Computers and Chemical Engineering 194 (2025) 108982

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Knowledge-enhanced data-driven modeling of wastewater treatment processes for energy consumption prediction

Louis Allen, Joan Cordiner *

Department of Chemical and Biological Engineering, The University of Sheffield, S10 2TN, Sheffield, United Kingdom

ARTICLE INFO

Keywords:
Wastewater treatment
Knowledge graphs
Disentangled representation learning
Graph convolutional networks
Timeseries forecasting

ABSTRACT

Rising energy usage in wastewater treatment processes (WWTPs) poses pressing economic and environmental challenges. Machine learning approaches to model these complex systems have been limited by highly non-linear processes and high dataset noise. To address this, we introduce a novel Knowledge-Enhanced Graph Disentanglement framework for Energy Consumption Prediction (KEGD-EC) that leverages causal inference and graph neural networks. This work combines specific knowledge of causal relationships with a disentangled graph convolutional network architecture to facilitate accurate predictions. In a study on a WWTP in Melbourne, we demonstrate a 59.7% reduction in root mean squared error in energy consumption prediction using KEGD-EC compared to the next best model. We show that causal models built using domain knowledge outperform data-driven causal discovery models for complex systems. These results signify a step forward in applying machine learning to complex manufacturing processes, with the integration of causal knowledge into deep learning architectures posing a promising area of research for predictive analytics in manufacturing.

1. Introduction

Secure and plentiful access to clean water underpins human life, enabling essential activities like drinking, irrigation, hygiene, and industrial processes. Under this vital framework, wastewater treatment is crucial in safeguarding human health. By treating and purifying contaminated water sources, wastewater treatment processes remove harmful pollutants and bacteria, thereby mitigating significant health risks for populations (Alali et al., 2022). The rapid growth of urban populations requires that wastewater treatment processes be highly efficient and robust to handle the demands of the expanding urban masses (Haase et al., 2018). To add to the challenge, the performance of wastewater treatment plants (WWTPs) intricately depends on climatic factors like rainfall, ambient temperature, and seasonal fluctuations in demand (Ahmad and Chen, 2018). These factors are becoming far more important with the intensification of climate change causing conditions towards the extremes of what would traditionally be expected. The result is pressure on WWTPs to remain reliable and efficient. Addressing challenges posed by increased urbanization and climate change is imperative for the sustainability of wastewater treatment moving forward.

WWTPs are highly energy-intensive owing to the transport and processing of large volumes of wastewater. The treatment of wastewater is the largest consumer of electricity in urban dwellings, and contributes to 25%–40% of the total energy consumption (EC) (Maslon et al., 2018). With the rising demand for these processes, the associated rise in EC is a pressing concern. It is expected that consumption figures for WWTPs will increase between 60%–100% over the next 15 years as demand for the services rises with population (Hamawand, 2023). This raises concern in terms of both the environmental sustainability of the processes, as well as the economic viability of operating plants at such scale. Since electricity costs account for as much as 40% of the operating costs of water companies, a rise in consumption would likely necessitate a rise in the water bills of consumers (Agency, 2021). With many places in the world already struggling to cope with an escalating cost of living, this leaves water companies with tremendous pressure to alleviate costs. Furthermore, with 123 countries pledging to double their energy efficiency improvements between now and 2030, a global imperative emerges; the need to reduce the energy requirements of wastewater treatment processes on a worldwide scale (Bansard et al., 2023).

Abbreviations: WWTP, Wastewater Treatment Plant; EC, Energy Consumption; KEGD-EC, Knowledge-Enhanced Graph Disentanglement for Energy Consumption Prediction; FCI, Fast Causal Inference; GBM, Gradient Boosting Machine; RF, Random Forest; kNN, k-Nearest Neighbors; LSTM, Long Short-Term Memory; ANN, Artificial Neural Network; GRU, Gated Recurrent Unit; DisenConv, Disentangled Convolutional; DisenGRU, Disentangled Gated Recurrent Unit; BiLSTM, Bidirectional Long Short-Term Memory; ASM, Activated Sludge Model

* Corresponding author.
E-mail address: j.cordiner@sheffield.ac.uk (J. Cordiner).

https://doi.org/10.1016/j.compchemeng.2024.108982
Received 29 April 2024; Received in revised form 6 December 2024; Accepted 11 December 2024
Available online 17 December 2024

By predicting energy demands based on incoming weather patterns, expected influent characteristics, and seasonal information, operators can proactively adjust treatment processes, particularly aeration rates, which often account for the largest portion of a WWTP's energy consumption (Longo et al., 2016). This predictive capability enables plants to participate more effectively in demand response programs, shifting energy-intensive operations to off-peak hours and potentially supporting grid stability (Kirchem et al., 2020). Furthermore, energy consumption predictions can inform maintenance schedules, allowing for equipment servicing during periods of lower predicted demand and helping to identify unexpected increases in energy use that may indicate equipment issues (Torregrossa et al., 2017). Long-term predictions facilitate capacity planning and technology assessment, guiding investment decisions for plant upgrades. Integration with renewable energy sources, such as optimizing the use of on-site solar or wind generation and biogas utilization, can be enhanced through accurate energy demand forecasts (Gandiglio et al., 2017). Additionally, these predictions can improve hydraulic management, optimizing pump operations and storage utilization to minimize energy use during peak demand periods. By implementing these strategies based on accurate energy consumption predictions, WWTPs can significantly reduce their energy footprint while maintaining or even improving treatment efficiency, aligning with broader sustainability goals in water management and contributing to the reduction of greenhouse gas emissions from the wastewater treatment sector (Wang et al., 2016).

We look to address the need to reduce energy consumption by first understanding the causal factors driving energy consumption through a knowledge-enhanced modeling approach that combines data with vital cause-and-effect relationships. The addition of causal knowledge can better aid the prediction of energy consumption and therefore aid in understanding how to reduce the overall consumption of wastewater treatment processes. This is significant since giving operational teams insight into EC ahead of time allows for improvements in load balancing, process scheduling, and resource allocation. The overall impact of this is reduced costs, enhanced process sustainability, and an increase in process robustness (Ahmad and Chen, 2018).

2. Background and relevant literature

To effectively reduce the energy consumption of WWTPs, it is essential to first identify and understand the factors contributing to high consumption. Producing mathematical models to describe the influence of these factors is notoriously difficult for wastewater treatment for several reasons. Firstly, WWTPs are subject to large fluctuations in influent properties including the flow rates and compositions. This is a challenge not seen by many other processes since these variations are inherently uncontrollable owing to the waste nature of the influent. The variability in the feed composition means that the potential number of reactions and organism species involved in the processes is very high. The result of this is that any mathematical model that can accurately describe a wastewater treatment process must be extremely large and overly complex, which may hinder their use from an operational standpoint (Jeppsson, 1996). Further complications come when considering the dynamic nature of WWTPs. Processes are highly non-linear, making it difficult to produce a system of equations that accurately defines the systems and captures the systematic complexity. The result of this is that standard lumped parameter modeling approaches cannot be applied to WWTPs easily (Lessard and Beck, 1991). Further to this, WWTPs are inherently non-stationary and contain several time-varying parameters meaning any attempt at modeling must be updateable dynamically (Newhart et al., 2019).

Predictive modeling in this sector typically uses mechanistic models built from domain knowledge and empirical data. The activated sludge model (ASM) is a well-established example of this and has been employed usefully both in the design and operation of WWTPs (Quaghebeur et al., 2022). The ASM incorporates mass and energy balances with reaction kinetics into a mathematical model that can be used to describe the process and predict the effect of disturbances. These models require extensive domain knowledge and experience which is often hard to come by. Furthermore, the underlying dynamics are often simplified, or asterisked with assumptions that cause uncertainty in the model's ability to predict (Duarte et al., 2024). More recently, and with the advent of Industry 4.0 (I4.0) bringing a proliferation of sensor data, researchers have explored the use of machine learning (ML) to model the behavior of wastewater treatment processes.

2.1. Machine learning for WWTP energy prediction

One of the benefits of the adoption of ML for WWTP modeling is the ability to directly predict variables that exist outside of the mechanistic profiles of the mass and energy balance. One such property is the EC. In previous work EC has been inferred from key performance indicators (KPIs) relating to energy such as volume of water treated, chemical oxygen demand (COD) or total oxygen demand (TOD) removed (Huang et al., 2023). This enables an understanding of the directionality or trend of consumption but does not directly address the magnitude of the EC since it neglects a wealth of information in other measured parameters. Machine learning can address this by providing models that directly predict the absolute value of energy consumption from measured WWTP historical data. Further advantages of ML lie in its inherent flexibility and scalability. By leveraging historical data alone, ML models can be readily constructed and adapted to evolving situations. This ability has made it a popular choice for predicting energy consumption (EC) in wastewater treatment plants (WWTPs), as evidenced by a vast array of relevant applications documented in the literature.

Highlighting the importance of ML for modeling WWTP operations due to the nonlinearity of variable relationships, Zhang et al. developed a Random Forest (RF) model to predict energy consumption for 2472 WWTPs in China (Zhang et al., 2021). This model selection stemmed from the RF's known stability and resilience to missing data, a crucial factor considering the frequent noise and incompleteness inherent in WWTP datasets (Zhang et al., 2021). While their work yielded positive results, it did not factor in external influences like climate or meteorological effects, which demonstrably impact WWTP operations. The model's long-term viability demands the inclusion of these factors, as their omission could render it less effective in adapting to shifting conditions (Zhang et al., 2019).

Alali et al. provide a comprehensive evaluation of machine learning methods applied to EC prediction in wastewater treatment. Despite finding the k-nearest neighbors (kNN) model the most effective, Alali et al. noted that deep learning models, in particular, recurrent neural networks (RNNs) mark an intriguing avenue for this type of research due to their inherent ability to handle time-series data (Alali et al., 2022). It was also found that model predictions can be improved through consideration of variables outside of the dataset since these latent variables could have a profound impact on the results of the prediction.

Boncescu et al. applied logistic regression models to the same problem of EC forecasting in WWTPs from Romania (Boncescu et al., 2021). This work was done in isolation of factors that make up water quality, for example, nitrogen concentration or ammonia which other authors found to be important when conducting feature selection (Bagherzadeh et al., 2021).

Bagherzadeh et al. took the work further by combining water quality information with meteorological data for a WWTP in Melbourne, Australia. The authors compared the effectiveness of different ML models for the prediction of EC, finding gradient-boosted machines (GBM) to be the best-performing model for the task followed by RF regression (Bagherzadeh et al., 2021). Interestingly, this work showed a better performance of GBMs over deep learning frameworks such as artificial neural networks (ANNs) and RNNs on the testing data. Bagherzadeh et al. also begin to explore the explainability of the model by considering the importance of each of the features to the predicted variable. This approach relies on the assumption that the dataset exhibits all possible variance for each feature to make an informed decision on which should be included in the modeling. It may therefore lack the information to explain all perturbations in the energy consumption trend of the WWTP which limits the operational ability of the model to advise on reducing EC in the future.

Torregrossa et al. found similar results in which RF regression performed similarly to ANNs with the author commenting that the overfitting of neural networks to the training data is a significant problem due to the highly volatile nature of the data coming from WWTPs (Torregrossa et al., 2018). This is a major concern for applying deep learning frameworks, however has largely been overcome in other fields with the use of regularization methods (Romero-Güiza et al., 2022).

Picos-Benítez et al. were able to demonstrate the effective use of ANNs to capture the non-linearities and auto-correlations from data in a raw sewage treatment plant to be able to optimize operational conditions. In particular, the authors demonstrated the use of ANNs to produce fast and accurate predictions about operational parameters to mitigate issues caused by long experimental wait times (Picos-Benítez et al., 2020). This has been verified by other researchers who used neural networks to predict critical WWTP parameters such as biological oxygen demand (BOD) (Ahmadi et al., 2018). Typically, ANNs are considered ‘black box’ methods which do not allow for interpretable predictions, leading to questions about their application for decision-making. Furthermore, deep learning methods such as ANNs are extremely ‘data-hungry’. Due to the offline nature of the sampling of water quality attributes, the level of usable data samples to train ML models is often insufficient for deeper architectures which could be why authors have seen lower performances from neural networks than from other ML models.

This factor could also contribute to the reason why few authors have experimented with RNNs to forecast energy consumption in WWTPs, despite their high performance in time-series forecasting tasks. In cases where this has been done, they too have shown to be less effective than classical ML techniques such as RF and GBM (Bagherzadeh et al., 2021). ML techniques, however, struggle with maintaining performance when testing noise increases and therefore are not appropriate for modeling WWTP data which is characterized by high variability and high noise.

The culmination of the above literature review goes to show that no purely statistical or data-driven techniques enable the accurate modeling of WWTPs without a compromise in flexibility, accuracy, interpretability or robustness. The most applicable models concern the combination of mechanistic and empirical domain knowledge with lumped parameter models, such as the ASMs, however, this is limited by assumptions and simplifications.

The combination of empirical, or mechanistic knowledge concerning cause-and-effect relationships with machine learning for modeling WWTPs is a promising research avenue. This approach, referred to here as knowledge-enhanced machine learning, leverages the accurate predictions and robustness of machine learning, with the interpretability of knowledge-based modeling. Further to this, learning from historical data alleviates the need for detailed mapping of all complex relationships within the process reducing the reliance on simplified equations or biased assumptions.

Recent literature has begun to explore this direction, albeit not framed as knowledge-enhanced approaches. For instance, Karadimos and Anthopoulos developed NN models to predict the energy consumption of WWTPs in Greece, incorporating both quantitative and qualitative variables. Their method, while not directly integrating causal knowledge, used feature selection methods to incorporate specific domain knowledge into the modeling approach (Karadimos and Anthopoulos, 2023). Both Zhang et al., Bagherzadeh et al. also implicitly include knowledge through a feature engineering approach to enhance the modeling capabilities of machine learning. The following subsection will explore in more detail the concept of knowledge-enhanced data-driven modeling and its applications in the context of WWTPs.

