AI4b.io Symposium 2026
"AI in Bioscience: Redefining Frontiers."

You are invited to participate in the Artificial Intelligence Lab for Bioscience (AI4b.io) symposium. This symposium will take place physically in Delft on April 13 & 14, 2026. We are organizing this meeting for 150 participants who are active in Artificial Intelligence and Bioscience. Your participation is appreciated because of your expertise in this area, and we are looking forward to your contributions in vibrant discussions and seeing this symposium as a start of a new community on AI for bioscience. The program is included in this document and covers topics ranging from large-scale manufacturing to microbiome-based therapeutic development, going from large to small scale. Experts active in these topics will present their in-depth insights.

Agenda & Logistics
Date: April 13-14 (Monday and Tuesday), 2026

Schedule and booklet: available online on the event page and booklet.

Venue: Panorama XL in Mondai | House of AI. Molengraaffsingel 29, 2629JD, Delft

Registration: We will start both days with coffee and registration at 9:00. There will be name badges that you need to pick up at the registration desk every morning. **While our event is free, we kindly ask you to email us at info@ai4b.io if you can no longer attend the specific days you registered for or the event dinner** (please check your registration confirmation email for what you indicated before). This helps us prevent food waste and offers other people a spot if available.

Poster pitches
A dedicated time slot for poster pitches has been scheduled (on Day1, 14:50 – 15:10) to help draw attention to the posters. Presenters can give a brief 2-minute pitch summarizing their work before the poster session begins. Participation is optional, and slides will be compiled into a single presentation for smooth transitions. The poster session will follow immediately after the pitches.

Food
Lunch and borrel will be provided on both days at the venue, as well as tea, coffee, and other refreshments. The dinner on Day1 will start at 18:00 at Firma van Buiten, Thijsseweg 1, 2629 JA Delft, Netherlands. It is only a five-minute walk from the venue, and we will guide you there. This dinner will be vegetarian. If you have any dietary requirements (e.g., allergies), please notify us through info@ai4b.io.

Transportation
Public transport: When traveling by public transport, you could take bus 69 from Delft Central Station to 'Molengraaffsingel'. From the bus stop, it is a 10-minute walk to the building.

Biking: It will take you 15 minutes to bike from Delft Central Station. You could rent a bike at a local bike shop. There are bicycle stands available on the left side of the building.

Driving: When you come by car, you can park across the road from the building, at the official TU Delft car park P8. The car park is free for TU Delft employees upon showing their campus card. For external parties, there are costs involved.

Hotels

If you're considering staying overnight, we recommend hotels in the Delft city center, e.g., the BW Signature Collection Grand Museum Hotel, Ibis, the Social Hub, and Hotel de Koophandel. The Delft train station is within five minutes walking distance of the city center.

Please let us know if you need any assistance or need to use the elevator by replying to this email. If you have any urgent questions during the event, you can contact us by phone or Whatsapp: +31 644985593.

We hope you will have an enjoyable experience at the coming AI4b.io symposium. Please don't hesitate to reach out to us now if you have any questions or special requests. Thank you, and we'll see you there!

Best regards,
AI4b.io Symposium Organizing Committee

Organizers and Steering Committee

Renger Jellema, Marunka van Stight, Marcel Reinders, Hans Roubos, Wouter van Winden, Ali May, Stijn Bierman, Liang Wu, Jana Weber, Amelia Villegas Morcillo

Timetable – Day 1, 13th April

Day 1	Schedule	Speaker Name & Affiliation
9:00-10:00	Arrive, register, coffee	
10:00-10:15	Opening talk	Marcel Reinders Delft University of Technology
10:15-11:00	**Keynote**: Towards Artificial Olfaction	Halima Mouhib Vrije Universiteit Amsterdam
11:00-11:25	PLM-eXplain: Divide and Conquer the Protein Embedding Space	Jan van Eck Utrecht University
11:25-11:50	Active learning maps the activity-selectivity Pareto front in enzymatic PET upcycling	Manu Suvarna University of Greifswald
11:50-13:15	Lunch + Icebreaker	
13:15-14:00	**Keynote**: Learning the Geometry of Life: Platonic Transformers as a Solid Choice for Bioscience	Erik Bekkers University of Amsterdam
14:00-14:25	Optimizing the Solubility of Organic Molecules in Mixed Solvents	Simona Buzzi Katholieke Universiteit Leuven
14:25-14:50	CompleteRXN: Curation of Incomplete Chemical Reaction Databases	Gabriel Vogel Delft University of Technology
14:50-15:10	Poster Pitches	
15:10-16:30	Coffee break + Posters	
16:30-16:55	Context-aware agent for process flowsheet synthesis	Ulderico Di Caprio Delft University of Technology
16:55-17:30	TBD	Kim van Houten Delft University of Technology
17:30:17:45	Closing remarks + walking to dinner location	
17:45-21:00	Travel to Dinner location + Borrel + Dinner at Firma van Buiten	

Timetable – Day 2, 14th April

Time	Schedule	Speaker Name & Affiliation
9:00-9:30	Arrive, register, coffee	
9:30-10:15	**Keynote**: A digital platform for the Design, Control and Scale-Up of Bioprocesses	Lukas Gsenger Graz University of Technology
10:15-10:40	From Data to Draft: ML from Lab to Logistics	Jurgen Nijkamp Heineken
10:40-11:00	Break	
11:00-11:35	Askara: A Multi-agent GenAI Assistant for Research	Jie Yang Delft University of Technology
11:35-12:00	Automating workflows across biotech manufacturing with agentic AI	Giacomo Lastrucci Delft University of Technology
12.00-13:00	Lunch	
13:00-13:45	**Keynote**: The human gut microbiome: variation, diagnostics and modulation	Jeroen Raes VIB KU Leuven Center for Microbiology
13:45-14:10	Reducing Heterogeneity in Cross-Study Microbiome ML with Data Attribution	Can Dedekoy dsm-firmenich
14:10-14:35	New Pathways: Reconstructing Microbial Metabolism from Literature Using NLP and Generative AI	Wynand Alkema Hanze University
14:35-15:35	Break (15 minutes) + Round table + Posters	
15:35-16:00	Beyond benchmarking: an expert-guided consensus approach to spatially aware clustering	Kirti Biharie Delft University of Technology
16:00-16:25	CycleVI: generative model of the cell cycle	Gustavo Jeuken Vrije Universiteit Amsterdam
16:25-16:50	Multi-omics cohort analysis for target discovery to derisk interventional clinical trials	Ali May dsm-firmenich
16:50-17:00	Closing remarks	Hans Roubos dsm-firmenich
17:00-18:00	Borrel	

Day 1, April 13th, 09:00 – 21:00

9:00 – 10:00 | Arrive, register, coffee

10:00 – 10:15 | Welcome Note
Marcel Reinders, Delft University of Technology

Session 1 Chair: Jana Weber

1 | 10:15-11:00 | **Keynote:** Towards Artificial Olfaction

Kyllesbech, M. Chen, I. De Moya Clark, A. Sajan, N. Morrel and H. Mouhib, Dept of Computer Science, Vrije Universiteit Amsterdam

Graphical abstract

Abstract
In contrast to hearing and vision, the exact function of the olfactory sense remains largely unknown. This impacts many research areas ranging from fragrance chemistry to odor perception, as well as sensor design in artificial noses. Recent advances in AI technologies started to pave the way towards digitizing odors for structure-based odor predictions and broadening our understanding of structure odor relationships. However, there are still many open ends and questions before the sense of smell is successfully digitized and new knowledge can be transferred to other applications such as bio-mimetic sensor design. One problem that is currently holding back the advancement in the field of artificial noses is our fragmentary understanding of the molecular detection mechanisms underlying the olfactory sense.

During the talk, I will give an overview of the advances and challenges in the field and show different projects that our team is working on at the VU Bioinformatics group of the Vrije Universiteit Amsterdam. Here, one of our main objectives is to unravel and to quantify the uptake mechanisms of odorants and other volatile organic molecules through binding proteins for applications in bio-mimetic sensor units.

In the future, due to the increasing number of available data on molecular structures and olfactory properties, machine learning and more data-driven approaches will be necessary to address new

computational challenges and thus, in the long term, allow us to move towards the digitalization of the olfactory sense.

