Chemical Engineering Research and Design 232 (2026) 302–318

Contents lists available at ScienceDirect

Chemical Engineering Research and Design

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

Federated learning from molecules to processes: A perspective ☆

Jan G. Rittig *, Clemens Kortmann

Process Systems Engineering (AVT.SVT), RWTH Aachen University, Aachen, Germany

ARTICLE INFO

Keywords:
Machine learning
Data privacy
Graph neural networks
Chemical industry
Collaboration
Federated learning

ABSTRACT

We present a perspective on federated learning (FL) in chemical engineering that envisions collaborative efforts in machine learning (ML) developments within the chemical industry. Large amounts of chemical and process data are proprietary to chemical companies and are therefore locked in data silos, hindering the training of ML models on large data sets in chemical engineering. Recently, the concept of FL has gained increasing attention in ML research, enabling organizations to jointly train ML models without disclosure of their individual data. We discuss potential applications of FL in several fields of chemical engineering, from the molecular to the process scale. In addition, we apply FL in two exemplary case studies that simulate practical scenarios of multiple chemical companies holding proprietary data sets: (i) prediction of binary mixture activity coefficients with graph neural networks and (ii) system identification of a distillation column with autoencoders. Our results indicate that ML models jointly trained with FL yield significantly higher accuracy than models trained by each chemical company individually and can perform similarly to models trained on combined datasets from all companies. FL has therefore great potential to advance ML models in chemical engineering while addressing corporate data privacy. Future research in pilot projects will be essential on the way towards practical application within the chemical industry.

1. Introduction

Machine learning (ML) methods have led to remarkable progress in diverse fields of chemical engineering (ChemE) research (Schweidtmann et al., 2021; Strieth-Kalthoff et al., 2020; Daoutidis et al., 2024; Cheng et al., 2024), but data scarcity remains a key limitation. Scaling up data set sizes has been a key factor for recent advances in ML applications such as natural language processing and computer vision, cf. neural scaling laws (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022). While large data sets of texts and images with billions of data points can be collected from the internet, ChemE applications typically face smaller data sets of hundreds to thousands of experimental data points and data sets with low variance, e.g., chemical process data at similar operating points, cf. Qin and Chiang (2019) and Thebelt et al. (2022). The vast majority of this highly valuable chemical and process data has been collected by chemical enterprises over the last decades, and is therefore locked in their data silos (Zhu et al., 2022). In fact, the companies have a strong interest in keeping their data private because they have invested a substantial amount of money and effort in obtaining it, and because it contains confidential information about the chemical space and the processes they are working on Dutta et al. (2024). As a result, only a few datasets are publicly available for ChemE applications, with notable exceptions in molecular property prediction (Geballe et al., 2010; Mobley and Guthrie, 2014; Ramakrishnan et al., 2014), reaction prediction and optimization (Felton et al., 2021; Wigh et al., 2024), and process fault detection (Chiang et al., 2012). The generally low public availability of large chemical and process datasets, and consequently missing benchmarks, limits ML research and the development of large-scale models for ChemE.

Several approaches have been considered to mitigate the problem of scarce chemical data. These include the incorporation of domain knowledge into the model, e.g., in hybrid architectures (Bradley et al., 2022; Schweidtmann et al., 2024) such as physics-informed ML (Karniadakis et al., 2021) and agentic architectures coupled with domain-specific tools (Boiko et al., 2023; Bran et al., 2024), which can reduce the amount of data needed for model training. Another way is to artificially increase the amount of data available for model training through approaches like multi-task, multi-fidelity, self-supervised, and transfer learning. For example, in transfer learning, data from related ChemE tasks is combined, whereas in multi-fidelity learning, high-fidelity experimental data is augmented with low-fidelity data from simulations, see, e.g., Vermeire and Green (2021a), Wu and Zhao

☆ This article is part of a Special issue entitled: 'Emerging Stars II' published in Chemical Engineering Research and Design.

* Corresponding author.
E-mail address: jan.rittig@rwth-aachen.de (J.G. Rittig).

https://doi.org/10.1016/j.cherd.2026.07.011
Received 8 July 2025; Received in revised form 29 June 2026; Accepted 5 July 2026
Available online 10 July 2026
0263-8762/© 2026 The Authors. Published by Elsevier Ltd on behalf of Institution of Chemical Engineers. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

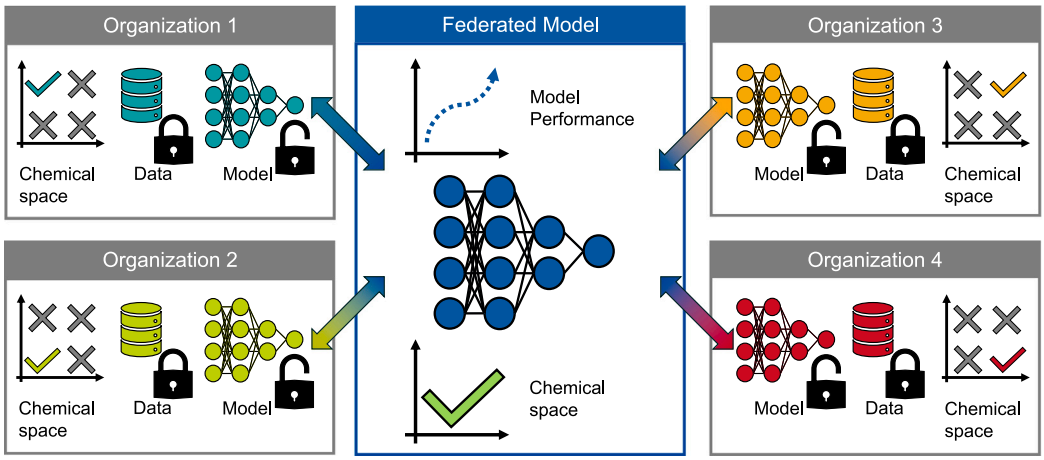


Fig. 1. Schematic illustration of four chemical companies working on different regions in the chemical space and using federated learning to jointly advance a machine learning model while reducing the need to share raw corporate data.

(2020), Li et al. (2020), Savage et al. (2023) and Nevolianis et al. (2024). Although these approaches have been shown to increase the performance of ML models, many ML applications in ChemE still lack sufficient experimental data for training or rely solely on simulated data.

We argue that leveraging private chemical and process data sets to increase the quantity and quality of the available data used for model training will be crucial to catalyze ML for ChemE applications. Therefore, an approach is needed that (a) leverages existing data from chemical enterprises without harming their data privacy and (b) enables new kinds of research cooperation between multiple industrial and possibly academic partners to value unexplored chemical data on a larger scale. To this end, federated learning (FL) is particularly promising.

FL enables multiple entities to jointly train ML models without disclosure of their individual data (McMahan et al., 2017; Kairouz et al., 2021; Daly et al., 2024). The FL approach has gained increasing attention in recent years for applications of ML where data privacy is a main concern (B. Liu et al., 2024; Wen et al., 2023). Fig. 1 illustrates the concept of FL for the example of four chemical companies that have private data on different regions of the chemical space and jointly train a neural network. Specifically, each company trains a ML model on its local data and only exchanges the trained model parameters, so that the individual models can be aggregated into a global model, which is ideally more accurate and has a wider applicability range than the individual ones. Hence, private data can be utilized for collaborative model development without leaving the individual entities.

FL has been utilized by both academia and industry in various domains, including natural language processing (Li et al., 2024), finance (Wen et al., 2023), and healthcare (Rieke et al., 2020; Nguyen et al., 2023; Rauniyar et al., 2024; Heyndrickx et al., 2024). One of the first FL applications points to Google setting up a federation of millions of individual mobile phone users to train recurrent neural networks for next-word prediction in text messages on Google keyboards – without the need that users share private text message data (Yang et al., 2018). Another example from natural language processing is the use of FL to pre-train billion-scale large language models (LLMs) (Sani et al., 2024). Furthermore, FL has been applied in healthcare for training convolutional neural networks in detecting cancer from medical images based on patient data locally stored in numerous healthcare institutions (Lu et al., 2022; Guan et al., 2024). On an industry scale, the MELLODDY project serves as a prime FL project (Heyndrickx et al., 2023): ten pharmaceutical companies teamed up with academia to develop quantitative structure–activity relationships (QSARs) in collaborative efforts. Specifically, the companies jointly trained neural networks with extended connectivity fingerprints as input for drug property prediction, demonstrating improved prediction accuracy and applicability range compared to the models available at the individual companies (Heyndrickx et al., 2023; Oldenhof et al., 2023; Heyndrickx et al., 2024). More recently, Hanser et al. demonstrated improved bioactivity prediction capabilities of ML models by combining FL with knowledge distillation (cf. Wu et al. (2022)) in a collaboration with eight pharmaceutical companies (Hanser et al., 2025). Overall, these examples showcase that FL has already demonstrated great potential in many domains and practical applications.

Given the promising results achieved with FL in various domains, we study FL as an approach to address data scarcity and advance ML in ChemE.

Reviewing existing FL works in ChemE engineering, we only find a few recent applications: Hayer investigated FL to jointly train neural networks to predict process variables of aluminum electrolysis based on process data stored at different chemical producers, showing slight improvements in predictive accuracy (Hayer, 2021). Furthermore, Zhu et al. (2022) developed FL approaches to train graph neural networks for molecular property prediction and propose FedChem, a benchmark for predicting a variety of physicochemical and biological properties when data is heterogeneously distributed across different entities. Recently, Xu and Wu (2024) proposed FL of a data-driven model for distributed model predictive control of two chemical reactors in series without the need to exchange data between the two reactor models, thereby increasing resilience against cyber attacks. Notably, Dutta et al. (2024) provide a tutorial of FL in relation to ChemE and present pharmaceutical examples. Applications of FL also emerge in other domains strongly related to ChemE, e.g., in energy systems for power production forecasting (Cheng et al., 2022; Grataloup et al., 2024). In general, FL for ChemE is recently emerging but is largely underexplored.

We here present a perspective on FL in ChemE. Specifically, we first describe the concept of FL (Section 2). We then discuss the aspects of FL in the context of ChemE and highlight potential applications, such as flowsheet digitization and process design, and corresponding challenges and research directions in practical implementation (Section 3). Then, we present two exemplary case studies: We propose FL for prediction of infinite dilution activity coefficients of binary mixtures with graph neural networks (Section 4.1) and for system identification of a dynamic chemical process using autoencoders (Section 4.2). Both case studies show that collaboration of chemical companies through FL can increase predictive accuracies of ML models. While transferring FL to industrial settings will require further data and model privacy considerations, the case studies demonstrate the potential for addressing data scarcity in ChemE and thus advance ML on an industrial scale.

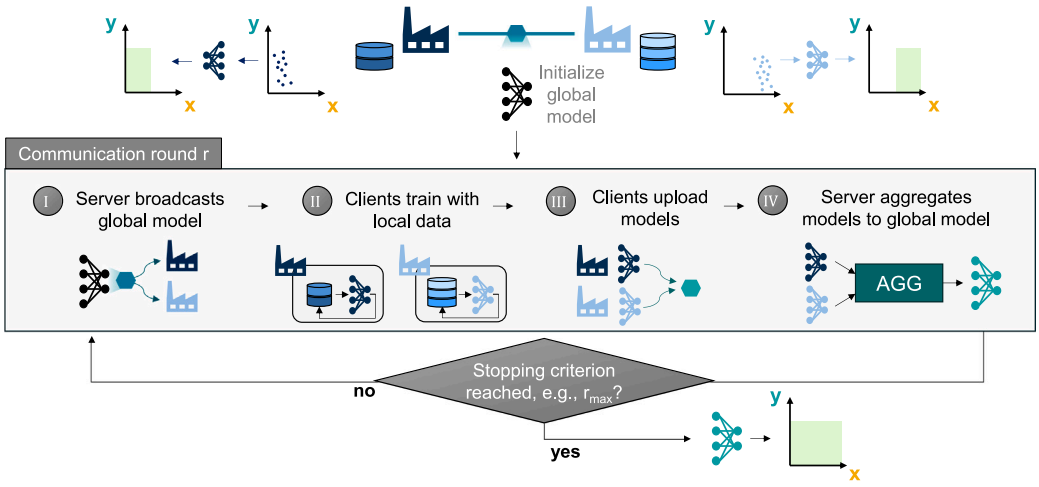


Fig. 2. Illustrative workflow of federated learning for two companies jointly training a machine learning model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Concept of federated learning

In this section, we provide an overview of the concept of FL. We focus on the scope of ChemE, where collaborating organizations, likely chemical companies that hold large data sets, jointly train a ML model while reducing the need to share raw data, as illustrated in Fig. 1 and referred to as *cross-silo* FL (Kairouz et al., 2021; Huang et al., 2022a). We further assume a *centralized* collaboration, i.e., a central server/framework is used for communication between companies during model training, which is common for cross-silo FL (Kairouz et al., 2021). For a thorough review of FL categorizations, we refer the interested reader to Kairouz et al. (2021), B. Liu et al. (2024) and Wen et al. (2023).

2.1. Fundamentals & workflow

The term FL was introduced by McMahan et al. (2017) and refined by Kairouz et al. (2021): Multiple entities, referred to as clients, jointly solve a ML task by model training without sharing their individual data. For joint training, the clients do not exchange raw data to address data privacy concerns. They rather exchange information about the model and corresponding updates during training. Therefore, FL shifts the focus from data sharing to model sharing.

We illustrate the workflow of FL in Fig. 2. Based on McMahan et al. (2017), Kairouz et al. (2021), Dutta et al. (2024) and Zhou et al. (2024), the setting can be formally described as follows: K clients aim to jointly develop a ML model with model parameters $\mathbf{w} \in \mathbb{R}^d$. Each client $k \in K$ holds a local data set $\mathcal{D}_k \sim \mathcal{P}_k(\mathbf{x}, \mathbf{y})$ with a data point containing features $\mathbf{x}$ and corresponding labels $\mathbf{y}$, and $\mathcal{P}_k$ being the client's local data distribution. Instead of each client training an individual model as would be the case without FL and is indicated in the top left and right corners of Fig. 2, a global model is trained jointly. Assuming that each client data distribution $\mathcal{P}_k$ is sampled independently and identically distributed (iid) from a global data distribution $\mathcal{P}$, the joint training of the ML model can be formulated as the following optimization problem (McMahan et al., 2017; Zhou et al., 2024):

$$\min_{\mathbf{w}} \ell(\mathbf{w}) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\mathbf{w}) = \sum_{k=1}^{K} \frac{n_k}{n} \sum_{i=1}^{n_k} \ell_i(\mathbf{w}; (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_k), \tag{1}$$

where $\ell(\mathbf{w}) = \sum_i \ell_i(\mathbf{w}, (\mathbf{x}_i, \mathbf{y}_i))$ is the global loss function for training, and $\mathcal{L}_k(\mathbf{w})$ is the expected local loss function of client k with $\mathbb{E}_{\mathcal{D}_k}[\mathcal{L}_k(\mathbf{w})] = \ell(\mathbf{w})$, in case of iid local data sets. Further, $n_k = \#\mathcal{D}_k$ denotes the number of samples in the local data set of client k, so the total number of samples is $n = \sum_{k=1}^{K} n_k$.

