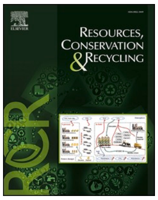
Resources, Conservation & Recycling 215 (2025) 108071

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

journal homepage: www.sciencedirect.com/journal/resources-conservation-and-recycling



Perspective

Breaking barriers to modelling biotechnologies with machine learning

Oliver J. Fisher [a,*], Michael Short [b], Dongda Zhang [c], Miao Guo [d], Rachel L. Gomes [a]

[a] *Food Water Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK*
[b] *School of Chemistry and Chemical Engineering, University of Surrey, Guildford, GU2 7XH, UK*
[c] *Department of Chemical Engineering, the University of Manchester, Manchester, M13 9PL, UK*
[d] *Department of Engineering, King's College London, London, WC2R 2LS, UK*

The chemical industry supply materials for 94 % of products produced today (Kähler et al., 2021). However, 88 % of the industry's carbon-based feedstocks are from fossil resources, posing a significant challenge in the context of climate change and the industry's net-zero targets (Kähler et al., 2021). To address this, it is essential to replace fossil-based resources with bio-based and/or recycled carbon feedstocks, thereby moving towards a circular economy.

Biotechnologies such as fermentation, anaerobic digestion, and bioelectrochemical systems show great promise in converting organic spent resources (waste) into platform chemicals and/or high-value products (e.g. proteins, pharmaceuticals) and bioenergy, removing our reliance on fossil-based resources. Despite their potential, development of these processes has lagged behind other circular economy technology (e.g. chemical recycling) (Schagen et al., 2023; Trump et al., 2023), due to the challenges of heterogeneous feedstocks, the need for flexible systems, supply chains, and regulations.

Machine learning (ML) is a transformative technology reshaping various fields of science and engineering, enabling advanced data analysis, automation, optimisation, prediction, and decision-making capabilities. ML refers to the ability of machines to learn from data, identify patterns, and make decisions with minimal human intervention. By analysing large datasets, ML algorithms can uncover new intricate patterns and relationships within biological systems that were previously difficult to detect. This capability is valuable for identifying the factors that are most conducive to accelerating the progression of biotechnologies up technology readiness levels (TRL), such as optimal microbial strains, feedstock characteristics, reactor conditions, and process efficiencies. By gaining these unprecedented insights, researchers and industry professionals can optimise conditions for biotechnologies, ultimately enhancing their economic feasibility for large-scale application. However, the pace of adoption of ML to model biotechnologies has been slower than other fields (Holzinger et al., 2023). Therefore, it is crucial to understand the modelling barriers (Table 1) that must be overcome.

1. Complexity and dynamic nature of biotechnologies and waste feedstocks

The complexity of microbial interactions and process dynamics presents challenges that can limit the accuracy, robustness, and generalisability of models. This complexity is particularly pronounced by biotechnologies utilising heterogeneous waste feedstocks, which often have complex and unknown chemical compositions. Reflecting the complexity of waste feedstocks requires understanding the chemical and composition makeup and which feedstock (and biotechnology / environmental) factors most influence process performance, yet this task is hindered by the temporal and spatial variability inherent in both feedstock/s and biotechnologies. The interplay of these factors underscores the need for innovative approaches to modelling and analysis to capture this complexity and variability.

2. Multi-omics data integration

Advancements in omics (e.g. metabolomics, metagenomics, transcriptomics) and bioinformatics have significantly enhanced our understanding of microbial systems and their functions, providing valuable insights for biotechnology development, and the influence of variable feedstocks and external environmental conditions on biotechnology performance. These tools and approaches when applied to biotechnologies for pollutant remediation, bioenergy production, platform or high value chemical production offer a detailed view of the metabolic activities within microbial systems. Despite the substantial information and hidden patterns present in high-dimensional data generated from these processes, fully leveraging these insights remains challenging. Effective utilisation of such high-dimensional datasets necessitates domain expertise to select appropriate ML algorithms and fine-tune model hyperparameters, ensuring accurate and meaningful analysis.

[*] Corresponding author.
E-mail address: oliver.fisher@nottingham.ac.uk (O.J. Fisher).

https://doi.org/10.1016/j.resconrec.2024.108071
Received 27 September 2024; Received in revised form 28 November 2024; Accepted 3 December 2024
Available online 7 December 2024

Table 1
Barriers to implementing machine learning (ML) in biotechnologies modelling with examples from authors' work.