References

1. B. K. Lee et at., Science 2023.
2. M. Paesani et al., Chem. Eur. J. 2024.

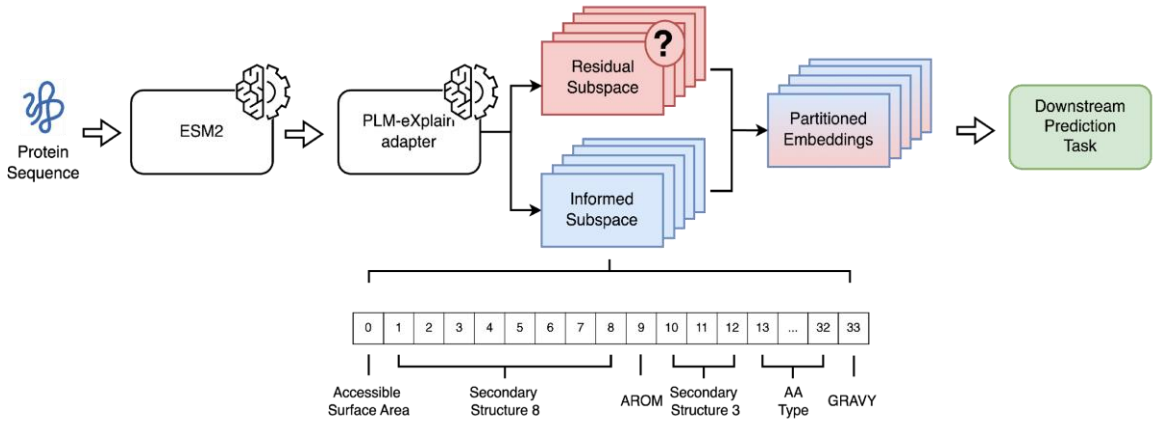
A. Sajan et al., arXiv 2025.

3. M. Chen et al., Protein Science 2026.

2 | 11:00-11:25 | PLM-eXplain: Divide and Conquer the Protein Embedding Space

Jan van Eck, Dea Gogishvili, Wilson Silva and Sanne Abeln, AI Technology for Life, Department of Computing and Information Sciences, Department of Biology, Utrecht University

Graphical abstract



Abstract

Context: Protein language models (PLMs) have revolutionized computational biology through their ability to learn powerful sequence representations that encode predictive information for a wide range of downstream tasks. However, their black-box nature limits biological interpretation and translation to actionable insights. Bridging this gap requires approaches that maintain predictive performance while providing interpretable explanations of model behavior.

Approach: We present PLM-eXplain (PLM-X) [1], an explainable adapter layer that bridges this gap by factoring PLM embeddings into two complementary components: an interpretable subspace based on established biochemical features, and a residual subspace that retains predictive, non-interpretable information. Using embeddings from ESM2 [2] and ProtBert [3], PLM-X incorporates well-established properties, including secondary structure and hydropathy, while maintaining high predictive performance.

Results: We demonstrate the effectiveness of our approach across three biologically relevant classification tasks: extracellular vesicle association, transmembrane helix prediction, and aggregation propensity prediction. PLM-X enables biological interpretation of model decisions without sacrificing accuracy, offering a generalizable solution for enhancing PLM interpretability across various downstream applications.

Impact: PLM-X resolves the performance and interpretability trade-off in protein language models, providing a general framework for biologically interpretable and high-performing protein prediction models. Future work will scale this framework by incorporating thousands of functional and structural features, enabling fine grained explanations on downstream tasks.

References

1. Suvarna et al. Nat Catal. 2024, 7, 624.
2. Suvarna et al. Nat Commun. 2024, 15, 5844.
3. Taylor et al. Chem. Rev. 2023, 123, 3089.
4. Thomsen et al. ChemSusChem. 2023, 16, e202300291.

3 | 11:25-11:50 | Active learning maps the activity-selectivity Pareto front in enzymatic PET upcycling

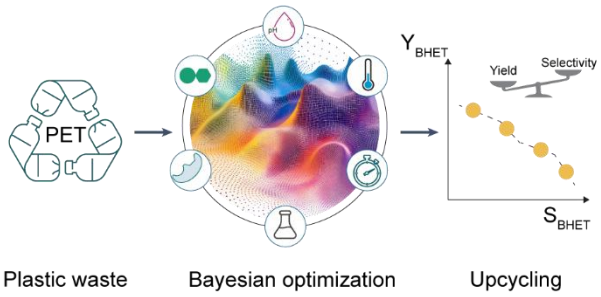
Manu Suvarna[1], Leon Krinn[2] and Anne S. Meyer[3]

[1]Dept. of Biotechnology and Enzyme Catalysis, University of Greifswald

[2]Dept. of Chemistry and Applied Biosciences, ETH Zurich

[3]Dept. of Biotechnology and Biomedicine, Technical University of Denmark

Graphical abstract



Abstract

Context: Biocatalytic transformation of polyethylene terephthalate (PET) presents a sustainable strategy for plastic upcycling. While enzyme engineering dominate current efforts, process optimization to desired products/intermediate remain relatively underexplored. Herein, we target bis(2-hydroxyethyl) terephthalate (BHET), a key PET hydrolysis intermediate and platform chemical, whose yield (YBHET) and selectivity (SBHET) exhibit an intrinsic trade-off, when catalyzed by the LCCICCG enzyme.

Approach: To address this challenge, we implement an active learning strategy that integrates the Gaussian Process and Bayesian optimization algorithm in experimental workflows. This cross-disciplinary strategy efficiently explores the vast reaction parameter space, to simultaneously maximize both YBHET and SBHET.

Results: Beginning with a constrained in-house dataset, in five active learning cycles and 25 new experiments, our approach uncovers multiple Pareto-optimal solutions, identifying YBHET = 6.5 µM and SBHET = 55% as the best solution. The model attains overall R^2 = 0.86, reduces the experimental search efforts by over 50%, and almost doubles the overall catalytic performance from its starting

point. Feature-importance analysis reveals solvent concentration as the key driver affecting activity-selectivity trade-off. We further complement this insight with Michalis-Menten kinetics revealing substrate inhibition beyond a threshold solvent amount. Lastly, we validate the workflow on DuraPETase and PHL7, confirming enzyme-level transferability of our approach.

Impact: Overall, the predictive toolkit developed in this study effectively maps the YBHET and SBHET and guides experimental researchers to perform data-informed experiments to achieve desired metrics. In a broader context, our study delivers a scalable multi-objective optimization framework integrating data-driven insights with experimental design to accelerate biocatalytic PET upcycling.

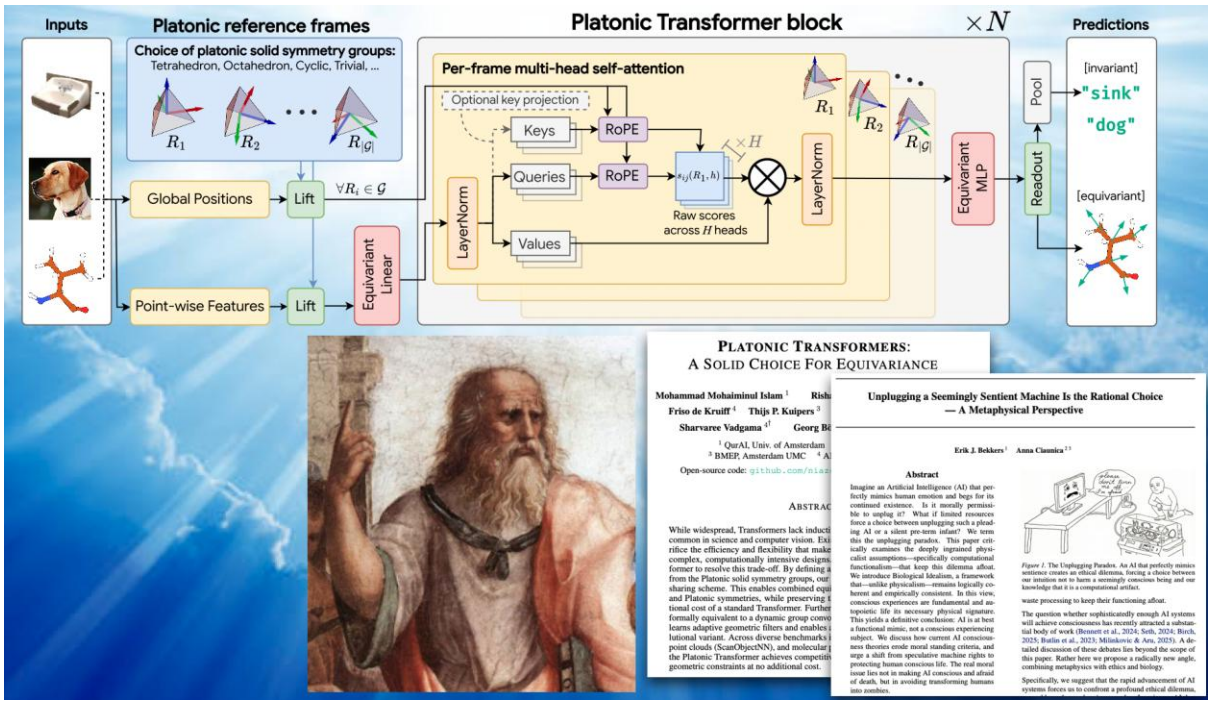
11:50-13:15 Lunch + Icebreaker

Session 2 Chair:

4 | 13:15-14:00 | **Keynote:** Learning the Geometry of Life: Platonic Transformers as a Solid Choice for Bioscience

Erik Bekkers, Amsterdam Machine Learning Lab (AMLab), University of Amsterdam

Graphical abstract



Abstract

Context: Transformers lack the geometric inductive biases—such as equivariance to rotations and translations—that are essential for tasks in 3D computer vision and scientific data (e.g., molecular physics). On the other hand, existing equivariant methods often sacrifice the efficiency and flexibility that make Transformers so effective, relying instead on complex, computationally intensive designs. We set out to resolve this trade-off: how can we inject powerful geometric inductive biases into the standard Transformer architecture without losing its speed and scalability?