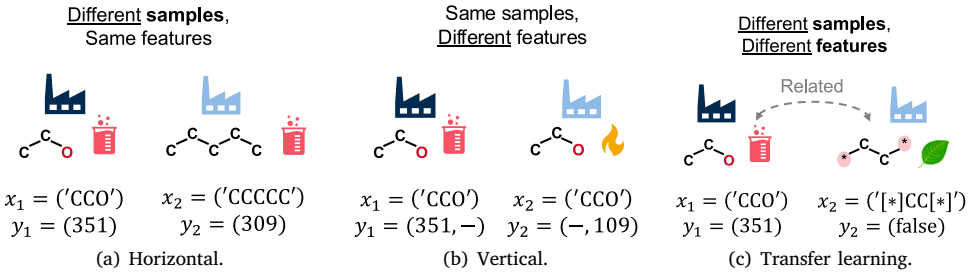
To solve the optimization problem in Eq. (1), an algorithm called Federated Averaging (FedAvg) (McMahan et al., 2017) is commonly used. In FedAvg, the training process starts by initializing the global model with parameters $\mathbf{w}_0$. Then, the so-called communication rounds $r \in \{1, 2, \dots, R\}$ are conducted. In each communication round r (Kairouz et al., 2021; Zhou et al., 2024),

(I) the parameters of the global model from the previous communication round, $\mathbf{w}_{r-1}$, are sent to the clients;
(II) the clients update the received model parameters $\mathbf{w}_{r-1}$ by training with their local data set $\mathcal{D}_k$, i.e., $\mathbf{w}_r^k \in \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n_k} \mathcal{L}_k(\mathbf{w})$;
(III) the clients upload the updated model parameters $\{\mathbf{w}_r^k\}_{k=1}^K$;
(IV) then the updated model parameters of all clients are aggregated by applying an aggregation function to obtain new parameters of the global model $\mathbf{w}_r = \text{AGG}(\mathbf{w}_r^1, \mathbf{w}_r^2, \dots, \mathbf{w}_r^{|K|})$, e.g., using weighted averaging: $\mathbf{w}_r = \sum_{k=1}^{K} \frac{n_k}{n} \mathbf{w}_r^k$, hence FedAvg (McMahan et al., 2017).

The new global model parameters are then passed to the clients as a starting point for a new communication round. This communication process is repeated until a stopping criterion is met, e.g., a predefined maximum number of communication rounds, $R = r_{\max}$, or convergence of the global model measured by changes in its parameters.

Joint training through FL promises higher model accuracies and applicability ranges than individual models. The idea behind multiple rounds of communication is that through iterative sharing and retraining, the model parameters will converge to values that provide a good fit for all of the clients' data sets, i.e., the updated models by the clients (step II) reach a consensus (Kairouz et al., 2021). In contrast to each client having an individual model with parameters fitted to its individual data set only, the models trained with FL are therefore influenced by more data and can result in higher accuracy and a wider applicability range (cf. green area in x-y plots in Fig. 2).

We note that, next to FedAvg, a variety of other algorithms have been proposed for FL over the last years. These algorithms typically share the goal of finding a parameter set that results in models with increased accuracy and larger applicability ranges for all clients. They can, for example, differ in the way the local models are aggregated and what kind of information is shared between the clients, e.g., sharing updates on model gradients instead of parameter values, which can result in improved convergence time and accuracy for different client and data distributions. We herein focus on FedAvg, which has been successfully applied in various FL settings. For an overview of different algorithms and convergence analysis, we refer the interested reader to Kairouz et al. (2021) and Zhou et al. (2024).



(a) Horizontal. (b) Vertical. (c) Transfer learning.

Fig. 3. Data distributions for two companies at the example of molecular property data.

2.2. Model performance indicators

To evaluate the model performance, e.g., in terms of accuracy and applicability range, the clients typically separate a test data set from their local data set $\mathcal{D}_k$, which is not used for training. So, standard metrics such as mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2) can be calculated for the global model that is passed to the individual clients. Also, after the joint training, the clients can fine-tune the global model on their local data (similar to step II). They can then compare the metrics for the global or fine-tuned model to a model trained only on their individual data. For this, Heyndrickx et al. (2024) proposed the relative improvement of proximity to perfection (RIPtoP), which measures how much FL closes the gap between the baseline performance and a perfect model:

$$\text{RIPtoP} = \frac{\text{metric}_{\text{MoI}} - \text{metric}_{\text{baseline}}}{\text{metric}_{\text{perfect}} - \text{metric}_{\text{baseline}}}, \tag{2}$$

where the metric could for example be the MAE, MoI is the model of interest, which is the model after federated learning (the global or fine-tuned model), *baseline* is the model trained only on client data set $\mathcal{D}_k$, and *perfect* stands for an artificial perfect model. For example, $\text{metric}_{\text{perfect}}$ would be 0 if the metric is MAE and 1 if the metric is R^2.

In research studies that explore FL in a simulated setting, the model FL performance is often compared to a *centralized* training. Centralized refers to training a model on the global data set $\mathcal{D}$, i.e., the aggregated set of the clients' local data sets. This is indeed not possible in practical scenarios, where data sets cannot be aggregated due to privacy concerns. Therefore, this comparison should be treated with caution as it simulates a scenario that is not practicable. Also, the model performance in a centralized setting should not be seen as an upper bound, as it has been shown that the model trained in a federated setting can outperform the centralized setting, even if the overall data set remains the same (Zhou et al., 2024; Dutta et al., 2024).

Another way to measure if clients benefit from FL can be to analyze the applicability domain of models (Heyndrickx et al., 2023). Further, it can be informative to perform sensitivity analyses with regard to several parameters, e.g., the number of clients, the level of data heterogeneity, and data quantity per client.

2.3. Data distribution & heterogeneity

The distributions of data sets across multiple companies typically differ in practice. The problem formulation in Eq. (1) assumes that all client data distributions $\mathcal{P}_k$ are iid samples from a global data distribution $\mathcal{P}$. However, several practical aspects typically lead to heterogeneous, non-iid data distributions, i.e., $\mathcal{P}_k \neq \mathcal{P}_l$ with $k \neq l$, also see Kairouz et al. (2021) and Zhou et al. (2024).

In FL, three types of data distributions are frequently considered, as shown with the example of molecular property data in Fig. 3. In *horizontal* distributions, different companies have data sets with the same features, e.g., pure component boiling points with molecules represented as SMILES, but different samples, i.e., different pure components. Vice versa, in *vertical distributions*, the companies have data sets with the same samples, the same pure components, but different features, e.g., different properties such as boiling points and autoignition temperatures. Lastly, the domains of the data sets can be related and enable *transfer learning*; typically, both samples and features are different, e.g., one company has boiling point data for small molecules represented as SMILES, whereas another company has biodegradability data on polymers represented as BigSMILES. Analogous examples can be derived for the process scale, e.g., (horizontal) two companies using a similar separation process for a specific mixture at different feed compositions and different operating conditions, (vertical) two chemical companies store material and process information along a product chain with one company being the raw material supplier and the other company being the manufacturer, and (transfer learning) two

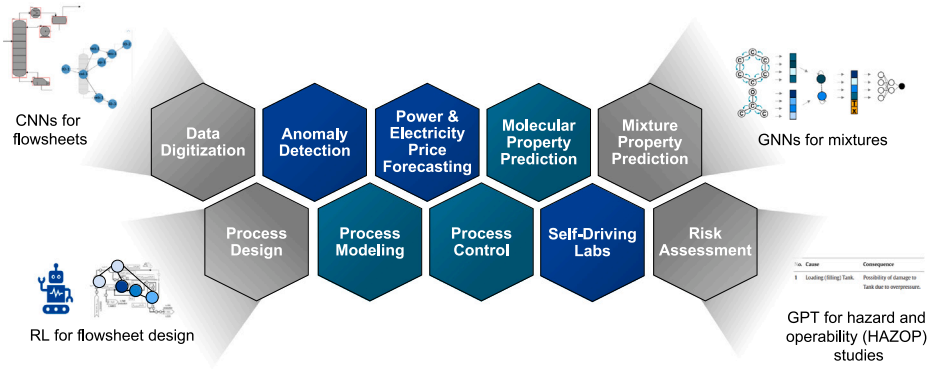


Fig. 4. Potential application fields of federated learning in chemical engineering. The colors indicate the current status in the literature: (green) has already been applied, (blue) has been proposed, or (gray) is rather unexplored. We further show potential applications within the unexplored fields. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: Images adapted from Rittig and Mitsos (2024), Theisen et al. (2023), Stops et al. (2023) and Ekramipooya et al. (2024).

Table 1
Decision impact of FL across the ChemE application domains, the metric by which impact can be assessed, and the practical, envisioned effect of FL relative to current practice.

Domain	Decision	Metric	Practical effect of FL
Data digitization	Which digitization model to deploy	Recognition accuracy; manual correction time	Legacy flowsheets/records across firms become machine-readable without disclosing any of them
Process modeling	Model structure; data to collect for a new unit	Model accuracy & extrapolation range	Model remains valid across wider operating ranges than one firm's data spans
Process design	Flowsheet topologies worth evaluating	Generalization to unseen design tasks	Generative models reused across design cases instead of retrained per study
Process control	Dynamic modeling; control strategy	Sample size; control performance	Control strategies with less data and identification effort
Anomaly detection	Fault flags; maintenance triggers	Detection vs. false-alarm rate	Detector generalizes across plants, avoiding per-unit retraining
Risk assessments	Hazard scenarios to flag in HAZOP	Hazard coverage; study time	Identification of hazards a single firm could miss
Power & Price forecasting	Scheduling; demand-side management	Forecast error; downstream scheduling cost	Scheduling and dispatch draw on price/generation while data stays on-site
Molecular property	Candidates to pursue in design/optimization	Prediction accuracy; chemical space coverage	Increased chemical design space; identification of promising molecules
Mixture property	Solvents or formulations to pursue	Prediction accuracy on new mixtures	More combinations of molecules are considered; identification of promising mixtures
Self-driving labs	Next experiment to run; planning	Sample efficiency	Fewer experiments to reach a target by reusing results from labs

companies employing different types of reactors for chemically related reactions.

The different data distributions result in heterogeneous, non-iid data. Next to different features and samples, the data sets of the clients can differ in the mean and variance of the feature and label distributions, also resulting in non-iid data. In fact, practical scenarios often face combinations of the heterogeneity sources listed above.

From an optimization perspective, the non-iid data sets can result in the local loss functions deviating from the global loss function, such that $\mathbb{E}_{D_k}[\mathcal{L}_k(\mathbf{w})] \neq \ell(\mathbf{w})$ (Yang et al., 2024). This can limit parameter convergence in training and thus have a negative impact on model performance. To address data heterogeneity, pre-training (Nguyen et al., 2023), multi-task, and transfer learning (Smith et al., 2017; Qi et al., 2024) can be used. Also, so-called *personalized* FL (Dinh et al., 2024), which balances the performance of local and global models, can be employed. Empirically, common FL algorithms such as FedAvg have often shown robust performance – even in heterogeneous data settings, cf. Kairouz et al. (2021), Shao et al. (2023), Wang et al. (2022) and Zhou et al. (2024).

3. Towards federated learning in chemical engineering

We now move to FL for ChemE. Having discussed existing applications in Section 1, we here focus on perspective FL applications (Section 3.1) and corresponding challenges (Section 3.2) within the chemical industry.

3.1. Perspective applications

We envision that FL enables chemical companies to collaborate in the development of ML models while respecting data privacy. We show perspective FL applications within the chemical industry in Fig. 4 and provide an overview of the potential practical effect of FL in Table 1. Note that we do not aim at a comprehensive review of ML applications in ChemE or chemistry, as such overviews can be found elsewhere (Schweidtmann et al., 2021; Strieth-Kalthoff et al., 2020; Daoutidis et al., 2024; Cheng et al., 2024). Rather, we highlight applications for which we think that FL can make a substantial impact based on readily available ML methods. We envision that this list will be extended in the future with further developments in ML for ChemE.

Data Digitization: Digitizing historical and newly obtained data is of major interest in the chemical industry. Many chemical enterprises hold large amounts of chemical and process data and are currently in the process of digitizing this data and corresponding knowledge. This includes the digitization of process descriptions and flowsheets, historical process data, and setups and results from lab experiments. To support this process, ML tools based on computer vision can be utilized, for example, for digitization of process flowsheets (Kang et al., 2019; Stürmer et al., 2024; Balhorn et al., 2022; Theisen et al., 2023). However, openly accessible flowsheets are mostly restricted to literature sources (Balhorn et al., 2022). Companies are typically not willing to share their flowsheets publicly. Therefore, FL could be utilized to enable joint training of such models. In general, using larger and more diverse data from the chemical process domain through FL will facilitate training ML models for digitizing chemical data without the need to share proprietary data.

Process Modeling: Going one step further, ML is greatly impacting the modeling of chemical processes in recent years (Schweidtmann et al., 2021; Daoutidis et al., 2024). Specifically, ML is often used in surrogate modeling (Bhosekar and Ierapetritou, 2018; Schweidtmann and Mitsos, 2019; McBride and Sundmacher, 2019; Esche et al., 2022; Misener and Biegler, 2023) and hybrid modeling (Sansana et al., 2021; Bradley et al., 2022; Schweidtmann et al., 2024) to capture the behavior of chemical phenomena, unit operations, or entire processes that are not yet fully understood mechanistically and/or have a high mathematical complexity. Furthermore, ML can guide the modeling of chemical processes. For example, ML models have been used to determine reaction networks (Meuwly, 2021; Margraf et al., 2023; Weber et al., 2021), rates (Jorner et al., 2021; Chung and Green, 2024), and kinetic models (Quaglio et al., 2020). However, the data available for process modeling is typically limited in terms of amount but also in terms of variation and diversity. Therefore, ML applications in process modeling typically face scarce data for similar conditions of a single phenomenon, unit operation, or entire process, and thus do not allow generalization beyond, which also hinders subsequent process optimization. The reasons for these data limitations can be manifold, including a low level of digitization, cost-intensive measurements and experiments, and narrow operation ranges of processes in practice, e.g., due to safety constraints. Here, FL has high potential to make use of data on similar applications from multiple companies, thereby allowing ML models to learn from larger and more diverse data sets. As highlighted in the introduction, Hayer has indicated this potential by using FL to train a neural network on aluminum electrolysis data from different chemical producers (Hayer, 2021). Another example for

future work would be a reaction kinetic classifier that is trained jointly by multiple companies on their proprietary kinetic data of different reactions (Quaglio et al., 2020). This will lead to enhanced robustness and generalization capabilities of ML-based chemical process models.

Process Design: Following the modeling stage, ML approaches have recently been transferred to process design (Schweidtmann, 2024). Specifically, generative ML has been utilized to design (Göttl et al., 2022; Vogel et al., 2023; Mann et al., 2024) and automatically correct (Schulze Balhorn et al., 2024) process flowsheets and predicting their control structures (Balhorn et al., 2025). For this, mainly transformer and reinforcement learning models have been proposed, cf. Stops et al. (2023), Vogel et al. (2023), Gao and Schweidtmann (2024) and Göttl et al. (2025). Such models can highly benefit from (pre-)training on large datasets. However, as mentioned above, public process data, e.g., in the form of flowsheets, is highly limited. This also results in the limitation of current models being typically trained for a specific process design case study (Gao and Schweidtmann, 2024). That is, the generative models hardly generalize and need to be retrained for new design tasks. Generalization in ML-based process design is thus an active research field, see, e.g., Göttl et al. (2025). Using FL for joint training of generative models for flowsheet design by several chemical companies enables learning from a more diverse set of processes, promising to advance generalization capabilities.