2.2. Knowledge-enhanced data-driven modeling of WWTPs

The topic of hybrid knowledge-enhanced data-driven modeling for WWTPs has been approached differently by several authors. Cheng et al. pose a hybrid system in which knowledge taken from the equations of the ASM2 model is combined with convolutional neural networks (CNNs) and LSTMs to create a hybrid network used to forecast water quality (Cheng et al., 2023). This interesting approach focuses on the industry’s reliance on ASM models and expands the capabilities through the use of deep learning to be able to incorporate hidden factors that may influence the WWTP system beyond those captured in the original ASM equations (Lindow et al., 2020). The method demonstrated by Cheng et al. marks a leap forward in WWTP modeling, showing improvement over classical ASM models and other deep learning architectures, pointing towards the promise of this avenue of research.

Henze et al. use the ASM1 model to augment sparse data sets from WWTP to improve the training of deep learning models. This works particularly well since the ASM1 data can act as a filter to remove some of the noise of the dataset, making it easier for the deep learning models to find appropriate correlations between variables (Henze et al., 2006). Heo et al. provide a multi-objective supervisory control strategy for wastewater treatment using a hybrid approach to determine optimal setpoints of controllers under varying inlet conditions. They first group the influent conditions into clusters before using the benchmark simulation model no.2 (BSM2), another semi-mechanistic model derivative of the ASMs, to generate the objective function outputs. Using this, deep neural networks are proposed to optimize the system and map the control set points to achieve the desired output. Koksal et al. investigated the use of physics-informed neural networks (PINNs) to enhance machine learning models for wastewater treatment plants (WWTPs). By incorporating simplified equations ASM1 into recurrent neural network (RNN) architectures, they aimed to improve predictions of key parameters like dissolved oxygen and chemical oxygen demand. Their approach showed promise, with physics-informed models often outperforming standard versions in offline validations. However, online testing revealed mixed results, highlighting the challenges of applying generalized models to specific industrial plants. Despite some limitations, the PINN approach demonstrated potential for maintaining long-term performance without frequent updates, suggesting a possible solution to model drift in certain scenarios (Koksal et al., 2024). While this method offers one path for integrating domain knowledge into WWTP modeling, researchers have also explored alternative knowledge incorporation strategies.

Spatiotemporal modeling is an approach to combine specific knowledge about the geographical layout of WWTPs to enhance model capabilities. Huang et al. developed a novel approach to predict the EC of WWTPs in China by using a ridge regression technique to extract spatial information, while applying RF regression to extract key temporal information. The authors showed that this approach enhanced model performance for prediction of EC compared to purely temporal methods. Spatiotemporal modeling of this kind has proved successful in modeling aspects of wastewater treatment where the geographically dispersed nature of the system is important. Guo et al. analyzed wastewater information from seven different areas in China to assess the variation in the wastewater composition. These factors can help to inform the operation of the WWTP.

Guo and Wang used graph neural networks for modeling urban wastewater treatment systems. The authors proposed HydroNet, using

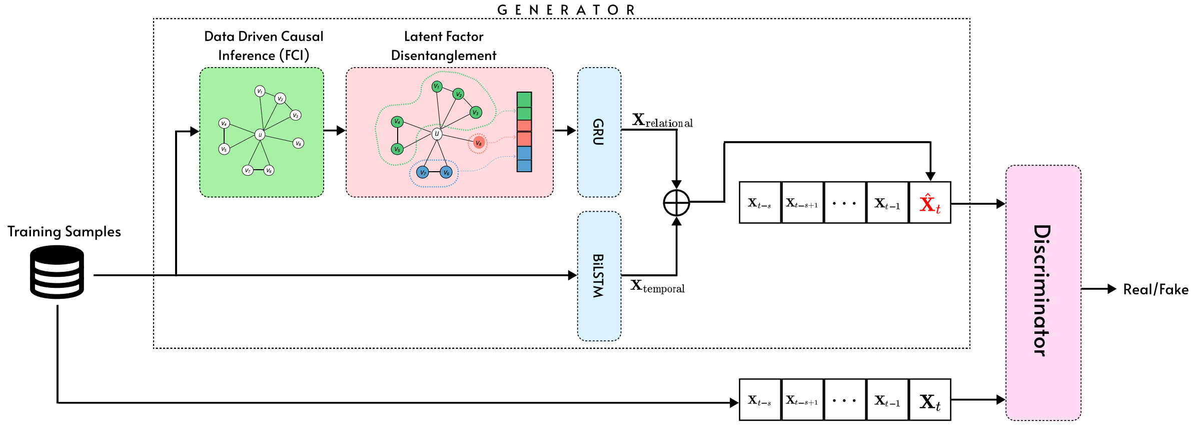


Fig. 1. Schematic for training the proposed KEGD-EC framework. The generator section trains to produce realistic forecasts for the system against a discriminator that attempts to differentiate real samples from the training data against fake samples from the generator. After training the generator is used to produce highly accurate predictions.

graphs to describe the layout of the wastewater network such that nodes represents junctions, manholes and other elements with edges weighted given the proximity of the network features. This information is processed using a graph convolutional network (GCN) to extract key features from both the graph and a historical dataset. It was shown that this method can help to accurately predict water infiltration (Guo and Wang, 2024).

Han et al. use an alternative approach for knowledge integration via networks in their work predicting and diagnosing faults in WWTPs. Their method, referred to as data-knowledge-driven (DKD) modeling, uses Bayesian Networks (BNs) to estimate the root cause of process disturbances (Han et al., 2021). Bayesian Networks are a probabilistic modeling technique to represent cause-and-effect relationships of complex processes in an acyclic graph (Scutari, 2017). Here, instead of using ASM models, the authors generated their causal graph using data-driven Granger causality (Granger, 1969). Representing domain knowledge as a network has distinct advantages since it allows the representation of directed cause-and-effect relationships within the system. Therefore, when a change is observed within the system it can be traced, with the network giving an understanding of why this change took place. Furthermore, the use of data-driven causal inference to model cause and effect in WWTP is effective in this application since it alleviates the requirement for an in-depth understanding of each process and removes the reliance on assumption and simplification seen in the ASMs.

While the work of Han et al. proved effective, representing manufacturing processes with BN is tricky since by nature a BN is an acyclic graph and many manufacturing processes inherently contain cycles due to recycling streams (Gharahbagheri et al., 2017). Furthermore, Granger causality assumes stationarity in the data, an assumption that does not hold for wastewater treatment due to the high variability in the waste and seasonal effects. This work marks an interesting avenue in the representation of WWTPs using knowledge graphs built from causal relationships within the process variables. The use of graphs grants inherent flexibility since fully defined mechanistic equations are not required to draw edges between graph nodes.

The field of graph machine learning is well-established in other domains within process manufacturing, however is yet to be applied to the modeling of WWTPs. Graph convolutional networks (GCNs) directly exploit the non-Euclidean relational structure of data, allowing them to learn complex dependencies (Hu et al., 2022). Graphs are diverse structures containing relational data that can refer to geospatial relations (Cao et al., 2020), plant layout information (Wu and Zhao, 2021), or relationships from mechanistic equations (Allen et al., 2024). This relational context facilitates more accurate predictions in complex systems where data is noisy, or relationships are highly non-linear. Liao et al. provide a comprehensive review of GCN applications in power systems. The authors pose that GCNs have high potential across this domain, particularly in exploring complex relationships in high dimensional data (Liao et al., 2022). However, some questions are raised around the efficacy of pure GCNs for time-series forecasting tasks, stating that GCNs may only be suitable for short-term forecasting tasks. Instead, researchers have highlighted opportunities in pairing GCNs with gated recurrent unit (GRU) architectures for the application of RNNs in the non-Euclidean domain (Liao et al., 2022).

There are relatively few works that have explored the practical application of combining GCNs with RNN architectures. There is little work showing the modeling of any manufacturing processes using GCN-RNN architectures to date. Allen et al. explored the integration of manufacturing domain knowledge for knowledge-enhanced spatiotemporal analysis (KESA) for fault detection and diagnosis (FDD) applied to a case study of the Tennessee Eastman Process (TEP) (Allen et al., 2024). Here it was found that the integration of knowledge graphs showing mechanistic causal relationships of manufacturing variables into the predictive framework increased the accuracy of predictions since the complex relationships present in manufacturing are difficult to learn in the absence of causal domain knowledge (Park et al., 2020). This phenomenon was also observed by authors Wu and Zhao who utilized a process topology convolutional network (PTCN) for FDD (Wu and Zhao, 2021). Further to the accuracy benefits, the authors found that the inclusion of knowledge graphs makes the results from predictions inherently more explainable (Allen et al., 2024). Providing explainable predictions is crucial for manufacturing decision-making, since being able to quickly and clearly understand a prediction leads to fast and effective decisions. However, it was also noted that the accuracy of GCN-RNN architectures is largely affected by the influence of external factors from the dataset. Given the nature of wastewater treatment and the impact that external factors have on the processes, some thought must go into considering the robustness of GCN-RNN algorithms. This could be one of the barriers to the adoption of this technology into the water treatment domain thus far. Other possible reasons for this include the highly non-linear relationships between variables, the complexity of the process itself, the influence of upstream effects on the processing leading to latent variables, or lack of knowledge of mechanistic models which make constructing knowledge graphs for this type of process tricky (Belia et al., 2009).

2.3. Addressing the challenges

The objective of this work is to address the difficulties of modeling and predicting wastewater treatment processes. This challenge is tackled through a hybrid knowledge-enhanced deep learning approach. Consideration is given to the current limiting factors for modeling WWTPs including the formulation of causal knowledge, and the high influence of external factors. Specifically, we present a framework to predict the energy consumption of wastewater treatment plants. Known as knowledge-enhanced graph disentanglement for energy consumption (KEGD-EC) prediction our framework disentangles latent factors from contextually rich datasets from WWTPs made up of knowledge graphs depicting cause-and-effect relationships, and historical data. This facilitates more accurate, robust and explainable predictions that can aid in resource planning, production scheduling and load balancing. The novel contributions are four-fold:

- Comparison of domain knowledge-based and data-driven causal discovery algorithms to uncover important causal relationships in WWTPs.
- Creation of a novel framework combining latent factor disentanglement with RNN architectures for time series prediction of complex WWTP systems.
- Comparison of model performance against established models documented in recent literature using a case study from a WWTP in Melbourne, Australia. Evaluated against model accuracy and model robustness.

The following Section 3 details the methodology employed, first outlining the mathematical background of each of the novel layers used followed by a description of the framework architecture. Then, the training of the model is described. Detail is given on the case study dataset in Section 4, including the construction of a process knowledge graph. Finally, Section 5 presents the results of the KEGD-EC framework in comparison to recent literature comparing both the accuracy and robustness of the models.

3. Methodology

Fig. 1 shows a machine learning framework in which historical plant data can be used to predict the future consumption of the plant (Newhart et al., 2019). Specifically, the approach described below concerns a graph-based ML approach in which the model is given the historical data, along with a knowledge graph depicting the cause-and-effect relationships between dataset variables. As stated, capturing cause-and-effect relationships from first principles for WWTPs is a difficult challenge due to highly variable processes, and a requirement for high amounts of domain expertise. Data-driven causal discovery solutions could pose a more realistic solution to obtain a causal knowledge graph directly from historical data (Han et al., 2021). This work aims to compare such methods with traditional causal-graph construction from domain knowledge. The methods employed in this work are explicitly described in Section 4.2. Once the cause-and-effect relationships have been mapped, they can be used in a graph-based deep-learning framework. The proposed framework adapts the KESA-AD algorithm for the unique challenges of modeling WWTPs (Allen et al., 2024). There are two clear distinctions between the initial KESA-AD model and the one presented in the following section.

Firstly, in this work, we have adapted the spatial dimension to instead represent the strength of cause-and-effect relationships in the graph rather than geometric proximity focusing while still maintaining the spatiotemporal forecasting principles. The dissemination of relational data from temporal data is a key factor in enabling accurate forecasts. It allows the model to capture the key drivers behind the data, independent of just analyzing historical trends. The expansion of the model from spatial information to relational information facilitates the expansion of the model to include variables external to the process such as the meteorological variables, which are important to the operations of WWTPs (Bagherzadeh et al., 2021).

Secondly, we propose the addition of disentanglement to the graph convolution operation. The goal of this is to produce a disentangled representation of the node embedding for the causal graph. This representation allows for an understanding of the impact of external factors on the dataset and therefore grants inherent robustness, and interpretability to the model (Alemi et al., 2016). This is important since WWTPs can be influenced by many factors outside of the measured dataset, for example by upstream events, local conditions or the skill of the operators. These factors are difficult to quantify but can have a substantial impact on the measured variables from the dataset.

The KEGD-EC framework's architecture is specifically designed to address three key challenges in WWTP energy consumption prediction: (1) complex interacting factors affecting energy use, (2) multi-scale temporal dependencies, and (3) the need for interpretable predictions. Each component of the framework addresses one or more of these challenges:

- The disentanglement mechanism enables separation of complex, interacting factors affecting energy consumption into interpretable latent representations. This is particularly crucial for WWTPs where multiple operational, environmental, and process parameters simultaneously influence energy use.
- The knowledge-enhanced graph structure explicitly captures known causal relationships between process variables, leveraging domain expertise to improve prediction accuracy and interpretability. This addresses the challenge of integrating expert knowledge with data-driven approaches.
- The temporal modeling components combine short-term dynamics through GRU layers with longer-term patterns via BiLSTM, specifically designed to capture the multi-scale temporal patterns characteristic of WWTP operations.