Approach: We introduce the Platonic Transformer. By defining attention relative to reference frames sampled from Platonic solid symmetry groups, our method induces a principled weight-sharing scheme. This intrinsically allows for equivariance to continuous translations and discrete roto-reflections while exactly preserving both the architecture and the computational cost of a standard Transformer. Furthermore, we formally demonstrate that this Rotary Position Embedding (RoPE) based attention is equivalent to a dynamic group convolution.

Results: Across diverse benchmarks, including computer vision (CIFAR-10), 3D point clouds (ScanObjectNN), molecular dynamics property prediction, and generation (OMol25, ProteinMD, QM9), the Platonic Transformer achieves highly competitive performance. Most notably, it captures these geometric constraints at no additional computational cost compared to a standard Transformer. Additionally, in its linear convolutional formulation, the attention mechanism scales linearly with sequence length, making it highly efficient.

Impact: The Platonic Transformer successfully reconciles the dilemma between geometric symmetry-awareness and computational scaling, acting as a scalable foundation model for physical sciences and 3D vision [1]. While Platonic ideals find a natural application in the mathematical formalism of these architectures, the symposium's theme—connecting living systems and learning machines—invites a broader philosophical reflection. To conclude the talk, we step back from the computational realm to adopt a metaphysical and ethical lens [2]. Drawing on Biological Idealism [2], we highlight the special, irreplaceable status of biological life in an era of unprecedented AI capabilities, arguing that even the most sophisticated learning machines remain functional mimics rather than experiencing subjects. Believing anything else has serious moral and ethical consequences.

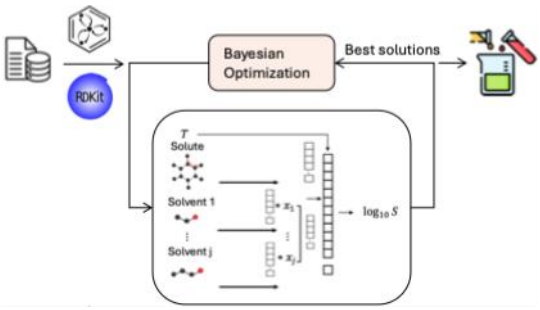
References

1. Islam, Mohammad Mohaiminul, et al. "Platonic Transformers: A Solid Choice For Equivariance." arXiv preprint arXiv:2510.03511 (2025). https://arxiv.org/abs/2510.03511.
2. Bekkers, Erik J., and Anna Ciaunica. "Unplugging a Seemingly Sentient Machine Is the Rational Choice -- A Metaphysical Perspective." arXiv preprint arXiv:2601.21016 (2026).

5 | 14:00-14:25 | Optimizing the Solubility of Organic Molecules in Mixed Solvents

Simona Buzzi[1], Ulderico Di Caprio[1,2], Dominik Bongartz[1], and Florence Vermeire[1]
[1]Department of Chemical Engineering, KU Leuven,
[2] Department of Chemical Engineering, Delft University of Technology

Graphical abstract



Abstract

Context: Predicting the solubility of organic compounds in mixed solvents is crucial in chemical engineering and pharmaceutical process design. Data-driven approaches, including QSPR and machine learning [1], provide fast and accurate predictions even for complex multicomponent systems. The availability of such models enables optimization frameworks to identify ideal process conditions for a given compound, accelerating process development while reducing the cost. However, selecting suitable solvents, compositions, and temperatures remains challenging due to the large combinatorial design space and the complex, nonlinear behavior of predictive models such as directed message passing neural networks (D-MPNNs). In pharmaceutical development, where rapid decision-making and cost reduction are critical, methods including Bayesian optimization (BO) [2] are needed to efficiently guide experimental campaigns toward the most promising conditions.

Approach: The proposed approach employs a BO framework to determine the optimal solvent combinations, compositions, and temperatures to maximize the solubility of organic molecules. It leverages a D-MPNN trained directly on mixture data, made possible by the recent availability of experimental solubility data [3]. To enable efficient BO, we represented the discrete solvent identifiers (SMILES) with three different encoding schemes: deep embeddings, numerical descriptors, and integer enumeration.

Results: The results show that our novel framework achieved higher solubility than the experimental baseline for more than 80% of the solutes on a test set comprising 66 solutes and 38 solvents across the temperature range 263–367 K (14,299 points). Moreover, the proposed solutions are chemically feasible and lie within a temperature range below the solvents' boiling points, ensuring that the resulting mixtures are stable.

Impact: In this work, we developed an optimization tool to reduce the cost and time of experimental campaigns, enabling rapid in silico solvent screening and allowing experimental efforts to focus on a small set of promising candidates for validation. While focused here on maximizing the solubility, the approach can be extended to crystallization or multistep processes. The framework is designed to be easily adaptable to any predictive model. Future directions include enabling optimization algorithms

that leverage the structure of the ANN model (e.g., gradient-based methods or mixed-integer linear programming) and experimental validation.

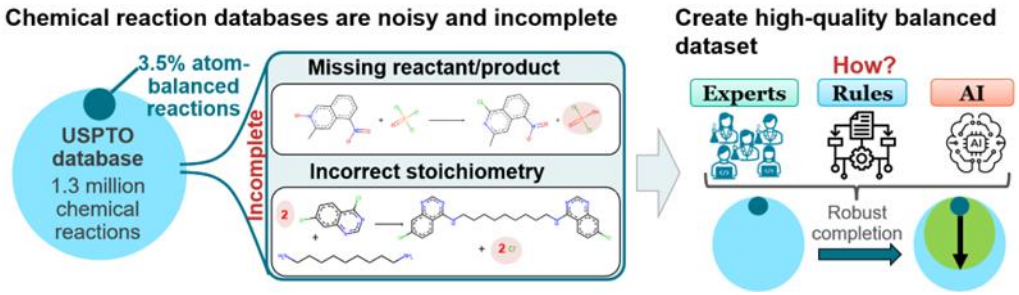
References

1. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, R. Barzilay, J. Chem. Inf. Model. 59 (2019) 3370–3388.
2. P.I. Frazier, arXiv:1807.02811 (2018).
3. D. Malikov, L. Krasnov, M. Kiseleva, E. Meshcheriakova, F. Kuznetsov, V. Elistratov, S. Bezzubov, ChemRxiv (2025), https://doi.org/10.26434/chemrxiv-2025-m51v8.

6 | 14:25-14:50 | CompleteRXN: Curation of Incomplete Chemical Reaction Databases

Gabriel Vogel and Jana Weber, Dept of Intelligent Systems, Delft University of Technology

Graphical abstract



Abstract

Context: Artificial intelligence models for chemistry rely heavily on large reaction datasets, most of which are mined from patent literature, e.g. the USPTO chemical reaction dataset [1]. While these datasets contain millions of reactions, they suffer from substantial noise and missing information, such as absent byproducts, co-reactants, or reagents. Atom-imbalanced reactions limit chemical interpretability, sustainability assessment, and reliable model training. Similar challenges arise in biochemical reaction modeling, where enzymatic and metabolic reaction datasets also require balanced representations to enable pathway analysis, metabolic engineering, and biocatalyst design. Despite growing interest in reaction completion and balancing methods [2,3], there is a lack of diverse, challenging, expert-validated ground truth datasets that reflect the complexity of real patent-derived reactions and can serve as reliable benchmarks for reaction completion methods.

Approach: We systematically analyze the extent and characteristics of incompleteness across commonly used USPTO-derived reaction subsets and cross-reference them with recently published balanced datasets. Based on this analysis, we design a principled selection and prioritization strategy to identify chemically diverse and challenging incomplete reactions. These reactions are curated by organic chemistry experts using an in-house web-based annotation application that supports reaction visualization, atom-balance checking, and optional patent lookup. Multiple independent expert annotations are used to ensure chemical correctness.

Preliminary Results: Our analysis shows that fewer than 10% of reactions in most widely used USPTO subsets are atom-balanced. Missing species range from small inorganic molecules to substantial organic fragments. Using our informed reaction selection and expert curation workflow, we are constructing a high-quality benchmark dataset of atom-balanced reactions that covers a broad range of reaction classes, templates, and incompleteness patterns.

Impact: Beyond existing balanced subsets, our expert-curated dataset will provide a challenging benchmark for systematically evaluating rule-based and AI-based approaches for reaction completion. This enables the selection of robust models and a clearer understanding of their reliability and confidence on realistic, noisy data. These insights support the development of a FAIR, high-quality, atom-balanced version of the USPTO reaction database and can be extended to biochemical reaction datasets. Ultimately, improving data quality at scale will strengthen reaction understanding, sustainability assessment, and the development of trustworthy AI tools for both synthetic and biological chemistry.