Process Control: Using ML for control of chemical processes has been an active research field in recent years (Tang and Daoutidis, 2022; Daoutidis et al., 2024). ML applications in control include system identification by capturing non-linear dynamics through neural networks (Chiuso and Pillonetto, 2019; Esche et al., 2022; Ahmed et al., 2025), and estimation of optimal control policies by reinforcement learning (Hoskins and Himmelblau, 1992; Badgwell et al., 2018; Nian et al., 2020; Devarakonda et al., 2025). As dynamic system responses and corresponding optimal control highly depend on the specifics of the process at hand, jointly training ML models based on different processes is challenging. Yet, in recent years, transfer learning has gained increasing attention, that is, to transfer knowledge from (multiple) source processes to new, related processes, see, e.g., Xiao et al. (2023), Arce Munoz et al. (2024) and Xiao et al. (2024). Such knowledge transfer promises to reduce modeling and system identification efforts for control, e.g., by reducing required data amounts. While companies presumably cannot share dynamic process data and control strategies due to privacy concerns, FL provides an opportunity to jointly train ML models for transfer learning purposes.

Furthermore, Xu and Wu (2024) recently demonstrated the use of FL for distributed model predictive control of a reaction network, in which information exchange between reactors is limited. In general, control in network structures, where multiple entities with proprietary process information participate and could benefit from cooperation, such as in distributed energy resource management systems for utilities, power and microgrids (Morstyn et al., 2016; Otashu et al., 2021; Dobbe et al., 2019), bear large potential for FL, with applications currently emerging (Lee and Choi, 2020; Li et al., 2023; Joshi et al., 2023). An interesting aspect here is to consider the close relationship of FL with algorithms for distributed optimization in process networks (Allman and Zhang, 2022; Li, 2025).

Anomaly Detection: Data on anomalies in real processes is naturally scarce, limiting ML applications. ML approaches have been actively explored to detect anomalies and faults in chemical processes and beyond (Qin and Chiang, 2019; Choi et al., 2021; Hartung et al., 2023). These approaches cover classification methods, such as support vector machines (SVMs), to distinguish normal process conditions from anomalies, and sequence models, such as recurrent neural networks, to analyze historical process time-series data, cf. overviews in Arunthavanathan et al. (2021) and Hartung et al. (2023). In particular, the Tennessee Eastman process (Downs and Vogel, 1993) serves as a benchmark for anomaly detection on process time-series data, which has been tackled by various ML models including generative adversarial networks (Yang and Feng, 2019), recurrent neural networks (Lomov et al., 2021), GNNs (Jia et al., 2023), and transformers (Schoch et al., 2024), also cf. Hartung et al. (2023) for a comprehensive comparison. However, the corresponding data set is synthetic and limited to the Tennessee Eastman process. The detection of anomalies in real process data is presumably more challenging (Hartung et al., 2023). Further, it is desirable to apply anomaly detection methods across a variety of different processes. It has recently been shown that transfer learning, using data from related tasks, can be beneficial in ML applications for anomaly detection in industrial time-series data (Yan et al., 2024). Having more real-world data from related processes usable for training through FL could lead to anomaly detection algorithms that have higher detection ratios and generalize across multiple process types. Thus, we envision that FL can enhance predictive maintenance and process safety.

Risk Assessments: The creation of risk assessments of chemical processes and plants, especially for hazard and operability (HAZOP) studies, is a critical but time-consuming task. Over recent years, data-driven approaches have been explored to support risk assessment studies. This includes the aforementioned ML models to predict failures and anomalies from process time-series data. Furthermore, ML approaches from natural language processing, such as large language models (LLMs), have been used to extract information from chemical accident databases (Single et al., 2020) and facilitate HAZOPs (Feng et al., 2021; Ekramipooya et al., 2024). A particular challenge is that relevant data for HAZOPs typically consists of different modalities, mainly text, images, and tables (Niu et al., 2024). Recent advances in ML provide methods that can utilize such multiple modalities, cf. Baltrušaitis et al. (2018), Liang et al. (2024) and S. Wu et al. (2024). The bottleneck for training associated models on chemical processes is the insufficient amount of publicly available data, as HAZOP studies from chemical enterprises are protected due to intellectual property and safety-critical information (Niu et al., 2024). We envision the training of multimodal ML models for risk assessments that can deal with tabular, text, and image data. As such, multimodal models need vast amounts of training data – even when the training is started from pre-trained LLMs–, it is desirable to leverage the process safety data from several chemical companies for training. FL presents a way to train such models, making use of industrial risk assessments without exposure of the underlying data.

Power and Electricity Price Forecasting: The future energy system will be decentralized and based on renewable, fluctuating energy sources (Scholtz et al., 2024). This gives rise to the field of forecasting time-series data like energy prices and power generation, which are highly relevant for the chemical industry. Due to the decentralized nature of future energy systems, FL can be a key technology to combine energy data from distributed sources while addressing data privacy (Zhang et al., 2025). In a recent review, ElRobrini et al. discuss such FL-based approaches in the light of photovoltaic and wind power applications (ElRobrini et al., 2024). We envision that enhancements in power and electricity price forecasting through FL will be beneficial for data-driven methods used in tackling flexibilisation of the chemical industry, like planning and scheduling, demand-side management, and economic nonlinear model predictive control (Tsay and Baldea, 2019).

Molecular Property Prediction: Moving from the process and systems scale to the molecular scale, we see large potential of FL to develop ML models for molecular property prediction, such as graph neural networks (Coley et al., 2017; Rittig et al., 2022; Reiser et al., 2022; Felton et al., 2024; Heid et al., 2024) and transformers (Chithrananda et al., 2020; Chen et al., 2023; Winter et al., 2025). In fact, the application of FL for learning tasks related to molecular data is increasingly targeted in pharmaceutical research (S. Chen et al., 2021; Huang et al., 2023; Zhu et al., 2022). Specifically, the idea is that multiple companies join a collaboration in training ML models to predict drug properties, whereas the proprietary molecular property data that has been gathered over the last decades by each company stays private. This

allows for the enlargement of both the coverage of the chemical space and the property range used for training beyond publicly available datasets or individual datasets owned by the companies. In doing so, the aforementioned MELLODDY project has demonstrated a successful application of FL on an industrial scale with multiple pharmaceutical companies (Heyndrickx et al., 2024). With the MELLODDY project serving as a blueprint, we advocate for similar collaborative efforts in FL projects for molecular property prediction by the chemical industry. As an example, we envision that ML models will be able to predict a wide range of thermodynamic properties for a wide spectrum of molecular classes that are relevant for chemical process design and optimization.

In conjunction with the improved prediction of properties, FL has recently been studied for computer-aided molecular design (Manu et al., 2024; Guo et al., 2024) and multimodal molecular representation learning (Feng et al., 0000). These works can serve as starting points for the collaborative development of data-driven molecular design models (Gómez-Bombarelli et al., 2018; Bilodeau et al., 2022; Decardi-Nelson et al., 2024a), which can, for example, be used for solvent (J. Zhang et al., 2023; Pirnay et al., 2025), fuel (Rittig et al., 2023d), and catalyst (dos Passos Gomes et al., 2021; Ishikawa, 2022; Schilter et al., 2023) design, in the chemical industry.

Mixture Property Prediction: Building on ML-based property prediction for pure components, great efforts have recently been dedicated to predicting properties of mixtures, which are highly relevant for chemical processes. Here, ML models, such as graph neural networks (Sanchez Medina et al., 2022; Rittig et al., 2023a; Vermeire and Green, 2021b; Qin et al., 2023; Leenhouts et al., 2025), matrix completion methods (Jirasek et al., 2020; G. Chen et al., 2021), and transformers (Winter et al., 2022), have been trained on quantum mechanical data, publicly available datasets, or commercial databases. Beyond these data, chemical companies have created private property databases of molecules and their mixtures over the last decades. Analogously to pure component properties, we envision that FL is used by academia, database organizations, and chemical companies to jointly develop ML models, without the need to share their property data. This will allow to utilize additional combinations of molecules and corresponding mixture properties during training and thereby enhancing predictive capabilities.

Self-Driving Labs: The advent of self-driving labs promises to catalyze chemical discovery and reaction optimization (Abolhasani and Kumacheva, 2023; Tom et al., 2024). The parallel execution of lab tasks by multiple self-driving labs located in different regions of the world has been actively studied (Skilton et al., 2015; Fitzpatrick et al., 2018; Bai et al., 2024). As data-driven models become an integral part of such self-driving labs, e.g., for automated synthesis prediction, spectra analysis, and experimental planning, we see that FL provides a way to use larger datasets for model training, especially if standardized protocols for self-driving labs mature, also cf. Tom et al. (2024).

Overall, we find great potential of FL for numerous applications in ChemE, ranging from the molecular to the process scale. We advocate for collaborative efforts between the chemical industry, database organizations, and academia to adapt and apply FL, thereby advancing ML models without the need for conducting expensive experiments to obtain new data – all while addressing data privacy. Ultimately, we envision using FL to leverage the substantial data and knowledge repositories of chemical companies to develop and fine-tune multi-purpose ML models (Bommasani et al., 2021; Das et al., 2024; Decardi-Nelson et al., 2024b; Li et al., 2024; Ren et al., 2024; Bran et al., 2023; Choi et al., 2025), also known as foundation models, for the field of ChemE, cf. Sani et al. (2024).

3.2. Challenges & research directions

Initiating and implementing FL in projects within the chemical industry comes with several challenges. Next, we discuss these challenges and provide related research directions.

Incentives: The basic premise for FL and corresponding potential improvement of ML models is that chemical companies and further data holders enter into collaborations. However, these companies are competitors – typically with conflicting interests –, so in most cases, they are not willing to openly share their proprietary data with others. Consequently, the initiation of FL projects inherently poses a trade-off between cooperation and competition (Huang et al., 2022c) and is thus challenging. A common approach to analyze this trade-off is game theory, which has already been applied to deduce incentives for FL projects in general (Tu et al., 2022; Huang et al., 2022b, 2024b). Here, a major aspect is to value the data that clients contribute to increase the model performance (Z. Wu et al., 2024) because the core principle of FL is that clients' data cannot be accessed by other clients. For example, one client might contribute more data than another client and will therefore expect a higher reward than the other client for taking part in the training process (Sim et al., 2024). Thus, ensuring a fair FL process with corresponding incentives is challenging (Huang et al., 2024). Game theoretical approaches in FL are so far restricted to a few competitors (Huang et al., 2024a) and have not been investigated for specific characteristics of the chemical industry, therefore requiring further research. Most importantly, demonstrating the benefits of FL for exemplary ChemE tasks, also in simulated settings, is needed to incentivize chemical companies to initiate FL projects.

Adversarial Attacks: Potential data privacy threats can discourage clients from participating in FL projects. Adversaries might attempt to reverse engineer data samples based on the shared global model parameters and corresponding model updates, sometimes referred to as gradient or model inversion (Geiping et al., 2020). For example, Krüger et al. recently investigated the possibility of extracting information on molecular structures present in the data sets of individual clients from a model trained through FL for property prediction in drug applications (Krüger et al., 2025). Further, clients could try to introduce bias into the model to deteriorate the predictive performance of other clients, referred to as data and model poisoning (Bagdasaryan et al., 2020). Therefore, data privacy and security are an active and crucial field of research in FL (Zhao et al., 2024), including approaches such as secure aggregation (Bonawitz et al., 2016) or differential privacy (Wei et al., 2020; El Ouadrhiri and Abdelhadi, 2022). While recent works have investigated adversarial attacks on ML models in ChemE by external parties, e.g., for process control (Koay et al., 2023) and demand-side management (Cramer and Gao, 2024), such investigations are lacking for FL setups, particularly at the process scale and require further research for industrial applications.

Data Heterogeneity & Uncertainty: The majority of chemical and process data will naturally show some level of heterogeneity and uncertainty. This originates from chemical companies working on different regions of the chemical space, operating processes with different raw materials and at different conditions in varying demographic locations, and using different types and quality of equipment, e.g., for sensors. Such heterogeneous data can result in performance degradation and global model biases, and also to lower prediction accuracy, e.g., for chemical species from regions of the chemical space that are underrepresented compared to species that are frequently used by multiple chemical companies. In addition, results from lab experiments will differ by experimental uncertainty, and process data will show measurement noise, referred to as aleatoric noise, cf. Heid et al. (2023). In such cases, data samples might share the same features but have different labels, also cf. Section 2.3, which can lead to slower convergence and instability in global model training. We refer to overviews in Zhao et al. (2018), Li et al. (2022) and Lu et al. (2024) for further effects of heterogeneous, *non-iid* data on FL applications. Generally, dealing with data heterogeneity is an inherent challenge in FL research for chemistry and ChemE. We anticipate that data availability from various sources will enable the refinement of aleatoric uncertainties in chemical and process data and of corresponding uncertainty quantification methods in a federated setting, similar to Linsner et al. (2021) and Y. Zhang et al.

(2023). Furthermore, research on FL has recently addressed the field of heterogeneity and label noise across clients, e.g., in Tuor et al. (2021), Fang and Ye (2022), Yang et al. (2022) and Tsouvalas et al. (2024), which can be a starting point for further investigations in ChemE.

Furthermore, we emphasize that data formats can differ greatly across chemical companies. This poses the challenge of deciding on input formats of the data used for training an ML model with FL. However, since data is not collected and stored centrally, but rather remains within individual companies, they can continue working with their existing data formats and implement parsers within the FL framework. As such, real-world industrial collaborations should prepare centralized protocols and corresponding implementations that standardize data handling and preprocessing, including secure aggregation of data statistics. For example, the MELLODDY project provided each participating company with an open-source Python package for data standardization (Heyndrickx et al., 2024).

Benchmarks: To drive research in FL for ChemE, we advocate to create open-source benchmarks. This will require the collection and publication of data sets by experts from academia and industry in order to cover realistic data distribution scenarios across chemical companies – also including heterogeneous, *non-iid* data. Based on these benchmarks, leaderboards should be introduced to make different FL approaches comparable in a straightforward manner. Developed FL models could also be evaluated by individual companies on their proprietary test data sets, with only the evaluation metrics shared publicly. As the research for FL in ChemE is at an early stage, designing such benchmarks can facilitate establishing data characteristics and research needs relevant in the chemical industry.

Overall, FL comes with an interplay of various aspects, such as data heterogeneity, privacy concerns, and incentives, and thus requires further research for ChemE applications in joint efforts of academia and industry. Practical implementations of FL projects also require setting up computational infrastructures and coding frameworks under consideration of cyber security and legal aspects, cf. Woisetschläger et al. (2024) and Chik and Gamper (2024). For this, the chemical industry can use successful large-scale industrial projects in related domains (Heyndrickx et al., 2023; Oldenhof et al., 2023; AISB Consortium, 0000).

4. Exemplary case studies

In this section, we demonstrate the application of FL in ChemE. As illustrated in Fig. 5, we consider two case studies motivated by separation processes, covering the molecular scale and the process scale. In the first case study, we investigate FL for the molecular scale by training graph neural networks that predict mixture properties (Section 4.1). In the second case study, we explore FL for the process scale by training a data-driven model that predicts dynamic system responses of a distillation column with varying feed composition (Section 4.2).

We stress that these case studies simulate FL scenarios, i.e., we take publicly available data sets and emulate data distributions across chemical companies, making the case studies openly accessible and reproducible. In industrial settings, additional challenges, such as computational infrastructure for communication and model sharing, client dropout and connection issues, cyber security, and adversarial attacks, need to be considered and require further research, see Section 3.2.