Barrier	Examples from authors work
Complexity and dynamic nature of biotechnologies and waste feedstocks	**Context**: Biofilms are aggregates of microorganisms embedded in a three-dimensional matrix of extracellular polymeric substances. They are increasingly valued for their applications in wastewater treatment, bioremediation, and the production of valuable substances like organic acids, alcohols, and proteins **Challenge**: Biofilm properties, including microstructure and composition can significantly impact process outcomes (e.g. pollutant removal rate), yet their inherent complexity and heterogeneity make modelling of biofilms highly challenging. **Solution**: ML-enhanced sensor fusion system to combine data from multiple sensors, to provide real-time insights and predictions on biofilm properties beyond traditional methods.
Multi-omics data integration	**Context**: Multiple omics data layers including genomics, transcriptomics, proteomics, metabolomics are generated by advanced sequencing and high-throughput technologies; each omics layer provides unique biological perspectives. **Challenge**: A major challenge, however, remains to integrate multi-omics data to provide system-level multi-layer views and enable decision-making on biological pathways. Firstly noisy and highly dimensional datasets and omics data heterogeneity complicate their integration. Secondly, omics data are context-dependent and complex with diverse metabolic pathways (linear, circular, convergent or divergent). **Solution**: Novel retrobiosynthesis method proposed and developed to integrate multi-omics.
Data quantity, quality, and availability	**Context:** Mammalian cell culture systems, such as those involving Chinese Hamster Ovary cells, are vital for the production of biopharmaceuticals like monoclonal antibodies. Efficient process monitoring and optimisation are critical to meet industrial demands for quality and yield. **Challenge:** Developing accurate models for mammalian cell culture systems is hindered by the limited availability of high-quality datasets and the lack of detailed mechanistic understanding. Traditional models often require extensive experimental data or rely heavily on assumptions, leading to challenges in model reliability and predictive capabilities under new conditions. **Solution:** We proposed a hybrid modelling framework that integrates domain knowledge with ML. This approach enables accurate process simulation and uncertainty estimation even in small-data scenarios, and allows the model to dynamically adapt to new data while maintaining high predictive accuracy. The methodology also facilitates the development of robust digital twins for optimising mammalian cell culture processes.
Model scalability and transferability	**Context**: There are thousands of anaerobic digesters worldwide, producing green gas and organic fertiliser, while treating agricultural wastes. The AI for Net Zero 'AI4AD' project is bringing together several industrial and academic partners to develop whole-site digital twins that combine different model types and scales built on data from across a variety of different sites to enhance whole site systems decision-making. **Challenge**: Many biogas sites have varying system layouts, different reactor configurations, varying feedstocks, and reaction conditions, making model transferability across sites challenging. **Solution**: Using combinations of physics-based models with artificial intelligence help to transfer modelling results to new systems. Uncertainty quantification helps to provide confidence and inform modelling approaches. Working closely between different companies and academia allows for solutions that leverage knowledge and know-how across the sector to build more transferable and general solutions.
Interpretability and trust	**Context**: In the UK, Lindhurst Engineering Ltd., in collaboration with the University of Nottingham, developed "H^{2}AD," a technology combining bioelectrochemical systems (BES) and anaerobic digestion to treat diverse wastewaters (e.g., agricultural, brewing, and biomanufacturing residues). This system reduces pollutant loads, enhances water reuse quality, and generates bioenergy. **Challenge**: When modelling H^{2}AD, the high variability in waste composition due to temporal and geographical factors undermines trust in model outputs. **Solution**: Novel data visualisation techniques developed to help assess whether new data points fall within the model's boundaries, boosting confidence in its predictions despite variability.
Accessibility and uncertainty	**Context**: Industrial crops serve as essential feedstocks for biotechnologies in developing countries, with their quality and composition often assessed manually through labour-intensive and subjective inspection processes. **Challenge**: Access to advanced and expensive technologies for characterising industrial crop feedstocks poses a significant barrier, particularly in resource-constrained settings. **Solution**: In partnership with the University of Alexandar, a low-cost quality evaluation system of Egyptian cotton was developed, integrating accessible imaging tools and ML.

3. Data quantity, quality, and availability

The accuracy and predictive capability of models in biotechnologies relies on the quality and quantity of data, including how representative is the data to the system being modelled. It is crucial to recognise that simply having a large quantity of data does not guarantee quality outcomes. Many ML studies on biotechnologies often deal with small-sized datasets that may not fully capture the variability within a system or extend to its boundaries. This can be compounded with biotechnologies on sites operating within a small window or range, and thereby deriving data that may not define the system and/or true optimal performance. Additionally, the variables that best define biotechnology performance may not or only occasionally be measured, potentially at great expense and time, or may not even be known]. When data is limited to specific areas of the design space, ML models cannot be effectively utilised for optimisation or control without extrapolation, which introduces uncertainty. Data sharing is hindered by confidentiality barriers, though federated learning presents a potential solution to share key model parameters while respecting data confidently. Alternatively, incorporating prior knowledge, such as physical principles, can enhance model reliability. By adopting different strategies like hybrid modelling and data-driven modelling, reliable ML models for biotechnology can be developed, addressing the challenges associated with data quantity, quality, and availability.