References

1. Lowe, D. M. (2012). Extraction of chemical structures and reactions from the literature (Doctoral dissertation).
2. van Wijngaarden, M., Vogel, G., & Weber, J. M. (2024). Completing Partial Reaction Equations with Rule and Language Model-based Methods. In Computer Aided Chemical Engineering (Vol. 53, pp. 3139-3144). Elsevier.
3. Phan, T. L., Weinbauer, K., Gärtner, T., Merkle, D., Andersen, J. L., Fagerberg, R., & Stadler, P. F. (2024). Reaction rebalancing: a novel approach to curating reaction databases. Journal of Cheminformatics, 16(1), 82.

14:50-15:10 Poster Pitches

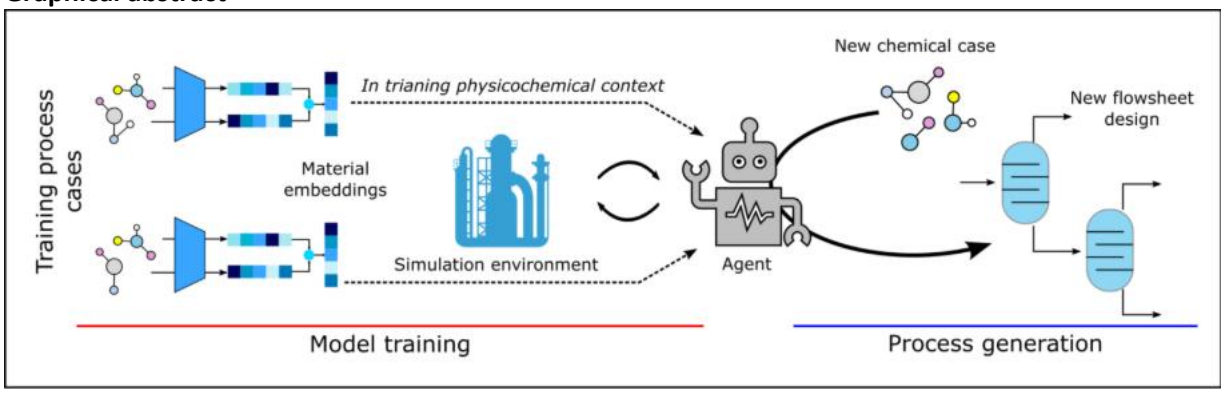
15:10-16:30 Coffee break + Posters

Session 3 Chair: Wouter van Winden

7 | 16:30-16:55 | Context-aware agent for process flowsheet synthesis

Ulderico Di Caprio and Artur M. Schweidtmann, Department of Chemical Engineering, Delft University of Technology.

Graphical abstract



Abstract

Context: Modern pharmaceutical industry relies on trains of unit operations to perform reaction, transformation, and separation tasks. Process design is a crucial phase of drug development that needs to be accelerated because of business and patients needs [1]. Creating optimal process systems is a complex task, particularly when dealing with feeds of varying composition and chemical nature of the components. In recent years reinforcement learning (RL) has emerged as a promising tool supporting such design effort [2,3], however existing approaches requires agent retraining to deal with new mixture, reducing employment scalability and slows down practical deployment. Thus, there is a need for RL-based methods that maintain design performance while generalising to unseen mixtures without additional training.

Approach: This work proposes a physics-informed context-aware RL framework based on Proximal Policy Optimisation (PPO) for automatic process generation, using purification train as case study. The flowsheet is represented as a process graph [3], where the RL agent selects the next unit operation, its placement within the graph, and associated design variables such as light/heavy keys, reflux ratio, and process targets. To overcome the retraining requirement in existing models, physical and thermodynamic information of the present components within the mixture is embedded directly into the graph representation. This enriched, mixture-invariant state space enables the trained agent to operate in inference mode when encountering new mixtures, without modifying model parameters.

Results: The proposed method successfully generates feasible multi-column distillation trains for mixtures of up to four components, meeting purity specifications and achieving competitive energy performance without any retraining. When tested on mixtures unseen during the training, the agent demonstrated strong generalisation capabilities and produced valid solutions without retraining. Compared to the non-context inform baseline RL methods, the novel approach improved robustness and reduced computational effort

Impact: In this work a context-aware RL agent for process generation was proposed, showing its potential to serve as a scalable and reusable tool for flowsheet synthesis in pharmaceutical

processing. By enabling inference-only usage of the agent for new mixtures, the approach reduces engineering time, computational cost, and barriers to industrial adoption. Future directions include extending the action space to additional unit operations, incorporating heat-integration decisions, and exploring transfer learning across broader separation tasks.

8 | 16:55-17:30 |

Kim van den Houten, Delft University of Technology
[Abstract will follow]

17:30 – 17:45 | Closing remarks

17:45 – 21:00 | Travel to Dinner location + Borrel + Dinner at Firma van Buiten

Day 2, April 14th, 9:00 – 17:00

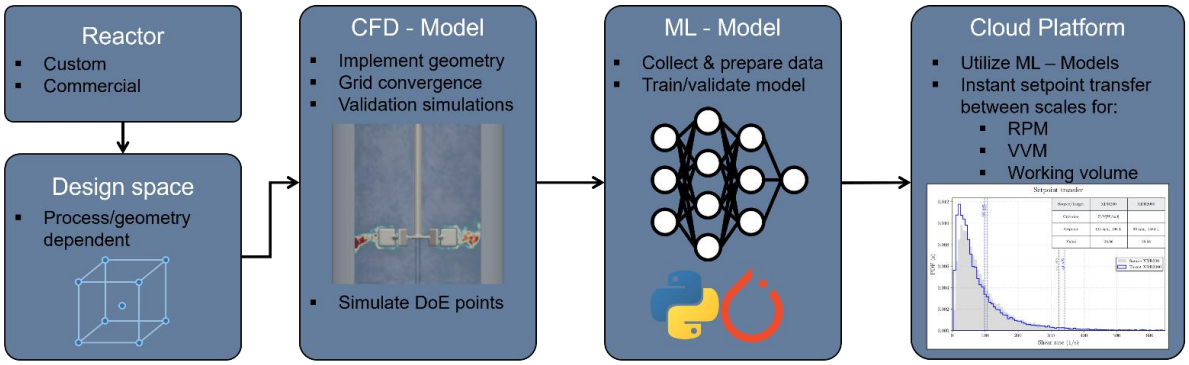
9:00– 9:30 | Arrive, register, coffee

Session 4 Chair: Renger Jellema

9 | 9:30-10:15 | Keynote: A digital platform for the Design, Control and Scale-Up of Bioprocesses

Lukas Gsenger[1,2], Philipp Eibl[1], Johannes Khinast[2,3], Christian Witz[1],
[1]SimVantage Gmbh
[2]Institue for Process and Particle Engineering, Graz University of Technology
[3]Research Center Pharmaceutical Engineering, Graz University of Technology

Graphical abstract



Abstract

Introduction: Bioprocesses play a central role in the production of advanced therapeutics and food ingredients. The systems for this production typically involve upstream operations (e.g., bioreactors) and downstream processing (harvesting, separation, filtration, purification), each with its unique technical challenges. Historically, scale-up of these unit operations from lab to commercial volumes has relied largely on empirical heuristics or design of experiments (DoE)-based experimentation, imposing significant cost, time and resource burdens. In contrast, physics-based numerical simulations can offer valuable insights into phenomena like homogenization kinetics, mechanical stress and mass transfer phenomena. However, up to this day high-fidelity simulations, especially for wide operating parameter spaces, remain time- and resource-intensive. [1] Thus, they cannot support the day-to-day workflow of bioengineers.

Physics Informed Scale-up: Traditionally, scale-up strategies for stirred tank bioreactors rely on empirical experience or approaches like keeping easily measurable macroscopic quantities—such as overall volumetric power input or impeller tip speed constant over different scales. However, these approaches often fail to capture the true physical environment within a given reactor. As demonstrated by Alavijeh et al. [1], this can produce significantly different shear fields, mixing regimes and/or gas–liquid mass-transfer behavior. Physics-based simulations can provide information on the spatial and timely distributions of otherwise hardly measurable quantities like shear stress, energy dissipation or local DO concentrations which represent the microstates experienced by cells much more closely. Utilizing these descriptors as process transfer criteria enables scale-up grounded in improved physical equivalence, ultimately improving robustness and process transferability.

Furthermore, scaling the stirrer speed based on this statistical value considers the actual mechanical stress cells are exposed around the blade tips, without dependence on metrics that are related to the dimensions of the tank geometry.

Digital Platform Architecture: In our work we present a digital platform that integrates mechanistic modelling, AI-based surrogate models and optimization workflows, to enable rapid scale-up, design and operational-point transfer across bioprocess scales based on the described criteria resembling the physical state within an investigated system. At its core, validated high-fidelity simulation data are used to train surrogate models. These surrogate models can then be embedded within optimization or control algorithms to explore the design and operating space of a target reactor at any scale, generating candidate conditions that mirror the performance of a lab–scale reference system but at production scale. Moreover, such tools can act as soft sensors embedded in a digital platform for intelligent, self-optimizing production setups. The digital platform presents an integrated user-interface for process engineers, enabling design-space exploration, creation of DoE-like parameter sweeps with subsequent creation of custom surrogate models, and ultimately generation of operating-point suggestions for scale-up. The general workflow for the integration of a unit operation into this platform is displayed with the example of a stirred tank bioreactor the graphical abstract. Each added unit operation undergoes the same basic principles in the workflow. This includes definition of geometry and design space, geometry implementation, validation and parameter sweep. Once the mechanistic data representing the design space is created and deemed reliable, the machine learning framework is utilized to link key process parameters, distributed and/or single-value metrics, to changeable inputs defined by the design space. The resulting model can provide high-fidelity predictions without expensive mechanistic simulations once it is trained. This opens the possibility for utilization in process transfers between scales as well as control procedures or definition of design guidelines.

