Activated sludge models at the crossroad of artificial intelligence—A perspective on advancing process modeling

Sin, Gürkan; Al, Resul

Published in:
npj Clean Water

Link to article, DOI:
10.1038/s41545-021-00106-5

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

npj | Clean Water

www.nature.com/npjcleanwater

PERSPECTIVE OPEN

Check for updates

Activated sludge models at the crossroad of artificial intelligence—A perspective on advancing process modeling

Gürkan Sin[1] ✉ and Resul Al[1]

The introduction of Activated Sludge Models No. 1 (ASM1) in the early 1980s has led to a decade-long experience in applying these models and demonstrating their maturity for the wastewater treatment plants' design and operation. However, these models have reached their limits concerning complexity and application accuracy. A case in point is that despite many extensions of the ASMs proposed to describe N_2O production dynamics in the activated sludge plants, these models remain too complicated and yet to be validated. This perspective paper presents a new vision to advance process modeling by explicitly integrating the information about the microbial community as measured by molecular data in activated sludge models. In this new research area, we propose to harness the synergy between the rich molecular data from advanced gene sequencing technology with its integration through artificial intelligence with process engineering models. This is an interdisciplinary research area enabling the two separate disciplines, namely environmental biotechnology, to join forces and work together with the modeling and engineering community to perform new understanding and model-based engineering for sustainable WWTPs of the future.

npj Clean Water (2021)4:16 ; https://doi.org/10.1038/s41545-021-00106-5

INTRODUCTION

Wastewater treatment is a complex process that employs a combination of physical, chemical, and biological unit operations to remove contaminants to sufficient quality before being discharged into the receiving environment. WWTPs have been around for over 100 years since the discovery of the activated sludge process[1,2], which resulted in many mature technologies and process concepts implemented in practice. Today the wastewater treatment sector is witnessing a growing number of initiatives (e.g., Digital Water, Water-Energy nexus, Circular economy, water scarcity and deteriorating water quality due to emerging contaminants such as micro-pollutants, climate change)[3–6]. These powerful initiatives are set to radically change the baseline concept of WWTPs, e.g. wastewater is no longer being perceived as a problem but increasingly as a potential resource to recover water, energy, and nutrients. Moreover, increasing sustainability awareness for more efficient use of energy, chemicals, and process-related greenhouse gas emissions (especially N_2O) in WWTPs, need to be considered. Currently, the design and operation of treatment plants rely on best practice and heuristic approaches, which are supplemented by using process models to simulate and evaluate a number of alternatives. In this regard, the introduction of Activated Sludge Models No. 1 (ASM1) in the early 1980s has led to a decade-long experience with calibrating and applying the models and demonstrating their maturity for application in the design and operation of the plants[1]. However, these models have reached their limits with respect to complexity and application accuracy and unable to comprehensively describe process performance parameters[2]. This is essential to realize model-based engineering and, consequently, the full potential of digitalization for achieving a sustainable WWTP operation. Therefore, we believe a steep change of foundational nature in advancing process modeling is needed in the wastewater treatment modeling and engineering community. The central hypothesis of this new vision is based on the following premise: (1) we have a strong conviction that data alone may not contain sufficient information to achieve useful models for digital applications. (2) Current mechanistic models alone are unable to describe newly emerging sustainability concerns of the plants, especially N_2O dynamics, among others. Indeed unlike social sciences/media where data is highly rich (high volume/high veracity)[7,8], data from engineering systems like WWTPs, which are designed and operated to deliver a steady and stable performance, has limited information (quality and quantity) compared to the scale of the volume of data in social media. Therefore, we need to make full use of prior scientific & engineering knowledge as nicely summarized in mechanistic models. Hence a multi-disciplinary approach where deep process knowledge is combined with deep learning from process data, is needed to generate advanced predictive models for digital applications in WWTPs. With this research concept, we propose, it will be possible for the first time to directly include molecular data about the microbiological community into the process models. Modern molecular tools measuring microbial community (relative abundance of species, their functions, using metagenomics and meta-transcriptomics analysis) have advanced very much and enabled identifying diverse microbial communities underpinning the biological transformation in the wastewater treatment plants from filamentous bulking phenomena to nitrogen removal and phosphorus removal, among others[9–11]. However, these valuable data have never been used directly in process modeling. Below we outline and expand the evidence that points out to current challenges and limitations of activated sludge models (ASMs), presents arguments why big-data analytics alone will not deliver, and the need for an interdisciplinary research area to advance the future of activated sludge modeling.

[1]Department of Chemical and Biochemical Engineering, Process and Systems Engineering Centre (PROSYS), Technical University of Denmark, Copenhagen, Denmark. ✉email: gsi@kt.dtu.dk