4.1. Case study I: Federated learning of activity coefficients in binary mixtures

We first investigate FL of GNNs for predicting activity coefficients in binary mixtures at infinite dilution, which are particularly relevant for chemical separation processes. FL has already been applied to GNNs, cf. R. Liu et al. (2024), also for prediction of pure component properties, e.g., phyiscochemical properties (Zhu et al., 2022) and drug properties (Krüger et al., 2024). However, the application to mixture properties relevant to ChemE is missing so far.

4.1.1. Model

The prediction of activity coefficients in binary mixtures with ML is an active area of research (Jirasek et al., 2020; Winter et al., 2022, 2023; Specht et al., 2024; Ottaiano and Martins, 2024), with GNNs being a prominent method, cf. Felton et al. (2022), Qin et al. (2023), Sanchez Medina et al. (2023), Rittig et al. (2023b,c) and Rittig and Mitsos (2024). As shown in Fig. 5, GNNs use graph convolutions to learn a fingerprint vector of a molecule based on its graph representations, which is then mapped to a property of interest by a multilayer perceptron (MLP) (Gilmer et al., 2017; Rittig et al., 2022; Reiser et al., 2022). For predicting activity coefficients of binary mixtures, we consider a GNN model from our previous work (Rittig et al., 2023c): To handle mixtures, we take the two molecular graphs of the solute and solvent as input and combine their fingerprints to a mixture fingerprint in an aggregation step, which is based on a mixture graph, cf. Qin et al. (2023) and Rittig et al. (2023c). The mixture fingerprint is then mapped to the activity coefficients.

4.1.2. Data set & federated learning scenarios

We use the infinite dilution activity coefficient dataset published by Brouwer and Schuur (2019) in the sanitized version by Winter et al. (2022) and Sanchez Medina et al. (2023). The data set consists of 18,016 activity coefficient samples of binary mixtures at infinite dilution at varying temperatures.

We use this dataset as a starting point to create two federated learning scenarios in a setting with four companies:

- **even-random (whole model shared)**: We randomly partition the data set into subsets of equal size (25%) and assign each to a company. Thus, the companies' data sets tend to be distributed identically and independently (*iid*), cf. Section 2.1, allowing us to validate FL in a rather ideal setting. Here, we assume that the companies share the whole model, and we use the FedAvg algorithm for training.
- **uneven-scaffold (partial model shared)**: We partition the data based on Bemis-Murcko scaffolds (Landrum, 0000; Bemis and Murcko, 1996) with uneven partition sizes of 40%, 30%, 20%, and 10% for the companies, respectively. Using scaffolds captures that different companies work on different regions of the chemical space. We partition the data based on scaffolds of the solvents. We find that acyclic molecules and benzene are the most common scaffolds. Clients 1 and 2 hold partitions of solely acyclic solvents, client 3's partition contains acyclic, benzene-based, and cyclohexane solvents, and client 4's partition consists of the remaining cyclohexane solvents together with a variety of additional scaffolds, cf. Bemis and Murcko (1996). We note that this split is a deliberately extreme, non-IID stress test rather than a realistic partition. Inspired by the MELLODDY project (Heyndrickx et al., 2023), we assume that the participating companies do not share their full models, but only the GNN layers up to the mixture embedding. This partial sharing serves as a heuristic privacy mitigation, limiting the exposure of both training data and downstream model components. The protective effect, however, is rather empirical and depends on the specific privacy threat considered, cf. Section 3.2 and Jourdan et al. (2021). That is, the prediction MLP is not shared and remains private, referred to as FedPer, see Arivazhagan et al. (2019) and Pillutla et al. (2022). This scenario covers several heterogeneity aspects, cf. Section 2.3, which we expect in practical data and thus enables evaluating FL in a more realistic setting.

As baselines, we consider (i) each company training a model only on their **private data** set and (ii) training a model on the **full data** set of activity coefficient, hence a centralized setting 2.2. In each case, we split the (partitioned) data sets into 70% training, 15% validation, and 15% test set.

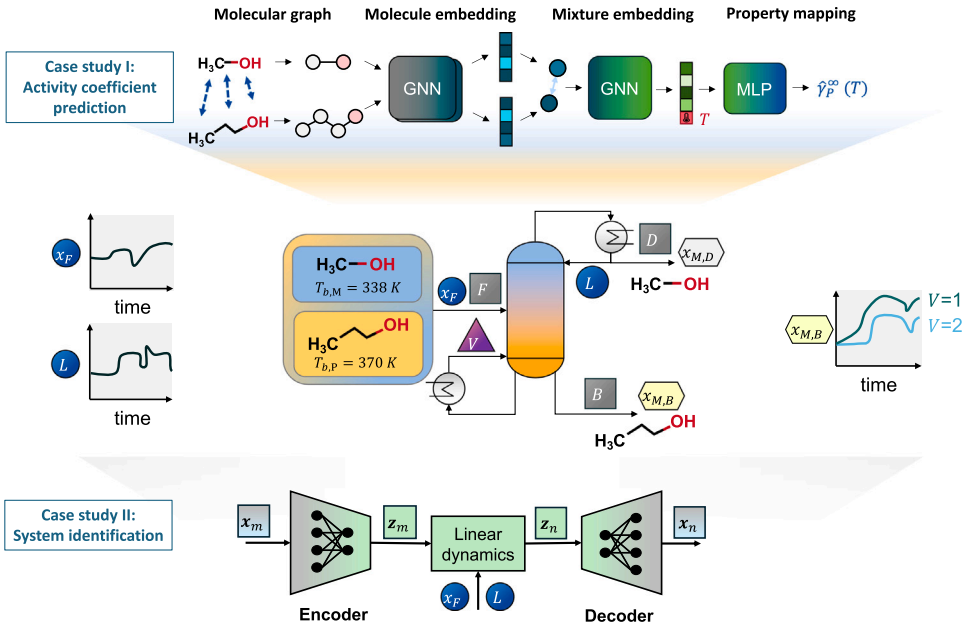


Fig. 5. Illustration of process for separating methanol and propanol by means of a distillation column with two related chemical engineering tasks: (I) predicting activity coefficients of binary mixtures by graph neural networks (GNNs), and (II) identifying a low-dimensional model for predicting system responses for dynamic input flows.

To assess the model quality, we report the following metrics. We compute the test set MSE after local training before FL aggregation, i.e., for the local models, and after FL aggregation, hence for the global model. Further, we compute the RIPtoP metric, cf. Eq. (2), where we use the test set MSE after the first round of local training as the *baseline* metric and the test set MSE after the last round of aggregation as *MoI* metric.

4.1.3. Implementation & hyperparameter

We use the GNN implementation of our MoLprop framework (Rittig et al., 2022), which is based on PyTorch Geometric (Fey and Lenssen, 2019) and the RDKit (Landrum, 0000). We use the default hyperparameters from MoLprop throughout all experiments and only remove the chirality atom feature because it has no variation in the data set.

For the implementation of the FL algorithms, we use Flower 1.7 (Beutel et al., 2020). Throughout all experiments, we use four clients, assume that all clients are connected at every communication round, and use the same hyperparameters for each client. We use 30 communication rounds and 150 local epochs with exponential learning rate decay and early stopping. We conduct each of our experiments ten times and report the arithmetic mean together with the standard error.

4.1.4. Results & discussion

We first report and discuss the results for the even-random scenario, where the whole model is shared among the clients. An important question is whether FL provides benefits for the clients in comparison to training just on their own data. In Fig. 6(a), the test MSEs for the models of the four individual clients are shown over the communication rounds. These models are evaluated after individual model training by the clients in each communication round, thus before model aggregation. The MSE values in the first round range from 0.123 to 0.146. These MSE values correspond to a setting in which each company simply trains a model on its own local data set without applying any FL, thus requiring only a single training run and considerably less computational effort than FL– an overhead we assume not to be a limiting factor here. In the following communication rounds, the prediction errors decrease significantly for all clients. Specifically, after 30 rounds, the MSE values range from 0.059 to 0.068; that is, the error range is more than halved. Therefore, FL yields a consistent performance gain for the clients in the even-random scenario.

In Fig. 6(b), we further show the average test MSE of the individual clients (in red) in comparison to the *global* model (in blue), which corresponds to the aggregated FL model after each communication round. After the first communication round, the test MSE of the global model is very high with a value of 2.770, which can be explained by the fact that the models trained by the individual clients are not calibrated to each other, so aggregating them highly decreases the prediction performance. Over the subsequent communication and retraining rounds, the test MSE of the global model decreases drastically and starts to converge, reaching a value of 0.050 after 30 rounds, yielding a RIPtoP (cf. Section 2.2) of 62%, which is calculated based on the average client model after round 1 with a MSE of 0.130. Interestingly, the MSE of the global model is below the MSE of 0.064 achieved by the average client models. This indicates that the individual training of the clients on their individual data sets in the continued FL communication rounds leads to local overfitting (client drift), whereas aggregation acts as a regularizer and lets the global model generalize better.

Another interesting aspect is the comparison to the *centralized* model, which represents the hypothetical, but practically infeasible case of training on the entire dataset. Fig. 6(b) shows the performance for the centralized model (in black and dashed), which does not depend on the communication rounds as it is just trained once without a FL setup. The centralized model exhibits a test MSE of 0.046 and a test MAE of 0.11, cf. Fig. 6(b), which is on par with the GNN developed by Sanchez Medina et al. (2023). Thus, the global model obtained through FL, exhibiting an MSE of 0.050, achieves a very similar prediction quality. In the even-random scenario, FL therefore enables almost full utilization of the molecular property information from the individual clients without the need to share proprietary data, demonstrating great potential for collaborative model training.

Next, we consider the results for the uneven-scaffold scenario, where clients only share a part of the model. The corresponding key question is whether the obtained benefits through FL can also be translated to more challenging and practical settings that come with data heterogeneity.

Considering the test MSE of the individual client models before aggregation, shown in Fig. 7, we find similar results to the even-random split: the test MSE values decrease over the communication

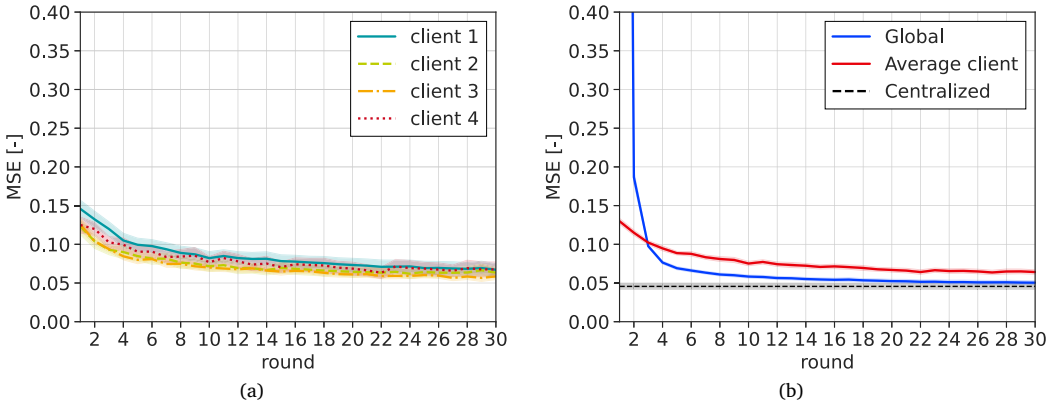


Fig. 6. The test mean squared error (MSE) values for the **even-random (whole model shared)** scenario: (a) results for the respective model of the *individual clients*; (b) aggregated results for the *global* model that is obtained after each round through aggregation of the individual clients' models, the *average client*, i.e., the averaged performances of the individual clients' models from (a), and *centralized* model that is trained on the whole data set data and serves as a baseline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

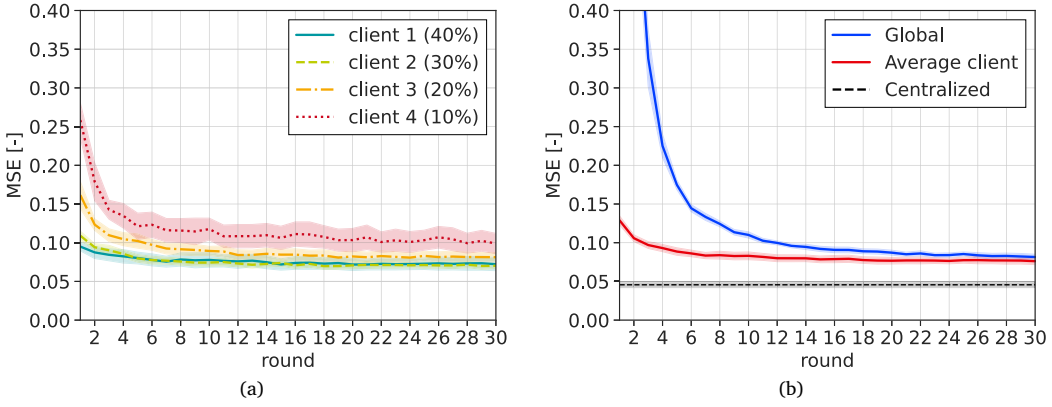


Fig. 7. The test mean squared error (MSE) values for the **uneven-scaffold (partial model shared)** scenario: (a) results for the respective model of the *individual clients* with different data set fractions; (b) aggregated results for the *global* model that is obtained after each round through aggregation of the individual clients' models, the *average client*, i.e., the averaged performances of the individual clients' models from (a), and *centralized* model that is trained on the whole data set data and serves as a baseline.

rounds; all clients benefit from FL. As expected, client 1 with the largest share of the data (40%) shows the smallest decrease in MSE (0.095 to 0.072), whereas client 4 with the smallest share of data (10%) obtains the largest improvement of the MSE (0.258 to 0.099). The different performance increases highlight the challenge of generating incentives and valuating contributions to a FL collaboration, cf. Section 3.2.

We further investigate the global model performance in Fig. 7. It is important to note that, in the uneven-scaffold scenario, the global model refers to a partly shared ML model between the clients, cf. 4.1.2. Hence for model evaluation, the globally shared part, i.e., the GNN layers, are combined with the private MLP heads of the clients, thereby representing the performance the clients obtain by using the received global model parameters; we refer to the Flower framework for implementation details (Beutel et al., 2020). Again, we find a drastic decrease in the test MSE in the first communication rounds, followed by rather slight decreases in later rounds. In contrast to the even-random case, we find that the global model and average client models are on a similar level after 30 rounds, with test MSE values of 0.082 and 0.076, respectively. We suspect that both the increased data heterogeneity of the individual clients and the fact that only a part of the model is shared make it more difficult to balance the individual model contributions. This is also reflected in the comparison to the centralized model, which is the same model as in the even-random scenario, and thus exhibits a test MSE of 0.046. Still, the final global model outperforms models that are trained by each client individually without FL, which have an average MSE of 0.156, resulting in a RIPtoP of 47%. In fact, this includes both the average client with a test MSE of 0.129 and the data-richest client (i.e., client 1) with a test MSE of 0.095. Overall, our results show that FL provides all clients with an enhanced model performance, even in more challenging settings of data heterogeneity. Future work should further investigate the effect of FL on the applicability domain of ML-based property prediction models, as in Heyndrickx et al. (2024).

4.2. Case study II: Federated learning for system identification of chemical processes

As a second case study, we explore FL for system identification of chemical processes, which is an active field in ChemE and process control (Z. Wu et al., 2024). While transfer learning has been proposed to reduce the amount of data needed to fit data-driven models for system identification and modeling (Niu et al., 2022; Xiao et al., 2023, 2024), it requires dataset aggregation for pretraining, hence does not preserve data privacy. Thus, we propose to utilize FL as a privacy-increasing alternative. Specifically, we investigate the training of a data-driven model through FL to capture the dynamics of a high-purity distillation column.