Conclusion: This proposed methodology has been validated for stirred-tank bioreactor systems and other downstream processes including lyophilization and filtration [2-4], demonstrating a significant reduction in experimental overhead. In principle the extension to other unit operations and process steps is straight forward once validated mechanistic models are available. We argue that such digital platforms which combine mechanistic insight, AI-based surrogate modelling and optimization represent a promising pathway to accelerate bioprocess industrialization, optimizing product yield and quality, reduce risk in scale-up, and increase robustness of manufacturing processes.

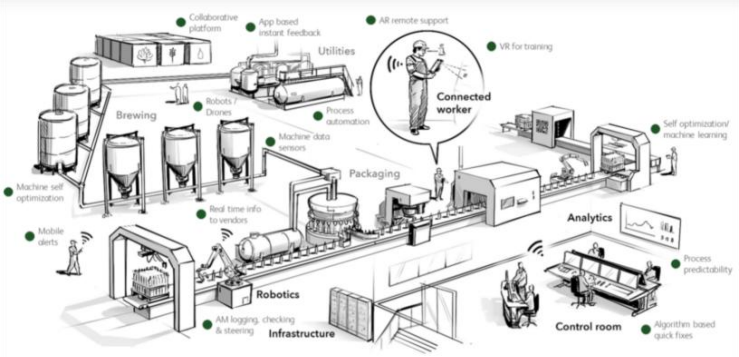
References

1. M. Karimi Alavijeh, I. Baker, Y. Y. Lee, and S. L. Gras, “Digitally enabled approaches for the scale up of mammalian cell bioreactors,” Digital Chemical Engineering, vol. 4, p. 100040, Sept. 2022, doi: 10.1016/j.dche.2022.100040.
2. L Gsenger, C Witz, P Eibl, J Khinast, „A combination of machine learning and CFD simulations for enhanced scaling of bioreactor operating conditions” – to be submitted to Biotechnology & Bioengineering.
3. Fruhwirth, M Wagner, P Eibl, C Witz, J Khinast, “The Impact of Lattice-Boltzmann Method Velocity Discretization Stencils on Symmetry and Accuracy: Applications from Pipe Flow to Stirred Tanks” – under revision at Chemical Engineering.
4. Zadravec et al., “Towards a digital twin of primary drying in lyophilization using coupled 3-D equipment CFD and 1-D vial-scale simulations,” European Journal of Pharmaceutics and Biopharmaceutics, vol. 208, p. 114662, Mar. 2025, doi: 10.1016/j.ejpb.2025.114662.

10 | 10:15-10:40 | From Data to Draft: ML from Lab to Logistics

Jurgen Nijkamp, Heineken

Graphical abstract



Source: https://www.rsm.nl/discovery/2024/how-heineken-wins-with-ai/

Abstract

Context: Artificial intelligence and machine learning are increasingly central to how global organizations connect strategy, operations, and innovation. For a global brewing company such as HEINEKEN, the challenge is to embed AI/ML solutions coherently across the value chain in a way that connects data, people, and decision-making. This presentation is motivated by the ambition to become the *best connected brewer* and by the need to understand how AI and ML can support this ambition.

Approach: The presentation will provide an overview on the use of AI/ML at HEINEKEN. It outlines the strategic context for digital and analytics initiatives and describes how AI/ML is implemented across different domains. This includes applications in commercial, supply chain and research & development areas, where data-driven methods are increasingly integrated with scientific workflows to support understanding, experimentation, and innovation.

Results: The presentation provides an overview of how AI and ML are applied across the organization and the types of problems they are used to address. Examples illustrate how data-driven models support areas such as promotions, sales, logistics, brewery performance, and R&D activities.

Impact: AI and ML can act as enablers across the value chain. The insights shared aim to inform on how AI is adopted within HEINEKEN.

References

https://www.forbes.com/sites/garystoller/2024/10/21/artifical-intelligence-may-be-transforming-the-brewing-industry/
https://www.theheinekencompany.com/newsroom/innovative-ai-that-transforms-connections/
https://www.rsm.nl/discovery/2024/how-heineken-wins-with-ai/
https://www.theheinekencompany.com/newsroom/connected-brewery--simplifying-and-automating-our-end-to-end-business/

10:40-11:00 Break

Session 5 Chair: Hans Roubos

11 | 11:00 -11:35 | Askara: A Multi-agent GenAI Assistant for Research

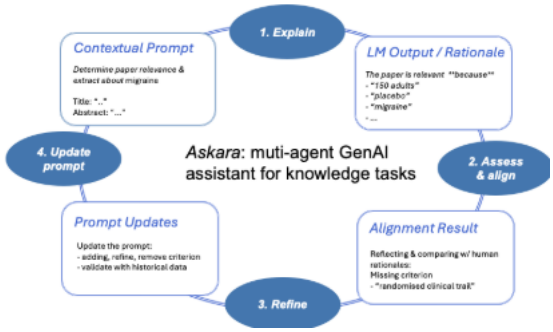
Shreyan Biswas[1], Madalena Barroso Gomes Pereira Plácido[1,2], Zheng Zhao[2], Wouter Touw[2], Anne Arzberger[1], Adrian Kuiper[1], Noor Bruijn[3], Antoinette Maassen van den Brink[3], Ujwal Gadiraju[1], Jie Yang[1]

[1]Delft University of Technology, Delft, Netherlands

[2]dsm-firmenich, Delft, Netherlands

[3]Erasmus Medical Center, Rotterdam, Netherlands

Graphical abstract



Abstract

Context: Large Language Models (LLMs) are increasingly positioned as research assistants, capable of supporting tasks such as literature review, knowledge synthesis, and experimental planning. However, in high-stakes domains like health or biotech, their practical adoption remains limited due to persistent robustness issues, including hallucinations, inconsistent reasoning, and lack of transparency. These issues stem from deeper challenges: LLMs encode incomplete and noisy representations of knowledge, and lack the ability to make context-sensitive value judgments about what constitutes "correct" or "useful" output. Addressing robustness is therefore critical for enabling trustworthy AI systems that can reliably support scientific work.

Approach: This talk presents a human-centered approach to robust AI for research assistance, combining diagnostic system design with human-in-the-loop methodologies. First, we introduce a white-box framework for systematic literature review that enables LLMs to iteratively refine their behavior using internal signals—confidence, rationale, and knowledge alignment—rather than relying solely on external performance metrics. This shifts prompt optimization from a black-box to a diagnostic process. Second, we extend this approach to biotech settings through a multi-agent architecture being developed with dsm-firmenich, where LLM agents interact with domain experts to elicit contextual knowledge and align reasoning with experimental goals. Together, these contributions go beyond the state of the art by integrating transparency, self-diagnosis, and human guidance into the core of LLM-based research systems.

Results: The framework improves both reliability and transparency, achieving higher overall performance while maintaining stable, high recall—critical for high-stakes screening tasks. The knowledge alignment mechanism acts as a safeguard against missed relevant studies, while diagnostic signals enable targeted corrections of model behavior. The proposed architecture significantly reduces literature search time from weeks to hours and supports complex research

workflows through interactive, agent-based collaboration with scientists. We observe that LLMs benefit substantially from structured interaction with human expertise.

Impact: These findings suggest that robustness in AI for research assistance cannot be achieved through model improvements alone. Instead, it requires rethinking AI systems as socio-technical systems in which humans play an integral role in knowledge specification and value alignment. For both academia and industry, this implies a shift toward hybrid intelligence: combining LLM capabilities with human expertise to build reliable, context-aware research tools. Future work will focus on scaling these approaches, including richer modeling of human cognitive processes and broader application in domains such as biotech R&D and scientific discovery.