Notation: Unless stated otherwise, matrices will be depicted with bold uppercase letters (e.g., $\mathbf{A}$), vectors will be described with lowercase bold letters (e.g., $\mathbf{x}$), uppercase italic letters for sets (e.g., G), and lowercase italic letters for scalar values (e.g., k). We use $^\top$ for matrix transpose and^{-1} for matrix inversion. The subscript $\mathbf{A}_{i,j}$ is used to represent the value on the ith row and the jth column of the matrix $\mathbf{A}$.

3.1. Mathematical background

In the following, we express a WWTP as a directed knowledge graph $G = (V, E)$ where V is a finite set of nodes such that $n = |V|$ corresponds to the number of features describing the wastewater treatment process in a dataset $\mathbf{X} \in \mathbb{R}^{m\times n}$. For a node $u \in V$, there is a corresponding feature vector $\mathbf{x}_u \in \mathbb{R}^m$. E is a set of edges such that $e_{u,v} \in E$ shows the existence of an edge between node u and node v where each edge represents a causal relationship. Therefore, given a graph G of a process and a corresponding dataset $\mathbf{X} \in \mathbb{R}^{m\times n}$, can we accurately learn the relationships between the feature variables to predict the energy consumption of the WWTP at a future timestep?

3.1.1. Disentangled graph convolution

Knowledge graphs are used in this work to capture key causal relationships that can aid in accurate forecasting. Graphs, however, cannot be directly implemented into machine learning frameworks since they do not follow the rules of Euclidean geometry (Kipf and Welling, 2017). Instead, to use graphs we must first apply a convolution operation to process the graph and unlock the relational information. Researchers have successfully applied both spectral and spatial graph convolution operations for ML (Bruna et al., 2014) and spatial graph convolutions (Shuman et al., 2013). In both cases, a global embedding of the graph nodes is generated where the representation for a node is learned from its neighborhood. Despite this being successful in many applications, some researchers have shown that this approach fails to recognize latent factors that may be driving change in the process resulting in non-robust and non-explainable results (Guo et al., 2022). Applied to WWTP, using global node embeddings could result in prediction uncertainty which since these processes can be influenced by a range of external factors (Li et al., 2022). The uncertainty arising from neglecting underlying latent factors has been addressed in the literature with disentangled convolutions that partition node neighborhoods into channels or capsules representing the latent factors which can be concatenated to get a disentangled node representation (Liu et al., 2020).

To perform the disentanglement of the process knowledge graph, G, we utilize a disentangled graph convolutional layer (DisenConv). The goal of the DisenConv layer is to output a disentangled node representation, $\mathbf{y}_u = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K] \in \mathbb{R}^d$ of node u into K channels where each channel describes an independent latent factor such that each graph feature is composed of K. $\mathbf{c}_k \in \mathbb{R}^{\frac{d}{K}}$ describes the channel for the individual node u relating to factor k. This is done by projecting the feature vector $\mathbf{x}_u$ into separate subspaces. Each channel k has its parameters: a weight matrix $\mathbf{W}_k \in \mathbb{R}^{m \times \Delta d}$ and a bias vector $\mathbf{b}_k \in \mathbb{R}^{\Delta d}$:

$$\mathbf{z}_{u,k} = \frac{\sigma(\mathbf{W}_k^\top \mathbf{x}_u + \mathbf{b}_k)}{\|\sigma(\mathbf{W}_k^\top \mathbf{x}_u + \mathbf{b}_k)\|_2}, \tag{1}$$

where $\Delta d = \frac{d}{K}$ is the output size of each channel and σ is a non-linear ReLU activation function.

In real-world scenarios, the feature vector $\mathbf{x}_u$ often contains missing values or may be incomplete, so the global network structure cannot be preserved. In this case, it becomes necessary to explore the second-order proximity of each node, determined through the shared neighborhood structures (Tang et al., 2015). Ma et al. propose a neighborhood routing mechanism to assess the likelihood of the edge between two nodes being explained by a latent factor k. It is assumed that a factor k is likely to be the reason why node u is connected to a subset of its neighbors if *(i)* the subset is large *and* the nodes of the subset are similar in aspect k *(ii)* if node u is similar in aspect k to each neighbor in the subset. Mining information from the neighborhood of each node in this way provides robustness to model predictions in cases where the feature vector of a node $\mathbf{x}_u$ is incomplete or noisy (Ma et al., 2019). This is important in the case of wastewater treatment where data is often noisy and regularly incomplete (Ba-Alawi et al., 2022).

The algorithm calculates the probability that an edge exists between node u and node v due to factor k. Let this probability be represented by $p_{v,k}$ where $p_{v,k} \geq 0$ and $\sum_{k'=1}^{K} p_{v,k'} = 1$. The process of neighborhood routing employs an iterative approach to estimate the value of $p_{v,k}$ (Eq. (3)) prior to generating c_k (Eq. (2)). Since nodes u and v are assumed to be connected *because* of factor k it stands that $\mathbf{z}_{v,k}{}^\top \mathbf{z}_{u,k}$ will provide information on the edge between them. Therefore, the iterations start by initializing $p_{v,k}$ as $p_{v,k}^{(1)} \propto \exp(\mathbf{z}_{v,k}{}^\top \mathbf{z}_{u,k}/\tau)$. The parameter τ governs the rigidity or leniency of the assignment process (Ma et al., 2019). This routing mechanism identifies the most substantial cluster within each subspace, adhering to the restriction that any given neighbor is confined to a single subspace. The allocation of distinct neighbor subsets to each channel ensures that every channel embodies an independent factor. These procedures are executed as such:

$$\mathbf{c}_k^{(i)} = \frac{\mathbf{z}_u^k + \sum_{v:e_{u,v}\in G} p_{v,k}^{i-1} \mathbf{z}_{v,k}}{\|\mathbf{z}_u^k + \sum_{v:e_{u,v}\in G} p_{v,k}^{i-1} \mathbf{z}_{v,k}\|_2}, \tag{2}$$

$$p_{v,k}^{(i)} = \frac{\exp(\mathbf{z}_{v,k}{}^\top \mathbf{c}_k^{(i)}/\tau)}{\sum_{k'=1}^{K} \exp(\mathbf{z}_{v,k'}{}^\top \mathbf{c}_{k'}^{(i)}/\tau)}, \tag{3}$$

for each iteration where $i = 2, \ldots, I$.

The DisenConv layer is used to produce a disentangled node embedding for a graph G for each node where $\mathbf{y}_u \in \mathbb{R}^{K\Delta d}$. This embedding allows us to gain better information about the relationships between dataset features, including an understanding of how latent factors might be influencing the outcome of the process. The goal of this work, however, is not just to understand the relationships but to be able to predict the future state of the system. To do this, we look to pair the disentangled convolutional layer above with specific RNN architectures that are adept at capturing temporal dynamics, using the disentangled representation $\mathbf{y_u}$ as an input. This enables us to capture both the relational and temporal information to create accurate predictions.

3.1.2. DisenGRU layer

Gated recurrent units (GRU) are RNN architectures adept at handling temporal data to forecasting time-series representations, in many cases outperforming other common RNN architectures such as LSTMs (Dey and Salemt, 2017). Combining this with the disentangled representations allows for accurate forecasting in dynamic and noisy systems, such as wastewater treatment processes.

A gating mechanism, reminiscent of that used in LSTMs, is utilized by GRUs to regulate the flow of information within the unit (Chung et al., 2014). This helps to learn which information is temporally significant and should be retained, and which should be discarded (Cho et al., 2014). The variant of GRU employed in this work differs slightly from the original architecture proposed by Cho et al. since we are using the disentangled node embeddings instead of the raw historical data as input. The DisenConv layer generates a disentangled embedding $\mathbf{y}_u(T) = [\mathbf{c}_1(T), \mathbf{c}_2(T), \ldots, \mathbf{c}_K(T)] \in \mathbb{R}^{K\Delta d}$ at a time stamp t, $\mathbf{y}_{u,t}$.

The GRU uses the previous information and the current input to determine how much its hidden state, denoted by h_t, should change. The hidden state acts like a memory and can be thought of as a blend between the previous hidden state ($\mathbf{h}_{t-1}$) and the newly proposed candidate activation function ($\tilde{\mathbf{h}}_t$). This blending is controlled by an update gate ($\mathbf{z}_t$) The update gate assigns weights between 0 and 1 to each element of the previous and candidate states using element-wise multiplication:

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t. \tag{4}$$

A value of 1 in $\mathbf{z}_t$ means the GRU relies more on the new information ($\tilde{\mathbf{h}}_t$), while a value of 0 indicates it favors the previous state ($\mathbf{h}_{t-1}$). The update gate is calculated as per (Chung et al., 2014):

$$\mathbf{z}_t = \sigma_z(\mathbf{W}_z \mathbf{y}_{u,t} + \mathbf{U_z} \mathbf{h}_{t-1} + \mathbf{b_z}). \tag{5}$$

where σ_r is a sigmoid activation function, and $\mathbf{W_z}, \mathbf{U_z}, \text{and}, \mathbf{b_z}$ are all learnable parameters.

The candidate activation function is computed as in Bahdanau et al. (2014):

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W_h} \mathbf{y}_{u,t} + \mathbf{U_h}(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b_h}), \tag{6}$$

where $\mathbf{r}_t$ is a reset gate where $\mathbf{W_h}, \mathbf{U_h}, \text{and}, \mathbf{b_h}$ are learnable parameters. The reset gate is computed as such:

$$\mathbf{r}_t = \sigma_r(\mathbf{W_r} \mathbf{y}_{u,t} + \mathbf{U_r} \mathbf{h}_{t-1} + \mathbf{b_r}), \tag{7}$$

where $\mathbf{W_r}, \mathbf{U_r}, \text{and}, \mathbf{b_r}$ are learnable parameters.

The above modifications to the traditional GRU enable the disentangled representation of the process graph to be integrated into a disentangled gated recurrent unit (DisenGRU) layer. This DisenGRU layer enables us to include the spatial dynamics of the system, including potential latent factors that might be influencing the process, when forecasting the process time series enabling more accurate and explainable predictions. This layer has been employed in the wider framework below to predict the energy consumption of a WWTP.

3.2. KEGD-EC framework

3.2.1. Overview

The framework adopted here follows a broadly similar structure to the KESA framework demonstrated in Allen et al. (2024) which has been shown to accurately capture manufacturing dynamics. This framework is structured as a generative adversarial network (GAN). First proposed by Goodfellow et al., GANs comprise two elements; a generator and a discriminator. The discriminator function evaluates the probability of a sample's origin being the empirical training distribution rather than the generated distribution, whereas the generator function endeavors to approximate the underlying data distribution and produce accurate forecasts (Goodfellow et al., 2014). The two models are pitted against one another in training, with the generator attempting to fool the discriminator by producing a lifelike forecast, and the discriminator training to determine which samples are from the training data and which are from the generator. The result is a model that can accurately replicate the dynamics of a manufacturing system, including the complex non-linearities to produce accurate forecasts.

3.2.2. Generator

The generator creates accurate predictions of manufacturing systems. It is crucial to consider both the short and long-term dynamics of the system so the generator is divided into two separate modules. The *relational* module aims to capture short-term changes stemming from variations in the feed to the process, changing ambient conditions or from step changes to process parameters during operation. By utilizing the DisenGRU layers described in Section 3.1.2 we also aim to capture the effects of latent factors such as operator shift change, or upstream events. Since these events occur without precursor in many cases, we only assess the relational dynamics across a shortened period s. The disentangled representation is taken from the DisenConv layer and input to the GRU which gives a final prediction in the form $\mathbf{X}_{\text{relational}} \in \mathbb{R}^{s\times K\Delta d}$.

It is also important to consider the impact of longer-term fluctuations in system performance due to seasonal effects, or the effects of process degradation and wear. These types of changes occur more slowly and are less likely to produce a fast reaction in multiple process variables, but instead produce a trend for each feature. Therefore, we include a *temporal* module consisting of a bidirectional LSTM (BiLSTM) model has been included to capture and forecast the dynamics of each graph feature across a period l where $s < l$. Bidirectional LSTMs excel in capturing temporal dynamics within sequences and have been readily applied to complex scenarios where spatial factors come into play (Xie et al., 2022). BiLSTMs comprise two LSTM architectures with one LSTM layer processing the sequence forwards, and one LSTM layer processing the sequence in reverse. This increases the context available to the model and therefore increases the performance of the model, with Siami-Namini et al. demonstrating a 37.78% uplift in performance when using a BiLSTM in comparison to a traditional LSTM model (Siami-Namini et al., 2019). This is done for each feature such that $\mathbf{X}_{\text{temporal}} \in \mathbb{R}^{l\times n}$.

The final prediction of the generator is gained by concatenating the relational and temporal module predictions, and passing it through a fully connected layer with a non-linear activation function:

$$\hat{\mathbf{X}} = \tanh(\mathbf{W}_f[\mathbf{X}_{\text{relational}}, \mathbf{X}_{\text{temporal}}] + \mathbf{b}_f), \tag{8}$$

where $\mathbf{W}_f$ and $\mathbf{b}_f$ represent the weight matrix and the bias vector of the fully connected layer.

3.3. Discriminator

The goal of the discriminator is to decide whether a sequence has been taken from the training data, or generated by the above generator sequence. The discriminator comprises a temporal feature extraction step, in which a DisenGRU extracts features from the final timestep of the sequence. A relational feature extraction occurs using a DisenConv layer. The relational and temporal features are extracted and concatenated. This is passed through a linear layer with a sigmoid activation function which predicts the authenticity of the sequence as a probability between zero and one.

