University of Nottingham
UK | CHINA | MALAYSIA

EBNet
Environmental Biotechnology

Unlocking AI and Machine Learning's Potential for Environmental Biotechnology

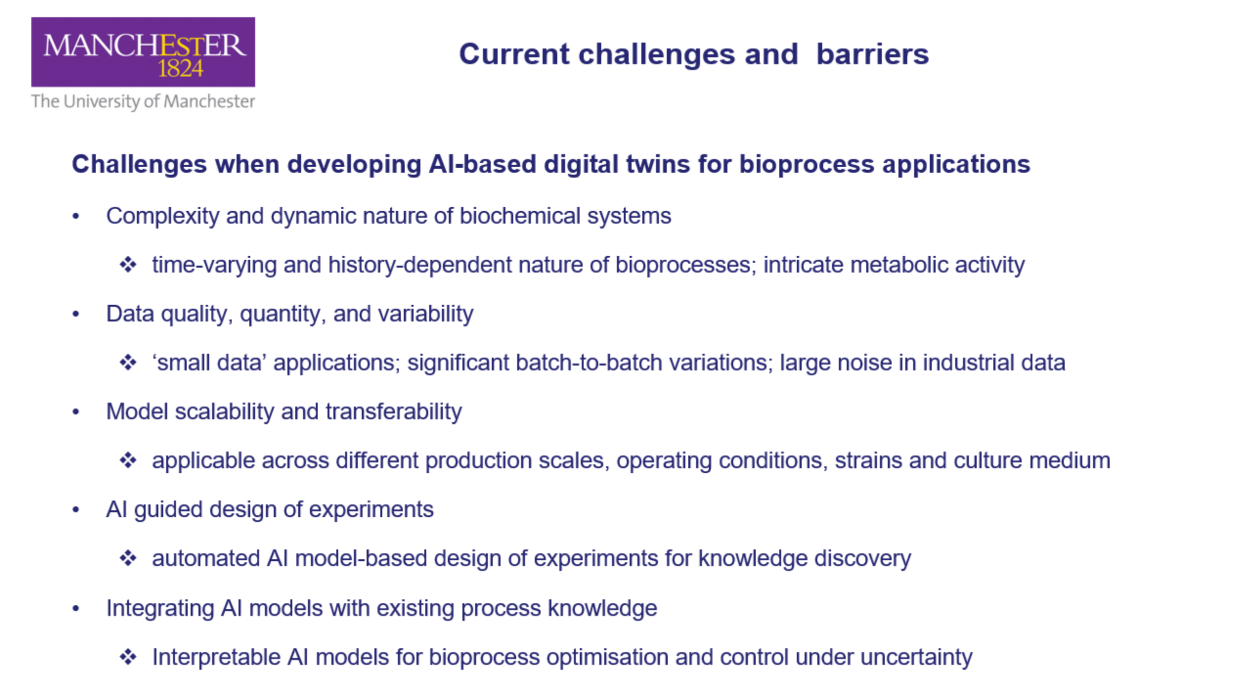
Oliver Fisher

Rachel Louise Gomes

Food Water Waste Research Group

Faculty of Engineering

University of Nottingham

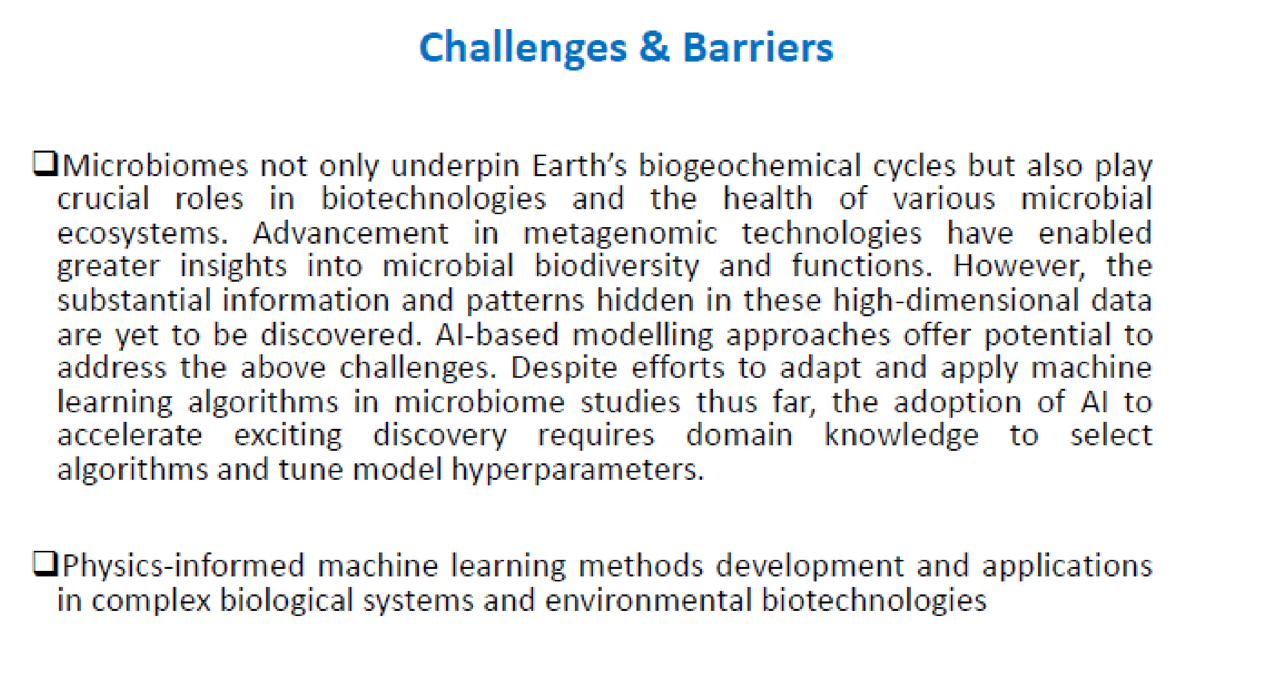


MANCHESTER 1824
The University of Manchester