Published in partnership with King Fahd University of Petroleum & Minerals

npj nature partner journals

SCIENTIFIC PERSPECTIVES

The activated sludge process is the most commonly used technology in wastewater treatment plants. Research into the design and operation of activated sludge systems are supported traditionally by two communities: (a) on the one hand, the environmental biotechnology[10] community with the advances in gene sequencing technology and molecular probes (such as qPCR, FISH, phylogenetic trees, operational taxonomic units (OTUs), meta-transcriptomics, proteomics, metabolomics, etc.) studies underpinning microbiological community (who are they and what they do) responsible for nutrient removal in the plant, and (b) process design, modeling and control community[12–16] that works with process models and uses traditional process data (such as influent COD their fractions (SS, XS, SI, and XI), nitrogen (NH_4, NO_3, NO_2-N, TKN), PO_4, TP, MLSS, VSS, BOD, etc.) to develop and validate new models and techniques to support design and operation of the plants. While clearly these multidimensional data collected at different scales (micro-scale at microbial community level versus process data at macro/full-scale at plant level) are complementary to understand the process, to date these two multiscale and diverse range of process data is yet to be integrated and interpreted jointly. For example, while microbial community underpins the biological conversions in WWTPs and therefore key to the performance of the plant (from effluent nitrogen, phosphorus, and COD quality to process-related GHG emissions), they have not been explicitly/directly included in the models. On the other hand, the in-situ techniques to identify microbial community structure and functions in WWTPs have advanced and matured very nicely in the last decades in parallel to advances in gene sequencing technology[17]. These advancements resulted in more in-depth insights into understanding the fundamental role and functions of microbial community both at the laboratory but also at full-scale WWTPs[18] when studying novel processes from anammox to understanding pathways responsible for N_2O emissions. However, these valuable data have not been integrated into process engineering applications, which remains a major gap still in the 2020s. One of the reasons is that the currently used modeling framework is not flexible to integrate such heterogeneous molecular data about different microbial communities.

ASMS AS ENABLING TECHNOLOGY FOR THE DESIGN OF WWTPS

The current process design and operation paradigm is highly process-expert knowledge-driven and supported by commercial process simulators, which allows evaluating and simulating a range of process configurations. Indeed to support engineering solutions currently, process models and simulations are indispensable tools widely used by the community. For process modeling, ASMs have been widely successful and extensively used. These models (e.g., ADM1, ASM1, ASM2d, Biowin® model, SUMO® model, etc.) are mechanistic and have yielded significant benefits to design and operation problems. Moreover, there are knowledge-based, and model-based environmental decision support tools have been proposed[19–22] to assist design engineers for process design/retrofitting to improved operation. However, there are still two fundamental shortcomings: (1) limit of current mechanistic models: The process models that underpin evaluations of WWTP solutions are not able to describe important sustainability metrics/performance of the plant, namely the N_2O emissions[2,19,23,24], (2) while the biological community is responsible for the major transformation of contaminants and their removal, these are not included/integrated in the process engineering practice from operation to design tasks.

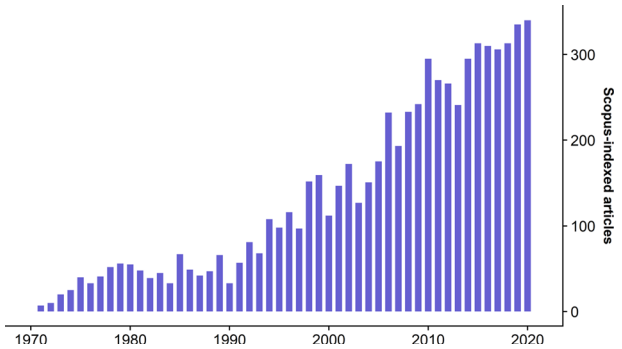

Fig. 1 Activated sludge modeling studies in literature. The data generated using Scopus database with the following specifications: Search Date: February 2021. TITLE-ABS-KEY ("activated sludge" AND "model").

Validation and complexity issues of ASMs

In particular, in the process modeling field, the current academic research is generating more and more complex and specialized models[1,25,26]. Figure 1 shows a continued interest in studies using these models. These models, which are mostly based on an extension of ASM models to describe nitrogen, phosphorus, and COD removal performance from the plant, are not fit for process design and operation applications due to numerical complexities and validation issues. The validation issue comes partly from the available process data used for model parameter estimation (which is limited) as well as the given structure of the model. Several studies have systematically analyzed the identifiability of such models[1,27,28] including our own studies, considering typical plant data collected from intensive measurement campaigns, which showed that among 60 plus model parameters, only a handful of them (6–10 parameter subsets) could be uniquely estimated from the data. These issues has been recognized already by calibration protocols in fact[29,30]. The rest of the model parameters need to be fixed or assumed when applying these models to simulate the activated sludge plants. While one can account for the uncertainty in the model parameters and perform design and operation evaluation[31–33], however, the key issues remain with respect to defining the range of uncertainties for the model parameters, which displays a wide variability as examined by Sin et al.[34] for nitrite models.

One of the underlying reasons for this model validation and uncertainty issues is that these models employ a Monod kinetics to describe microbial growth. Theoretical identifiability studies[35–37], performed as early as 1982 by Holmberg[35], already found out that given perfect measurements (no noise) on a simple batch reactor used to measure biomass activity (e.g., substrate measurements in time), even then the unique estimation of the yield, maximum growth rate, and biomass concentration is not possible. Later on, Petersen et al.[37] used respirometric and titrimetric measurements of the activity of nitrifying activated sludge samples confirmed the same conclusion. Namely, instead of unique parameters, only a combination of the parameters is uniquely identifiable, for example ((4.57−YA)/YA*μmax*X). An important observation about this conclusion is that X (mgCOD/L) is a lumped parameter defined to represent an active fraction of the microbial group involved in the experiment. For example, in nitrifying activity studies, X would be classified as X_{AOO} and X_{NOO}, referring to ammonia-oxidizing organisms and nitrite-oxidizing organisms, respectively. These examples can be extended to other microbial groups in activated sludge, e.g., denitrifying heterotrophic organisms, phosphorus-accumulating organisms, glycogen accumulation organisms, all hypothetically modeled with a unit mg COD/L as a state variable. In the end, these different fractions of the biomass will be inferred indirectly from a set of corresponding batch activity tests or using