3.4. Adversarial training

The generator is trained against the discriminator to enable the most accurate prediction possible. Each network has a different loss function. The generator loss is made up of a forecasting error, i.e. the difference between the predicted sequence and the real value, and a realism loss which is minimized to fool the discriminator. The overall loss function for the discriminator is therefore given as:

$$L_G(\theta) = \lambda_G \sum_{t\in\text{batch}} \|G_\theta(\mathbf{X}_{t-s}, \dots, \mathbf{X}_{t-1}) - \mathbf{X}_t\|_2 - \log(D_\phi(\hat{\mathbf{X}}_t)), \tag{9}$$

where the generator function, denoted by $G_\theta(\cdot)$, is parameterized by $\boldsymbol{\theta}$, while $D_{\boldsymbol{\phi}}(\cdot)$ signifies the discriminator function with parameter $\boldsymbol{\phi}$. A hyperparameter λ_G is employed to equilibrate the loss between forecast accuracy and realism. The discriminator loss function comprises a realism loss between the data taken from the training data and the generated data:

$$L_D(\boldsymbol{\phi}) = \sum_{t\in\text{batch}} -\log(D_\phi(\hat{\mathbf{X}}_t)) - \log(D_\phi(\mathbf{X}_t)). \tag{10}$$

Once the model has been trained, the generator can be used to accurately forecast the dynamics of the plant. The purpose of the discriminator is to ensure the generator is producing the most accurate possible predictions.

4. Case Study - Melbourne Eastern Treatment Plant

The efficacy of the above framework from Fig. 1 has been demonstrated below on a dataset taken from the Eastern Treatment Plant (ETP) of Melbourne Water. The ETP treats approximately 449 megalitres of sewage per day on average from around half the population of Melbourne, producing 13022 megalitres of recycled, safe water in 2022–23 (Melbourne Water, 2023). The process follows a three-stage treatment process. Primary treatment involves removing physical pollutants from sewage using a combination of screening processes, sedimentation and grit removal. Secondary treatment uses aerobic and anaerobic digestion to break down organic material in the sewage before it is passed through clarifiers where more sediment is settled. The final tertiary treatment disinfects the water before it can be released to the outfall pump stations.

Given the scale of the treatment that occurs at the ETP, understanding and being able to accurately forecast the energy consumption is extremely important to maintaining operational efficiency, allowing operators to adjust treatment processes to factors that might affect the sustainability of the plant (Alali et al., 2023).

4.1. Dataset description

For an accurate energy consumption forecast at a WWTP, a diverse dataset reflecting influential factors is essential. This includes characteristics of the wastewater, hydraulic variables as well as meteorological variables. This study uses a dataset compiled from both the Eastern Treatment plant and Melbourne airport weather station, compiled initially by Bagherzadeh et al. (2021). The total dataset comprises samples collected between 2014 and 2019. Wastewater characteristics, such as the Ammonia concentration (NH_4-N), biological oxygen demand (BOD), total nitrogen (TN), and chemical oxygen demand (COD) were sampled daily from the ETP. Meteorological data was taken from the weather station at Melbourne Airport since it is the closest available weather data to the ETP. Finally, EC data was collected using revenue quality meters with a frequency of every 15 min which was averaged daily and joined to the overall dataset using an inner-join operation (Bagherzadeh et al., 2021). In preprocessing, samples containing null data were removed, equating to approximately 5% of the total dataset leaving 1382 samples. The data has been summarized in Table 1. The data was normalized since features differ in scale:

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X} - \mathbf{X}_{\text{min}}}{\mathbf{X}_{\text{max}} - \mathbf{X}_{\text{min}}}. \tag{11}$$

The raw dataset can be used to extract the temporal relationships for each of the features but offers little to explain the interactions between the variables, and how a change in one might impact the rest of the plant. Therefore, we also formulate a graph $G_{\text{ETP}} = (V, E)$ where $n = |V|$ is the number of sensors equal to the 16 variables described in Table 1.

In our previous KESA-AD work, we explored the construction of knowledge graphs using mechanistic profiles of the process. Wu and Zhao used the physical layout of the process to dictate the graph where

Table 1
Description of the variables in the ETP dataset.

Parameter (Abbreviation)	Unit	Mean	std	Min	Max
Hydraulic Parameters					
Average Outflow (Q_{out})	m^3/s	3.93	1.23	0.00	7.92
Average Inflow (Q_{in})	m^3/s	4.51	1.44	2.59	18.97
Wastewater Parameters					
Ammonia ($NH_4 - N$)	mg/L	39.22	7.76	13.00	93.00
Biological Oxygen Demand (BOD)	mg/L	382.06	86.00	140.00	850.00
Chemical Oxygen Demand (COD)	mg/L	845.96	145.42	360.00	1700.00
Total Nitrogen (TN)	mg/L	62.74	3.57	40.00	92.00
Climate Parameters					
Average Temperature T_{av}	°C	15.04	5.40	0.00	35.50
Maximum Temperature T_{max}	°C	20.53	7.10	0.00	43.50
Minimum Temperature T_{min}	°C	10.04	4.66	−2.00	28.50
Atmospheric Pressure (AP)	hPa	3.68	61.01	0.00	1022.00
Average Humidity (H)	%	63.56	14.53	0.00	97.00
Total Precipitation (Pr)	mm	0.22	1.31	0.00	18.03
Average Visibility (VIS)	Km	9.10	16.32	0.00	512.00
Average Wind Speed (WS_{av})	Km/h	19.48	7.14	0.00	49.10
Maximum Wind Speed (WS_{max})	Km/h	35.38	11.63	0.00	83.50
Energy Consumption					
Energy Consumption (EC)	MWh	275.16	44.64	116.64	398.33

units were represented by nodes and edges represent the physical connection between the units. While these methods provide good results in their applications, they do not directly apply to this case study since either the mechanistic relationships between variables are difficult to derive, or there are no physical connections to relate to edges (for example between T_{max} and Q_{in}). Therefore, alternative methods must be established to construct the edges, E between nodes in G_{ETP}.

4.2. Graph building

Data-driven causal discovery is an interesting approach that looks to recover the underlying causal structure of variables from observed data. This process is often far quicker and more scalable than deriving a causal graph from experiments or domain expertise. However, it comes at the cost of accuracy since it is limited to the information present in the dataset and the assumptions that constrain the algorithm (Molak, 2023). Here we compare two separate methods, one data-driven and one from domain expertise, for deriving the causal structure between the dataset variables.

While the intention is to see if the data-driven algorithm can uncover the correct causal structure to match the domain expert, only data that is relevant to the cause of energy consumption prediction should be considered. In the case of the data presented in Table 1 some of the environmental variables have little bearing on the overall energy consumption (Wiesmann et al., 2007). Furthermore, when analyzing this data set Bagherzadeh et al. found that there is little correlation between EC, and the variables for wind speed, visibility and atmospheric pressure, and therefore these variables were removed from the dataset before the causal discovery and model training.

4.2.1. Data driven causal discovery

When building graphs for manufacturing purposes it is important to identify links that represent *causation* as opposed to *correlation*. The former occurs when a node v has a direct effect on the state of node u. Alternatively, a correlation may occur when a third variable w directly causes v and u. In this case, while nodes u and v may be highly correlated, a change in one will not affect the other and therefore drawing an edge here would be misleading. Node w here is known as a confounder. Identifying correlation instead of causation is problematic for the explainability of the model since simply noting the correlation between variables could lead to false conclusions about factors driving energy consumption. To solve this, there are dedicated algorithms for determining the causal structure from observed data. Here, we employ a constraint-based causal discovery technique known as fast causal inference (FCI) (Spirtes et al., 1999).

FCI constructs causal structures based on conditional independence constraints. Conditional independence occurs when a 'Y' shaped structure is formed. In this scenario, both node u and node w are conditionally independent of node y since y is independent of u and w conditional on knowing the state of v. This reduces concerns about hidden factors influencing both node v and node y (Mani et al., 2012). The FCI algorithm starts by assuming all variables are directly connected in a fully connected, undirected graph. It then iteratively removes edges between conditionally independent variables. Finally, it analyzes the specific patterns in the remaining edges to determine the direction of the causal relationships (Shen et al., 2020). The result is a directed, unweighted graph that represents causal relationships between observed variables. While other methods for data-driven causal discovery exist, FCI does not overlook the existence of latent common causes and is therefore appropriate for use here (Lee et al., 2023).

Applied to the adjusted ETP dataset, the FCI algorithm identifies a graph with 11 nodes relating to each of the dataset variables and 16 edges as visualized in Fig. 2(a). The FCI algorithm derived causal relationships between some of the nodes, for example between average temperature (T_{av}) and energy consumption (EC). However, the algorithm has also identified edges likely to be influenced by latent confounders. Therefore, it is vital to perform validation of the connections against the written knowledge of WWTP operation, and against the knowledge of expert practitioners (Hagedorn et al., 2022).

4.3. Domain knowledge graph

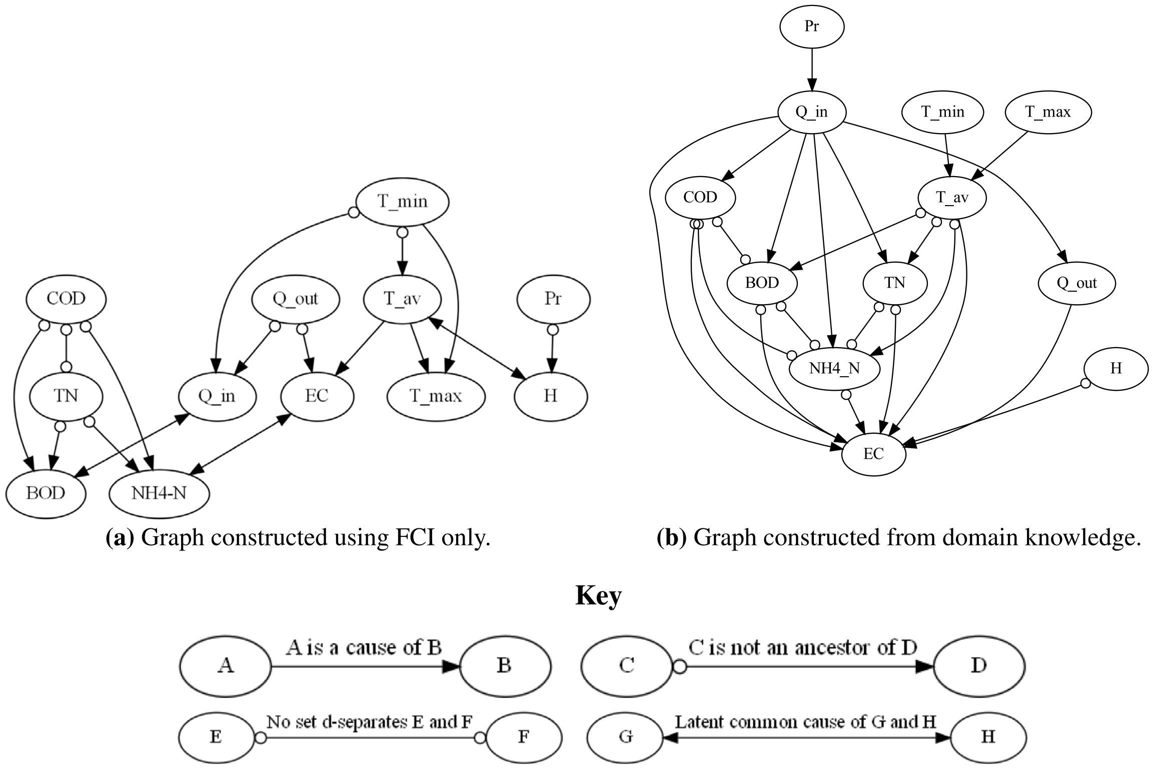
Traditionally, causal knowledge is drawn from practitioner expertise. Here, we use a variety of literature sources, engineering knowledge, and industrial experience to produce a second graph Fig. 2(b), with a justification of each edge explained in Table 2. By assessing the FCI graph against the explanations in the table, we can gauge the success of the data-driven causal discovery on the ETP dataset.

4.3.1. Discussion of graph structures

There are some key structural differences between Figs. 2(a) and 2(b) that represent differences in the causal discovery approaches. The expert knowledge-derived graph exhibits a higher degree of directional certainty in its causal relationships compared to the FCI-generated graph which shows a greater amount of bidirectional edges. This is particularly evident for the treatment of EC which appears as a sink node in Fig. 2(b), however shows bidirectional relationships in the FCI graph. This difference indicates a more conservative approach to causal attribution using FCI, which potentially indicates an introduction of bias when using expert knowledge - a known risk of causal graphs built from expert systems.

There are several notable differences in the graph structures which highlight some important aspects of the two approaches. The difference in the relationship between inflow (Q_{in}) and outflow (Q_{out}) is particularly striking. While the expert graph shows a logical $Q_{in} \rightarrow Q_{out}$ relationship, the FCI algorithm suggests the opposite direction. This discrepancy suggests that the FCI algorithm may be misinterpreting temporal data, highlighting the challenges in inferring causal direction from time-series data.

Treatment of wastewater parameters (COD, BOD, NH_4-N, TN) also differs between the graphs. The expert graph acknowledges some confounded relationships, while the FCI graph suggests a more intricate web of interactions. This complexity in the FCI graph could offer valuable insights into the biochemical processes within WWTPs, potentially revealing subtle interactions that expert knowledge might overlook.



(a) Graph constructed using FCI only.

(b) Graph constructed from domain knowledge.

Fig. 2. Causal knowledge graph of the Eastern Treatment Plant of Melbourne Water using both data-driven causal discovery (a) and domain knowledge (b).