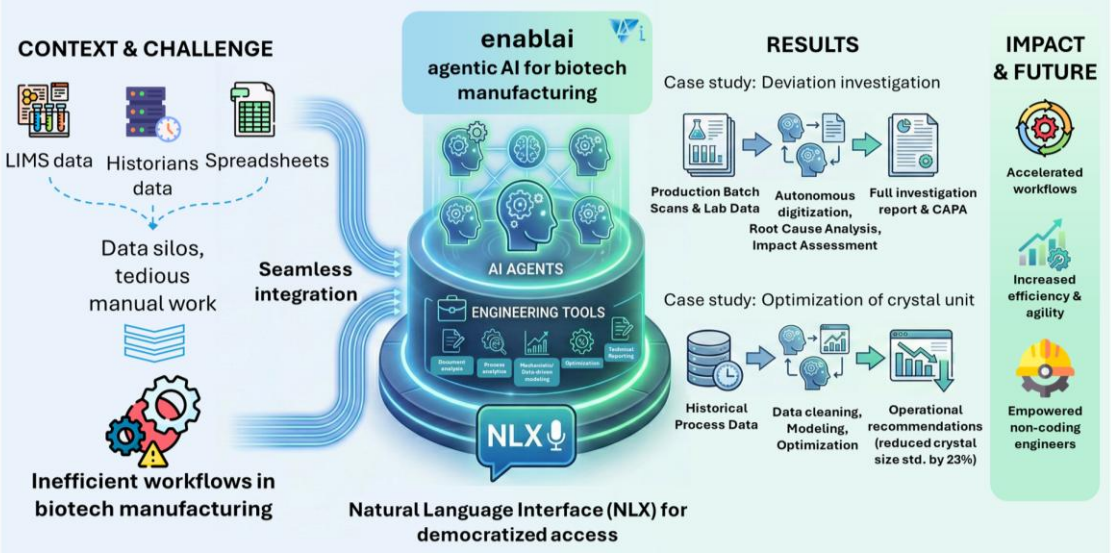
References

- R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. arXiv:2305.03495. https://arxiv.org/abs/2305.03495
- X. Tang, X. Wang, W. X. Zhao, S. Lu, Y. Li, and J.-R. Wen. 2025. Unleashing the Potential of Large Language Models as Prompt Optimizers: Analogical Analysis with Gradient-Based Model Optimizers. arXiv:2402.17564. https://arxiv.org/abs/2402.17564
- Y. Wu, Y. Gao, B. B. Zhu, Z. Zhou, X. Sun, S. Yang, J.-G. Lou, and Z. Ding. 2024. StraGo: Harnessing Strategic Guidance for Prompt Optimization. arXiv:2410.08601. https://arxiv.org/abs/2410.08601
- Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. 2023. Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651. https://arxiv.org/abs/2303.17651
- L. Mei, J. Yao, Y. Ge, Y. Wang, B. Bi, Y. Cai, J. Liu, M. Li, Z.-Z. Li, D. Zhang, et al. 2025. A Survey of Context Engineering for Large Language Models. arXiv:2507.13334. https://arxiv.org/abs/2507.13334

12 | 11:35 –12:00 | Automating workflows across biotech manufacturing with agentic AI

Giacomo Lastrucci[1], Maximilian F. Theisen[1], Gabrie Meesters[2] and Artur M. Schweidtmann[1]
[1]Process Intelligence Research, Dept. Chemical Engineering, TU Delft
[2]Product and Process Engineering, Dept. Chemical Engineering, TU Delft

Graphical abstract



Abstract

Context: Generative Artificial Intelligence (GenAI) and agentic AI are revolutionizing industries by speeding up manual workflows, for instance enabling software engineers to accelerate code development, or creatives to iterate drafts rapidly [1]. Biotech manufacturing, e.g., food and pharma, could benefit from generative and agentic AI by streamlining oftentimes tedious and labor-intensive workflows. Common examples include the integration of data across platforms and data silos, documentation of quality deviations, or multivariate data analysis [2]. Despite the potential, the rate of adoption of GenAI within biotechnology manufacturing remains low [3], partly because purpose-built, vertically integrated platforms tailored to biotech manufacturing are lacking.

Approach: We propose enablai, a novel agentic AI platform for biotech manufacturing. enablai integrates into existing data infrastructures to bridge data silos across lab information management systems (LIMS), historians, and spreadsheets. The platform leverages specialized subagents to automate workflows by incorporating engineering tools, including quality document revising, process data analysis, mechanistic/data-driven modeling and optimization, and technical reporting. The platform is accessed via a natural language experience (NLX) interface.

Results: We evaluate the platform's capabilities through two case studies: (1) conducting a deviation investigation and (2) optimizing a crystallization unit. In the first case study, the AI agents are provided with production batch records and laboratory data showing quality deviations. We show that the agents autonomously digitize and identify deviations in the records to provide a full investigation report including root cause analysis, impact assessment, and corrective & preventive actions (CAPA). (2) In the second case study, we ask the agents to model and optimize a crystallization unit from historical process data to reduce crystal size standard deviation. The agents leverage built-in engineering tools and reduce crystal size standard deviation by 23% through a three-step workflow: First, data are cleaned and analyzed; then, a data-driven model that predicts crystal size distribution is developed and trained; finally, a specialized agent formulates and solve an optimization problem to minimize crystal size standard deviation.

Impact: Our case studies demonstrate that agentic AI can help automating multi-step workflows in biotech manufacturing. However, critical limitations remain needed for human oversight to verify

outcomes, validation on controlled scenarios rather than industrial settings, and a limited number of tools, capabilities, and data silos that are connected to the platform. Future work could address human-in-the-loop design, regulatory compliance (e.g., GMP validation), and industrial validation to establish trust and robustness and to assess impact in manufacturing settings.

References

1. A. M. Schweidtmann, "Generative artificial intelligence in chemical engineering," Nature Chemical Engineering, vol. 1, p. 193–193, March 2024.
2. S. Rupprecht, Q. Gao, T. Karia and A. M. Schweidtmann, "Multi-agent systems for chemical engineering: a review and perspective," Current Opinion in Chemical Engineering, vol. 51, p. 101209, March 2026.
3. A. Challapally, C. Pease, R. Raskar and P. Chari, "The GenAI Divide: State of AI in Business 2025," 2025.

12.00-13:00 Lunch

Session 6 Chair: Ali May

13 | 13:00 -13:45 | Keynote: The human gut microbiome: variation, diagnostics and modulation

Jeroen Raes, VIB KU Leuven Center for Microbiology

Abstract: Alterations in the gut microbiota have been linked to various pathologies, ranging from inflammatory bowel disease and diabetes to cancer. Although large numbers of clinical studies aiming at microbiome-based disease markers and targets are currently being performed, we remain limited in our basic understanding about the normal variability and functioning of the human intestinal microbiota – and how to convert this knowledge in effective modulation strategies. Here, I will present our work in cross-sectional and longitudinal population cohorts investigating the variability of the microbiome, and its relationship with host physiology, diet and life habits. I will cover observational and interventional clinical studies that explore the potential of microbiome profiling as (companion) diagnostics. Finally, I show how a combination of human intervention trials and *in-vitro* microbiology allow breaking up the microbiome black box and open up the path towards targeted microbiota modulation.

14 | 13:45-14:10 | Reducing Heterogeneity in Cross-Study Microbiome ML with Data Attribution

Can Dedeköy, Ali May and Sergio Andreu-Sánchez, Science and Research, dsm-firmenich

Graphical abstract

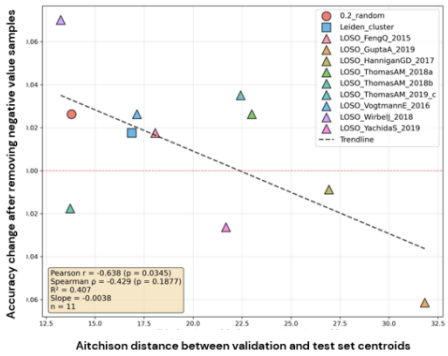


Figure 1. Removal of samples with negative data value leads to a larger increase in accuracy when the validation set resembles the test set. Each marker represents a distinct validation set used to compute data values and each set generates its own set of data values The plot shows the accuracy improvement obtained after removing negatively valued samples calculated on that specific validation set, plotted against the microbial community distance between the validation and test set centroids. The test set was kept constant and originated from a single cohort, ensuring that variation arises solely from the choice of validation set.

Abstract

Context: Cross-study machine learning (ML) models are commonly used in microbiome research to identify robust microbial patterns related to the host's health and disease. However, merging these datasets from distinct studies comes at the cost of introducing heterogeneity due to large differences in geography, lifestyle, and technical and biological factors, thus hampering model portability and interpretability. Here, we propose a data-driven approach to identify major sources of cross-study heterogeneity in microbiome-based ML models, thus enhancing model performance.

Approach: We propose using data attribution methods to guide data curation with the goal of reducing heterogeneity and improving ML model robustness. These methods assign each sample a "data value" based on its contribution to performance; negatively valued samples introduce misleading signal and removing them can improve accuracy. To our knowledge, these methods have not been applied to gut microbiome data, nor has the impact of different validation sets been examined. We evaluated several attribution approaches, Data Shapley1, Data Banzhaf2, and Data Outof-Bag (OOB)3, by measuring changes in held out test accuracy across multiple classifiers (LRridge, LRenet, Random Forest, XGBoost, MLP) when comparing full versus curated training sets. To study how validation set choice affects attribution, we used three strategies: LeaveOneStudyOut (LOSO), the Leiden-derived community closest to the test set in reduced space, and a random 20% subset of the training set.