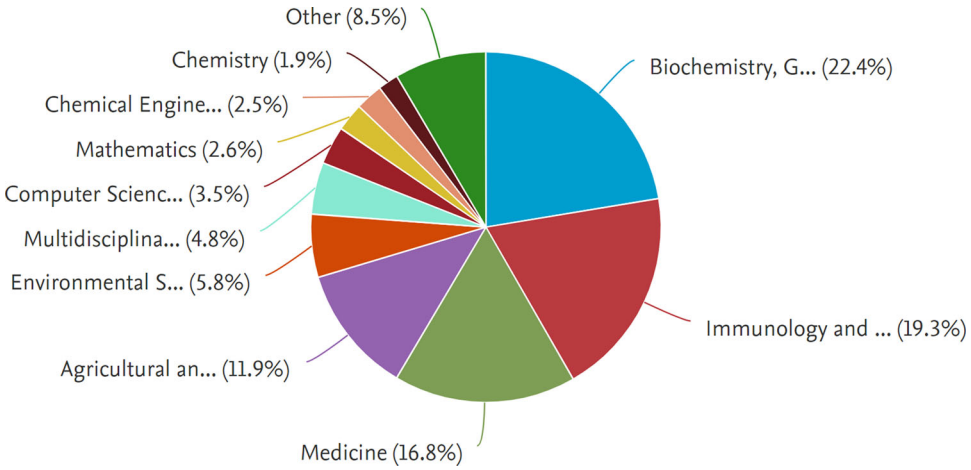

Fig. 2 Metagenomics studies by subject area. The data generated using Scopus database with the following specifications: Search Date: February 2021. TITLE-ABS-KEY ("metagenomics").

full-scale measurements from steady-simulations with the model. Notice the irony here that a biological community is represented by a pseudo-parameter in the model, which has no direct way of measuring it in reality. Therefore, there is no independent experimental procedure to verify the simulated values of these fractions of biomass responsible for different functions in the plant without making assumptions and conversion factors (e.g. VSS to COD ratio, etc.). Instead, these fractions of biomass are inferred indirectly from model-based fitting to measured activity of the biomass (e.g., through depletion of NH_4-N rate during a batch test with nitrifying activated sludge). It turns out even if you had perfect activity measurements, the estimated values of these fractions are still coupled to yield and maximum growth rate parameters in the model (as discussed above by Holmberg[35] and Petersen et al.[37]). While the modeling community uses *X* and attempts to describe corresponding activity in the WWTPs, the biological community that studies this process uses modern molecular probe technologies (e.g., metagenomics, qPCR, FISH, etc.) to identify which organisms (phylogenetic tree), their relative abundance and their activity (e.g., expression of protein genes in meta-transcriptomics analysis) that metabolizes many of the pollutants present in the influent (from NH_4-N to COD and others).

The rise of metagenomics data and what to do about it

The research field of metagenomics investigates the genomic analysis of microbial DNA from environmental communities, and it has become one of the hottest fields of science by rapidly growing over the last 5–10 years—more than 16,000 research papers indexed in Scopus (Figs. 1 and 2)—yielding substantial advances in microbial ecology, evolution, and diversity[17]. The field provides scientists with sequencing-based metagenome examination tools that enable them to identify microbes in a sample without a priori knowledge of what that sample contains—opening up new doors in many disciplines, such as medicine, environmental sciences, microbial ecology, microbiology, and wastewater engineering (Fig. 2). Among these techniques is fluorescence in situ hybridization (FISH), which can be used to find specific genes of interest in DNA so as to identify specific microorganisms, but it is a low-throughput technology[9]. Quantitative PCR (qPCR) is sensitive and quantitative, and it monitors the amplification of a targeted DNA using fluorescent dyes. Therefore, it can only evaluate a few microorganisms at a time. On the other hand, 16S ribosomal RNA sequencing targets 16S rRNA genes that are highly species-specific and present in most microbes; thus it is a rapid and cheap alternative capable of both identifying and classifying bacteria. In addition to sequencing, the characterization of meta-transcriptomes (using mRNA sequencing) provides an ability to reflect the actual metabolic activity by differentiating between expressed and non-expressed genes, thereby overcoming the shortcomings of metagenomic DNA-based analyses. It comes with a higher price; however, the obtained data is more information-rich, allowing modeling of the quantified metabolic activities, which might explain process phenomena driven by such microbial communities.

In wastewater treatment, Seviour et al.[10] pioneered the application of these methods to activated sludge systems, inviting the research community to understand which organisms are present in activated sludge and what they might be doing there. Saunders et al.[18] identified a core community of microorganisms actively present in activated sludge using 16S rRNA gene sequencing technique along with microbial diversity analysis. By applying these methods, highly comprehensive information about this centuries-old activated sludge process is produced. Incorporating this new knowledge of the frequency and diversity of these microbial communities, as well as their spatial and time-dependent dynamic profiles, can support quantative modeling of the underlying phenomena in wastewater treatment processes such as the N_2O emissions. One potential way of integrating such molecular data (metagenomics) with ODE-based mathematical models is through leveraging artificial intelligence techniques such as deep learning. Recent applications of DNNs in diverse fields[38] from design of products/materials to property and process modeling, have shown that neural AI excels particularly when presented with diverse range/heterogenous source of data in the form of textual, image, spectral data[38–40].