Another notable difference lies in the treatment of precipitation across the two graphs. The expert-derived graph establishes a direct link between precipitation and inflow, reflecting a strong assumed relationship. In contrast, the FCI-generated graph assigns a more isolated role to precipitation, suggesting a weaker or more complex relationship with other variables. This discrepancy potentially reveals limitations in either the data or the FCI algorithm, but it also highlights a possible bias introduced through expert knowledge. The UK-based practitioner who contributed to the expert graph may place a higher weight on the influence of precipitation, reflecting the climate conditions typical of the United Kingdom. However, this assumption may not hold for a WWTP located in Australia, where average monthly rainfall is considerably lower. This geographical disparity underscores the importance of contextualizing expert knowledge and reveals how data-driven approaches like FCI might help identify and mitigate regional biases in causal assumptions.

These structural differences have significant implications for predicting energy consumption in WWTPs. While the expert graph's directional certainty might lead to more straightforward predictive models, it could overlook subtle relationships captured by the FCI graph. Conversely, the FCI graph's complex treatment of wastewater parameters might capture subtle interactions affecting energy use, potentially improving prediction accuracy.

The subsequent sections of this paper will leverage both of these graphs, ultimately comparing the success of the causal discovery by assessing the impact of the graph structure on the accuracy of energy consumption prediction. This comparative analysis aims to evaluate the predictive power of expert knowledge versus data-driven causal discovery in the context of WWTP energy consumption, identify areas where the FCI algorithm provides novel insights that could enhance expert understanding of WWTP dynamics, and assess the practical implications of using different causal structures in graph neural network models for energy prediction.

4.4. Model training & evaluation

4.4.1. Training data

The total ETP comprises 1382 samples and each sample has 12 features including energy consumption. A training–testing split of 75/25% is applied to the dataset leaving 1036 samples for training, and 346 samples for testing. Since the data is a time series, the temporal order is maintained instead of shuffling the dataset during the split. Splitting the data allows objective model evaluation by testing a trained model on data it has never seen before. The testing dataset remains untouched during model training and tuning.

The training data is further divided into a validation dataset of 259 samples. This dataset is used to test model configurations and tune hyperparameters without revealing the final test set which is held for unbiased evaluation. Table 3 shows the range of hyperparameters tested using a grid search approach. This involves training a model using each configuration and validating the result using the validation dataset. The best model configuration is selected and the model is trained, and final results are collected on the testing dataset.

Each of the models is trained over 5 epochs, in which we divide the data into mini-batches of 16 samples. The Adam optimizer was used, with a learning rate of 0.001 (Kingma and Ba, 2014).

4.5. Model evaluation

Since the purpose of the model developed in this paper is to predict the consumption of energy in a particular WWTP in Melbourne, the metrics used to evaluate the model should reflect how closely the predicted time series reflects the true values of the data. Therefore, we select metrics that capture the closeness of the fit of the predicted data to the real data. The coefficient of determination R^2 measures how well the variance of the true data is captured by the model, with values closer to one representing a better fit, and is calculated as follows:

$$\mathrm{R}^2 = 1 - \frac{\sum_{t=1}^{T} (x_{t,\mathrm{EC}} - \hat{x}_{t,\mathrm{EC}})^2}{\sum_{t=1}^{T} (x_{t,\mathrm{EC}} - \bar{x})^2}, \tag{12}$$

Table 2
Explanation for causal graph edges shown in Fig. 2(b) based on domain knowledge (see Refs. Carrera et al. (2004), de Almeida Fernandes et al. (2018), Dubber and Gray (2010), Henze et al. (2008), Metcalf et al. (1991), Xie et al. (2024), Zhou and Xu (2019)).

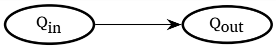
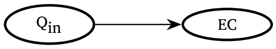
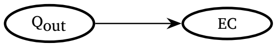
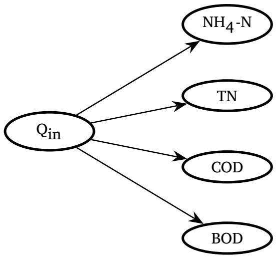
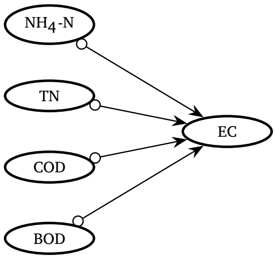
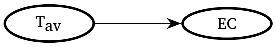
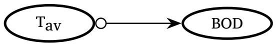
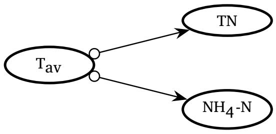
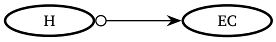
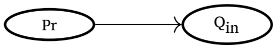
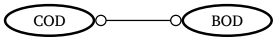
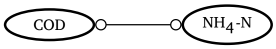
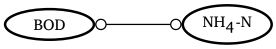
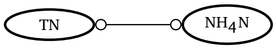
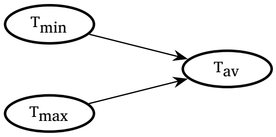
Edge	Justification
	Inflow directly affects the outflow in wastewater treatment plants.
	Higher influent flow rates will result in higher energy consumption.
	Outflow rates affect the pumping energy requirement.
	High inflow rates affect the concentration of key wastewater constituents through dilution or loading (Henze et al., 2008). High inflow rates can also affect scouring rates in some systems, again affecting concentrations (Xie et al., 2024).
	The energy demand for wastewater treatment is closely linked to the removal of organic matter and nutrients. Therefore, the concentration of these components has a direct causal link to the energy consumption (Hamawand, 2023).
	Temperature affects biological processes and oxygen solubility, influencing the energy requirements (Metcalf et al., 1991).
	Temperature can affect microbial activity, potentially influencing BOD (de Almeida Fernandes et al., 2018).
	Temperature affects nitrification rates (Zhou and Xu, 2019).
	Humidity may affect evaporation rates and potentially influence energy consumption. The two features have shown a correlation in previous works and therefore this is considered a confounded relationship (Bagherzadeh et al., 2021).
	Significant precipitation events result in higher inflows to wastewater treatment plants.
	COD and BOD are related measures of organic matter. The relationship is confounded by the biodegradability of the organic matter present (Dubber and Gray, 2010).
	High COD can inhibit nitrification, affecting NH_4-N removal, but this relationship is uncertain and confounded by factors like carbon-to-nitrogen ratio and oxygen availability (Carrera et al., 2004).
	Like COD, high BOD can affect nitrification, but this relationship is uncertain and confounded by factors such as oxygen competition between heterotrophs and nitrifiers (Henze et al., 2008).
	NH_4N is a component of TN, but their relationship can be influenced by various factors in the treatment process. For example, nitrification and denitrification can affect total nitrogen concentration without changing ammonia concentration (Henze et al., 2008).
	Average temperature is a function of both the maximum and minimum temperatures.

Table 3
Shows hyperparameters for grid search. Bold values indicate the final model hyperparameters.

Parameter	Values
Number of Hidden Nodes (DisenConv)	8, 16, **32**, 64
Size of period s (number of steps)	**3**, 5, 7
Size of period l (number of steps)	**30**, 45, 60
Number of neighborhood routing iterations	3, **5**, 7
Number of Channels K	2, 3, **4**
Number of layers (LSTM)	**2**, 3
Number of Hidden Nodes (RNNs)	16, **32**, 64

where $x_{t,\mathrm{EC}}$ represents the true value of the energy consumption at time t, $\hat{x}_{t,\mathrm{EC}}$ is the model predicted value, and $\bar{x}$ is the mean value of energy consumption across the period.

Along with the coefficient of determination, we use root mean squared error (RMSE) to measure the magnitude of the difference between the predicted and actual values. Lower values of RMSE are better. RMSE is calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_{t,\mathrm{EC}} - \hat{x}_{t,\mathrm{EC}})^2}. \tag{13}$$

Mean absolute percentage error (MAPE) and mean absolute error (MAE) are two common metrics used to assess how well a model performs. MAPE expresses the difference between predicted and actual values as a percentage, while MAE focuses on the average absolute difference in their units. They are calculated as follows:

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|x_{t,\mathrm{EC}} - \hat{x}_{t,\mathrm{EC}}|, \tag{14}$$

Table 4
Results comparison between other ML models used for EC prediction and the KEGD-EC framework. Bold results show the best performance.

Model	RMSE	R^2	MAE	MAPE	Source
GBM	26.78	−0.41	20.27	0.07	Bagherzadeh et al. (2021)
RF	26.06	−0.34	19.48	0.07	Torregrossa et al. (2018)
kNN	32.90	−1	26.67	0.1	Alali et al. (2023)
LSTM (RNN)	33.25	−0.74	24.39	0.09	Harrou et al. (2023)
ANN	29.27	−0.36	21.13	0.08	Picos-Benítez et al. (2020)
Transformer	29.14	−0.63	22.85	0.08	–
KEGD-EC (FCI)	14.23	0.65	10.91	0.04	–
KEGD-EC (Domain Knowledge)	**10.50**	**0.81**	**7.52**	**0.03**	–

$$\text{MAPE} = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{x_{t,\text{EC}} - \hat{x}_{t,\text{EC}}}{x_{t,\text{EC}}}\right|\%. \tag{15}$$

5. Results & discussion

In this work, we assess the ability of the KEGD framework proposed in Fig. 1 for forecasting the energy consumption of a wastewater treatment plant in Melbourne, Australia. To evaluate our approach comprehensively, we conducted comparisons against both traditional machine learning methods and state-of-the-art deep learning architectures. The traditional baseline models, identified as effective in recent literature, include gradient-boosted machine (GBM), random forest (RF), k-Nearest-Neighbors (kNN), LSTM, and ANN. To ensure comparison against current state-of-the-art approaches, we implemented a Transformer model incorporating multi-head self-attention mechanisms and positional encoding, which has shown remarkable success in temporal modeling tasks (Farahani et al., 2024). This architecture was specifically adapted for time-series forecasting through temporal attention masking and specialized positional embeddings (Zhou et al., 2022). Furthermore, the KEGD framework was trained separately using two different causal graph structures – one derived from expert domain knowledge and another from Fast Causal Inference (FCI) – to demonstrate the importance of graph structure selection on forecasting accuracy.

Table 4 compares the model results from each model. All models were trained, validated and tested on the same datasets for transparency. The trained models were tested on three subsets of the testing data and the results averaged between the results for each model to get an understanding of the consistency of the model n. This ensures that the results presented are consistently achieved across different training data and not perchance results from a single testing set. The models were all trained on the same system equipped with an AMD Ryzen 7 5800H with Radeon Graphics, 3201 Mhz, 8 Cores, 16 Logical Processors, 32 GB RAM, with access to an NVIDIA GeForce RTX 3080 16GBGDDR6 GPU.

5.1. Impact of disentanglement

In a previous ablation study, the authors examined the impact of including the disentanglement mechanism in the graph convolution (Allen and Cordiner, 2024). Using a reduced network pairing the different graph convolution methods to an LSTM architecture, the authors showed that the inclusion of the disentanglement reduced the RMSE on a subset of the energy consumption data by 61.5%. These results underscore the importance of the overall disentanglement approach in capturing complex relationships between process variables. The study demonstrated that standard graph convolution methods struggle to identify latent factors driving energy consumption. The significant improvement suggests that the disentangled representation enables the model to better capture and separate the various factors influencing WWTP energy consumption, such as operational parameters, environmental conditions, and process dynamics.

5.2. Model comparison

Table 4 showcases the superior performance of our KEGD-EC model compared to established models documented in the literature for similar datasets. KEGD-EC, using either the domain knowledge graph or the FCI graph, demonstrates significant improvement across all relevant metrics. Most notably, the results demonstrate an average reduction in RMSE of 59.7% between KEGD-EC using domain knowledge (10.50MWh), and the next best model (RF) (26.06 MWh). Furthermore, KEGD-EC stands alone as the only model with a positive coefficient of determination. This is a significant result, showing that the inclusion of causal knowledge in machine learning facilitates better learning than purely data-driven algorithms despite a relatively limited dataset. This presents an opportunity for recently digitized industries that may not have access to sufficient historical data to train larger models.

The method of knowledge graph construction significantly influenced model performance in our study. The Knowledge-Enhanced Graph Deep Learning (KEGD) model using domain knowledge outperformed the one utilizing Fast Causal Inference (FCI), highlighting the value of domain expertise in constructing causal process graphs for complex systems like WWTPs. However, both KEGD models surpassed other machine learning approaches, with the FCI-based model offering greater scalability.

While domain knowledge proved superior for our 12-node graph, its time and expertise requirements could hinder implementation in larger systems with hundreds or thousands of nodes. Future research should therefore focus on hybrid approaches that combine FCI's scalability with targeted domain knowledge refinement. This strategy could yield models that balance accuracy and scalability, making them suitable for larger, more complex systems.

Fig. 3 visually corroborates KEGD-EC's superiority by plotting predicted energy consumption against actual values (blue line). As evident, the KEGD-EC with domain knowledge (KEGD-EC (DK)) predictions (orange line) closely align with the actual trend, accurately capturing the data's shape and overall movement. In contrast, other models consistently deviate, either overestimating or underestimating true consumption. Such inconsistencies, with significant over- or under-predictions, can lead to operational challenges due to unreliable estimates of energy use, potentially resulting in inefficiencies and costly overconsumption.

An examination of Fig. 3 reveals a limitation in the model's predictive capability during periods of atypically high energy consumption. Specifically, the model struggled to capture the shape of the data between April 2018, when the energy consumption of the ETP was unusually elevated. This sudden spike in energy usage suggests the influence of external factors on the water treatment process. Upon investigation of local events, we found that this anomalous increase in energy consumption coincides with bushfires in the Thompson and Upper Yarra catchments of Melbourne. These fires, which began on March 1st, 2018, and were not contained until March 25th, 2018 (Thwaites et al., 2018), likely contributed significantly to the observed spike. The aftermath of such events can substantially impact water quality, as rainfall washes ash and sediment into waterways, necessitating more intensive treatment processes. Consequently, the production of clean water requires more energy than usual, as evidenced by the increased consumption shown in Fig. 3.