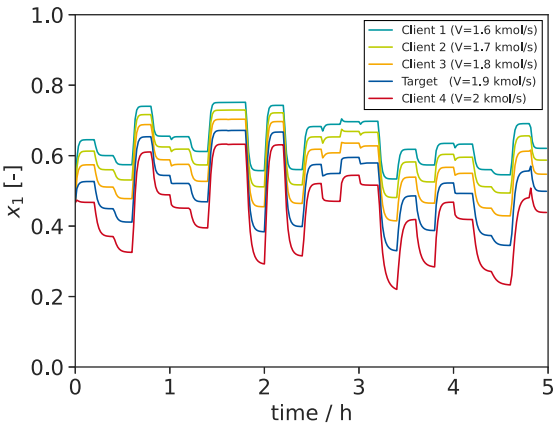


Fig. 8. The evolution of the composition in the condenser, $x_1(t)$, in the test sets for different values of the vapor flow rate V. All simulations start from the same initial state and are subject to the same input signals $\mathbf{u}(t)$.

4.2.1. Model

We use a multi-input multi-output (MIMO) Wiener-type model that was developed for system identification and model reduction by Schulze et al. (2022). Based on Koopman theory, the model architecture consists of a linear state-space model sandwiched by an autoencoder, as schematically illustrated in Fig. 5. That is, the physical states $\mathbf{x}(t)$ are first mapped to a lower-dimensional space, the latent space $\mathbf{z}(t)$, by a neural network, referred to as encoder. Then, the linear state-space model is applied to capture the system dynamics. Subsequently, the variables in the latent space are decoded back to the physical states by a second neural network, referred to as decoder. The model is trained on full-state trajectories of length p that are simulated responses to steps in the system inputs $\mathbf{u}(t)$. For more details on the model architecture, the training process, and the loss function, we refer to the original work (Schulze et al., 2022).

As a chemical process, we consider a distillation column, as illustrated in Fig. 5. Specifically, we choose the methanol-propanol high-purity distillation column model studied in Jacobsen and Skogestad (1991), Pearson and Pottmann (2000) and Schulze et al. (2022) and reimplemented in Schulze et al. (2022). The column consists of eight trays, a reboiler, and a condenser. As in Schulze et al. (2022), the feed composition $x_F(t)$ and the liquid flow rate $L(t)$ are inputs of the system $\mathbf{u}(t)$ that vary over time.

4.2.2. Data set & federated learning scenario

We simulate five different column models and create heterogeneity with different values of the vapor flow rate $V(t)$, cf. Fig. 5. Specifically, we generate data sets of 192 trajectories assuming a constant $V \in \{1.6, 1.7, 1.8, 1.9, 2.0\}$ kmol/s. Exemplary trajectories for different vapor flow rates are shown in Fig. 8.

For testing FL, we consider a scenario in which five different companies operate the same distillation column but each at different vapor flow rates $V \in \{1.6, 1.7, 1.8, 2.0\}$ kmol/s: Four companies have extensively collected system response data, i.e., a *rich* data set of 192 trajectories, which is the same amount of data as used by Schulze et al. (2022). A fifth company, the *target* company, which operates the same distillation column at a different vapor flow rate ($V = 1.9$ kmol/s), has collected only a *scarce* data set with 2 trajectories so far. We are thus interested in whether FL provides a benefit in training a MIMO Wiener-type model for system identification for the company with fewer data. For this, we assume that the companies apply FL with FedAvg as the aggregation algorithm and share the whole model.

We compare the federated learning scenario to the fifth company (i) training a model only on their **private scarce data** set of 2 trajectories and (ii) training a model on a **full data** set of 192 trajectories for $V = 1.9$ kmol/s, i.e., the hypothetical scenario of having the same amount of data available for training as the other four companies.

All models are evaluated in test sets of five hours in length, see Fig. 8. For testing, we pass only the initial state $\mathbf{x}_0$ and the inputs $\mathbf{u}(t)$ to the model. Then, we perform a multistep prediction for the entire five hours and compute the multi-step prediction MSE, cf. Schulze et al. (2022).

4.2.3. Implementation & hyperparameters

For the MIMO Wiener-type model, we use the implementation by Schulze et al. (2022), which is in Python and uses TensorFlow (TensorFlow Developers, 2024). The mechanistic column models are implemented in Pyomo (Hart et al., 2011; Bynum et al., 2021) and simulated using CasADi (Andersson et al., 2019), analogously to Schulze et al. (2022).

We scale all data linearly to the range between zero and one based on the maximum and minimum of each state in the source data, such that we apply the same scaling to all column models. For the autoencoder, we use the same hyperparameters as in Schulze et al. (2022) and map the ten physical states to a latent space of dimension two. We repeat all computational runs with ten different seeds.

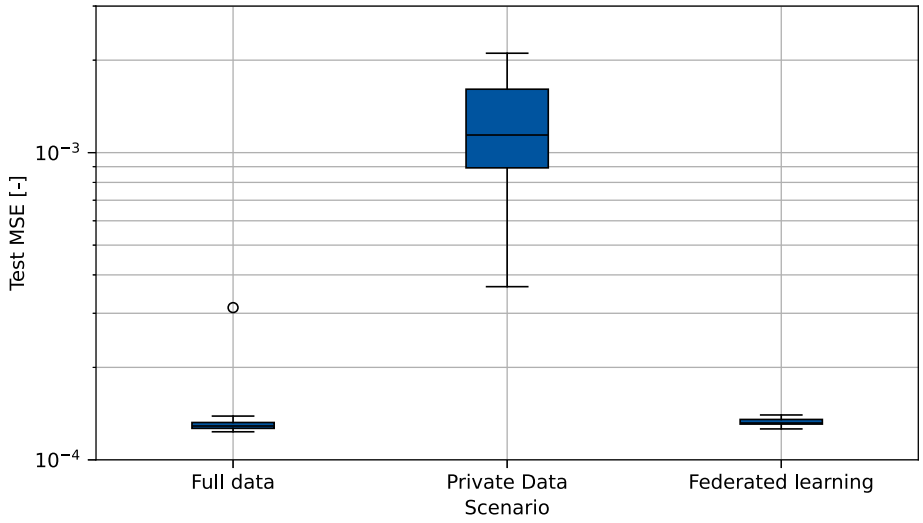


Fig. 9. Test mean squared error (MSE) for two data availability scenarios, full data set with 192 trajectories and private data set with 2 trajectories, in comparison to federated learning in collaboration with multiple companies. Each boxplot is based on 10 runs with different seeds.

4.2.4. Results & discussion

We show the results for FL in comparison to the two baselines, i.e., training on the private scarce data set and on the full data set, in Fig. 9. The test MSE for the full data set baseline is 1.47×10^{-4} (and 1.16×10^{-4} excluding the outlier, which still exhibits an accuracy in the same order of magnitude, and we can attribute to the stochastic training of the autoencoder, see Fig. 9), which is similar to the errors reported in Schulze et al. (2022). The private data set baseline results in an MSE of 1.44×10^{-3}, hence a significant increase of about one order of magnitude compared to the full data set case. Thus, the small amount of data available for the target company does not suffice to train a model with similar accuracy as the other companies with more historical system response data. We note that such an accuracy might still suffice for practical application in process control; yet, we want to investigate if this accuracy can be improved with FL.

Considering the FL scenario, we observe an MSE of 1.33×10^{-4}, which corresponds to a RIPtoP of 91%, cf. Fig. 9. FL thus enables the target company to decrease the error by an order of magnitude to a level similar to the other companies, without having to collect any additional process data. Therefore, FL enables leveraging knowledge from similar source processes without exchange of underlying process data, as for example required in transfer learning when a model is pretrained on the data of the four companies without FL (Niu et al., 2022; Xiao et al., 2023).

Notably, in our case study, the system responses for different vapor flow rates show similar dynamics, as can be seen in Fig. 8. Future work could thus investigate whether FL can also provide accuracy improvements for more heterogeneous dynamics. It would also be interesting to test FL for PINNs (Velioglu et al., 2025; Chen et al., 2024) and transfer learning in process control applications (Xiao et al., 2024; Arce Munoz et al., 2024). In summary, FL shows promise for system identification, with many scenarios yet to be explored.

5. Conclusion

FL allows to jointly advance ML models without sharing data and is thus highly interesting for the chemical industry, where companies store large amounts of proprietary chemical and process data. We identify numerous promising applications of FL in ChemE, including digitization, modeling, and design methods from the molecular up to the process scale. We showcase in two such applications, namely mixture property prediction and system identification, that chemical companies can use FL in collaborations to greatly improve their ML models without the need to share proprietary data, making it highly promising for practical applications. We anticipate and advocate for further FL applications in ChemE, e.g., in process flowsheet digitization and anomaly detection.

It will be of major importance to define benchmark data sets, raise incentives, elucidate data heterogeneity, and discuss potential risk through adversarial attacks in strong collaboration of academia, database organization, and chemical companies. Pilot projects with practical applications of FL will further elucidate challenges of data and collaboration within the chemical industry – with the goal of addressing data scarcity and advancing ML in ChemE.

CRediT authorship contribution statement

Jan G. Rittig: Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Clemens Kortmann:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This project was partly funded by the Werner Siemens Foundation within the WSS project of the century “catalaix”. This work was also performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). Simulations were performed with computing resources granted by RWTH Aachen University under project “thes1813”. We sincerely thank Alexander Mitsos for his valuable support and feedback within this project.

References

Abolhasani, M., Kumacheva, E., 2023. The rise of self-driving labs in chemical and materials sciences. Nat. Synth. 2 (6), 483–492.

Ahmed, A., del Rio-Chanona, E.A., Mercangöz, M., 2025. Comparative study of machine learning and system identification for process systems engineering dynamics. Ind. Eng. Chem. Res.

AISB Consortium, AISB Consortium: https://www.apheris.com/industries/aisb. URL https://www.apheris.com/industries/aisb.

Allman, A., Zhang, Q., 2022. Distributed fairness-guided optimization for coordinated demand response in multi-stakeholder process networks. Comput. Chem. Eng. 161, 107777. http://dx.doi.org/10.1016/j.compchemeng.2022.107777, URL https://www.sciencedirect.com/science/article/pii/S0098135422001181.

Andersson, J.A.E., Gillis, J., Horn, G., Rawlings, J.B., Diehl, M., 2019. CasADi: a software framework for nonlinear optimization and optimal control. Math. Program. Comput. 11 (1), 1–36. http://dx.doi.org/10.1007/s12532-018-0139-4, URL https://link.springer.com/article/10.1007/s12532-018-0139-4.

Arce Munoz, S., Pershing, J., Hedengren, J.D., 2024. Physics-informed transfer learning for process control applications. Ind. Eng. Chem. Res. http://dx.doi.org/10.1021/acs.iecr.4c02781.

Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S., 2019. Federated learning with personalization layers. arXiv Preprint arXiv:1912.00818v1. URL http://arxiv.org/pdf/1912.00818.

Arunthavanathan, R., Khan, F., Ahmed, S., Imtiaz, S., 2021. An analysis of process fault diagnosis methods from safety perspectives. Comput. Chem. Eng. 145, 107197.

Badgwell, T.A., Lee, J.H., Liu, K.-H., 2018. Reinforcement learning–overview of recent progress and implications for process control. Comput. Aided Chem. Eng. 44, 71–85.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V., 2020. How to backdoor federated learning. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2938–2948.

Bai, J., Mosbach, S., Taylor, C.J., Karan, D., Lee, K.F., Rihm, S.D., Akroyd, J., Lapkin, A.A., Kraft, M., 2024. A dynamic knowledge graph approach to distributed self-driving laboratories. Nat. Commun. 15 (1), 462. http://dx.doi.org/10.1038/s41467-023-44599-9, URL https://www.nature.com/articles/s41467-023-44599-9.

Balhorn, L.S., Degens, K., Schweidtmann, A.M., 2025. Graph-to-sfiles: Control structure prediction from process topologies using generative artificial intelligence. Comput. Chem. Eng. 109121.

Balhorn, L.S., Gao, Q., Goldstein, D., Schweidtmann, A.M., 2022. Flowsheet recognition using deep convolutional neural networks. In: Yamashita, Y., Kano, M. (Eds.), 14th International Symposium on Process Systems Engineering. In: Computer-Aided Chemical Engineering, vol. 49, Elsevier, Amsterdam and Boston and Heidelberg and London, pp. 1567–1572. http://dx.doi.org/10.1016/B978-0-323-85159-6.50261-X.

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2018. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443.

Bemis, G.W., Murcko, M.A., 1996. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 39 (15), 2887–2893. http://dx.doi.org/10.1021/jm9602928.

Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., de Gusmão, P.P.B., Lane, N.D., 2020. Flower: A friendly federated learning research framework. arXiv Preprint arXiv:2007.14390v5. URL http://arxiv.org/pdf/2007.14390.

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. Comput. Chem. Eng. 108, 250–267.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., Jensen, K.F., 2022. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 12 (5), e1608. http://dx.doi.org/10.1002/wcms.1608.

Boiko, D.A., MacKnight, R., Kline, B., Gomes, G., 2023. Autonomous chemical research with large language models. Nature 624 (7992), 570–578.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T.,

Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P., 2021. On the opportunities and risks of foundation models. arXiv Preprint arXiv:2108.07258v3. URL http://arxiv.org/pdf/2108.07258.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K., 2016. Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482.

Bradley, W., Kim, J., Kilwein, Z., Blakely, L., Eydenberg, M., Jalvin, J., Laird, C., Boukouvala, F., 2022. Perspectives on the integration between first-principles and data-driven modeling. Comput. Chem. Eng. 166, 107898.

Bran, A.M., Cox, S., Schilter, O., Baldassari, C., White, A.D., Schwaller, P., 2023. Chemcrow: Augmenting large-language models with chemistry tools. arXiv preprint arXiv:2304.05376.

Bran, A.M., Cox, S., Schilter, O., Baldassari, C., White, A.D., Schwaller, P., 2024. Augmenting large language models with chemistry tools. Nat. Mach. Intell. 6 (5), 525–535.

Brouwer, T., Schuur, B., 2019. Model performances evaluated for infinite dilution activity coefficients prediction at 298.15 K. Ind. Eng. Chem. Res. 58 (20), 8903–8914. http://dx.doi.org/10.1021/acs.iecr.9b00727.

Bynum, M.L., Hackebeil, G.A., Hart, W.E., Laird, C.D., Nicholson, B.L., Siirola, J.D., Watson, J.-P., Woodruff, D.L., 2021. Pyomo - optimization modeling in Python, third ed. In: Springer eBook Collection, vol. 67, Springer, Cham, http://dx.doi.org/10.1007/978-3-030-68928-5.

Chen, H., Flores, G.E.C., Li, C., 2024. Physics-informed neural networks with hard linear equality constraints. Comput. Chem. Eng. 189, 108764. http://dx.doi.org/10.1016/j.compchemeng.2024.108764.

Chen, G., Song, Z., Qi, Z., Sundmacher, K., 2021. Neural recommender system for the activity coefficient prediction and UNIFAC model extension of ionic liquid-solute systems. AIChE J. 67 (4), e17171.

Chen, G., Song, Z., Qi, Z., Sundmacher, K., 2023. Generalizing property prediction of ionic liquids from limited labeled data: a one-stop framework empowered by transfer learning. Digit. Discov. 2 (3), 591–601.

