network at the steady state. At this rate, the algal biodiesel network will sequester 2.60×10^8 kg CO_2 per year. Table 3 provides spatial accumulation of the US states into 9 major geographical regions,[52] where the time (in years) to sequester the energy related CO_2 emissions present for a particular year from 1970 to 2050 has been detailed for different numbers of the algal biodiesel network. These three cases involve 50, 75, and 100 of the same bio-diesel network giving a total algal biodiesel throughput of 6.273 MMT per year, 9.41 MMT per year, and 12.546 MMT per year respectively. These values are close to the existing capacity of soybean biodiesel production with 68 plants operating at a capacity of about 10.009 MMT per year,[53] and hence the proposed expansion for biodiesel production capacity seems reasonable.

Table 3 shows the time required to sequester energy related emissions in various regions for different years and future projections using this algal biodiesel network. As can be seen, increasing the number of plants allows the mitigation of the carbon emissions sooner, and hence rapid efficient investments into algae based technology can provide sustainable solution to both the energy crisis and goal of removing atmospheric carbon emissions. From Table 3, it can be inferred that the Mountain and New England region has the shortest set of time frames among other regions required to neutralize the CO_2 footprint, with the future outlook for 50 biodiesel networks being in the order of 35 years and 11 years respectively. In contrast, the West South Central and the East Midwest regions requires a longer time to neutralize among other regions, where the future outlook for 50 biodiesel networks has the time magnitude of the order of 90 years and 68 years respectively. This kind of analytical inference provides insight into the spatial distribution of such algal biodiesel networks that can be established in the entire US region for the sequestration of energy related CO_2 in an optimal timeframe; for instance, more biodiesel network plants in the West South Central and East Midwest regions while fewer plants in the Mountain and New England regions can be established from a set of N such networks. The amount of N here has been hypothetically varied among three different values of 50, 75, and 100 networks; however, from the viewpoint of realistic establishments and set-ups of these plants it is important to set the limit of such networks in accordance with the availability of the chemical precursors and supply of the upstream materials required to be processed into algal biodiesel.

In our work, the mechanistic model has been designed based on the plant situated at NREL in Colorado. Using the energy-related CO_2 emission in Colorado for the year 2020 of 79.9×10^9 kg,[54] the time taken to completely neutralize the emission in the presence of 10 algal biodiesel carbon sinks was calculated using eqn (5), which gives a value of approximately 36 years. This implies that if there exist 10 such algal biodiesel networks operating with the same control technology in the state of Colorado, the atmospheric energy-related CO_2 accumulated in the year 2020 will be neutralized completely in 36 years. The high values of years shown in Table 3 needed to neutralize energy related emissions in different US regions indicate that more aggressive policies need to be implemented for achieving the goal of creating net zero systems.

4 Conclusions

The hybrid mechanistic machine learning approach proposed in this work provides a robust computational approach for studying the overall dynamics of industrial networks. Utilizing the surrogate models built for the dynamics of each individual node and coupling these models to study the overall dynamics of resource consumption and emissions *etc.*, help in the evaluation of the industrial network to meet the sustainability goals over time. As industrial networks are large complex systems, modeling the dynamics of a whole integrated industrial network with a large number of nodes proves to be challenging and computationally prohibitive. Hence, this data driven approach to build surrogate models overcomes a major challenge to study the overall dynamics of industrial networks by preserving the essential dynamics of each node in the surrogate models. The nonlinear mathematical surrogate models that are obtained capture the precise functional relationship among several parameters existing in the state space and control space of the industrial network, thus giving insight into key variables effecting the overall dynamics of the network.

A crucial aspect of sustainability assessment of such complex networks is the temporal behavior of the overall system under various external forcing functions and accounting for the nonlinear interactions between the nodes. Existing sustainability assessment methods such as process life cycle assessment (LCA) or dynamic LCAs do not address this aspect and focus mainly on the linear extrapolation of relationships between two nodes of the system, thus missing a true dynamic evaluation of the overall system. Here, as the surrogate model coupling allows for accounting for nonlinear interactions between nodes, this approach is able to capture the dynamic behaviour of the overall system for the evaluation of resource consumption and emissions *etc.* Utilizing this strength of the proposed approach, we demonstrate calculation of the net carbon footprint for industrial networks. It is more precise to analyze the carbon footprint using the mathematical model, where the precision can be attributed to capturing the nonlinear relationship of the flow of CO_2 with other industrial variables existing in the state space and control space. In our case study, the algal bio-diesel network is a carbon sink as the net carbon footprint comes out to be negative. This provides a way to accurately evaluate the net carbon footprint of emerging industrial networks based on carbon capture technologies

accounting for complex dynamics of interactions between different industries in the network. Hence, this proposed approach can be applied to quantify the carbon footprint of various industrial networks utilizing the data collected from each industrial node and developing dynamic models for each node. Such data collection will become easier as more industries adopt automated manufacturing processes and sensor data are readily available.