The model's inability to predict this spike can be attributed to the absence of relevant data in our dataset. Information about bushfire occurrences or the quantity of solids in the water stream was not included, leaving the model unable to account for these factors in its

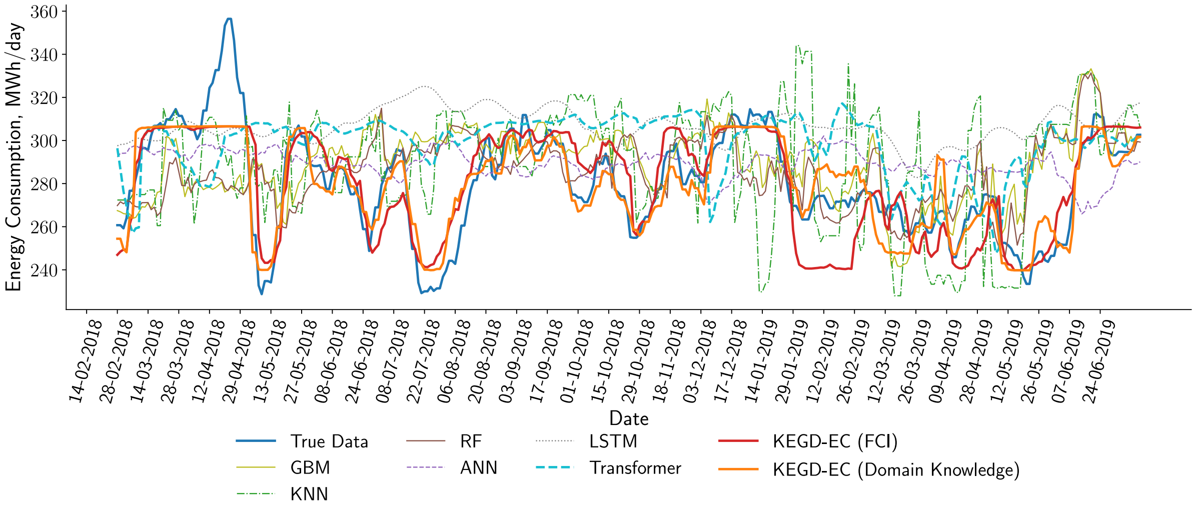


Fig. 3. Comparison of KEGD-EC prediction on testing data, with the FCI graph model shown in red and the domain knowledge-based model shown in orange, compared with other model performance including GBM (olive), kNN (green), RF (brown), ANN (purple), Transformer (cyan) and LSTM (gray) with the true data plotted (blue).

predictions. This limitation underscores the importance of incorporating area-specific practitioner knowledge in the construction of causal knowledge graphs. It is worth noting that the expert knowledge used to construct Fig. 2(b) was based on experience managing WWTPs in the UK. Consequently, the effects of bushfires, which are more common in Australia, were not considered in the original model. This oversight resulted in the model's inability to anticipate this particular deviation in energy consumption. Had there been consideration for region-specific environmental factors such as bushfires in the causal graph, the model might have been better equipped to predict or at least account for such anomalies. This observation highlights the need for adaptable, region-specific models that can incorporate local environmental factors and potential extreme events in their predictive frameworks.

The inclusion of domain knowledge in the formation of the knowledge graph is an important step for future work since an experienced practitioner with knowledge of the local area would have been able to highlight the effect of bushfires. Therefore, when creating the knowledge graph a binary node to represent bushfires could have been included, or the dataset could have been expanded to include reading the percentage of solids in the water supply which would give the model the context required to make accurate predictions.

5.3. Model robustness

Alongside performance metrics, a model applied to manufacturing systems must demonstrate resilience to noisy data, as this is a hallmark of industrial processes. In this work, the robustness of a model is tested by adding increasing amounts of Gaussian noise to the testing dataset. The performance is recorded for each model for each noise level and plotted in Fig. 4. We add noise to each testing dataset using NumPy where the noise has the same shape as the data, but the noise intensity is increased by altering the standard deviation starting at $\sigma = 0$ (no noise) through to $\sigma = 10$ (Harris et al., 2020).

What is clear is that the traditional ML methods struggle to cope with increasing noise in the testing dataset. For the GBM, RF and KNN models, the performance significantly decreases as σ increases, shown by the increase in RMSE, MAE and MAPE values, as well as the decrease in the coefficient of determination. Deep learning models, on the other hand, show a far greater resilience to the noise in the testing data. The ANN, LSTM and Transformer-based models show little change in performance as the noise increases, in some cases. This shows the flexibility of these models to adapt to noise levels. Deep learning models have a far deeper structure compared to ML models and have a larger number of parameters that allow them to capture complex non-linearities providing them with greater flexibility to respond to noise. In particular, we see the LSTM model, while initially exhibiting worse performance than the ANN as per Table 4, responds better to increased noise owing to its noted ability to handle temporal data.

Since KEGD-EC also makes use of RNN deep learning architectures (LSTM, GRU) we see it maintains a far higher performance than other models, even despite increased noise levels in the testing data. It is interesting to note that KEGD-EC has a much higher performance than traditional LSTM models despite integrating RNN architectures. This improvement must be put down to the inclusion of causal knowledge. Causal knowledge included in this away enables the model to go beyond learning temporal information as with the tested RNN architectures, facilitating a deeper understanding of the cause-and-effect relationships to gain a more accurate and meaningful prediction.

The KEGD-EC models, using both FCI and domain knowledge graphs, demonstrate superior resilience to noise compared to other models. While there is a slight difference in their performance under increasing noise, both versions maintain high accuracy across all noise levels. The FCI graph-based model shows more consistent performance, while the domain knowledge graph version experiences a minor increase in error around $\sigma = 5$ before improving again. This subtle difference could be attributed to variations in graph structures stemming from data-driven (FCI) versus expert-based (domain knowledge) approaches. For instance, the domain knowledge graph might overemphasize certain factors like precipitation, based on UK-centric expertise, which may not perfectly align with Melbourne's WWTP conditions. Despite this, both KEGD-EC versions significantly outperform all other models across all noise levels, highlighting the robustness gained from combining causal knowledge with deep learning architectures in noisy industrial environments.

6. Conclusions

In this study, we present KEGD-EC, a novel knowledge-enhanced graph disentanglement network for predicting energy consumption in wastewater treatment plants (WWTPs). Our approach leverages the power of fast causal inference (FCI) to generate a knowledge graph, extracting and encoding causal relationships between features like hydraulics, water quality, and even meteorological conditions. By separating the underlying cause-and-effect relationships from temporal data

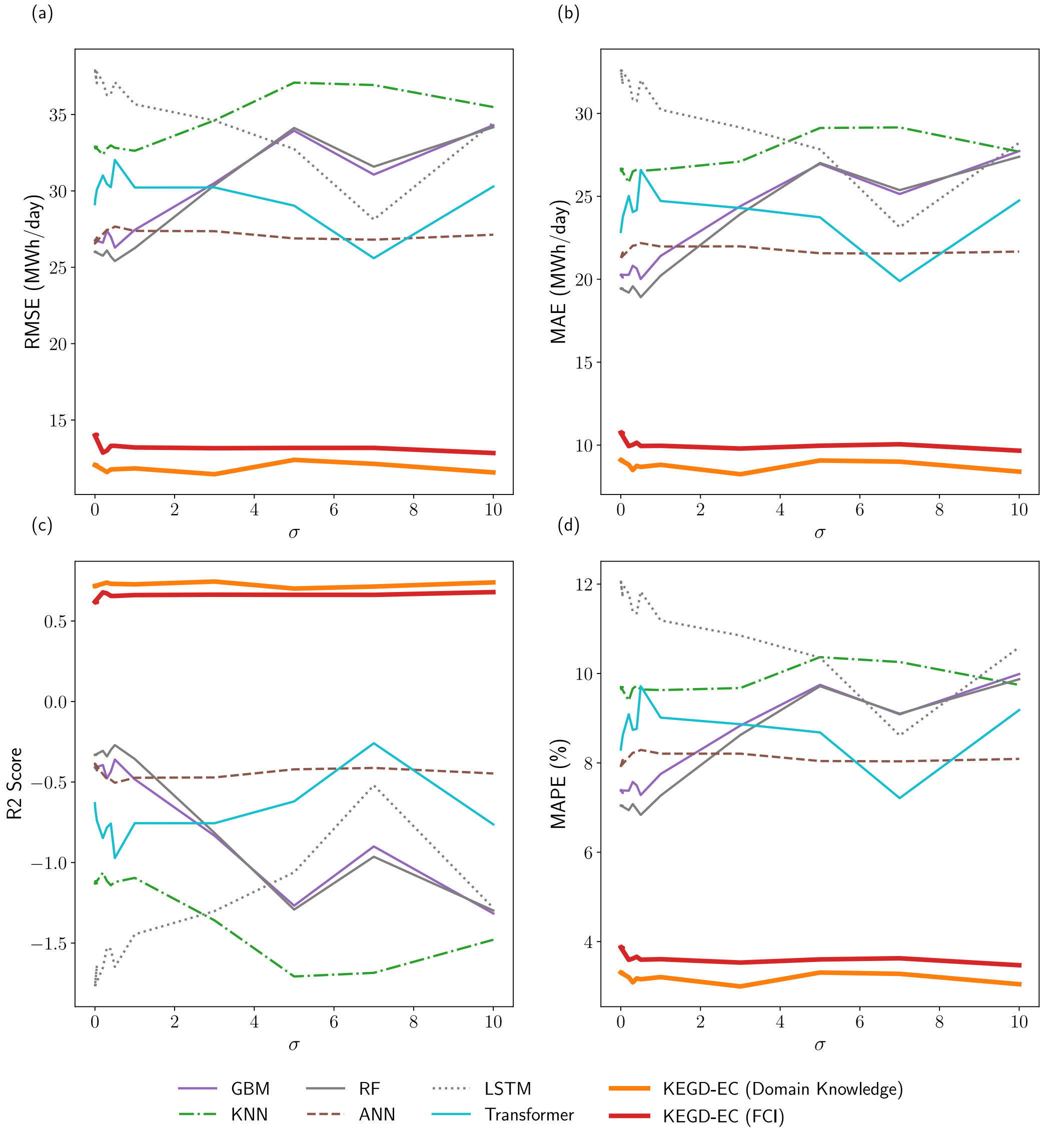


Fig. 4. Comparison of model robustness for each model. Plots show the change in performance for each metric with an increasing noise of the testing data where the *x*-axis represents the standard deviation of the noise applied to the testing data. **(a)** RMSE, **(b)** MAE, **(c)** R^2 Score, **(d)** MAPE.

patterns, we overcome challenges posed by noisy data streams and highly correlated variables in these intricate systems.

The disentanglement proves crucial, leading to significant improvements in prediction accuracy compared to other top-performing machine learning models, including deep learning architectures like ANNs and RNNs and other graph-based methods using spectral convolutional approaches. Notably, we achieved a **59.7% reduction in RMSE compared to the next best model**, demonstrating exceptional accuracy. Moreover, KEGD effectively captures 81% of the variance in the energy consumption data.

Interestingly, KEGD-EC outperforms even similar RNN architectures like LSTMs. This success, despite limited training data, suggests that the **inclusion of causal knowledge and disentanglement mechanism together significantly reduce training data requirements**. This opens possibilities for newly digitized industries, where traditional deep learning methods struggle with insufficient historical data.

The effectiveness of this architectural design is demonstrated through ablation studies. Removing the disentanglement mechanism results in a 61.5% increase in RMSE, while using a standard graph structure without knowledge enhancement reduces model performance by 26%. These results validate that each component makes a necessary contribution to the framework's overall performance.

An improvement in EC prediction enables WWTPs to optimize their operations through proactively adjusting treatment processes to minimize energy consumption, effectively participating in demand response programs to support grid stability, scheduling energy-intensive operations during off-peak hours, and strategically planning maintenance activities. Accurate prediction also enables plants to better integrate renewable energy sources, optimize hydraulic management, and make informed long-term decisions on capacity planning and technology investments, ultimately leading to substantial cost savings and improved environmental sustainability. Future work should consider the integration of process parameters into the dataset, allowing online optimization of plant parameters in response to changing external conditions that may affect energy consumption.

Importantly, our study revealed that while both data-driven (FCI) and domain knowledge-based approaches outperformed traditional methods, the model using domain expertise for graph construction showed superior performance. This finding underscores the critical importance of incorporating specific domain knowledge in modeling complex systems like WWTPs. However, the FCI-based model still outperformed other machine learning approaches, offering a more scalable solution for larger systems where comprehensive domain expertise might be challenging to obtain.

The robustness of KEGD-EC was evident in its resilience to increasing levels of noise in the testing data, with both versions significantly outperforming other models across all noise levels. This demonstrates the value of integrating causal knowledge with deep learning architectures in noisy industrial environments.

Despite these advances, our model showed limitations in predicting anomalous events, such as the energy consumption spike related to bushfires. This highlights the need for incorporating more diverse data sources and region-specific expertise in future iterations of the model.

Looking ahead, future research should explore hybrid approaches that combine data-driven causal discovery with targeted domain knowledge, potentially offering a balance between accuracy and scalability. Expanding the dataset to include a greater range of operational scenarios and external factors could enhance the model's predictive accuracy and generalizability. This includes consideration of the online deployment of a trained KEGD model in a real WWTP environment. Such testing would validate the model's performance, and give insights into its long-term reliability.

CRediT authorship contribution statement

Louis Allen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Joan Cordiner:** Supervision, 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.

Data availability

Data will be made available on request.

References

Agency, U.S.E.P., 2021. Energy efficiency for water utilities. https://www.epa.gov/sustainable-water-infrastructure/energy-efficiency-water-utilities.