Current challenges and barriers

Challenges when developing AI-based digital twins for bioprocess applications

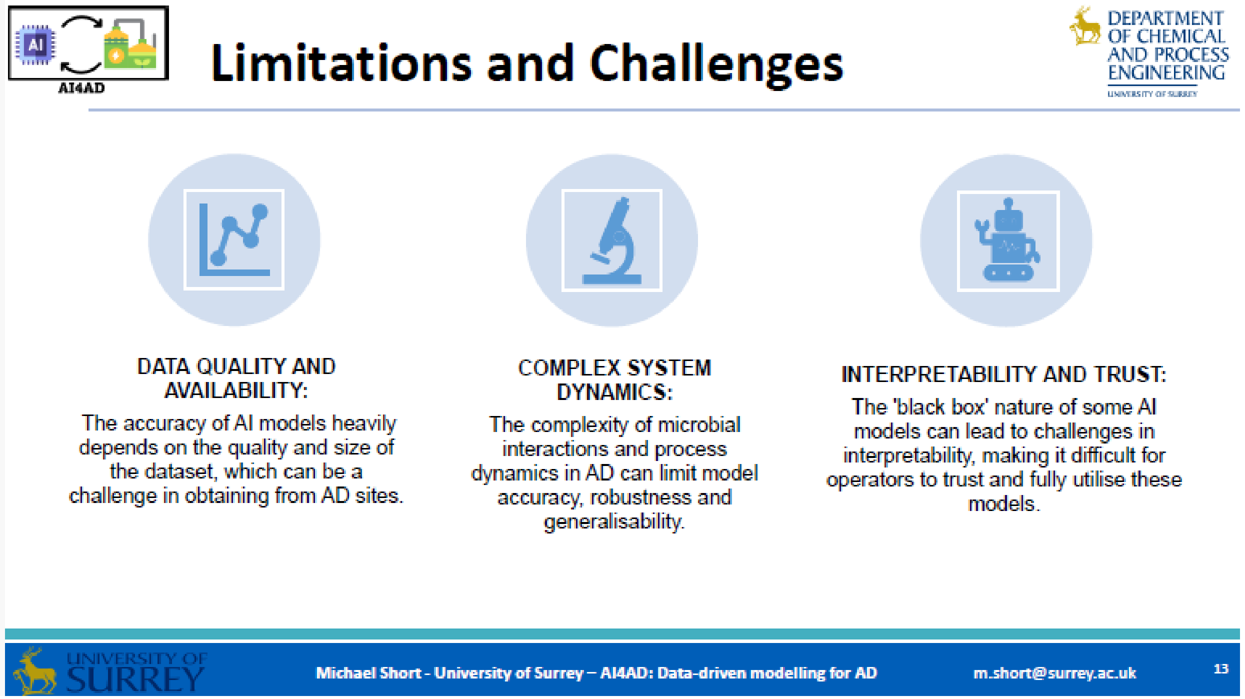
- Complexity and dynamic nature of biochemical systems
  - ❖ time-varying and history-dependent nature of bioprocesses; intricate metabolic activity
- Data quality, quantity, and variability
  - ❖ 'small data' applications; significant batch-to-batch variations; large noise in industrial data
- Model scalability and transferability
  - ❖ applicable across different production scales, operating conditions, strains and culture medium
- AI guided design of experiments
  - ❖ automated AI model-based design of experiments for knowledge discovery
- Integrating AI models with existing process knowledge
  - ❖ Interpretable AI models for bioprocess optimisation and control under uncertainty