Moving beyond: neither ASM nor AI models alone

In this new research area, we call for the study and synergistic integration of biological data through the machine-learning (ML) branch of AI with first-principles models of activated sludge systems. Indeed, it is duly noted that what we propose here is not just hybrid modeling, which in itself is nothing new and has been extensively studied for a variety of applications. For example, hybrid modeling in chemical engineering (crystallization, drying, milling, polymerization) and biochemical engineering (mainly different fermentation process modeling from fungi to bacteria, yeast, and mammalian cell culture) as well as water treatment[41]. The motivation for the earlier hybrid modeling is to improve the predictions of the first principles models, hence correct for the error/uncertainties present in the mass and energy balances as calculated by mechanistic models. A variety of parametric hybrid model designs are proposed, e.g., parallel, serial, and multiple

combinations[42,43]. Often the hybrid models are also used to predict complex process phenomena, which is otherwise very difficult to mechanistically describe, e.g., cake formation on the cross-filtration unit or formation rates/kinetics of products in fermentation processes. In wastewater, hybrid model applications have also been studied both for industrial[44,45] and domestic wastewater-treatment plants. For example, Anderson et al.[46] has integrated an ANN model to learn from the biological kinetic rates from the process data in the mechanistic (ASM2d) model using a parallel combination, while Fang et al.[47] used a serial combination in which the authors modeled the errors from a mechanistic model (ASM3g-Eawag) with a neural network model to improve predictions of effluent COD, NH_4, and PO_4^{-3} similar to the application of extended Kalman filter (EKF to learn from the errors of the model[8,48]). While these models presented the potential for improving the fit to the data (as measured by the model's R^2, coefficient of determination), however their application, for example, for process control and operation, proved to be problematic as demonstrated in Anderson et al.[46]. More importantly, they also failed to model, for example, kinetic rates related to more complex phenomena such as P-removal[46]. These previous studies show that just using process data alone and even hybrid modeling is not the answer by itself. We need to have comprehensive process data to describe it.

As regards modeling N_2O, which is the example we cite in this perspective paper, a number of extensions of ASM models with several different mechanisms (e.g., single pathway versus two pathway models—AOB denitrification pathway and incomplete hydroxylamine oxidation pathway—models, chemical conversion, etc.) have been proposed[49–54]. These models, thanks to the many parameters introduced, can be made to fit the N_2O data collected from a certain period during calibration (i.e., by fine-tuning or fitting a subset of model parameters). However, these models become easily falsified when confronted with another dataset not used in the calibration, e.g., underestimating N_2O emission rate by as much as 42% in Mampaey et al.[55], to name an example. In Fig. 3, we present a schematics to help visualization of model performance between calibration and validation datasets as demonstrated in the study of Mampaey et al.[55] as well as the importance of which pathway is modeled/considered in the model as demonstrated in Pocquet et al.[56] study. Indeed the current mathematical models do not a priori know which metabolic pathway is dominant contributing factor and therefore can fail to quantitatively describe the N_2O emission factors in plants. It is noted that most municipal wastewater treatment plants operate at lower NO_2 levels (much lower than 30 mgN/L) where especially the failure of the models is pronounced. In a way, the extended models just increased the number of parameters that need to be estimated from the same activity measurements, hence compounding the existing identifiability issues of these models. This makes transferability and general applicability for process design and operation difficult and not possible as they are not predictive. On the other hand, using ML of the process data has also been applied, which demonstrated that PCA-based clustering techniques[57] could be used for identifying operational situations causing N_2O emissions. Using support vector machines (SVM) as a ML technique, Vasilaki et al.[58] demonstrated that N_2O emissions could be described, although the cross-validation R^2 remains low even for describing a dataset from a relatively simple pilot-scale reactor. In our own work using a DNN[59], we have also demonstrated the potential of describing N_2O using a deep learning network (DNN) with high accuracy, R^2 up to 0.9 in cross-validation test data. While these models are useful for performing sensitivity analysis on the inputs, however, the main challenge here is that these purely data-driven models are not useful for process design and operation purposes. Simply on account of these data-driven models fails again to predict changing/seasonal variations in the N_2O emissions. In short, we argue that neither mechanistic nor AI (ML) models alone are able to predictively describe N_2O data.

Proposal of a solution: A multi-disciplinary research field to advance process modeling in WWTPs

Therefore, our opinion is that neither data alone nor the current mechanistic models are sufficient to develop predictive models for emerging sustainability concerns, as argued for N_2O emissions. It is our belief that these models fail to predictively describe the system due to the lack of incorporation of data directly related to the composition and activity of the microbial community. One possible strategy to solve this issue is to describe N_2O emissions with the help of ML models processing biological data (e.g., metagenomics) as input and other relevant process data (mass balances for NO_3, NO_2, and NH_4) through mechanistic process models. This strategy is sketched in Fig. 4. Here we emphasize the need to study different AI techniques to extract information from such untraditional data source, for example, to parameterize the gene sequence data and forward this as input to ML models (such as DNN (forward neural network), CNNs, and GANs, among others). Much research in AI techniques and graph theory has shown its potential in extracting information from 2-D and 3-D chemical structure (i.e., feature selection) and process in DNN to predict some property of interest (e.g., the biodegradability of different chemical compounds) or in synthesizing new materials such as zeolites. Indeed, an ambitious research effort is needed to develop such new AI-based techniques. Here we call for a community-wide and interdisciplinary collaboration to address several open and fundamental questions on how to achieve this thoughtful fusion and integration of data sources and knowledge competencies. Foremost, does integrating biological data (such as metagenomics, meta-transcriptomics, qPCR, FISH, etc.) through ML with mechanistic models help achieve predictive performance (not just calibration/training data but test data)? What is the optimal design of the hybrid integration scheme (parallel versus serial, multiplicative versus embedded combination in which rates of formation of N_2O, the active fraction of biomass of different groups at the genus level is linked with mechanistic models for mass balances, etc.), what is the efficient integration of data, ML and mechanistic models for digital applications, among many others? Which particular metagenomics data is useful for which modeling purpose? Should we use metagenomics or metabolomics (protein expression data), and for which modeling purpose?