Ahmad, T., Chen, H., 2018. Utility companies strategy for short-term energy demand forecasting using machine learning based models. Sustainable Cities Soc. 39, 401–417. http://dx.doi.org/10.1016/j.scs.2018.03.002.

Ahmadi, A., Fatemi, Z., Nazari, S., 2018. Assessment of input data selection methods for BOD simulation using data-driven models: a case study. Environ. Monit. Assess. 190, 1–17. http://dx.doi.org/10.1007/S10661-018-6608-4/TABLES/6, URL: https://link.springer.com/article/10.1007/s10661-018-6608-4.

Alali, Y., Harrou, F., Sun, Y., 2022. Predicting energy consumption in wastewater treatment plants through light gradient boosting machine: A comparative study. In: 10th International Conference on Systems and Control. ICSC, http://dx.doi.org/10.1109/icsc57768.2022.9993872.

Alali, Y., Harrou, F., Sun, Y., 2023. Unlocking the potential of wastewater treatment: Machine learning based energy consumption prediction. Water (Switzerland) 15, http://dx.doi.org/10.3390/W15132349.

Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K., 2016. Deep variational information bottleneck. In: 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. International Conference on Learning Representations, ICLR, URL: https://arxiv.org/abs/1612.00410v7.

Allen, L., Cordiner, J., 2024. Towards sustainable WWTP operations: Forecasting energy consumption with explainable disentangled graph convolutional networks. In: Proceedings of the 34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering.

Allen, L., Lu, H., Cordiner, J., 2024. Knowledge-enhanced spatiotemporal analysis for anomaly detection in process manufacturing. Comput. Ind. 161, 104111.

Ba-Alawi, A.H., Loy-Benitez, J., Kim, S.Y., Yoo, C.K., 2022. Missing data imputation and sensor self-validation towards a sustainable operation of wastewater treatment plants via deep variational residual autoencoders. Chemosphere 288, 132647. http://dx.doi.org/10.1016/J.CHEMOSPHERE.2021.132647.

Bagherzadeh, F., Nouri, A.S., Mehrani, M.J., Thennadil, S., 2021. Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach. Process Saf. Environ. Prot. 154, 458–466. http://dx.doi.org/10.1016/J.PSEP.2021.08.040.

Bahdanau, D., Cho, K.H., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR, URL: https://arxiv.org/abs/1409.0473v7.

Bansard, J., Allan, J., Eni-ibukun, T.A., Dubrova, A., Luomi, M., Tan, J.M., Mundin, C., Rosentreter, H., Wagner, L., 2023. Earth negotiations bulletin a reporting service for environment and development negotiations cop 28 final. International Institute for Sustainable Development, Dubai.

Belia, E., Amerlinck, Y., Benedetti, L., Johnson, B., Sin, G., Vanrolleghem, P.A., Gernaey, K.V., Gillot, S., Neumann, M.B., Rieger, L., Shaw, A., Villez, K., 2009. Wastewater treatment modelling: Dealing with uncertainties. Water Sci. Technol. 60, 1929–1941. http://dx.doi.org/10.2166/WST.2009.225.

Boncescu, C., Robescu, L.D., Bondrea, D.A., Macinic, M.E., 2021. Study of energy consumption in a wastewater treatment plant using logistic regression. IOP Conf. Ser.: Earth Environ. Sci. 664, 012054. http://dx.doi.org/10.1088/1755-1315/664/1/012054, https://iopscience.iop.org/article/10.1088/1755-1315/664/1/012054, https://iopscience.iop.org/article/10.1088/1755-1315/664/1/012054/meta.

Bruna, J., Zaremba, W., Szlam, A., Lecun, Y., 2014. Spectral networks and deep locally connected networks on graphs.

Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., Zhang, Q., 2020. Spectral temporal graph neural network for multivariate time-series forecasting. In: 34th Conference on Neural Information Processing Systems. NeurIPS, Vancouver.

Carrera, J., Vicent, T., Lafuente, J., 2004. Effect of influent COD/N ratio on Biological Nitrogen Removal (BNR) from high-strength ammonium industrial wastewater. Process Biochem. 39, 2035–2041. http://dx.doi.org/10.1016/J.PROCBIO.2003.10.005.

Cheng, X., Guo, Z., Shen, Y., Yu, K., Gao, X., 2023. Knowledge and data-driven hybrid system for modeling fuzzy wastewater treatment process. Neural Comput. Appl. 35, 7185–7206. http://dx.doi.org/10.1007/S00521-021-06499-1/FIGURES/18, URL: https://link.springer.com/article/10.1007/s00521-021-06499-1.

Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

de Almeida Fernandes, L., Pereira, A.D., Leal, C.D., Davenport, R., Werner, D., Filho, C.R.M., Bressani-Ribeiro, T., Chernicharo, C.A.d., de Araújo, J.C., 2018. Effect of temperature on microbial diversity and nitrogen removal performance of an anammox reactor treating anaerobically pretreated municipal wastewater. Bioresour. Technol. 258, 208–219. http://dx.doi.org/10.1016/J.BIORTECH.2018.02.083.

Dey, R., Salemt, F.M., 2017. Gate-variants of Gated Recurrent Unit (GRU) neural networks. In: Midwest Symposium on Circuits and Systems, vol. 2017-August, Institute of Electrical and Electronics Engineers Inc., pp. 1597–1600. http://dx.doi.org/10.1109/MWSCAS.2017.8053243.

Duarte, M.S., Martins, G., Oliveira, P., Fernandes, B., Ferreira, E.C., Alves, M.M., Lopes, F., Pereira, M.A., Novais, P., 2024. A review of computational modeling in wastewater treatment processes. ACS ES T Water 4, 784–804. http://dx.doi.org/10.1021/ACSESTWATER.3C00117.

Dubber, D., Gray, N.F., 2010. Replacement of Chemical Oxygen Demand (COD) with Total Organic Carbon (TOC) for monitoring wastewater treatment performance to minimize disposal of toxic analytical waste. J. Environ. Sci. Health - A Toxic/Hazardous Substances Environ. Eng. 45, 1595–1600. http://dx.doi.org/10.1080/10934529.2010.506116.

Farahani, M.A., Kalach, F.E., Harper, A., McCormick, M.R., Harik, R., Wuest, T., 2024. Time-series forecasting in smart manufacturing systems: An experimental evaluation of the state-of-the-art algorithms. URL: https://arxiv.org/abs/2411.17499v1.

Gandiglio, M., Lanzini, A., Soto, A., Leone, P., Santarelli, M., 2017. Enhancing the energy efficiency of wastewater treatment plants through co-digestion and fuel cell systems. Front. Environ. Sci. 5, 289034. http://dx.doi.org/10.3389/FENVS.2017.00070/BIBTEX, URL: www.frontiersin.org.

Gharahbagheri, H., Imtiaz, S.A., Khan, F., 2017. Root cause diagnosis of process fault using KPCA and Bayesian network. Ind. Eng. Chem. Res. 56, 2054–2070. http://dx.doi.org/10.1021/ACS.IECR.6B01916/ASSET/IMAGES/IE-2016-01916K_M048.GIF, URL: https://pubs.acs.org/doi/full/10.1021/acs.iecr.6b01916.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27.

Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: J. Econom. Soc. 424–438.

Guo, J., Huang, K., Yi, X., Zhang, R., 2022. Learning disentangled graph convolutional networks locally and globally. Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/TNNLS.2022.3195336,

Guo, S., Nkinahamira, F., Adyari, B., Zhang, Y., Hu, A., Sun, Q., 2023. Fate and spatial–temporal variation of 23 elements at 7 wastewater treatment plants in Southeast City of China. Water (Switzerland) 15, 1226. http://dx.doi.org/10.3390/W15061226/S1, https://www.mdpi.com/2073-4441/15/6/1226/htm, https://www.mdpi.com/2073-4441/15/6/1226.

Guo, Q., Wang, W., 2024. HydroNet: A spatio-temporal graph neural network for modeling hydraulic dependencies in urban wastewater systems. In: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, ACM, New York, NY, USA, pp. 717–718. http://dx.doi.org/10.1145/3678717.3695761, URL: https://dl.acm.org/doi/10.1145/3678717.3695761.

Haase, D., Güneralp, B., Dahiya, B., Bai, X., Elmqvist, T., et al., 2018. Global urbanization. In: The Urban Planet: Knowledge Towards Sustainable Cities, vol. 19, Cambridge University Press, Cambridge, pp. 326–339.

Hagedorn, C., Huegle, J., Schlosser, R., 2022. Understanding unforeseen production downtimes in manufacturing processes using log data-driven causal reasoning. J. Intell. Manuf. 33 (7), 2027–2043.

Hamawand, I., 2023. Energy consumption in water/wastewater treatment industry—Optimisation potentials. Energies 2023, Vol. 16, Page 2433 16, 2433. http://dx.doi.org/10.3390/EN16052433, https://www.mdpi.com/1996-1073/16/5/2433/htm https://www.mdpi.com/1996-1073/16/5/2433.

Han, H.G., Dong, L.X., Qiao, J.F., 2021. Data-knowledge-driven diagnosis method for sludge bulking of wastewater treatment process. J. Process Control 98, 106–115. http://dx.doi.org/10.1016/J.JPROCONT.2021.01.001.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. http://dx.doi.org/10.1038/s41586-020-2649-2.

Harrou, F., Dairi, A., Dorbane, A., Sun, Y., 2023. Energy consumption prediction in water treatment plants using deep learning with data augmentation. Results Eng. 20, 101428. http://dx.doi.org/10.1016/J.RINENG.2023.101428.

Henze, M., Gujer, W., Mino, T., Van Loosedrecht, M., 2006. Activated sludge models ASM1, ASM2, ASM2d and ASM3. IWA publishing.

Henze, M., van Loosdrecht, M.C., Ekama, G.A., Brdjanovic, D., 2008. Biological Wastewater Treatment. IWA publishing.

Heo, S.K., Nam, K.J., Tariq, S., Lim, J.Y., Park, J., Yoo, C.K., 2021. A hybrid machine learning–based multi-objective supervisory control strategy of a full-scale wastewater treatment for cost-effective and sustainable operation under varying influent conditions. J. Clean. Prod. 291, 125853. http://dx.doi.org/10.1016/J.JCLEPRO.2021.125853.

Hu, Y., Cheng, X., Wang, S., Chen, J., Zhao, T., Dai, E., 2022. Times series forecasting for urban building energy consumption based on graph convolutional network. Appl. Energy 307, 118231. http://dx.doi.org/10.1016/J.APENERGY.2021.118231.

Huang, R., Yu, C., Wang, H., Zhang, S., Wang, L., Li, H., Zhang, Z., Zhou, Z., 2023. Spatial and temporal modeling on energy consumption of wastewater treatment based on machine learning algorithms. ACS ES T Water 4, 1119–1130. http://dx.doi.org/10.1021/ACSESTWATER.3C00430/ASSET/IMAGES/LARGE/EW3C00430_0007.JPEG, URL: https://pubs.acs.org/doi/full/10.1021/acsestwater.3c00430.

Jeppsson, U., 1996. Modelling Aspects of Wastewater Treatment Processes. Citeseer.

Karadimos, P., Anthopoulos, L., 2023. Machine learning-based energy consumption estimation of wastewater treatment plants in Greece. Energies 2023, Vol. 16, Page 7408 16, 7408. http://dx.doi.org/10.3390/EN16217408, https://www.mdpi.com/1996-1073/16/21/7408/htm, https://www.mdpi.com/1996-1073/16/21/7408.

Kingma, D.P., Ba, J.L., 2014. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR, URL: https://arxiv.org/abs/1412.6980v9.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: ICLR 2017.

Kirchem, D., Lynch, M., Bertsch, V., Casey, E., 2020. Modelling demand response with process models and energy systems models: Potential applications for wastewater treatment within the energy-water nexus. Appl. Energy 260, 114321. http://dx.doi.org/10.1016/J.APENERGY.2019.114321.

Koksal, E.S., Asrav, T., Esenboga, E.E., Cosgun, A., Kusoglu, G., Aydin, E., 2024. Physics-informed and data-driven modeling of an industrial wastewater treatment plant with actual validation. Comput. Chem. Eng. 189, 108801. http://dx.doi.org/10.1016/J.COMPCHEMENG.2024.108801.

Lee, J.J., Srinivasan, R., Ong, C.S., Alejo, D., Schena, S., Shpitser, I., Sussman, M., Whitman, G.J., Malinsky, D., 2023. Causal determinants of postoperative length of stay in cardiac surgery using causal graphical learning. J. Thorac. Cardiovasc. Surg. 166, e446–e462. http://dx.doi.org/10.1016/J.JTCVS.2022.08.012, URL: https://pubmed.ncbi.nlm.nih.gov/36154975/.

Lessard, P., Beck, M., 1991. Dynamic modeling of wastewater treatment processes. Environ. Sci. Technol. 25 (1), 30–39.

Li, Y., Chen, Z., Zha, D., Du, M., Ni, J., Zhang, D., Chen, H., Hu, X., 2022. Towards learning disentangled representations for time series. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, http://dx.doi.org/10.1145/3534678.

Liao, W., Bak-Jensen, B., Pillai, R., Wang, Y., Wang, Y., 2022. A review of graph neural networks and their applications in power systems. J. Mod. Power Syst. Clean Energy 10, http://dx.doi.org/10.35833/MPCE.2021.000058.

Lindow, F., Muñoz, C., Jaramillo, F., Bishop, R.H., Proal-Nájera, J.B., Antileo, C., 2020. Active biomass estimation based on ASM1 and on-line OUR measurements for partial nitrification processes in sequencing batch reactors. J. Environ. Manag. 273, 111150. http://dx.doi.org/10.1016/J.JENVMAN.2020.111150.