Another crucial aspect for emerging industrial networks is identifying the design and optimal parameters for the most sustainable network. Most sustainability assessment methods provide a comparative analysis of different designs, thus showing if one design is better than the other. However, these methods do not help in identifying the most sustainable design by analyzing the dynamics of the overall system under consideration. In this hybrid mechanistic machine learning approach, the surrogate models not only inform the industries about the mathematical relationship between state variables and the underlying governing equation that exists behind an industry operation, but can also be used to inform the industries on how to effectively design the process and identify the optimal values of control variables of the overall network to meet the sustainability goals. This has been demonstrated in the undertaken case study where the surrogate model for the coupled algal biodiesel industrial network has successfully provided the information on the values of 4 principal control parameters *viz.* flowrate of the algal complex, flowrate of CO_2 required for algal growth, flowrate of water required for pretreatment, and flowrate of lipid impurities to anaerobic digestion. The operation of the industrial network at the optimal control values (identified from the dynamics of emissions and sequestration) sequesters the highest amount of CO_2 in the network at a rate of 29 750.34 kg h^{-1}. Such utility of surrogate models for the coupled dynamics of the overall network bridges an important research gap existing for the sustainable design of the industrial networks. Again, industries implementing the IoT can use sensor data to develop surrogate models that can be used in the proposed approach for informing the design of industrial networks towards net zero emission operation or net carbon negative operation.

Subsequently, with the control values set and the network operating at the best possible carbon sequestration rate, the time necessary for complete neutralization of energy-related CO_2 emissions in a particular region can be evaluated based on the available valid data for different years. Such a type of surrogate model facilitated assessment for a regional industrial network can aid policymakers and government organizations to effectively make decisions and implement the policy for the establishment of industrial carbon sinks at certain geospatial locations where the years for net neutralization of energy-related CO_2 are well optimized. In each of these applications, the hybrid mechanistic machine learning approach offers a unique advantage over traditional methods. *Via* a combination of the strengths of mechanistic modeling and machine learning, it can create accurate and reliable surrogate models of complex systems which can capture the dynamics of the system, allowing for the design of macroscale networks. Since this approach is flexible and modular as data from different systems can be integrated, it can therefore be applied to a wide range of systems and processes, from plant scale, supply chains, industrial symbiosis to waste management. Additionally, since the hybrid approach is data-driven, it can learn from the data as and when available from sensors, allowing it to continuously improve its models and predictions as more data becomes available. This makes it a powerful tool for dealing with the complexity and uncertainty of real-world manufacturing industrial networks.

However, the approach will face some challenges as the complexity and size of the industrial network increase. These challenges include handling of stiff nonlinear surrogate models, iterative training for varying dynamic regimes, and significant domain expertise to facilitate the coupling of ODEs. These challenges can easily be overcome with the increase in computational power, utilizing advanced numerical integration methods specifically designed for stiff systems such as LSODA used in this work, higher availability of sensor data from automated manufacturing systems, *etc.* The domain expertise is an invaluable skill and cannot be replaced in the foreseeable future. Furthermore, the non-availability of the temporal data that have been used in the algorithm can prove to be a big challenge in benchmarking and validation of the models since the data are often proprietary in nature. However, the models can be regenerated and updated accordingly upon the availability of the data, thus maintaining the novelty of the proposed methodology and assessment technique.

The proposed approach of amalgamating mechanistic modeling with the machine learning approaches set forth a novel way of the nonlinear dynamic assessment technique and design of industrial networks in the domain of sustainability science and industrial ecology. Resonating with the emphasis on the transition towards dynamic sustainability assessment,[55] this research study proves to be one of the first steps amidst upcoming novel techniques that exist enabling an efficient quantification and assessment of the sustainability metrics of the industries. As physics informed machine learning approaches have shown great promise in novel material discovery, chemical engineering design, reaction pathways *etc.*, this paper demonstrates that design of complex industrial networks can benefit significantly from this hybrid approach. While in this work we have only used the SINDy approach, numerous other algorithms can be used to improve the interpretable surrogate models for industrial systems. Furthermore, the approach is scalable to expand the size of the network, as new industries can be added into the system with their own surrogate model for the new node. For instance, in our case study of the algal biodiesel production the network consists of 5 industrial systems for which the surrogate models have been coupled. The networked surrogate model for these 5 industries can be changed with the addition of an extra industrial system of the hydrogen production process where the exchange of hydrogen gas takes place between the hydrogen production and the purification process. Thus, in this way any existing industrial network of interest can be expanded if a new industrial system is introduced with which there exists certain material

interdependency. If the regional industrial network needs to test a novel technology, it can also be inserted as a surrogate model to simulate and design the network. The technique can also be expanded to study other sustainability metrics such as water consumption, other resource consumption *etc.* that are relevant for the network, and thus a multivariable design space for sustainable operation of the industrial network can be explored. The assessment of the dynamic impacts of circular economy implementation *via* reuse of waste material (such as wastewater) is also feasible through this approach.