4. Model scalability and transferability

The ability to scale ML models across various production scales, operating conditions, external environments, strains, and feedstock is crucial for advancing biotechnologies TRL. However, the diversity in process design adds complexity and challenges with translating models, with differing plant designs spanning reactor geometries, mixing conditions, and flowsheet configurations. This is compounded with use of differing feedstocks and stakeholder expertise e.g. municipal wastewater treatment to waste food digestion. This diversity complicates the transfer of models and knowledge between locations. Moreover, the inherent time and spatial variability of waste between sites necessitates personalised solutions, further complicating standardisation efforts. To address these challenges, combining transfer learning with model structure identification techniques could facilitate the rapid transfer of existing models across different applications, enhancing scalability and adaptability of biotechnologies.

5. Interpretability and trust

The "black box" nature of ML models poses significant challenges in interpretability, making it difficult to fully trust and utilise these models. This lack of transparency also complicates the integration of ML models with existing process knowledge for biotechnology optimisation and control under uncertainty. To address these issues, future approaches should move to hybrid models incorporating mechanistic-based models alongside data-driven ML techniques. Enhancing human-machine interactions, such as allowing process operators to modify models through app-based interfaces based on their knowledge, is crucial for improving trust in ML models. Additionally, training industry personnel and implementing effective data visualisation are vital for making operators aware of and capable of interpreting these tools and techniques.

6. Accessibility and uncertainty

The adoption of ML in biotechnologies is hindered by limited digital infrastructure, particularly in low- and middle-income regions, despite their potential for bio-based innovation. The absence of regulations and standardised protocols for data collection and process optimisation further complicates the integration and scalability of ML models across biotechnological applications. Moreover, high costs of AI tools and lack of industry-specific solutions further limit their accessibility, necessitating innovative, cost-effective, and tailored solutions to enable AI-driven advancements in biotechnologies.

7. Looking forward

To fully harness ML in advancing industrial biotechnologies a collaborative, cross-sectoral approach is essential. This requires efforts across supply chains and industries, addressing challenges like waste feedstock traceability. However, success relies on multidisciplinary expertise beyond biology, incorporating engineering, computer science, and business. Moreover, targeted funding is essential to address the lack of industrial knowledge and resources in comparison to other industries like petrochemicals, pharmaceuticals, and renewables. Without substantial investment comparable to these sectors, progress in the bio-industry may lag further behind. Furthermore, there is a need for greater transparency, where industries utilising biotechnologies openly share their challenges to guide researchers in developing real-world solutions that meet industry needs.

CRediT authorship contribution statement

Oliver J. Fisher: Writing – original draft, Visualization, Project administration, Investigation, Funding acquisition, Conceptualization. **Michael Short:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Dongda Zhang:** Writing – review & editing, Conceptualization. **Miao Guo:** Writing – review & editing. **Rachel L. Gomes:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Funding acquisition, 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.

Acknowledgements

The authors (OJF and RLG) would like to acknowledge the support from the Environmental Biotechnology Network (EBNet) funded by Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/S009795/1]. The authors (MS and DZ) would also like to gratefully acknowledge the support Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/Y005600/1], Artificial Intelligence Enabling Future Optimal Flexible Biogas Production for Net-Zero.

Data availability

No data was used for the research described in the article.

References

Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., Müller, H., 2023. AI for life: trends in artificial intelligence for biotechnology. N. Biotechnol. 74, 16–24. https://doi.org/10.1016/j.nbt.2023.02.001.

Kähler, F., Carus, M., Porc, O., & vom Berg, C. (2021). *Turning off the tap for fossil carbon: future prospects for a global chemical and derived material sector based on renewable carbon.* www.renewable-carbon.eu/publications.

Schagen, O.M., Metze, T.A.P., de Olde, E.M., Termeer, C.J.A.M., 2023. Energizing a transformation to a circular bioeconomy: mechanisms to spread, deepen and broaden initiatives. Sustain. Sci. 18 (3), 1099–1115. https://doi.org/10.1007/s11625-022-01249-1.

Trump, B.D., Cummings, C.L., Loschin, N., Keisler, J.M., Wells, E.M., Linkov, I., 2023. The worsening divergence of biotechnology: the importance of risk culture. Front. Bioeng. Biotechnol. 11. https://doi.org/10.3389/fbioe.2023.1250298.