Liu, Y., Wang, X., Wu, S., Xiao, Z., 2020. Independence promoted graph disentangled networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, AAAI Press, pp. 4916–4923. http://dx.doi.org/10.1609/AAAI.V34I04.5929, URL: https://ojs.aaai.org/index.php/AAAI/article/view/5929.

Longo, S., d'Antoni, B.M., Bongards, M., Chaparro, A., Cronrath, A., Fatone, F., Lema, J.M., Mauricio-Iglesias, M., Soares, A., Hospido, A., 2016. Monitoring and diagnosis of energy consumption in wastewater treatment plants. a state of the art and proposals for improvement. Appl. Energy 179, 1251–1268. http://dx.doi.org/10.1016/J.APENERGY.2016.07.043.

Ma, J., Cui, P., Kuang, K., Wang, X., Zhu, W., 2019. Disentangled graph convolutional networks. In: Proceedings of the 36th International Conference on Machine Learning, PMLR. pp. 4212–4221.

Mani, S., Spirtes, P.L., Cooper, G.F., 2012. A theoretical study of y structures for causal discovery. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence. UAI2006, URL: https://arxiv.org/abs/1206.6853v1.

Maslon, A., Wojcik, M., Chmielowski, K., 2018. Efficient Use of Energy in Wastewater Treatment Plants. Energy Policy Institute, Krakow, pp. 12–23.

Melbourne Water, L., 2023. Melbourne Water Annual Report. Melbourne Water, Melbourne, URL: http://www.melbournewater.com.au.

Metcalf, L., Eddy, H.P., Tchobanoglous, G., 1991. Wastewater Engineering: Treatment, Disposal, and Reuse, vol. 4, McGraw-Hill, New York.

Molak, A., 2023. Causal Inference and Discovery in Python: Unlock the Secrets of Modern Causal Machine Learning with Dowhy, Econml, Pytorch and More. Packt Publishing Ltd.

Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: A review. Water Res. 157, 498–513. http://dx.doi.org/10.1016/J.WATRES.2019.03.030.

Park, Y.J., Fan, S.K.S., Hsu, C.Y., 2020. A review on fault detection and process diagnostics in industrial processes. Processes 2020, Vol. 8, Page 1123 8, 1123. http://dx.doi.org/10.3390/PR8091123, https://www.mdpi.com/2227-9717/8/9/1123/htm https://www.mdpi.com/2227-9717/8/9/1123.

Picos-Benítez, A.R., Martínez-Vargas, B.L., Duron-Torres, S.M., Brillas, E., Peralta-Hernández, J.M., 2020. The use of artificial intelligence models in the prediction of optimum operational conditions for the treatment of dye wastewaters with similar structural characteristics. Process Saf. Environ. Prot. 143, 36–44. http://dx.doi.org/10.1016/J.PSEP.2020.06.020.

Quaghebeur, W., Torfs, E., Baets, B.D., Nopens, I., 2022. Hybrid differential equations: Integrating mechanistic and data-driven techniques for modelling of water systems. Water Res. 213, 118166. http://dx.doi.org/10.1016/J.WATRES.2022.118166.

Romero-Güiza, M.S., Flotats, X., Asiain-Mira, R., Palatsi, J., 2022. Enhancement of sewage sludge thickening and energy self-sufficiency with advanced process control tools in a full-scale wastewater treatment plant. Water Res. 222, 118924. http://dx.doi.org/10.1016/J.WATRES.2022.118924.

Scutari, M., 2017. Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. J. Stat. Softw. 77 (2), 1–20. http://dx.doi.org/10.18637/jss.v077.i02.

Shen, X., Ma, S., Vemuri, P., Simon, G., Weiner, M.W., Aisen, P., Petersen, R., Jack, C.R., Saykin, A.J., Jagust, W., Trojanowki, J.Q., Toga, A.W., Beckett, L., Green, R.C., Morris, J., Shaw, L.M., Khachaturian, Z., Sorensen, G., Carrillo, M., Kuller, L., Raichle, M., Paul, S., Davies, P., Fillit, H., Hefti, F., Holtzman, D., Mesulam, M.M., Potter, W., Snyder, P., Schwartz, A., Montine, T., Thomas, R.G., Donohue, M., Walter, S., Gessert, D., Sather, T., Jiminez, G., Balasubramanian, A.B., Mason, J., Sim, I., Harvey, D., Bernstein, M., Fox, N., Thompson, P., Schuff, N., DeCArli, C., Borowski, B., Gunter, J., Senjem, M., Jones, D., Kantarci, K., Ward, C., Koeppe, R.A., Foster, N., Reiman, E.M., Chen, K., Mathis, C., Landau, S., Cairns, N.J., Franklin, E., Taylor-Reinwald, L., Lee, V., Korecka, M., Figurski, M., Crawford, K., Neu, S., Foroud, T.M., Potkin, S., Faber, K., Kim, S., Nho, K., Thal, L., Buckholtz, N., Albert, M., Frank, R., Hsiao, J., Kaye, J., Quinn, J., Silbert, L., Lind, B., Carter, R., Dolen, S., Schneider, L.S., Pawluczyk, S., Beccera, M., Teodoro, L., Spann, B.M., Brewer, J., Vanderswag, H., Fleisher, A., Heidebrink, J.L., Lord, J.L., Mason, S.S., Albers, C.S., Knopman, D., Johnson, K., Doody, R.S., Villanueva-Meyer, J., Pavlik, V., Shibley, V., Chowdhury, M., Rountree, S., Dang, M., Stern, Y., Honig, L.S., Bell, K.L., Ances, B., Carroll, M., Creech, M.L., Franklin, E., Mintun, M.A., Schneider, S., Oliver, A., Marson, D., Geldmacher, D., Love, M.N., Griffith, R., Clark, D., Brockington, J., Roberson, E., Grossman, H., Mitsis, E., Shah, R.C., deToledo Morrell, L., Duara, R., Greig-Custo, M.T., Barker, W., Onyike, C., D'Agostino, D., Kielb, S., Sadowski, M., Sheikh, M.O., Ulysse, A., Gaikwad, M., Doraiswamy, P.M., Petrella, J.R., Borges-Neto, S., Wong, T.Z., Coleman, E., Arnold, S.E., Karlawish, J.H., Wolk, D.A., Clark, C.M., Smith, C.D., Jicha, G., Hardy, P., Sinha, P., Oates, E., Conrad, G., Lopez, O.L., Oakley, M.A., Simpson, D.M., Porsteinsson, A.P., Goldstein, B.S., Martin, K., Makino, K.M., Ismail, M.S., Brand, C., Preda, A., Nguyen, D., Womack, K., Mathews, D., Quiceno, M., Levey, A.I., Lah, J.J., Cellar, J.S., Burns, J.M., Swerdlow, R.H., Brooks, W.M., Apostolova, L., Tingus, K., Woo, E., Silverman, D.H., Lu, P.H., Bartzokis, G., Graff-Radford, N.R., Parfitt, F., Poki-Walker, K., Farlow, M.R., Hake, A.M., Matthews, B.R., Brosch, J.R., Herring, S., van Dyck, C.H., Carson, R.E., MacAvoy, M.G., Varma, P., Chertkow, H., Bergman, H., Hosein, C., Black, S., Stefanovic, B., Caldwell, C., Hsiung, G.Y.R., Mudge, B., Sossi, V., Feldman, H., Assaly, M., Finger, E., Pasternack, S., Rachisky, I., Rogers, J., Trost, D., Kertesz, A., Bernick, C., Munic, D., Rogalski, E., Lipowski, K., Weintraub, S., Bonakdarpour, B., Kerwin, D., Wu, C.K., Johnson, N., Sadowsky, C., Villena, T., Turner, R.S., Johnson, K., Reynolds, B., Sperling, R.A., Johnson, K.A., Marshall, G., Yesavage, J., Taylor, J.L., Lane, B., Rosen, A., Tinklenberg, J., Sabbagh, M.N., Belden, C.M., Jacobson, S.A., Sirrel, S.A., Kowall, N., Killiany, R., Budson, A.E., Norbash, A., Johnson, P.L., Obisesan, T.O., Wolday, S., Allard, J., Lerner, A., Ogrocki, P., Tatsuoka, C., Fatica, P., Fletcher, E., Maillard, P., Olichney, J., DeCarli, C., Carmichael, O., Kittur, S., Borrie, M., Lee, T.Y., Bartha, R., Johnson, S., Asthana, S., Carlsson, C.M., Tariot, P., Burke, A., Milliken, A.M., Trncic, N., Fleisher, A., Reeder, S., Bates, V., Capote, H., Rainka, M., Scharre, D.W., Kataki, M., Kelly, B., Zimmerman, E.A., Celmins, D., Brown, A.D., Pearlson, G.D., Blank, K., Anderson, K., Flashman, L.A., Seltzer, M., Hynes, M.L., Santulli, R.B., Sink, K.M., Gordineer, L., Williamson, J.D., Garg, P., Watkins, F., Ott, B.R., Tremont, G., Daiello, L.A., Salloway, S., Malloy, P., Correia, S., Rosen, H.J., Miller, B.L., Perry, D., Mintzer, J., Spicer, K., Bachman, D., Pomara, N., Hernando, R., Sarrael, A., Schultz, S.K., Smith, K.E., Koleva, H., Nam, K.W., Shim, H., Relkin, N., Chaing, G., Lin, M., Ravdin, L., Smith, A., Raj, B.A., Fargher, K., 2020. Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. Sci. Rep. 2020 10:1 10, 1–12. http://dx.doi.org/10.1038/s41598-020-59669-x, URL: https://www.nature.com/articles/s41598-020-59669-x.

Shuman, D., Narang, S., Frossard, P., 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal URL: https://ieeexplore.ieee.org/abstract/document/6494675/, Discussions on grpah spectral theory and graph convolutions. Ideas used for GCGRU arhcitecture..

Siami-Namini, S., Tavakoli, N., Namin, A.S., 2019. The performance of LSTM and BiLSTM in forecasting time series. In: Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019. Institute of Electrical and Electronics Engineers Inc., pp. 3285–3292. http://dx.doi.org/10.1109/BIGDATA47090.2019.9005997.

Spirtes, P., Meek, C., Richardson, T., 1999. An Algorithm for Causal Inference in the Presence of Latent Variables and Selection Bias, vol. 1, MIT Press.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., 2015. LINE: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web. Association for Computing Machinery, Inc, pp. 1067–1077. http://dx.doi.org/10.1145/2736277.2741093.

Thwaites, J., Wandmaker, M., O'Shannessy, A., 2018. Melbourne Water Annual Report. Melbourne Water, Melbourne, URL: www.melbournewater.com.au.

Torregrossa, D., Hansen, J., Hernández-Sancho, F., Cornelissen, A., Schutz, G., Leopold, U., 2017. A data-driven methodology to support pump performance analysis and energy efficiency optimization in waste water treatment plants. Appl. Energy 208, 1430–1440. http://dx.doi.org/10.1016/J.APENERGY.2017.09.012.

Torregrossa, D., Leopold, U., Hernández-Sancho, F., Hansen, J., 2018. Machine learning for energy cost modelling in wastewater treatment plants. J. Environ. Manag. 223, 1061–1067. http://dx.doi.org/10.1016/J.JENVMAN.2018.06.092.

Wang, H., Yang, Y., Keller, A.A., Li, X., Feng, S., nan Dong, Y., Li, F., 2016. Comparative analysis of energy intensity and carbon emissions in wastewater treatment in USA, Germany, China and South Africa. Appl. Energy 184, 873–881. http://dx.doi.org/10.1016/J.APENERGY.2016.07.061.

Wiesmann, U., Choi, I.S., Dombrowski, E.-M., 2007. Fundamentals of Biological Wastewater Treatment. John Wiley & Sons.

Wu, D., Zhao, J., 2021. Process topology convolutional network model for chemical process fault diagnosis. Process Saf. Environ. Prot. 150, 93–109. http://dx.doi.org/10.1016/J.PSEP.2021.03.052, Added non-observed variables as nodes to graph initialised with random value filling..

Xie, Y., Liu, C., Zhou, C., Wei, H., Tao, Y., Zhou, J., 2024. Effects of flow rate and wastewater concentration on the transformation of nitrogen in sediment–water system of sewage pipelines. Water Environ. Res. 96, e10976. http://dx.doi.org/10.1002/WER.10976, https://onlinelibrary.wiley.com/doi/full/10.1002/wer.10976, https://onlinelibrary.wiley.com/doi/abs/10.1002/wer.10976 https://onlinelibrary.wiley.com/doi/10.1002/wer.10976.

Xie, C., Yao, R., Liu, Z., Zhu, L., Chen, X., 2022. Process performance prediction based on spatial and temporal feature extraction through bidirectional LSTM. 49, Elsevier, pp. 1615–1620. http://dx.doi.org/10.1016/B978-0-323-85159-6.50269-4,

Zhang, Q., Li, Z., Snowling, S., Siam, A., El-Dakhakhni, W., 2019. Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. Water Sci. Technol. 80, 243–253. http://dx.doi.org/10.2166/WST.2019.263, URL: http://iwaponline.com/wst/article-pdf/80/2/243/621330/wst080020243.pdf.

Zhang, S., Wang, H., Keller, A.A., 2021. Novel machine learning-based energy consumption model of wastewater treatment plants. ACS EST Water 1, 2531–2540. http://dx.doi.org/10.1021/acsestwater.1c00283.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R., 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In: International Conference on Machine Learning. PMLR, pp. 27268–27286.

Zhou, H., Xu, G., 2019. Integrated effects of temperature and COD/N on an up-flow anaerobic filter-biological aerated filter: Performance, biofilm characteristics and microbial community. Bioresour. Technol. 293, 122004. http://dx.doi.org/10.1016/J.BIORTECH.2019.122004.