Challenges & Barriers

❑ Microbiomes not only underpin Earth's biogeochemical cycles but also play crucial roles in biotechnologies and the health of various microbial ecosystems. Advancement in metagenomic technologies have enabled greater insights into microbial biodiversity and functions. However, the substantial information and patterns hidden in these high-dimensional data are yet to be discovered. AI-based modelling approaches offer potential to address the above challenges. Despite efforts to adapt and apply machine learning algorithms in microbiome studies thus far, the adoption of AI to accelerate exciting discovery requires domain knowledge to select algorithms and tune model hyperparameters.

❑ Physics-informed machine learning methods development and applications in complex biological systems and environmental biotechnologies



AI4AD

DEPARTMENT OF CHEMICAL AND PROCESS ENGINEERING
UNIVERSITY OF SURREY

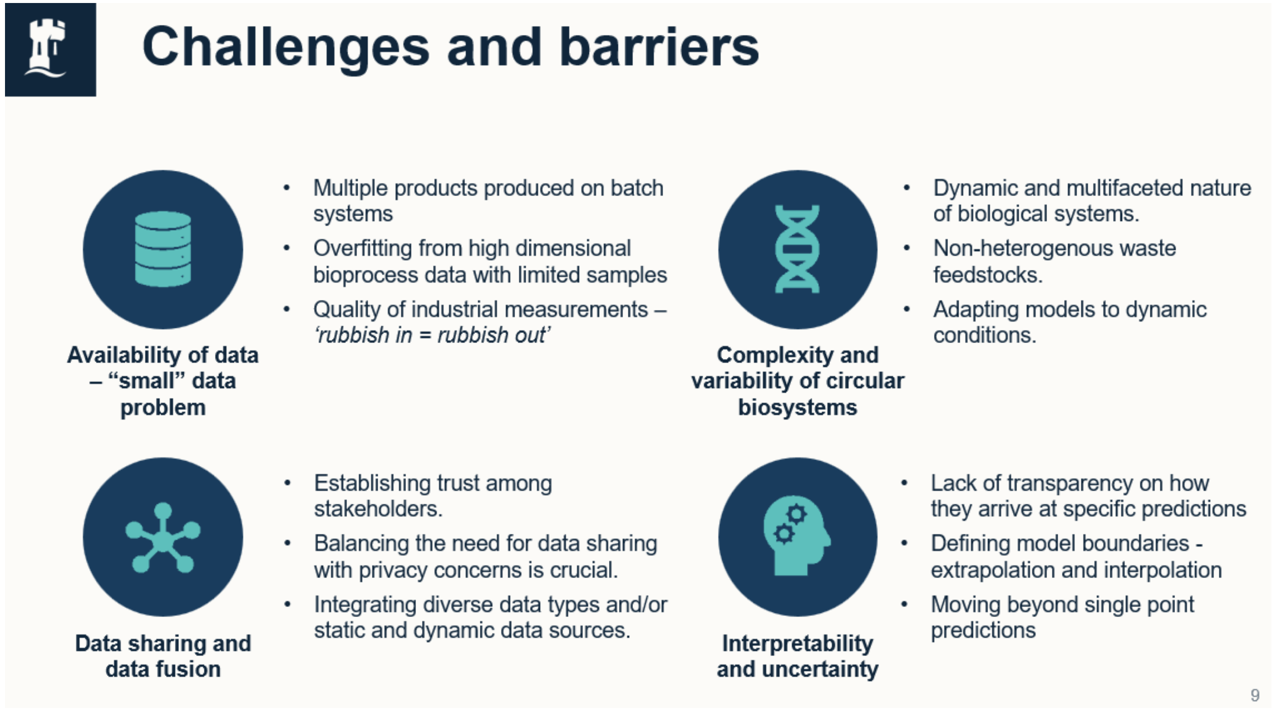
Limitations and Challenges

DATA QUALITY AND AVAILABILITY:
The accuracy of AI models heavily depends on the quality and size of the dataset, which can be a challenge in obtaining from AD sites.

COMPLEX SYSTEM DYNAMICS:
The complexity of microbial interactions and process dynamics in AD can limit model accuracy, robustness and generalisability.

INTERPRETABILITY AND TRUST:
The 'black box' nature of some AI models can lead to challenges in interpretability, making it difficult for operators to trust and fully utilise these models.

UNIVERSITY OF SURREY   Michael Short - University of Surrey – AI4AD: Data-driven modelling for AD   m.short@surrey.ac.uk   13



Challenges and barriers

Availability of data – "small" data problem

- Multiple products produced on batch systems
- Overfitting from high dimensional bioprocess data with limited samples
- Quality of industrial measurements – *'rubbish in = rubbish out'*

Complexity and variability of circular biosystems

- Dynamic and multifaceted nature of biological systems.
- Non-heterogenous waste feedstocks.
- Adapting models to dynamic conditions.