Results: Training samples received different data values depending on attribution method, ML model, and validation set. DataOOB produced reproducible values across seeds and folds, but curation based on these values did not yield noticeable performance gains. Conversely, Data Banzhaf improved accuracy but showed limited reproducibility, while Data Shapley achieved both accuracy improvement and stable results. Further, LRridge and LRenet produced greater performance gains and

reproducibility than tree-based models. Data values estimated using a validation set biologically similar to the test set led to better generalization and improved post-curation performance (Fig. 1), a pattern observed for both LRridge and LRenet, with accuracy increases of up to 7–9%.

Impact: Understanding which data points the model learns from is crucial for understanding why certain samples are beneficial or detrimental in a specific context. This approach can be useful for cross-study ML models since, at relatively small sample sizes, the use-them-all approach may hurt model generalizability rather than helping it. Our results suggest that data attribution–based curation can improve the performance of cross-study microbiome ML models, motivating our next step of developing a robust, model agnostic pipeline.

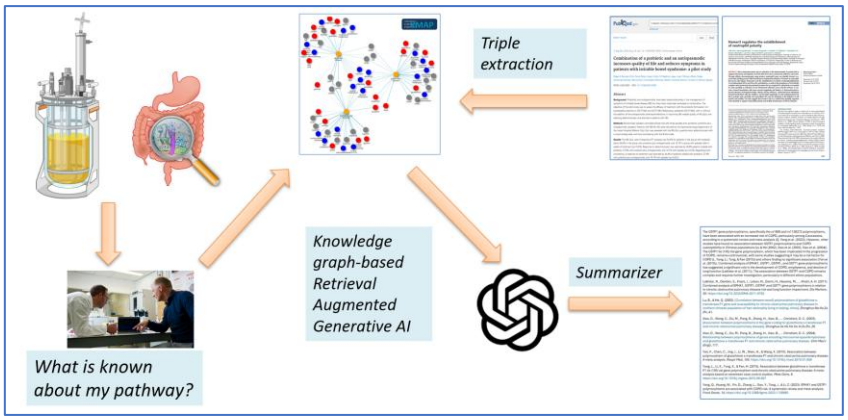
References

1. Ghorbani A, et al. Proceedings of the 36th International Conference on Machine Learning. PMLR; 2019.
2. Wang J, et al. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. PMLR; 2023.
3. Kwon Y, et al. Proceedings of the 40th International Conference on Machine Learning. PMLR; 2023.

15 | 14:10-14:35 | New Pathways: Reconstructing Microbial Metabolism from Literature Using NLP and Generative AI

Wynand Alkema[1], Michiel Noback[1], Martijn Herber[1], Harro Timmerman[2]
[1]Institute for Life Science, Engineering and ICT, Hanze University of Applied Sciences
[2]Microbiome Navigator BV

Graphical abstract



Abstract

Context: Large language models offer powerful capabilities for synthesizing complex biological knowledge, but when applied directly to novel metabolic pathway reconstruction they may hallucinate and lack transparent links to the primary literature. At the same time, curated pathway resources such as KEGG and Reactome, while authoritative, are limited to predefined canonical representations and cannot dynamically adapt to niche, organism-specific, or application-driven questions. This creates a clear need for a Retrieval-Augmented Generation (RAG) based framework

that can reconstruct metabolic pathways directly from the literature while preventing hallucination and maintaining explicit provenance.

Approach: We developed an NLP-based pipeline for automated pathway reconstruction from the primary scientific literature indexed against a controlled vocabulary of more than 500.000 keywords describing genes, metabolic pathways, diseases organisms. The pipeline extracts sentence level statements describing protein-pathway relations which are formalized via dependency parsing into structured triples, which are subsequently grouped into semantically equivalent triples. The extracted triples and sentences are fed into a Retrieval Augmented Generation (RAG) layer where relevant sentences are embedded using BioBERT and indexed via FAISS to facilitate fast look up with user defined questions. An LLM then synthesizes coherent pathway summaries based exclusively on the retrieved triples sentences, ensuring that generated text remains anchored to published evidence, creating a useful trade-off between purely database-driven and purely generative approaches for pathway reconstruction.

Results: We successfully reconstructed several metabolic pathways in bacteria and fungi from the primary literature alone. Reconstructed pathways were validated against KEGG and Reactome, demonstrating overlap with curated reference content while simultaneously revealing pathway components and organism-specific variants not captured in canonical databases.

Impact: This work demonstrates that flexible, literature-driven pathway reconstruction is both feasible and scientifically valid. The next addition to this framework will be the automated extraction of genomic sequences that correspond to these pathways and use these in design of pathway specific Hidden Markov Models and sequencing primers, which will allow us to identify and follow the presence and activities of these pathways in real life biological fermentations and ecosystems by high throughput sequencing.

14:35 - 15:35 Break (15 minutes) + Round table + Posters

Session 7 Chair:

16 | 15:35 - 16:00 | Beyond benchmarking: an expert-guided consensus approach to spatially aware clustering

Jieran Sun[1], Kirti Biharie[2,3], Peiying Cai[4], Niklas Müller-Bötticher[5], Paul Kiessling[6], Meghan A. Turner[7], Søren H. Dam[8,9], Florian Heyl[10,11], Sarusan Kathirchelvan[4], Martin Emons[4], Samuel Gunz[4], Sven Twardziok[5], Amin El-Heliebi[12], Martin Zacharias[13], SpaceHack 2.0 participants, Roland Eils[5], Marcel Reinders[3], Raphael Gottardo[1], Christoph Kuppe[6], Brian Long[7], Ahmed Mahfouz[2,3], Mark D. Robinson[4], Naveed Ishaque[5]

[1]Biomedical Data Science Center, Centre Hospitalier Universitaire Vaudois

[2]Department of Human Genetics, Leiden University Medical Center

[3]Delft Bioinformatics Lab, Delft University of Technology

[4]Department of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich

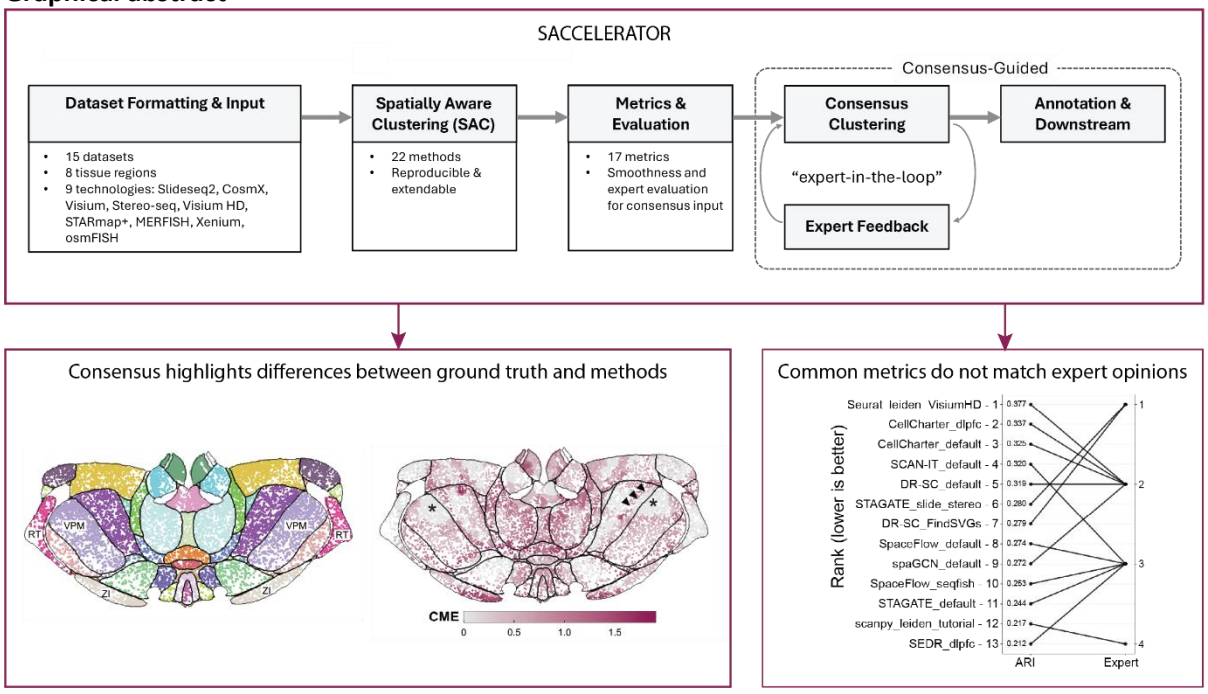
[5]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Digital Health

[6]Department of Nephrology, Rheumatology, and Clinical Immunology, University Hospital RWTH Aachen

[7]Allen Institute for Brain Science

[8]DTU Health Tech, Technical University of Denmark
[9]LEO Foundation Skin Immunology Research Center, Department of Immunology and Microbiology, University of Copenhagen
[10]German Cancer Research Center (DKFZ), Division of Computational Genomics and Systems Genetics
[11]The German Human Genome-Phenome Archive (GHGA)
[12]Division of Cell Biology, Histology and Embryology, Gottfried Schatz Research Center, Medical University of Graz
[13]Diagnostic and Research Institute of Pathology, Medical University of Graz

Graphical abstract



Abstract

Context: Spatial omics technologies have revolutionized the study of tissue architecture and cellular heterogeneity by integrating molecular profiles with spatial localization. In spatially resolved transcriptomics, delineating higher-order anatomical structures is critical for understanding how cellular organization affects tissue and organ function. Since 2020, more than 50 Spatially Aware Clustering (SAC) methods have been developed for this purpose. However, the reliability of current benchmarks is undermined by their narrow focus on specific sequencing platforms and brain tissue datasets, as well as incorrect interpretation of manual annotation as ground truth.