In conclusion, to meet the future goals of sustainable development, a macroscale design assessment of industrial networks is necessary accounting for the dynamics of overall networks. This hybrid mechanistic approach helps in dynamic coupling between several individual industrial nodes, thus helping to evaluate design space for using emerging technologies in a synergistic manner with existing industries. The approach also provides modularity to study the overall dynamics of industrial networks, as the recovered surrogate models can be coupled with many systems. The coupled dynamics of the overall network(s) can be used to inform the most sustainable design for emerging industrial networks in terms of overall resource requirement, carbon capture potential and emissions by accounting for the non-linear interactions between different nodes of the networks over time.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors are grateful for support from the U.S. National Science Foundation CBET-1805741 and FMRG ECO-2229250. We are also grateful to the Purdue Undergraduate Research Experience (PURE) program to support Raghav Rajesh Moar as an Undergraduate Research Student.

Notes and references

1 *I. P. on Climate Change, The evidence is clear: the time for action is now. We can halve emissions by 2030*, 2022, **https://www.ipcc.ch/2022/04/04/ipcc-ar6-wgiii-pressrelease/**.
2 U. S. E. P. Agency, *Greenhouse Gas Inventory Data Explorer*, 2022, **https://cfpub.epa.gov/ghgdata/inventoryexplorer/#industry/entiresector/allgas/category/all**.
3 A. J. L. Angela and C. Jones, *Carbon Capture and Sequestration (CCS) in the United States (R44902)*, 2022.
4 G. Laufenberg, B. Kunz and M. Nystroem, *Bioresour. Technol.*, 2003, **87**, 167–198.
5 A. K. Panda, R. K. Singh and D. Mishra, *Renewable Sustainable Energy Rev.*, 2010, **14**, 233–248.
6 R. A. Muhlack, R. Potumarthi and D. W. Jeffery, *Waste Manage.*, 2018, **72**, 99–118.
7 A. Shekhar, M. Parekh and V. Pol, *J. Power Sources*, 2022, **523**, 231015.
8 J. B. Guinee, R. Heijungs, G. Huppes, A. Zamagni, P. Masoni, R. Buonamici, T. Ekvall and T. Rydberg, *Life Cycle Assessment: Past, Present, and Future*, 2011.
9 M. Fischer-Kowalski, F. Krausmann, S. Giljum, S. Lutter, A. Mayer, S. Bringezu, Y. Moriguchi, H. Schütz, H. Schandl and H. Weisz, *J. Ind. Ecol.*, 2011, **15**, 855–876.
10 S. C. Bankes, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 7199–7200.
11 J. Bebbington, J. Brown and B. Frame, *Ecol. Econ.*, 2007, **61**, 224–236.
12 M. Shamsuzzaman, A. Shamsuzzoha, A. Maged, S. Haridy, H. Bashir and A. Karim, *Applied Energy*, 2021, **300**, 117352.
13 C. J. Corbett and J.-N. Pan, *Eur. J. Oper. Res.*, 2002, **139**, 68–83.
14 R. G. Coyle, *J. Oper. Res. Soc.*, 1997, **48**, 544.
15 G. Radons and R. Neugebauer, *Nonlinear Dynamics of Production Systems*, Wiley Online Library, 2004.
16 G. Leonov, A. Pogromsky, K. Starkov, B. Andrievsky, N. Kuznetsov and I. Adan, *IFAC Proceedings Volumes*, 2013, **46**, 33–42.
17 T. Shi, W. Yang and J. Qiao, *J. Phys.: Conf. Ser.*, 2021, 012037.
18 B. J. Angerhofer and M. C. Angelides, *2000 Winter Simulation Conference Proceedings (Cat. No. 00CH37165)*, 2000, pp. 342–351.
19 N. Bichraoui, B. Guillaume and A. Halog, *Procedia Environ. Sci.*, 2013, **17**, 195–204.
20 K. J. Keesman and K. J. Keesman, *System Identification: an Introduction*, Springer, 2011, vol. 2.
21 L. Ljung, *Annu. Rev. Control*, 2010, **34**, 1–12.
22 S. L. Brunton, B. R. Noack and P. Koumoutsakos, *Annu. Rev. Fluid. Mech.*, 2020, **52**, 477–508.
23 Y. J. Cho, N. Ramakrishnan and Y. Cao, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 142–150.
24 A. V. Karnaukhov, E. V. Karnaukhova and J. R. Williamson, *Biophys. J.*, 2007, **92**, 3459–3473.
25 G. Craciun and C. Pantea, *J. Math. Chem.*, 2008, **44**, 244–259.
26 A. Bachnas, R. Tóth, J. Ludlage and A. Mesbah, *J. Process Control*, 2014, **24**, 272–285.
27 R. Subramanian, R. R. Moar and S. Singh, *Machine Learning with Applications*, 2021, **3**, 100014.
28 W. Farlessyost and S. Singh, *Nonlinear Dyn.*, 2022, **110**, 1613–1631.
29 J.-P. Noël and G. Kerschen, *Mech. Syst. Signal. Process.*, 2017, **83**, 2–35.
30 G. Sirca Jr and H. Adeli, *Sci. Iran.*, 2012, **19**, 1355–1364.
31 M. Sorokina, S. Sygletos and S. Turitsyn, *2017 19th International Conference on Transparent Optical Networks (ICTON)*, 2017, pp. 1–4.
32 J.-C. Loiseau, *Theor. Comput. Fluid Dyn.*, 2020, **34**, 339–365.
33 M. Hoffmann, C. Fröhner and F. Noé, *J. Chem. Phys.*, 2019, **150**, 025101.
34 Z. Lai, C. Mylonas, S. Nagarajaiah and E. Chatzi, *J. Sound Vib.*, 2021, **508**, 116196.
35 Y. Sun, L. Zhang and H. Schaeffer, *Mathematical and Scientific Machine Learning*, 2020, pp. 352–372.
36 L. Zhang and H. Schaeffer, *Multiscale Model. Simul.*, 2019, **17**, 948–972.