Data sharing and data fusion

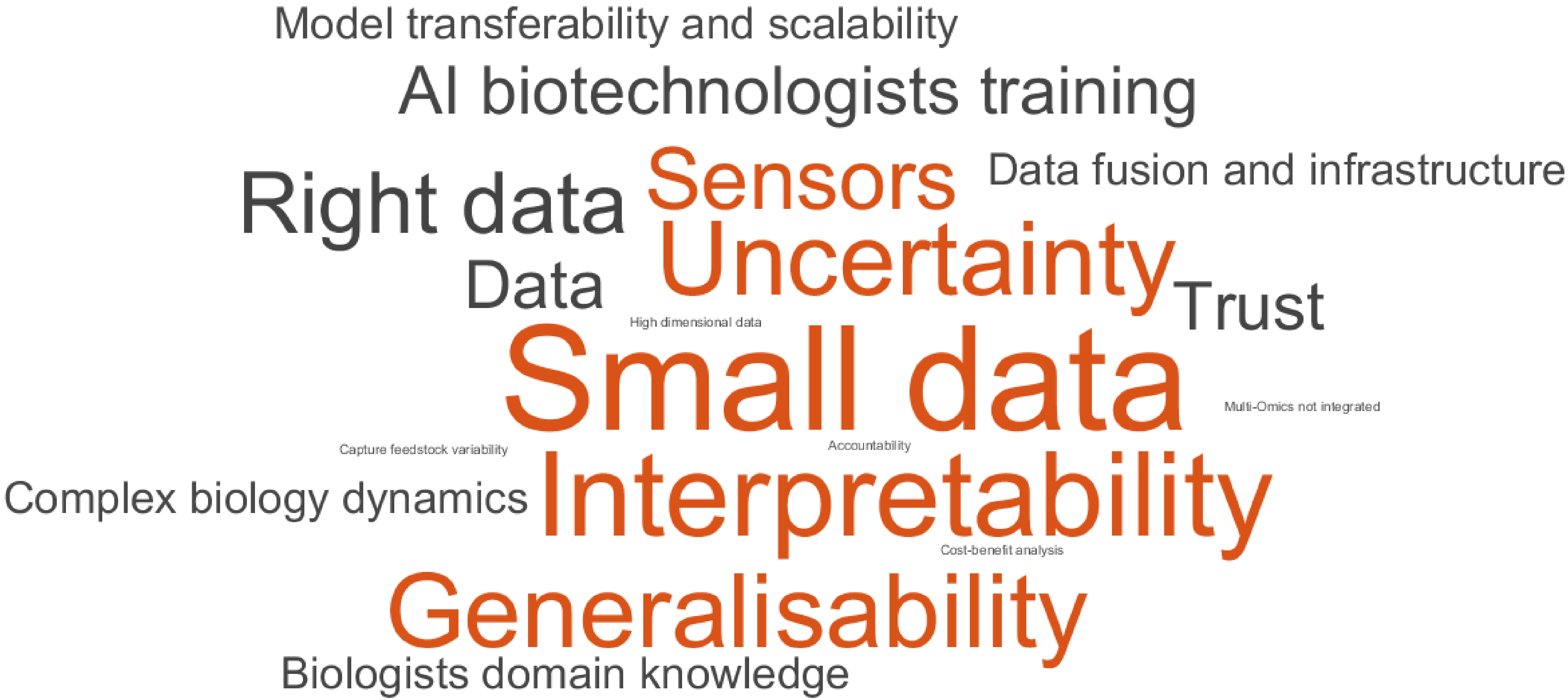
- Establishing trust among stakeholders.
- Balancing the need for data sharing with privacy concerns is crucial.
- Integrating diverse data types and/or static and dynamic data sources.

Interpretability and uncertainty

- Lack of transparency on how they arrive at specific predictions
- Defining model boundaries - extrapolation and interpolation
- Moving beyond single point predictions

9

What do you think are the main challenges in using AI and machine learning in environmental biotechnology?



Model transferability and scalability

AI biotechnologists training

Sensors

Data fusion and infrastructure

Right data

Uncertainty

Data

Trust

High dimensional data

Small data

Multi-Omics not integrated

Accountability

Capture feedstock variability

Complex biology dynamics

Interpretability

Cost-benefit analysis

Generalisability

Biologists domain knowledge

Dongda Zhange (DZ), University of Manchester

- These models often become much more complicated at production scales. Have you looked at taking models from bench reactors to larger scales? How did that go?

DZ: Thanks, yes. Based on our experience, data-driven models are not adept at scaling up because industrial data often contains more noise, and additional physical phenomena (e.g., mass transfer) begin to affect the process behaviour. Consequently, directly improving data-driven models is not straightforward. However, hybrid models are more robust when it comes to upscaling predictions, as they possess a physical model structure that can incorporate additional terms to account for transport phenomena. In practice, when applying hybrid models for larger-scale reactor predictions, we would still require a few experimental data points to fine-tune the models' parameters.

- Can you talk about the analytics, connectivity and data management infrastructure you had to build to get to modelling?

DZ: Sure. We would like to mention that building a model is quite flexible. Depending on the types of analytics available (e.g., in-line/on-line/at-line), there are different models that can be constructed, each serving distinct purposes. In case studies with only off-line measurements, we typically construct either a hybrid model or a pure physical model for process optimisation. For processes involving in-line/on-line analytics, we either develop a pure data-driven model or a hybrid model for process monitoring and control.