Chen, S., Xue, D., Chuai, G., Yang, Q., Liu, Q., 2021. FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery. Bioinform. (Oxf. Engl.) 36 (22–23), 5492–5498. http://dx.doi.org/10.1093/bioinformatics/btaa1006.

Cheng, X., Li, C., Liu, X., 2022. A review of federated learning in energy systems. In: 2022 IEEE/IAS Industrial and Commercial Power System Asia. I&CPS Asia, IEEE, http://dx.doi.org/10.1109/icpsasia55496.2022.9949863.

Cheng, A., Skreta, M., Ser, C.T., Guzman-Cordero, A., Thiede, L., Burger, A., Pablo-García, S., Aldossary, A., Leong, S.X., Strieth-Kalthoff, F., Aspuru-Guzik, A., 2024. How to do impactful research in artificial intelligence for chemistry and materials science. Faraday Discuss. http://dx.doi.org/10.1039/D4FD00153B, URL https://pubs.rsc.org/en/content/articlehtml/2024/fd/d4fd00153b.

Chiang, L.H., Russell, E.L., Braatz, R.D., 2012. Fault Detection and Diagnosis in Industrial Systems. Springer Science & Business Media.

Chik, W., Gamper, F., 2024. Chapter 21 - Ethical considerations and legal issues relating to federated learning. In: Nguyen, L.M. (Ed.), Federated Learning. Elsevier Science & Technology, San Diego, pp. 369–391. http://dx.doi.org/10.1016/B978-0-44-319037-7.00032-6, URL https://www.sciencedirect.com/science/article/pii/B9780443190377000326.

Chithrananda, S., Grand, G., Ramsundar, B., 2020. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

Chiuso, A., Pillonetto, G., 2019. System identification: A machine learning perspective. Annu. Rev. Control Robot. Auton. Syst. 2 (1), 281–304.

Choi, J., Nam, G., Choi, J., Jung, Y., 2025. A perspective on foundation models in chemistry. JACS Au 5 (4), 1499–1518.

Choi, K., Yi, J., Park, C., Yoon, S., 2021. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. IEEE Access 9, 120043–120065.

Chung, Y., Green, W.H., 2024. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. Chem. Sci. 15 (7), 2410–2424.

Coley, C.W., Barzilay, R., Green, W.H., Jaakkola, T.S., Jensen, K.F., 2017. Convolutional embedding of attributed molecular graphs for physical property prediction. J. Chem. Inf. Model. 57 (8), 1757–1772. http://dx.doi.org/10.1021/acs.jcim.6b00601.

Cramer, E., Gao, J., 2024. A black-box adversarial attack on demand side management. Comput. Chem. Eng. 186, 108681. http://dx.doi.org/10.1016/j.compchemeng.2024.108681.

Daly, K., Eichner, H., Kairouz, P., McMahan, H.B., Ramage, D., Xu, Z., 2024. Federated learning in practice: Reflections and projections. arXiv Preprint arXiv:2410.08892v1.

Daoutidis, P., Lee, J.H., Rangarajan, S., Chiang, L., Gopaluni, B., Schweidtmann, A.M., Harjunkoski, I., Mercangöz, M., Mesbah, A., Boukouvala, F., Lima, F.V., Del Rio Chanona, A., Georgakis, C., 2024. Machine learning in process systems engineering: Challenges and opportunities. Comput. Chem. Eng. 181, 108523. http://dx.doi.org/10.1016/j.compchemeng.2023.108523, URL https://www.sciencedirect.com/science/article/pii/S0098135423003939.

Das, A., Kong, W., Sen, R., Zhou, Y., 2024. A Decoder-Only Foundation Model for Time-Series Forecasting.

Decardi-Nelson, B., Alshehri, A.S., Ajagekar, A., You, F., 2024a. Generative AI and process systems engineering: The next frontier. Comput. Chem. Eng. 108723.

Decardi-Nelson, B., Alshehri, A.S., Ajagekar, A., You, F., 2024b. Generative AI and process systems engineering: The next frontier. Comput. Chem. Eng. 187, 108723. http://dx.doi.org/10.1016/j.compchemeng.2024.108723, URL https://www.sciencedirect.com/science/article/pii/S0098135424001418.

Devarakonda, V.S., Sun, W., Tang, X., Tian, Y., 2025. Recent advances in reinforcement learning for chemical process control. Processes 13 (6), http://dx.doi.org/10.3390/pr13061791, URL https://www.mdpi.com/2227-9717/13/6/1791.

Dinh, C.T., Vu, T.T., Tran, N.H., 2024. Chapter 7 - Personalized federated learning: theory and open problems. In: Nguyen, L.M. (Ed.), Federated Learning. Elsevier Science & Technology, San Diego, pp. 125–141. http://dx.doi.org/10.1016/B978-0-44-319037-7.00010-7, URL https://www.sciencedirect.com/science/article/pii/B9780443190377000107.

Dobbe, R., Sondermeijer, O., Fridovich-Keil, D., Arnold, D., Callaway, D., Tomlin, C., 2019. Toward distributed energy services: Decentralizing optimal power flow with machine learning. IEEE Trans. Smart Grid 11 (2), 1296–1306.

dos Passos Gomes, G., Pollice, R., Aspuru-Guzik, A., 2021. Navigating through the maze of homogeneous catalyst design with machine learning. Trends Chem. 3 (2), 96–110.

Downs, J.J., Vogel, E.F., 1993. A plant-wide industrial process control problem. Comput. Chem. Eng. 17 (3), 245–255. http://dx.doi.org/10.1016/0098-1354(93)80018-i.

Dutta, S., Leal de Freitas, I., Maciel Xavier, P., Miceli de Farias, C., Bernal Neira, D.E., 2024. Federated learning in chemical engineering: A tutorial on a framework for privacy-preserving collaboration across distributed data sources. Ind. Eng. Chem. Res.

Ekramipooya, A., Boroushaki, M., Rashtchian, D., 2024. Predicting possible recommendations related to causes and consequences in the HAZOP study worksheet using natural language processing and machine learning: BERT, clustering, and classification. J. Loss Prev. Process. Ind. 89, 105310. http://dx.doi.org/10.1016/j.jlp.2024.105310.

El Ouadrhiri, A., Abdelhadi, A., 2022. Differential privacy for deep and federated learning: A survey. IEEE Access 10, 22359–22380.

ElRobrini, F., Bukhari, S.M.S., Zafar, M.H., Al-Tawalbeh, N., Akhtar, N., Sanfilippo, F., 2024. Federated learning and non-federated learning based power forecasting of photovoltaic/wind power energy systems: A systematic review. Energy AI 18, 100438. http://dx.doi.org/10.1016/j.egyai.2024.100438.

Esche, E., Weigert, J., Rihm, G.B., Göbel, J., Repke, J.-U., 2022. Architectures for neural networks as surrogates for dynamic systems in chemical engineering. Chem. Eng. Res. Des. 177, 184–199.

Fang, X., Ye, M., 2022. Robust federated learning with noisy and heterogeneous clients. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10072–10081.

Felton, K.C., Ben-Safar, H., Alexei, A.A., 2022. DeepGamma: A deep learning model for activity coefficient prediction. In: 1st Annual AAAI Workshop on AI To Accelerate Science and Engineering. AI2ASE.

Felton, K.C., Raßpe-Lange, L., Rittig, J.G., Leonhard, K., Mitsos, A., Meyer-Kirschner, J., Knösche, C., Lapkin, A.A., 2024. ML-SAFT: a machine learning framework for PCP-SAFT parameter prediction. Chem. Eng. J. 151999. http://dx.doi.org/10.1016/j.cej.2024.151999.

Felton, K.C., Rittig, J.G., Lapkin, A.A., 2021. Summit: benchmarking machine learning methods for reaction optimisation. Chemistry-Methods 1 (2), 116–122.

Feng, X., Dai, Y., Ji, X., Zhou, L., Dang, Y., 2021. Application of natural language processing in HAZOP reports. Process. Saf. Environ. Prot. 155, 41–48.

Feng, J., Wang, Z., Wei, Z., Li, Y., Ding, B., Xu, H., Federated heterogeneous contrastive distillation for molecular representation learning. In: Serra, Spezzano (Ed.) 2024 – Proceedings of the 33rd ACM, pp. 1038–1048. http://dx.doi.org/10.1145/3627673.3679725.

Fey, M., Lenssen, J.E., 2019. Fast graph representation learning with PyTorch geometric. arXiv Preprint arXiv:1903.02428v3.

Fitzpatrick, D.E., Maujean, T., Evans, A.C., Ley, S.V., 2018. Across-the-world automated optimization and continuous-flow synthesis of pharmaceutical agents operating through a cloud-based server. Angew. Chem. (Int. Ed. Engl.) 57 (46), 15128–15132. http://dx.doi.org/10.1002/anie.201809080.

Gao, Q., Schweidtmann, A.M., 2024. Deep reinforcement learning for process design: Review and perspective. Curr. Opin. Chem. Eng. 44, 101012.

Geballe, M.T., Skillman, A.G., Nicholls, A., Guthrie, J.P., Taylor, P.J., 2010. The SAMPL2 blind prediction challenge: introduction and overview. J. Comput. Aided Mol. Des. 24, 259–279.

Geiping, J., Bauermeister, H., Dröge, H., Moeller, M., 2020. Inverting gradients-how easy is it to break privacy in federated learning? Adv. Neural Inf. Process. Syst. 33, 16937–16947.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. In: International Conference on Machine Learning. pp. 1263–1272, URL https://proceedings.mlr.press/v70/gilmer17a.

Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A., 2018. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. 4 (2), 268–276.

Göttl, Q., Grimm, D.G., Burger, J., 2022. Automated synthesis of steady-state continuous processes using reinforcement learning. Front. Chem. Sci. Eng. 1–15.

Göttl, Q., Pirnay, J., Burger, J., Grimm, D.G., 2025. Deep reinforcement learning enables conceptual design of processes for separating azeotropic mixtures without prior knowledge. Comput. Chem. Eng. 194, 108975.

Grataloup, A., Jonas, S., Meyer, A., 2024. A review of federated learning in renewable energy applications: Potential, challenges, and future directions. Energy AI 100375.

Guan, H., Yap, P.-T., Bozoki, A., Liu, M., 2024. Federated learning for medical image analysis: A survey. Pattern Recognit. 110424.

Guo, Y., Gao, Y., Song, J., 2024. MolCFL: A personalized and privacy-preserving drug discovery framework based on generative clustered federated learning. J. Biomed. Inform. 157, 104712. http://dx.doi.org/10.1016/j.jbi.2024.104712.

Hanser, T., Ahlberg, E., Amberg, A., Anger, L.T., Barber, C., Brennan, R.J., Brigo, A., Delaunois, A., Glowienke, S., Greene, N., Johnston, L., Kuhn, D., Kuhnke, L., Marchaland, J.-F., Muster, W., Plante, J., Rippmann, F., Sabnis, Y., Schmidt, F., van Deursen, R., Werner, S., White, A., Wichard, J., Yukawa, T., 2025. Data-driven federated learning in drug discovery with knowledge distillation. Nat. Mach. Intell. 7 (3), 423–436. http://dx.doi.org/10.1038/s42256-025-00991-2.

Hart, W.E., Watson, J.-P., Woodruff, D.L., 2011. Pyomo: modeling and solving mathematical programs in Python. Math. Program. Comput. 3 (3), 219–260. http://dx.doi.org/10.1007/s12532-011-0026-8, URL https://link.springer.com/article/10.1007/s12532-011-0026-8.

Hartung, F., Franks, B.J., Michels, T., Wagner, D., Liznerski, P., Reithermann, S., Fellenz, S., Jirasek, F., Rudolph, M., Neider, D., Leitte, H., Song, C., Kloepper, B., Mandt, S., Bortz, M., Burger, J., Hasse, H., Kloft, M., 2023. Deep anomaly detection on Tennessee Eastman process data. Chem. Ing. Tech. 95 (7), 1077–1082. http://dx.doi.org/10.1002/cite.202200238.

Hayer, M., 2021. Decentralized and Incentivized Federated Learning for the Chemical Engineering Domain (Master's thesis). Tsinghua University, This thesis is not available online.

Heid, E., Greenman, K.P., Chung, Y., Li, S.-C., Graff, D.E., Vermeire, F.H., Wu, H., Green, W.H., McGill, C.J., 2024. Chemprop: A machine learning package for chemical property prediction. J. Chem. Inf. Model. 64 (1), 9–17. http://dx.doi.org/10.1021/acs.jcim.3c01250.

Heid, E., McGill, C.J., Vermeire, F.H., Green, W.H., 2023. Characterizing uncertainty in machine learning for chemistry. J. Chem. Inf. Model. 63 (13), 4012–4029. http://dx.doi.org/10.1021/acs.jcim.3c00373.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y., 2017. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Heyndrickx, W., Arany, A., Simm, J., Pentina, A., Sturm, N., Humbeck, L., Mervin, L., Zalewski, A., Oldenhof, M., Schmidtke, P., Friedrich, L., Loeb, R., Afanasyeva, A., Schuffenhauer, A., Moreau, Y., Ceulemans, H., 2023. Conformal efficiency as a metric for comparative model assessment befitting federated learning. Artif. Intell. Life Sci. 3, 100070. http://dx.doi.org/10.1016/j.ailsci.2023.100070, URL https://www.sciencedirect.com/science/article/pii/S2667318523000144.

Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., Schmidtke, P., Fechner, N., Simm, J., Arany, A., Drizard, N., Jabal, R., Afanasyeva, A., Loeb, R., Verma, S., Harnqvist, S., Holmes, M., Pejo, B., Telenczuk, M., Holway, N., Dieckmann, A., Rieke, N., Zumsande, F., Clevert, D.-A., Krug, M., Luscombe, C., Green, D., Ertl, P., Antal, P., Marcus, D., Do Huu, N., Fuji, H., Pickett, S., Acs, G., Boniface, E., Beck, B., Sun, Y., Gohier, A., Rippmann, F., Engkvist, O., Göller, A.H., Moreau, Y., Galtier, M.N., Schuffenhauer, A., Ceulemans, H., 2024. MELLODDY: Cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J. Chem. Inf. Model. 64 (7), 2331–2344. http://dx.doi.org/10.1021/acs.jcim.3c00799.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Hoskins, J.C., Himmelblau, D.M., 1992. Process control via artificial neural networks and reinforcement learning. Comput. Chem. Eng. 16 (4), 241–251.

Huang, C., Dachille, J., Liu, X., 2024a. Technical report: Coopetition in heterogeneous cross-silo federated learning. URL http://arxiv.org/pdf/2408.11355.

Huang, C., Dachille, J., Liu, X., 2024b. When federated learning meets oligopoly competition: Stability and model differentiation. IEEE Internet Things J. 11 (16), 27409–27420. http://dx.doi.org/10.1109/jiot.2024.3400216.

Huang, C., Huang, J., Liu, X., 2022a. Cross-silo federated learning: Challenges and opportunities. arXiv preprint arXiv:2206.12949.

Huang, C., Huang, J., Liu, X., 2022b. Cross-silo federated learning: Challenges and opportunities. arXiv Preprint arXiv:2206.12949v1. URL http://arxiv.org/pdf/2206.12949.

Huang, C., Ke, S., Kamhoua, C., Mohapatra, P., Liu, X., 2022c. Incentivizing data contribution in cross-silo federated learning. arXiv Preprint arXiv:2203.03885v2. URL http://arxiv.org/pdf/2203.03885v2.