Data extracted from molecular probe techniques are highly heterogeneous and expensive to gather, usually resulting in much smaller datasets than those available for other ML tasks. Such datasets often require featurization, alternatively defined as feature engineering which is a process of using domain knowledge of the data to create features that help ML algorithms to learn better. In our proposed vision for hybrid modeling in Fig. 4, therefore, the feature extraction will be the key step that needs to be researched and developed to extract useful and related features to transform molecular/metagenomic data from activated sludge plants into a form suitable for current machine/deep learning algorithms. In the wider literature, feature extraction techniques are fast developing and important field which have yielded several successful techniques already, such as extended-connectivity fingerprints, Coulomb matrix, weave featurization, and graph convolutions. Depending on the chosen featurization, different types of ML models are also proposed for molecular datasets, such as message passing neural networks (MPNN), deep tensor neural networks (DTNN), directed acyclic graphs (DAG), graph convolutional networks. More details of these different molecular featurizations and models can be found elsewhere[29]. The graph-based featurizations and neural networks have recently gained significant research attention in cheminformatics and bioinformatics due to their superior performances on molecular

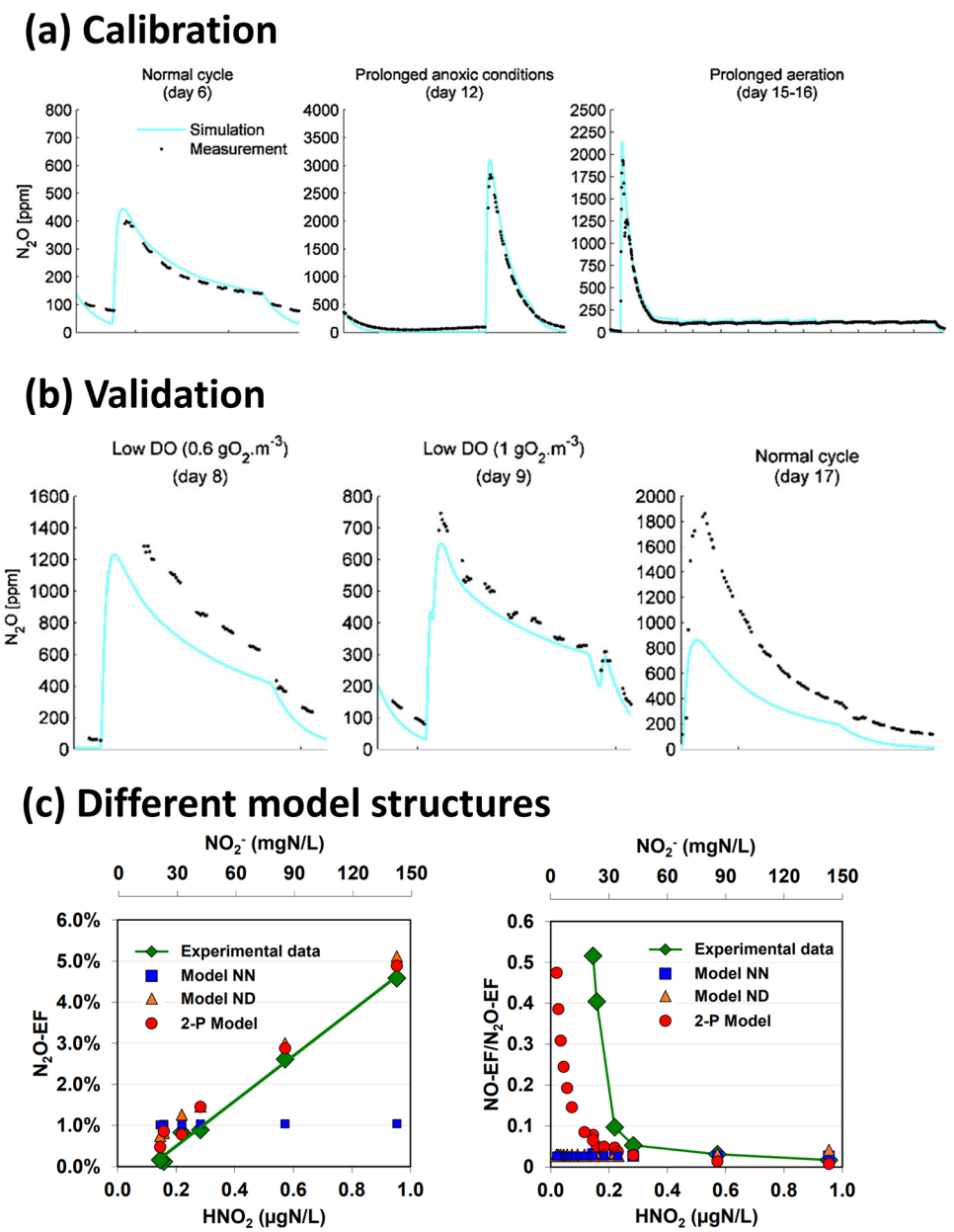

Fig. 3 Performance of some N_2O models in literature. Illustrations in **a** and **b** are from Mampaey et al.[55] that shows a two pathway N_2O model application to describe process data from a full-scale SHARON reactor: (top) the calibration results for gas-phase N_2O measurements using a calibration dataset and (middle) the model validation results from using a validation dataset. The illustration in **c** is from Pocquet et al.[56], presenting a comparison of single pathways versus two-pathway models to describe N_2O and NO emission factors from a lab-scale SBR study. It is noted that most municipal wastewater treatment plants work at lower NO_2 levels (lower than 30 mgN/L), where model deviations from the measurements are significant. Figures reused with permissions from Mampaey et al.[55] copyright (Elsevier, 2013), and Pocquet et al.[56] copyright (Elsevier, 2016), respectively.