- Have you used system identification (modulating inputs, logging outputs) on the bioreactors to pull more data out of each run?

DZ: Yes, system identification is crucial for us to identify the most influential parameters to include in model construction. Depending on whether the available data is already informative, sometimes we may need to gather more data to improve model accuracy.

Dongda Zhange (DZ), University of Manchester

- Scaling up is sometimes very challenging. For a new process, how to use AI to guide the scaling up?

DZ: We utilise lab-scale data to construct data-driven/hybrid models. Subsequently, we refine the model structure slightly based on prior knowledge when making upscaling predictions. Then, using the new data collected from the larger-scale reactor, we apply transfer learning to further refine the model structure and parameters. Additionally, the updated AI models can be employed to design an optimal experiment through model-based Design of Experiments (DoE), extracting the most information to further enhance their accuracy.

- Sensors are not widely used in wastewater and waste treatment processes, how to continuously collect data to train the model and/or control the processes?

DZ: In many cases, we are not seeking the data itself; rather, we are seeking the information embedded in the data. In my experience, we can frequently use data collected offline to build robust models. When employing these models to control the process, we can predict time-series control actions in advance and, based on newly collected data, update control actions after a certain time interval.

- Do you have examples of AI "unusual" predictions that revealed interesting or novel mechanisms?

DZ: Yes, when we were building different transfer learning AI models to predict the effects of two plant hormones on a new strain, we observed some unusual predictions while testing the topology of ANNs. These results helped us eventually identify the individual effects of the two hormones on the strain.

- Is there a clear strategy in place to overcome the common challenges/limitations across the presentations?

DZ: Applying AI/ML to biochemical engineering is still in its early stages and is largely experience-based. However, many strategies have been proposed in the literature, and they can be effective for many of the challenges we highlighted in the presentation.

Micheal Short (MS), University of Surrey

- Presumably the BMP (biomethane potential) labs tests/kit are without any pre-treatment methods/technology? So changes to conversion yield and rates (i.e. complex metabolic pathways) aren't yet being testing (for advancements)?

MS: Yes, this is usually the case. BMP tests usually just try to see what is the maximum gas yield and get some feeling for the first-order rate kinetics. It is a big issue for our modelling as these tests are limited in how they link to full-scale behaviour (microbial communities and reactor geometries are significantly different), and they do not account for different pre-treatments, nor are they usually performed for co-digestion. We often see that co-digestion can significantly alter the BMP (often through synergistic enhancement). In general, most in the AD industry are only looking at those BMP tests to predict output, without including anything other than first-order reaction rates fitted to the BMP curve

- Could a digital twin model for commercial scale projections be done to look at this, and also post-treatment methods/tech (often improved once pre-treatment included)?

MS: Ultimately, yes, this is something that can be done, although I wouldn't call this 'digital twinning' rather techno-economic modelling with enhanced predictive modelling for the AD unit, potentially using hybrid modelling. This can then be used to test, simulate and assess the commercial scale plants retrofitted with pre- and post-treatments.

- Struggle getting data outside of regions of comfort, so supplementing with bench/pilot scale. How does scale impact on this?

MS: Very nice question and something that is certainly a challenge, and not one only found in bioprocessing. Recently, this has been getting increasing attention in many areas of engineering, and multi-fidelity modelling is one way to approach this problem where we have some 'cheaper' and less accurate models or experiments, and some other more expensive and more accurate models/experiments. These new approaches have been shown to be quite effective in handling these differences in performance at different scales. Another approach is transfer learning and hybrid modelling to allow for learning to take place that can account for both the system behaviour that can be accounted for in both fidelities, as well as those behaviours which may differ between fidelities.

Micheal Short (MS), University of Surrey

- Spoke about AD differs at each site and for your project with multi partners able to see how move solutions between sites. What have you learnt from this so far in terms of transferring knowledge from modelling between sites? Especially given feedstock is changing and reactor design is differing.