Huang, W., Ye, M., Shi, Z., Wan, G., Li, H., Du, B., Yang, Q., 2024. Federated learning for generalization, robustness, fairness: A survey and benchmark. IEEE Trans. Pattern Anal. Mach. Intell. PP, 1–20. http://dx.doi.org/10.1109/tpami.2024.3418862.

Huang, D., Ye, X., Zhang, Y., Sakurai, T., 2023. Collaborative analysis for drug discovery by federated learning on non-IID data. Methods (San Diego Calif.) 219, 1–7. http://dx.doi.org/10.1016/j.ymeth.2023.09.001.

Ishikawa, A., 2022. Heterogeneous catalyst design by generative adversarial network and first-principles based microkinetics. Sci. Rep. 12 (1), 11657.

Jacobsen, E.W., Skogestad, S., 1991. Multiple steady states in ideal two–product distillation. AIChE J. 37 (4), 499–511. http://dx.doi.org/10.1002/aic.690370404.

Jia, M., Hu, J., Liu, Y., Gao, Z., Yao, Y., 2023. Topology-guided graph learning for process fault diagnosis. Ind. Eng. Chem. Res. 62 (7), 3238–3248.

Jirasek, F., Alves, R.A.S., Damay, J., Vandermeulen, R.A., Bamler, R., Bortz, M., Mandt, S., Kloft, M., Hasse, H., 2020. Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion. J. Phys. Chem. Lett. 11 (3), 981–985. http://dx.doi.org/10.1021/acs.jpclett.9b03657.

Jorner, K., Brinck, T., Norrby, P.-O., Buttar, D., 2021. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. Chem. Sci. 12 (3), 1163–1175.

Joshi, A., Capezza, S., Alhaji, A., Chow, M.-Y., 2023. Survey on AI and machine learning techniques for microgrid energy management systems. IEEE/CAA J. Autom. Sin. 10 (7), 1513–1529.

Jourdan, T., Boutet, A., Frindel, C., 2021. Privacy assessment of federated learning using private personalized layers. arXiv preprint arXiv:2106.08060.

Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Zhao, S., 2021. Advances and open problems in federated learning. Found. Trends® Mach. Learn. 14 (1–2), 1–210, URL https://www.researchgate.net/profile/mehdi-bennis-2/publication/337905781_advances_and_open_problems_in_federated_learning.

Kang, S.-O., Lee, E.-B., Baek, H.-K., 2019. A digitization and conversion tool for imaged drawings to intelligent piping and instrumentation diagrams (P&ID). Energies 12 (13), 2593.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nat. Rev. Phys. 3 (6), 422–440.

Koay, A.M., Ko, R.K.L., Hettema, H., Radke, K., 2023. Machine learning in industrial control system (ICS) security: current landscape, opportunities and challenges. J. Intell. Inf. Syst. 60 (2), 377–405.

Krüger, F.P., Östman, J., Mervin, L., Tetko, I.V., Engkvist, O., 2024. Publishing neural networks in drug discovery might compromise training data privacy. URL http://arxiv.org/pdf/2410.16975.

Krüger, F.P., Östman, J., Mervin, L., Tetko, I.V., Engkvist, O., 2025. Publishing neural networks in drug discovery might compromise training data privacy. J. Cheminformatics 17 (1), 38.

Landrum, G., RDKit: Open-Source Cheminformatics. URL http://www.rdkit.org.

Lee, S., Choi, D.-H., 2020. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. IEEE Trans. Ind. Inform. 18 (1), 488–497.

Leenhouts, R.J., Morgan, N., Al Ibrahim, E., Green, W.H., Vermeire, F.H., 2025. Pooling solvent mixtures for solvation free energy predictions. Chem. Eng. J. 513, 162232.

Li, C., 2025. Breaking data silos in drug discovery with federated learning. Nat. Chem. Eng. 2 (5), 288–289. http://dx.doi.org/10.1038/s44286-025-00213-x.

Li, Q., Diao, Y., Chen, Q., He, B., 2022. Federated learning on non-iid data silos: An experimental study. In: 2022 IEEE 38th International Conference on Data Engineering. ICDE, IEEE, pp. 965–978.

Li, W., Gu, S., Zhang, X., Chen, T., 2020. Transfer learning for process fault diagnosis: Knowledge transfer from simulation to physical processes. Comput. Chem. Eng. 139, 106904.

Li, Y., He, S., Li, Y., Shi, Y., Zeng, Z., 2023. Federated multiagent deep reinforcement learning approach via physics-informed reward for multimicrogrid energy management. IEEE Trans. Neural Netw. Learn. Syst.

Li, S., Ye, F., Fang, M., Zhao, J., Chan, Y.-H., Ngai, E.C.H., Voigt, T., 2024. Synergizing foundation models and federated learning: A survey. arXiv Preprint arXiv:2406.12844v1.

Liang, P.P., Zadeh, A., Morency, L.-P., 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Comput. Surv. 56 (10), 1–42.

Linsner, F., Adilova, L., Däubener, S., Kamp, M., Fischer, A., 2021. Approaches to uncertainty quantification in federated deep learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 128–145.

Liu, B., Lv, N., Guo, Y., Li, Y., 2024. Recent advances on federated learning: A systematic survey. Neurocomputing 597, 128019. http://dx.doi.org/10.1016/j.neucom.2024.128019.

Liu, R., Xing, P., Deng, Z., Li, A., Guan, C., Yu, H., 2024. Federated graph neural networks: Overview, techniques, and challenges. IEEE Trans. Neural Netw. Learn. Syst. PP, http://dx.doi.org/10.1109/TNNLS.2024.3360429.

Lomov, I., Lyubimov, M., Makarov, I., Zhukov, L.E., 2021. Fault detection in Tennessee Eastman process with temporal deep learning models. J. Ind. Inf. Integr. 23, 100216.

Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. Med. Image Anal. 76, 102298.

Lu, Z., Pan, H., Dai, Y., Si, X., Zhang, Y., 2024. Federated learning with non-iid data: A survey. IEEE Internet Things J. 11 (11), 19188–19209.

Mann, V., Sales-Cruz, M., Gani, R., Venkatasubramanian, V., 2024. eSFILES: Intelligent process flowsheet synthesis using process knowledge, symbolic AI, and machine learning. Comput. Chem. Eng. 181, 108505.

Manu, D., Yao, J., Liu, W., Sun, X., 2024. GraphGANFed: A federated generative framework for graph-structured molecules towards efficient drug discovery. IEEE/ACM Trans. Comput. Biol. Bioinform. 21 (2), 240–253. http://dx.doi.org/10.1109/TCBB.2024.3349990.

Margraf, J.T., Jung, H., Scheurer, C., Reuter, K., 2023. Exploring catalytic reaction networks with machine learning. Nat. Catal. 6 (2), 112–121.

McBride, K., Sundmacher, K., 2019. Overview of surrogate modeling in chemical process engineering. Chem. Ing. Tech. 91 (3), 228–239.

McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B., 2017. Communication-efficient learning of deep networks from decentralized data. Artif. Intell. Stat. 1273–1282, URL https://proceedings.mlr.press/v54/mcmahan17a?ref=https://githubhelp.com.

Meuwly, M., 2021. Machine learning for chemical reactions. Chem. Rev. 121 (16), 10218–10239.

Misener, R., Biegler, L., 2023. Formulating data-driven surrogate models for process optimization. Comput. Chem. Eng. 179, 108411.

Mobley, D.L., Guthrie, J.P., 2014. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J. Comput. Aided Mol. Des. 28, 711–720.

Morstyn, T., Hredzak, B., Agelidis, V.G., 2016. Control strategies for microgrids with distributed energy storage systems: An overview. IEEE Trans. Smart Grid 9 (4), 3652–3666.

Nevolianis, T., Rittig, J.G., Mitsos, A., Leonhard, K., 2024. Multi-fidelity graph neural networks for predicting toluene/water partition coefficients. http://dx.doi.org/10.26434/Chemrxiv-2024-3t818, ChemRxiv preprint.

Nguyen, D.C., Pham, Q.-V., Pathirana, P.N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., Hwang, W.-J., 2023. Federated learning for smart healthcare: A survey. ACM Comput. Surv. 55 (3), 1–37. http://dx.doi.org/10.1145/3501296.

Nian, R., Liu, J., Huang, B., 2020. A review on reinforcement learning: Introduction and applications in industrial process control. Comput. Chem. Eng. 139, 106886.

Niu, Y., Fan, Y., Ju, X., 2024. Critical review on data-driven approaches for learning from accidents: Comparative analysis and future research. Saf. Sci. 171, 106381. http://dx.doi.org/10.1016/j.ssci.2023.106381, URL https://www.sciencedirect.com/science/article/pii/S0925753523003235.

Niu, K., Zhou, M., Abdallah, C.T., Hayajneh, M., 2022. Deep transfer learning for system identification using long short-term memory neural networks. arXiv preprint arXiv:2204.03125.

Oldenhof, M., Ács, G., Pejó, B., Schuffenhauer, A., Holway, N., Sturm, N., Dieckmann, A., Fortmeier, O., Boniface, E., Mayer, C., Gohier, A., Schmidtke, P., Niwayama, R., Kopecky, D., Mervin, L., Rathi, P.C., Friedrich, L., Formanek, A., Antal, P., Rahaman, J., Zalewski, A., Heyndrickx, W., Oluoch, E., Stößel, M., Vančo, M., Endico, D., Gelus, F., de Boisfossé, T., Darbier, A., Nicollet, A., Blottière, M., Telenczuk, M., van Nguyen, T., Martinez, T., Boillet, C., Moutet, K., Picosson, A., Gasser, A., Djafar, I., Simon, A., Arany, Á., Simm, J., Moreau, Y., Engkvist, O., Ceulemans, H., Marini, C., Galtier, M., 2023. Industry-scale orchestrated federated learning for drug discovery. Proc. AAAI Conf. Artif. Intell. 37 (13), 15576–15584. http://dx.doi.org/10.1609/aaai.v37i13.26847.

Otashu, J.I., Seo, K., Baldea, M., 2021. Cooperative optimal power flow with flexible chemical process loads. AIChE J. 67 (4), e17159.

Ottaiano, G.Y., Martins, T.D., 2024. Machine learning models for vapor-liquid equilibrium of binary mixtures: State of the art and future opportunities. Chem. Eng. Res. Des.

Pearson, R.K., Pottmann, M., 2000. Gray-box identification of block-oriented nonlinear models. J. Process Control 10 (4), 301–315. http://dx.doi.org/10.1016/s0959-1524(99)00055-4.

Pillutla, K., Malik, K., Mohamed, A.-R., Rabbat, M., Sanjabi, M., Xiao, L., 2022. Federated learning with partial model personalization. In: International Conference on Machine Learning. pp. 17716–17758, URL https://proceedings.mlr.press/v162/pillutla22a.html.

Pirnay, J., Rittig, J.G., Wolf, A.B., Grohe, M., Burger, J., Mitsos, A., Grimm, D.G., 2025. GraphXForm: graph transformer for computer-aided molecular design. Digit. Discov. 4 (4), 1052–1065.

Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., Piccialli, F., 2024. Model aggregation techniques in federated learning: A comprehensive survey. Future Gener. Comput. Syst. 150, 272–293. http://dx.doi.org/10.1016/j.future.2023.09.008, URL https://www.sciencedirect.com/science/article/pii/S0167739X23003333.

Qin, S.J., Chiang, L.H., 2019. Advances and opportunities in machine learning for process data analytics. Comput. Chem. Eng. 126, 465–473.

Qin, S., Jiang, S., Li, J., Balaprakash, P., van Lehn, R.C., Zavala, V.M., 2023. Capturing molecular interactions in graph neural networks: a case study in multicomponent phase equilibrium. Digit. Discov. 2 (1), 138–151. http://dx.doi.org/10.1039/D2DD00045H.

Quaglio, M., Roberts, L., Jaapar, M.S.B., Fraga, E.S., Dua, V., Galvanin, F., 2020. An artificial neural network approach to recognise kinetic models from experimental data. Comput. Chem. Eng. 135, 106759.

Ramakrishnan, R., Dral, P.O., Rupp, M., von Lilienfeld, O.A., 2014. Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data 1 (1), 140022.

Rauniyar, A., Hagos, D.H., Jha, D., Håkegård, J.E., Bagci, U., Rawat, D.B., Vlassov, V., 2024. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J. 11 (5), 7374–7398. http://dx.doi.org/10.1109/JIOT.2023.3329061.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., Friederich, P., 2022. Graph neural networks for materials science and chemistry. Commun. Mater. 3 (1), 93. http://dx.doi.org/10.1038/s43246-022-00315-6.

Ren, C., Yu, H., Peng, H., Tang, X., Li, A., Gao, Y., Tan, A.Z., Zhao, B., Li, X., Li, Z., Yang, Q., 2024. Advances and Open Challenges in Federated Learning with Foundation Models.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al., 2020. The future of digital health with federated learning. NPJ Digit. Med. 3 (1), 119.

Rittig, J.G., Ben Hicham, K., Schweidtmann, A.M., Dahmen, M., Mitsos, A., 2023a. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. Comput. Chem. Eng. 171, 108153. http://dx.doi.org/10.1016/j.compchemeng.2023.108153.

Rittig, J.G., Ben Hicham, K., Schweidtmann, A.M., Dahmen, M., Mitsos, A., 2023b. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. Comput. Chem. Eng. 171, 108153. http://dx.doi.org/10.1016/j.compchemeng.2023.108153, URL https://www.sciencedirect.com/science/article/pii/S0098135423000224.

Rittig, J.G., Felton, K.C., Lapkin, A.A., Mitsos, A., 2023c. Gibbs–Duhem-informed neural networks for binary activity coefficient prediction. Digit. Discov. 2 (6), 1752–1767. http://dx.doi.org/10.1039/D3DD00103B.

Rittig, J.G., Gao, Q., Dahmen, M., Mitsos, A., Schweidtmann, A.M., 2022. Graph neural networks for the prediction of molecular structure–property relationships. Mach. Learn. Hybrid Model. React. Eng. R. Soc. Chem. 159–181. http://dx.doi.org/10.1039/BK9781837670178-00159, URL http://arxiv.org/pdf/2208.04852.

Rittig, J.G., Mitsos, A., 2024. Thermodynamics-consistent graph neural networks. arXiv preprint arXiv:2407.18372. URL http://arxiv.org/pdf/2407.18372.

Rittig, J.G., Ritzert, M., Schweidtmann, A.M., Winkler, S., Weber, J.M., Morsch, P., Heufer, K.A., Grohe, M., Mitsos, A., Dahmen, M., 2023d. Graph machine learning for design of high–octane fuels. AIChE J. 69 (4), e17971. http://dx.doi.org/10.1002/aic.17971.

Sanchez Medina, E.I., Linke, S., Stoll, M., Sundmacher, K., 2022. Graph Neural Networks for the prediction of infinite dilution activity coefficients. Digit. Discov. 1 (3), 216–225. http://dx.doi.org/10.1039/D1DD00037C.

Sanchez Medina, E.I., Linke, S., Stoll, M., Sundmacher, K., 2023. Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. Digit. Discov. 2, 781–798. http://dx.doi.org/10.1039/D2DD00142J.

Sani, L., Iacob, A., Cao, Z., Marino, B., Gao, Y., Paulik, T., Zhao, W., Shen, W.F., Aleksandrov, P., Qiu, X., Lane, N.D., 2024. The future of large language model pre-training is federated. arXiv Preprint arXiv:2405.10853v3.