ML tasks, as found in recent literature[29,30,60]. For instance, a spatial graph-based molecular representation of amino acid residue pairs has allowed AlfaFold[61], an AI program developed by DeepMind, to perform 3D protein structure predictions far more accurately than ever made. Similarly, the interactions and the metabolic functions of microorganisms present in activated sludge systems and how they affect treatment plant performance metrics, such as N_2O emissions can be studied using graph-based featurizations extracted from their metagenomics data and thereby used in, including but not limited to, convolutional type neural networks. Such networks have already found uses to classify metagenomics data using patristic distance defined on the phylogenetic tree as a proximity measure[62]. Therefore these models can be explored to establish a link between microorganisms' cellular level metabolic activities and the impact measured at the process plant level. Such an effort requires combining domain specific knowledge from activated sludge microbiology, e.g. genetic sequence for key enzymes involved in the metabolism of different N_2O production pathways (e.g. nitrite reduction to NO (mediated by NirK enzyme), reduction of NO to N_2O (mediated by Nor enzyme), etc.) to formulate/define unique features for the graph neural network models (GCNN) and extract relevant features/information as input to the hybrid activated sludge model concept given in Fig. 4. Precisely this step is the bridge to connect the domain knowledge and expertise of environmental biotechnology and its genomic data with process engineering/modeling community and their

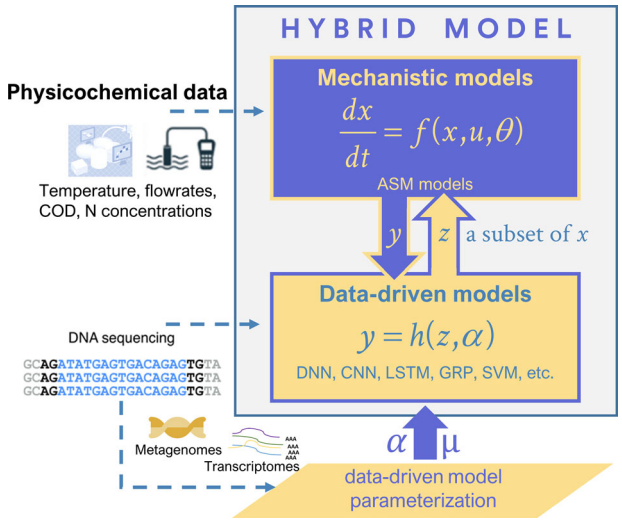

Fig. 4 Integration of physicochemical and molecular sequence data in hybrid modeling. A proposal to integrate explicitly biological data in process models.

models. In addition to offering a flexible modeling framework to include genomic data, this neural-based modeling approach address an important limitation of the current models. Indeed the main drawback of the current models is that they assume the metabolic pathways a priori and formulate a corresponding model structure, which is fixed and applied to all wastewater treatment plants by calibrating its model parameters. While in the new hybrid modeling approach, using metagenomics data one can extract features that can inform about which pathways are actually present/active and contributing to dominant microbial activity such as N_2O phenomena, which helps the model to be valid for each wastewater treatment plant applied and aligned with underlying microbial community composition and changes. Future research will therefore be needed to develop customized neural AI methodologies tailored for the needs and domain of metagenomics data for use in activated sludge modeling field.

Given these recent advances and successful applications of GCNN in chemical property prediction to generative adversial networks (GANs) in materials design and GANs/reinforcement learning for drug discovery in the wider literature, there is a rich and ample ground to explore these and many questions in such an interdisciplinary field to establish the foundation of a new research area. This presents a rich intellectual basis for improving modeling (dynamic, steady-state, and meta/surrogate models) and develop new model-based digital applications, especially for the sustainable operation of WWTP[63]. Certainly, there is a need to join forces with fellow scientists from environmental biotechnology (high throughput gene sequencing technology), wastewater engineering and modeling community, as well as computer science applications for AI/big data analytics.

CONCLUDING

ASMs have been an invaluable tool to help conceive, design and operate many wastewater treatment plants. We argue that the time is opportune to take the field to its next step by leveraging a multi-disciplinary research approach: combining emerging AI techniques to extract feature and information from non-traditional heterogeneous sources of data, as well as increasing availability and diversity of big data, especially metagenomics data which is hitherto never been used in the process modeling. However, much research is needed to exploit the ML approach for integrating biological molecular data. Thanks to this enhanced biological phenomena-based approach, we anticipate that the new research area will be able to generate previously unknown design, operation, and control solutions to meet the increasing demand from WWTPs: climate change neutrality, sustainability, circularity, etc.

Received: 23 November 2020; Accepted: 11 February 2021;
Published online: 08 March 2021