MS: So far, this has been quite challenging. Due to data sharing concerns, we have generally been able to only move data-driven solutions between sites owned by the same company. This has made using data-driven only solutions challenging, as the data availability is often insufficient, and the best solutions have employed significant data augmentation which is time-consuming. We are therefore developing more accurate hybrid models, relying more on mechanistic insights, that we believe will be able to have improved predictive capabilities between sites.

- Scaling up is sometimes very challenging. For a new process, how to use AI to guide the scaling up?

MS: See comments regarding multi-fidelity approaches. There are several new and exciting approaches using surrogates to assist in experimental designs for scale-up.

- Sensors are not widely used in wastewater and waste treatment processes, how to continuously collect data to train the model and/or control the processes?

MS: Agreed. We can only use data-driven approaches when we have data! Sometimes fairly cheap sensors can be used to collect sufficient data, and remaining data could be collected/simulated from high-fidelity simulations that can be fairly accurate should the model be well-designed and fitted to existing data. It depends on the purpose of any model ('all models are wrong…', etc.) and what needs to be controlled.

- Is there a clear strategy in place to overcome the common challenges/limitations across the presentations?

MS: There is a lot to be learnt in the bioprocessing space from other AI communities. Increasingly, we are seeing advances from computer science and other fields impacting engineering, and I think that leveraging advances in omics and hybrid modelling are 2 such solutions. One of the biggest barriers to progress in this space is that the bioprocessing community lacks widespread awareness of the newer modelling tools and techniques. Increasing the acceptance and use of modelling throughout the sector is important. Delivering value is required by the modelling community in this space to gain resources to increase research here. We will also find that with more value delivered that we are likely to see more data being collected as its value will be appreciated to improve models and decisions.

Oliver Fisher (OF), University of Nottingham

- Does pump speed mean inflow?

OF: Good question. I didn’t have time to explain in the webinar, but the H^2AD reactor system recirculates the effluent through a series of microbial fuel cells. The pump speed controls the recirculation rate through the reactor.

- Scaling up is sometimes very challenging. For a new process, how to use AI to guide the scaling up?

OF: AI can be a powerful tool to help scale up emerging technologies. For instance, experiments and data collection to increasing our knowledge of a new biotechnology operates at scale may be expensive, meaning that we need to prioritise the measurement of informative data. AI can help to design the best possible sequence of experiments to observe the best possible data and to quickly discover new knowledge. Additionally, we can use techniques, such as transfer learning, to translate our knowledge of how these biotechnologies work at lab or bench scale to larger scales.

- Sensors are not widely used in wastewater and waste treatment processes, how to continuously collect data to train the model and/or control the processes?

OF: I agree it can be a challenge. However, a model built from static data can still be useful to generate predictive insights on a how a process works to improve it KPIs. Semi-regular sampling can then inform if the model is still representative of the system over time or requires retraining with newly collected data. Additionally, when real-time control is paramount, AI and ML models can improve existing methods like soft-sensors to estimate currently not measured variables from cheap available sensor readings.

Oliver Fisher (OF), University of Nottingham

- Do you have examples of AI "unusual" predictions that revealed interesting or novel mechanisms?

OF: Not for a bioprocess application. But when creating a computer vision model for classifying Egyptian cotton samples by grade (with scores ranging from 1 to 9), we employed unsupervised machine learning techniques to identify substantial overlaps in cotton grades based on colour features. This analysis demonstrated potential sources of error arising from human graders performing this task and reinforced the need for an automated system to replace manual grading processes.

- Is there a clear strategy in place to overcome the common challenges/limitations across the presentations?

OF: As an emerging field there is certainty lots we can learn from the application of AI and ML to other fields (e.g., finance, aerospace) that have been employing these techniques for longer. As a working group, it is our aim to bring together those working on AI and ML to model environmental biotechnologies to co-create solutions to these challenges.

Tell us your wants from the working group by contacting us at

https://forms.office.com/e/Bq20wyjCty

or by scanning the QR code