Sansana, J., Joswiak, M.N., Castillo, I., Wang, Z., Rendall, R., Chiang, L.H., Reis, M.S., 2021. Recent trends on hybrid modeling for Industry 4.0. Comput. Chem. Eng. 151, 107365.

Savage, T., Basha, N., McDonough, J., Matar, O.K., del Rio Chanona, E.A., 2023. Multi-fidelity data-driven design and analysis of reactor and tube simulations. Comput. Chem. Eng. 179, 108410.

Schilter, O., Vaucher, A., Schwaller, P., Laino, T., 2023. Designing catalysts with deep generative models and computational data. A case study for Suzuki cross coupling reactions. Digit. Discov. 2 (3), 728–735.

Schoch, F., Graf, P., Schmieg, T., Wittenberg, C., Lanquillon, C., Stache, N.C., 2024. Deep anomaly detection with extended transformer-based model on Tennessee Eastman process dataset. In: 2024 IEEE 19th International Conference on Computer Science and Information Technologies. CSIT, IEEE, pp. 1–4.

Scholtz, E., Oudalov, A., Harjunkoski, I., 2024. Power systems of the future. Comput. Chem. Eng. 180, 108460. http://dx.doi.org/10.1016/j.compchemeng.2023.108460.

Schulze, J.C., Doncevic, D.T., Mitsos, A., 2022. Identification of MIMO Wiener-type Koopman models for data-driven model reduction using deep learning. Comput. Chem. Eng. 161, 107781. http://dx.doi.org/10.1016/j.compchemeng.2022.107781.

Schulze Balhorn, L., Caballero, M., Schweidtmann, A.M., 2024. Toward autocorrection of chemical process flowsheets using large language models. In: Manenti, F., Reklaitis, G.V. (Eds.), Computer Aided Chemical Engineering : 34 European Symposium on Computer Aided Process Engineering / 15 International Symposium on Process Systems Engineering. Vol. 53, Elsevier, pp. 3109–3114. http://dx.doi.org/10.1016/B978-0-443-28824-1.50519-6, URL https://www.sciencedirect.com/science/article/pii/B9780443288241505196.

Schweidtmann, A.M., 2024. Generative artificial intelligence in chemical engineering. Nat. Chem. Eng. 1 (3), 193.

Schweidtmann, A.M., Esche, E., Fischer, A., Kloft, M., Repke, J.-U., Sager, S., Mitsos, A., 2021. Machine learning in chemical engineering: A perspective. Chem. Ing. Tech. 93 (12), 2029–2039. http://dx.doi.org/10.1002/cite.202100083.

Schweidtmann, A.M., Mitsos, A., 2019. Deterministic global optimization with artificial neural networks embedded. J. Optim. Theory Appl. 180 (3), 925–948.

Schweidtmann, A.M., Zhang, D., von Stosch, M., 2024. A review and perspective on hybrid modeling methodologies. Digit. Chem. Eng. 10, 100136.

Shao, J., Li, Z., Sun, W., Zhou, T., Sun, Y., Liu, L., Lin, Z., Mao, Y., Zhang, J., 2023. A survey of what to share in federated learning: Perspectives on model utility, privacy leakage, and communication efficiency. arXiv preprint arXiv:2307.10655.

Sim, R.H.L., Tay, S.S., Xu, X., Zhang, Y., Wu, Z., Lin, X., Ng, S.-K., Foo, C.-S., Jaillet, P., Hoang, T.N., Low, B.K.H., 2024. Chapter 16 - Incentives in federated learning. In: Nguyen, L.M. (Ed.), Federated Learning. Elsevier Science & Technology, San Diego, pp. 299–309. http://dx.doi.org/10.1016/B978-0-44-319037-7.00026-0, URL https://www.sciencedirect.com/science/article/pii/B9780443190377000260.

Single, J.I., Schmidt, J., Denecke, J., 2020. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. Saf. Sci. 129, 104747.

Skilton, R.A., Bourne, R.A., Amara, Z., Horvath, R., Jin, J., Scully, M.J., Streng, E., Tang, S.L.Y., Summers, P.A., Wang, J., Pérez, E., Asfaw, N., Aydos, G.L.P., Dupont, J., Comak, G., George, M.W., Poliakoff, M., 2015. Remote-controlled experiments with cloud chemistry. Nat. Chem. 7 (1), 1–5. http://dx.doi.org/10.1038/nchem.2143.

Smith, V., Chiang, C.-K., Sanjabi, M., Talwalkar, A.S., 2017. Federated multi-task learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc, URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6211080fa89981f66b1a0c9d55c61d0f-Paper.pdf.

Specht, T., Nagda, M., Fellenz, S., Mandt, S., Hasse, H., Jirasek, F., 2024. HANNA: hard-constraint neural network for consistent activity coefficient prediction. Chem. Sci. 15 (47), 19777–19786.

Stops, L., Leenhouts, R., Gao, Q., Schweidtmann, A.M., 2023. Flowsheet generation through hierarchical reinforcement learning and graph neural networks. AIChE J. 69 (1), e17938.

Strieth-Kalthoff, F., Sandfort, F., Segler, M.H.S., Glorius, F., 2020. Machine learning the ropes: principles, applications and directions in synthetic chemistry. Chem. Soc. Rev. 49 (17), 6154–6168. http://dx.doi.org/10.1039/c9cs00786e.

Stürmer, J.M., Graumann, M., Koch, T., 2024. Transforming engineering diagrams: A novel approach for P&ID digitization using transformers. arXiv Preprint arXiv:2411.13929v1.

Tang, W., Daoutidis, P., 2022. Data-driven control: Overview and perspectives. In: 2022 American Control Conference. ACC, IEEE, pp. 1048–1064.

TensorFlow Developers, 2024. TensorFlow. http://dx.doi.org/10.5281/ZENODO.4724125.

Thebelt, A., Wiebe, J., Kronqvist, J., Tsay, C., Misener, R., 2022. Maximizing information from chemical engineering data sets: Applications to machine learning. Chem. Eng. Sci. 252, 117469.

Theisen, M.F., Flores, K.N., Balhorn, L.S., Schweidtmann, A.M., 2023. Digitization of chemical process flow diagrams using deep convolutional neural networks. Digit. Chem. Eng. 6, 100072.

Tom, G., Schmid, S.P., Baird, S.G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E.M., Skreta, M., Yoshikawa, N., Corapi, S., Akkoc, G.D., Strieth-Kalthoff, F., Seifrid, M., Aspuru-Guzik, A., 2024. Self-driving laboratories for chemistry and materials science. Chem. Rev. 124 (16), 9633–9732. http://dx.doi.org/10.1021/acs.chemrev.4c00055.

Tsay, C., Baldea, M., 2019. 110Th anniversary: Using data to bridge the time and length scales of process systems. Ind. Eng. Chem. Res. 58 (36), 16696–16708. http://dx.doi.org/10.1021/acs.iecr.9b02282.

Tsouvalas, V., Saeed, A., Ozcelebi, T., Meratnia, N., 2024. Labeling chaos to learning harmony: Federated learning with noisy labels. ACM Trans. Intell. Syst. Technol. 15 (2), 1–26.

Tu, X., Zhu, K., Luong, N.C., Niyato, D., Zhang, Y., Li, J., 2022. Incentive mechanisms for federated learning: From economic and game theoretic perspective. IEEE Trans. Cogn. Commun. Netw. 8 (3), 1566–1593. http://dx.doi.org/10.1109/tccn.2022.3177522.

Tuor, T., Wang, S., Ko, B.J., Liu, C., Leung, K.K., 2021. Overcoming noisy and irrelevant data in federated learning. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 5020–5027.

Velioglu, M., Zhai, S., Rupprecht, S., Mitsos, A., Jupke, A., Dahmen, M., 2025. Physics-informed neural networks for dynamic process operations with limited physical knowledge and data. Comput. Chem. Eng. 192, 108899. http://dx.doi.org/10.1016/j.compchemeng.2024.108899.

Vermeire, F.H., Green, W.H., 2021a. Transfer learning for solvation free energies: From quantum chemistry to experiments. Chem. Eng. J. 418, 129307.

Vermeire, F.H., Green, W.H., 2021b. Transfer learning for solvation free energies: From quantum chemistry to experiments. Chem. Eng. J. 418, 129307. http://dx.doi.org/10.1016/j.cej.2021.129307.

Vogel, G., Schulze Balhorn, L., Schweidtmann, A.M., 2023. Learning from flowsheets: A generative transformer model for autocompletion of flowsheets. Comput. Chem. Eng. 171, 108162. http://dx.doi.org/10.1016/j.compchemeng.2023.108162, URL https://www.sciencedirect.com/science/article/pii/S0098135423000315.

Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., Zhang, T., 2022. On the unreasonable effectiveness of federated averaging with heterogeneous data. arXiv preprint arXiv:2206.04723.

Weber, J.M., Guo, Z., Zhang, C., Schweidtmann, A.M., Lapkin, A.A., 2021. Chemical data intelligence for sustainable chemistry. Chem. Soc. Rev. 50 (21), 12013–12036.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q., Poor, H.V., 2020. Federated learning with differential privacy: Algorithms and performance analysis. IEEE Trans. Inf. Forensics Secur. 15, 3454–3469.

Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., Zhang, W., 2023. A survey on federated learning: challenges and applications. Int. J. Mach. Learn. Cybern. 14 (2), 513–535. http://dx.doi.org/10.1007/s13042-022-01647-y.

Wigh, D.S., Arrowsmith, J., Pomberger, A., Felton, K.C., Lapkin, A.A., 2024. Orderly: data sets and benchmarks for chemical reaction data. J. Chem. Inf. Model. 64 (9), 3790–3798.

Winter, B., Rehner, P., Esper, T., Schilling, J., Bardow, A., 2025. Understanding the language of molecules: Predicting pure component parameters for the PC-SAFT equation of state from SMILES. Digit. Discov.

Winter, B., Winter, C., Esper, T., Schilling, J., Bardow, A., 2023. SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients. Fluid Phase Equilib. 568, 113731. http://dx.doi.org/10.1016/j.fluid.2023.113731.

Winter, B., Winter, C., Schilling, J., Bardow, A., 2022. A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing. Digit. Discov. 1 (6), 859–869. http://dx.doi.org/10.1039/d2dd00058j.

Woisetschläger, H., Erben, A., Marino, B., Wang, S., Lane, N.D., Mayer, R., Jacobsen, H.-A., 2024. Federated learning priorities under the European Union artificial intelligence act. arXiv Preprint arXiv:2402.05968v1. URL http://arxiv.org/pdf/2402.05968.

Wu, Z., Christofides, P.D., Wu, W., Wang, Y., Abdullah, F., Alnajdi, A., Kadakia, Y., 2024. A tutorial review of machine learning-based model predictive control methods. Rev. Chem. Eng. http://dx.doi.org/10.1515/revce-2024-0055.

Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.-S., 2024. Next-gpt: Any-to-any multimodal llm. In: Forty-First International Conference on Machine Learning.

Wu, C., Wu, F., Lyu, L., Huang, Y., Xie, X., 2022. Communication-efficient federated learning via knowledge distillation. Nat. Commun. 13 (1), 2032.

Wu, Z., Xu, X., Sim, R.H.L., Shu, Y., Lin, X., Agussurja, L., Dai, Z., Ng, S.-K., Foo, C.-S., Jaillet, P., Hoang, T.N., Low, B.K.H., 2024. Chapter 15 - Data valuation in federated learning. In: Nguyen, L.M. (Ed.), Federated Learning. Elsevier Science & Technology, San Diego, pp. 281–296. http://dx.doi.org/10.1016/B978-0-44-319037-7.00024-7, URL https://www.sciencedirect.com/science/article/pii/B9780443190377000247.

Wu, H., Zhao, J., 2020. Fault detection and diagnosis based on transfer learning for multimode chemical processes. Comput. Chem. Eng. 135, 106731.

Xiao, M., Hu, C., Wu, Z., 2023. Modeling and predictive control of nonlinear processes using transfer learning method. AIChE J. 69 (7), e18076.

Xiao, M., Vellayappan, K., P S, P., Gudena, K., Wu, Z., 2024. Optimization-based multi-source transfer learning for modeling of nonlinear processes. Chem. Eng. Sci. 295, 120117. http://dx.doi.org/10.1016/j.ces.2024.120117, URL https://www.sciencedirect.com/science/article/pii/S0009250924004172.

Xu, Z., Wu, Z., 2024. Privacy-preserving federated machine learning modeling and predictive control of heterogeneous nonlinear systems. Comput. Chem. Eng. 187, 108749. http://dx.doi.org/10.1016/j.compchemeng.2024.108749, URL https://www.sciencedirect.com/science/article/pii/S0098135424001674.

Yan, P., Abdulkadir, A., Luley, P.-P., Rosenthal, M., Schatte, G.A., Grewe, B.F., Stadelmann, T., 2024. A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. IEEE Access 12, 3768–3789. http://dx.doi.org/10.1109/ACCESS.2023.3349132.

Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., Beaufays, F., 2018. Applied federated learning: Improving google keyboard query suggestions. arXiv preprint arXiv:1812.02903.

Yang, X., Feng, D., 2019. Generative adversarial network based anomaly detection on the benchmark tennessee eastman process. In: 2019 5th International Conference on Control, Automation and Robotics. ICCAR, IEEE, pp. 644–648.

Yang, S., Park, H., Byun, J., Kim, C., 2022. Robust federated learning with noisy labels. IEEE Intell. Syst. 37 (2), 35–43.

Yang, S., Zhao, F., Zhou, Z., Shi, L., Ren, X., Xu, Z., 2024. Review of mathematical optimization in federated learning. arXiv Preprint arXiv:2412.01630v1.

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., 2022. Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12104–12113.

Zhang, X., Glaws, A., Cortiella, A., Emami, P., King, R.N., 2025. Deep generative models in energy system applications: Review, challenges, and future directions. Appl. Energy 380, 125059. http://dx.doi.org/10.1016/j.apenergy.2024.125059, URL https://www.sciencedirect.com/science/article/pii/S0306261924024437.

Zhang, J., Wang, Q., Eden, M., Shen, W., 2023. A deep learning-based framework towards inverse green solvent design for extractive distillation with multi-index constraints. Comput. Chem. Eng. 177, 108335.

Zhang, Y., Xia, T., Ghosh, A., Mascolo, C., 2023. Uncertainty quantification in federated learning for heterogeneous health data. In: International Workshop on Federated Learning for Distributed Data Mining.

Zhao, J.C., Bagchi, S., Avestimehr, S., Chan, K.S., Chaterji, S., Dimitriadis, D., Li, J., Li, N., Nourian, A., Roth, H.R., 2024. Federated learning privacy: Attacks, defenses, applications, and policy landscape - A survey. arXiv Preprint arXiv:2405.03636v1. URL http://arxiv.org/pdf/2405.03636.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V., 2018. Federated learning with non-iid data. arXiv preprint arXiv:1806.00582.

Zhou, T., Lin, Z., Zhang, J., Tsang, D.H., 2024. Understanding and improving model averaging in federated learning on heterogeneous data. IEEE Trans. Mob. Comput. 1–16. http://dx.doi.org/10.1109/TMC.2024.3406554.

Zhu, W., Luo, J., White, A.D., 2022. Federated learning of molecular properties with graph neural networks in a heterogeneous setting. Patterns 3 (6), 100521. http://dx.doi.org/10.1016/j.patter.2022.100521, URL https://www.cell.com/patterns/fulltext/s2666-3899(22)00118-0.