REFERENCES

1. Gujer, W. Activated sludge modelling: past, present and future. *Water Sci. Technol.* **53**, 111–119 (2006).
2. Eddy, M. et al. *Wastewater Engineering: Treatment and Resource Recovery* (McGraw Hill Education; 2014).
3. OECD. *OECD Environmental Outlook to 2050* (OECD, 2012).
4. Water Europe. *Water Europe's Vision "The Value of Water"*. https://watereurope.eu/wp-content/uploads/2020/04/WE-Water-Vision-english_online.pdf (2018).
5. Dutch Foundation for Applied Water Research (STOWA), *Dutch roadmap for WWTP of 2030*. https://www.stowa.nl/publicaties/news-dutch-roadmap-wwtp-2030 (2010).
6. IWA. *Digital Water*. https://iwa-network.org/wp-content/uploads/2019/06/IWA_2019_Digital_Water_Report.pdf (2019).
7. Venkatasubramanian, V. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE J.* **65**, 466–478 (2019).
8. Qin, S. J. & Chiang, L. H. Advances and opportunities in machine learning for process data analytics. *Comput. Chem. Eng.* **126**, 465–473 (2019).
9. Xu, W. et al. Community members in activated sludge as determined by molecular probe technology. *Water Res.* **168**, 115104 (2020).
10. Seviour, R., Halkjær, P. & Nielsen, R. *Microbial Ecology of Activated Sludge*, Vol. 9 (Water Intelligence Online, IWA Publishing, 2010).
11. Nielsen, P. H., Kragelund, C., Seviour, R. J. & Nielsen, J. L. Identity and ecophysiology of filamentous bacteria in activated sludge. *FEMS Microbiol. Rev.* **33**, 969–998 (2009).
12. Brdjanovic, D., Meijer, S. C. F., Lopez-Vazquez, C. M., Hooijmans, C. M. & van Loosdrecht, M. C. M. *Applications of Activated Sludge Models* (Iwa Publishing, 2015).
13. Rieger, L. et al. *Guidelines for Using Activated Sludge Models* (IWA Publishing; 2012).
14. IWA. *Benchmarking*. http://iwa-mia.org/benchmarking/ (2020).
15. IWA. *Task Group on Physicochemical Framework*. https://iwa-connect.org/group/task-group-on-generalized-physicochemical-framework (2020).
16. IWA. *Task Group on Design and Operations Uncertainty*. https://iwa-connect.org/group/task-group-on-design-and-operations-uncertainty-dout (2020).
17. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics—a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**, 3 (2012).
18. Saunders, A. M., Albertsen, M., Vollertsen, J. & Nielsen, P. H. The activated sludge ecosystem contains a core community of abundant organisms. *ISME J.* **10**, 11–20 (2016).
19. Bozkurt, H., Quaglia, A., Gernaey, K. V. & Sin, G. A mathematical programming framework for early stage design of wastewater treatment plants. *Environ. Model. Softw.* **64**, 164–176 (2015).
20. Castillo, A. et al. An integrated knowledge-based and optimization tool for the sustainable selection of wastewater treatment process concepts. *Environ. Model. Softw.* **84**, 177–192 (2016).
21. Rodriguez-Garcia, G. et al. Environmental and economic profile of six typologies of wastewater treatment plants. *Water Res.* **45**, 5997–6010 (2011).
22. Molinos-Senante, M., Hernández-Sancho, F., Sala-Garrido, R. & Cirelli, G. Economic feasibility study for intensive and extensive wastewater treatment considering greenhouse gases emissions. *J. Environ. Manag.* **123**, 98–104 (2013).
23. Daigger, G. T. & Crawford, G. V. Wastewater treatment plant of the future—Decision analysis approach for increased sustainability. In (eds van Loosdrecht, M. C. M. & Clement, J.), *2nd IWA Leading-Edge Conference on Water and Wastewater Treatment Technology, Water and Environment Management Series*. 361–369 (IWA Publishing, 2005).
24. Garrido-Baserba, M., Reif, R., Rodriguez-Roda, I. & Poch, M. A knowledge management methodology for the integrated assessment of WWTP configurations during conceptual design. *Water Sci. Technol.* **66**, 165–172 (2012).
25. Flores-Alsina, X. et al. Modelling phosphorus (P), sulfur (S) and iron (Fe) interactions for dynamic simulations of anaerobic digestion processes. *Water Res.* **95**, 370–382 (2016).
26. Henze, M., Gujer, W., Mino, T. & van Loosdrecht, M. C. M. *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*, Vol. 121 (IWA Publishing, 2000).