37 L. Boninsegna, F. Nüske and C. Clementi, *J. Chem. Phys.*, 2018, **148**, 241723.
38 K. J. Åström and P. Eykhoff, *Automatica*, 1971, **7**, 123–162.
39 J. Bongard and H. Lipson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 9943–9948.
40 S. L. Brunton, J. L. Proctor and J. N. Kutz, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 3932–3937.
41 M. Raissi, P. Perdikaris and G. E. Karniadakis, *J. Comput. Phys.*, 2017, **348**, 683–693.
42 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 083802.
43 R. T. Chen, Y. Rubanova, J. Bettencourt and D. K. Duvenaud, *Advances in Neural Information Processing Systems*, 2018, vol. 31.
44 S. L. Kukreja, J. Löfberg and M. J. Brenner, *IFAC Proceedings Volumes*, 2006, **39**, 814–819.
45 A. Cortiella, K.-C. Park and A. Doostan, *Comput. Methods Appl. Mech. Eng.*, 2021, **376**, 113620.
46 D. W. Marquardt and R. D. Snee, *Am. Stat.*, 1975, **29**, 3–20.
47 P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz and A. Y. Aravkin, *IEEE Access*, 2018, **7**, 1404–1423.
48 H. Zou and T. Hastie, *J. R. Stat. Soc. Series B Stat. Methodol.*, 2005, **67**, 301–320.
49 B. M. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. N. Kutz and S. L. Brunton, arXiv, preprint, 2020, arXiv:2004.08424, DOI: **10.48550/arXiv.2004.08424**.
50 R. Davis, C. Kinchin, J. Markham, E. Tan, L. Laurens, D. Sexton, D. Knorr, P. Schoen and J. Lukas, *Process Design and Economics for the Conversion of Algal Biomass to Biofuels: Algal Biomass Fractionation to Lipid-And Carbohydrate-Derived Fuel Products*, technical report, National Renewable Energy Lab, Golden, Co, United States, 2014.
51 J. D. Lambert, *et al.*, *Numerical Methods for Ordinary Differential Systems*, Wiley, New York, 1991, vol. 146.
52 U. C. Bureau, *Census Bureau Regions and Divisions with State FIPS Codes*, 2019, **http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm**.
53 B. International, *U.S. Biodiesel Plants*, 2023, **https://biodieselmagazine.com/plants/listplants/USA/**.
54 Independent Statistics and Analysis, U.S. E. I. A., *U.S. Energy-Related Carbon Dioxide Emissions 2021*, 2022, **https://www.eia.gov/environment/emissions/carbon/pdf/2021_co2analysis.pdf**.
55 Y. Huang, *Toward Dynamic Sustainability Assessment in the Digital Age*, 2022.