27. Brun, R., Kühni, M., Siegrist, H., Gujer, W. & Reichert, P. Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets. *Water Res.* **36**, 4113–4127 (2002).
28. Sin, G., Vanhulle, S., Depauw, D., Vangriensven, A. & Vanrolleghem, P. A critical comparison of systematic calibration protocols for activated sludge models: a SWOT analysis. *Water Res.* **39**, 2459–2474 (2005).
29. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
30. Peng, Y. et al. Enhanced graph isomorphism network for molecular ADMET properties prediction. *IEEE Access* **8**, 168344–168360 (2020).
31. Sin, G., Gernaey, K. V., Neumann, M. B., van Loosdrecht, M. C. M. & Gujer, W. Uncertainty analysis in WWTP model applications: a critical discussion using an example from design. *Water Res.* **43**, 2894–2906 (2009).
32. Sin, G., Gernaey, K. V., Neumann, M. B., van Loosdrecht, M. C. M. & Gujer, W. Global sensitivity analysis in wastewater treatment plant model applications: prioritizing sources of uncertainty. *Water Res.* **45**, 639–651 (2011).
33. Flores-Alsina, X., Rodriguez-Roda, I., Sin, G. & Gernaey, K. V. Uncertainty and sensitivity analysis of control strategies using the benchmark simulation model No1 (BSM1). *Water Sci. Technol.* **59**, 491–499 (2009).
34. Sin, G. et al. Modelling nitrite in wastewater treatment systems: a discussion of different modelling concepts. *Water Sci. Technol.* **58**, 1155–1171 (2008).
35. Holmberg, A. On the practical identifiability of microbial growth models incorporating Michaelis–Menten type nonlinearities. *Math. Biosci.* **62**, 23–43 (1982).
36. Chappell, M. J. & Godfrey, K. R. Structural identifiability of the parameters of a nonlinear batch reactor model. *Math. Biosci.* **108**, 241–251 (1992).
37. Petersen, B., Gernaey, K., Devisscher, M., Dochain, D. & Vanrolleghem, P. A. A simplified method to assess structurally identifiable parameters in Monod-based activated sludge models. *Water Res.* **37**, 2893–2904 (2003).
38. Chui, M. et al. *Notes from the AI Frontier: Applications and Value of Deep Learning* McKinsey Global Institute Discussion Paper. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning# (2018).
39. Babi, D. K. et al. Sustainable process synthesis–intensification. *Comput. Chem. Eng.* **81**, 218–244 (2015).
40. Hwangbo, S. & Sin, G. Design of control framework based on deep reinforcement learning and Monte–Carlo sampling in downstream separation. *Comput. Chem. Eng.* **140**, 106910 (2020).
41. Conlin, J., Peel, C. & Montague, G. A. Modelling pressure drop in water treatment. *Artif. Intell. Eng.* **11**, 393–400 (1997).
42. von Stosch, M., Oliveira, R., Peres, J. & Feyo de Azevedo, S. Hybrid semi-parametric modeling in process systems engineering: past, present and future. *Comput. Chem. Eng.* **60**, 86–101 (2014).
43. Bikmukhametov, T. & Jäschke, J. Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Comput. Chem. Eng.* **138**, 106834 (2020).
44. Lee, D. S., Jeon, C. O., Park, J. M. & Chang, K. S. Hybrid neural network modeling of a full-scale industrial wastewater treatment process. *Biotechnol. Bioeng.* **78**, 670–682 (2002).
45. Lee, D. S., Vanrolleghem, P. A. & Park, J. M. Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *J. Biotechnol.* **115**, 317–328 (2005).
46. Anderson, J. S., McAvoy, T. J. & Hao, O. J. Use of hybrid models in wastewater systems. *Ind. Eng. Chem. Res.* **39**, 1694–1704 (2000).
47. Fang, F. et al. An integrated dynamic model for simulating a full-scale municipal wastewater treatment plant under fluctuating conditions. *Chem. Eng. J.* **160**, 522–529 (2010).
48. Prunescu, R. M., Blanke, M., Jakobsen, J. G. & Sin, G. Dynamic modeling and validation of a biomass hydrothermal pretreatment process—a demonstration scale study. *AIChE J.* **61**, 4235–4250 (2015).
49. Boiocchi, R., Gernaey, K. V. & Sin, G. A novel fuzzy-logic control strategy minimizing N2O emissions. *Water Res.* **123**, 479–494 (2017).
50. Spérandio, M. et al. Evaluation of different nitrous oxide production models with four continuous long-term wastewater treatment process data series. *Bioprocess Biosyst. Eng.* **39**, 493–510 (2016).
51. Boiocchi, R., Gernaey, K. V. & Sin, G. Understanding N2O formation mechanisms through sensitivity analyses using a plant-wide benchmark simulation model. *Chem. Eng. J.* **317**, 935–951 (2017).
52. Mampaey, K. E., Spérandio, M., van Loosdrecht, M. C. M. & Volcke, E. I. P. Dynamic simulation of N2O emissions from a full-scale partial nitritation reactor. *Biochem. Eng. J.* **152**, 107356 (2019).
53. Ni, B.-J., Ye, L., Law, Y., Byers, C. & Yuan, Z. Mathematical modeling of nitrous oxide (N2O) emissions from full-scale wastewater treatment plants. *Environ. Sci. Technol.* **47**, 7795–7803 (2013).
54. Domingo-Félez, C. & Smets, B. F. Modelling N2O dynamics of activated sludge biomass: uncertainty analysis and pathway contributions. *Chem. Eng. J.* **379**, 122311 (2020).
55. Mampaey, K. E. et al. Modelling nitrous and nitric oxide emissions by autotrophic ammonia-oxidizing bacteria. *Environ. Technol.* **34**, 1555–1566 (2013).
56. Pocquet, M., Wu, Z., Queinnec, I. & Spérandio, M. A two pathway model for N2O emissions by ammonium oxidizing bacteria supported by the NO/N2O variation. *Water Res.* **88**, 948–959 (2016).
57. Bellandi, G., Weijers, S., Gori, R. & Nopens, I. Towards an online mitigation strategy for N2O emissions through principal components analysis and clustering techniques. *J. Environ. Manag.* **261**, 110219 (2020).
58. Vasilaki, V. et al. A knowledge discovery framework to predict the N2O emissions in the wastewater sector. *Water Res.* **178**, 115799 (2020).
59. Hwangbo, S., Al, R. & Sin, G. An integrated framework for plant data-driven process modeling using deep-learning with Monte-Carlo simulations. *Comput. Chem. Eng.* 107071 (2020). https://doi.org/10.1016/j.compchemeng.2020.107071
60. Schweidtmann, A. M. et al. Graph neural networks for prediction of fuel ignition quality. *Energy Fuels* **34**, 11395–11407 (2020).
61. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
62. Fioravanti, D. et al. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinforma.* **19**, 49 (2018).
63. van Loosdrecht, M. C. M. & Brdjanovic, D. Anticipating the next century of wastewater treatment. *Science (80-.)* **344**, 1452–1453 (2014).

ACKNOWLEDGEMENTS

This work was supported by Technical University of Denmark.

AUTHOR CONTRIBUTIONS

G.S. and R.A. performed the initial literature survey and prepared an initial draft of the manuscript. G.S. conceived and designed the idea for the work; Both G.S. and R.A. contributed to the acquisition/collection, analysis and interpretation of data; G.S. has written the original draft and R.A. has contributed to the draft and revised it.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to G.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.