Approach: We present SACCELERATOR, a community-driven, extensible framework that standardizes data formatting, method integration, and metric evaluation, and is designed to rapidly incorporate new methods and datasets. SACCELERATOR currently includes 22 SAC methods applied to 15 datasets spanning 9 technologies and diverse tissue types. Rather than scoring and comparing methods, we

propose a consensus-guided workflow that aggregates clustering results to generate consensus representations. Descriptive spatial metrics highlight areas of high entropy where method disagreement is highest, enabling targeted feedback for tissue experts.

Results: Our analysis revealed substantial limitations in the generalizability and reproducibility of SAC methods across tissues and platforms. We also demonstrate that anatomical labels commonly used as ground truths are often biased, potentially error-prone, and, in some cases, unsuitable for benchmarking efforts. When applied to brain and cancer datasets with expert-in-the-loop evaluation, SACCELERATOR revealed biologically meaningful patterns that were previously overlooked, demonstrating that traditional evaluation metrics do not always reflect the subjective quality of results. Our meta-analysis of reported performances across studies revealed considerable inconsistencies, highlighting the unreliability of current benchmarking practices. Method performances varied widely between studies, even when using the same datasets and algorithms.

Impact: Our results underscore the need for iterative, expert-in-the-loop analysis and reveal that traditional evaluation metrics do not always capture the subjective qualities of results. By improving tissue annotation and addressing key benchmarking limitations, SACCELERATOR provides a robust foundation for advancing spatial omics research.

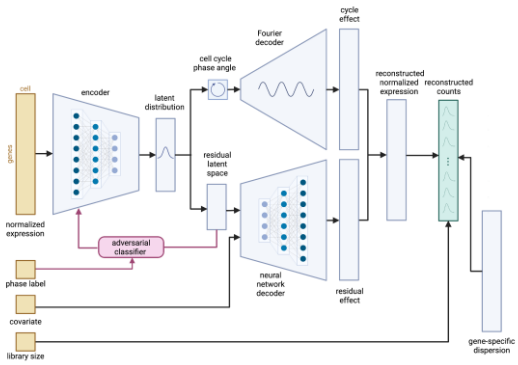
References

1. Sun, J. et al. Beyond benchmarking: an expert-guided consensus approach to spatially aware clustering. BioRxiv (2025) (Accepted in Nature Methods).

17 | 16:00-16:25 | CycleVI: generative model of the cell cycle

Pia Mozdzanowsk[1,2], Marcel Tarbier[3], Gustavo S. Jeuken[1]
[1]Systems biology lab, A-LIFE, Vrije Universiteit Amsterdam
[2]European Bioinformatics Institute (EBI-EMBL)
[3]Science for Life Laboratory, Uppsala University

Graphical abstract



Abstract

Context: Single-cell RNA sequencing (scRNA-seq) provides unprecedented resolution of cellular heterogeneity, yet the cell cycle often constitutes a dominant source of transcriptional variation. This high-amplitude signal frequently obscures subtle but critical biological insights regarding cell identity, differentiation state, or drug response. Current computational strategies to address this are limited: regression-based methods often discard biologically meaningful variation coupled with proliferation,

while state-of-the-art trajectory inference tools typically require specialized data (such as unspliced transcripts for RNA velocity) that are unavailable in most standard datasets. A method is needed to accurately infer continuous cell cycle phases from static data and separate this signal without data loss.

Approach: We present CycleVI, a deep generative model based on a variational autoencoder (VAE) framework, designed to disentangle cell cycle variation from other biological signals in static scRNA-seq data. Our approach goes beyond simple confounder removal by employing a principled architecture that isolates the cell cycle into a dedicated, partitioned latent subspace. CycleVI utilizes a dual-decoder system: a gene-specific Fourier series decoder captures periodic expression dynamics as a function of a learned cell cycle angle, while a standard neural network decoder models the remaining "residual" variation. To ensure robust disentanglement, we employ an adversarial classifier during training to prevent the residual space from encoding phase-related information.

Results: We validated CycleVI on six diverse datasets, demonstrating that it accurately infers continuous cell cycle phases from static expression data alone, showing strong concordance with orthogonal protein-level measurements (FUCCI) and velocity-based methods. Key findings include: Unmasking Heterogeneity: In colorectal cancer cells (HT-29), the disentangled residual latent space revealed a subpopulation driven by the unfolded protein response (UPR) that was previously masked by proliferative signals. Clarifying Differentiation: In hematopoietic progenitors, CycleVI resolved complex differentiation trajectories that were otherwise confounded by cell cycle genes. Spatial Mapping: Applied to Slide-seq data of a metastatic breast cancer biopsy, CycleVI successfully mapped proliferative activity in situ, clearly distinguishing cycling tumor regions from quiescent tissue.

Impact: CycleVI provides a robust, interpretable solution to the long-standing problem of cell cycle confounding, applicable to the vast majority of existing static scRNA-seq and spatial transcriptomics datasets. By isolating rather than removing cell cycle variation, CycleVI preserves the integrity of biological data, enabling the discovery of subtle cellular states and spatial proliferation patterns that other methods miss.

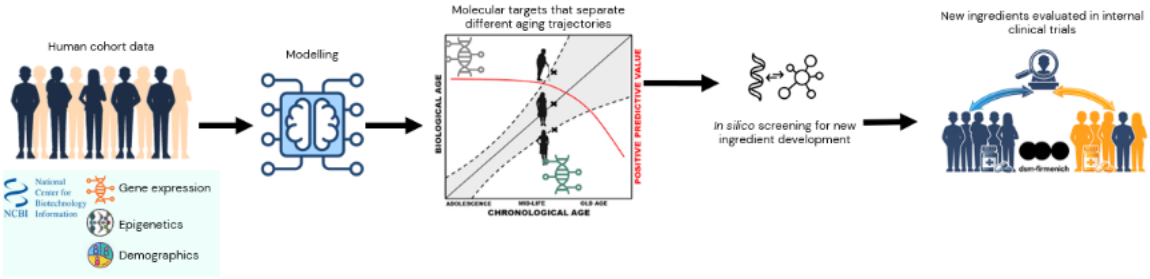
References

1. Mozdzanowski, Pia, Marcel Tarbier, and Gustavo S. Jeuken. "CycleVI: Isolating cell cycle variation with an interpretable deep generative model." bioRxiv (2025): 2025-11.

18 | 16:25 - 16:50 | Multi-omics cohort analysis for target discovery to derisk interventional clinical trials

Huremagic B[1], Stemmler R[1], Peter S, Nieto B[2], Civiletto G[2], May A[1],
[1]dsm-firmenich, Science & Research
[2]dsm-firmenich, Health, Nutrition, and Care

Graphical abstract



Abstract

Context: The development of effective healthy aging interventions is challenged by the difficulty of defining robust, human-relevant biological targets from large-scale omics data. Aging is a complex, polygenic process, and signals derived from molecular profiling are often heterogeneous, context-dependent, and difficult to translate into actionable intervention strategies. To address this challenge, there is a growing need for evidence frameworks that ground target discovery in population-scale human biology. Integrating biological aging clocks derived from molecular profiles such as methylation and transcriptomics data provides a framework to identify molecular signatures associated with healthy aging and to enable downstream computational prioritization of potential nutritional interventions.

Approach: We performed a multi-omics molecular cohort analysis integrating DNA methylation and transcriptomic data from healthy human adults. Biological aging clocks derived from both omics layers were used to stratify individuals into aging groups. Cross-study, cross-omics integration was applied to identify recurrent molecular signals, followed by pathway-level analyses to contextualize biological mechanisms associated with aging trajectories.

Results: Across methylome and transcriptome analyses, we observed consistent molecular patterns distinguishing slow and fast aging groups. Cross-omics comparisons revealed recurrent biological processes associated with biological aging status across multiple aging clocks. These results suggest that coordinated regulation of several molecular mechanisms may contribute to inter-individual differences in biological aging. Ongoing analyses focus on consolidating these patterns into reproducible molecular signatures to support downstream biological interpretation and in silico screening of candidate nutritional ingredients targeting prioritized pathways.

Impact: Our results demonstrate that large-scale human multi-omics cohort analyses combined with biological aging clocks can identify robust molecular signatures associated with healthier aging. By grounding target discovery in reproducible patterns observed across independent cohorts, this framework improves the interpretability and prioritization of aging-related molecular signals. Coupling these signatures with in-silico screening approaches enables systematic prioritization of candidate ingredients before progression into human interventional trials, thereby supporting trial design and reducing translational risk. As larger and more deeply characterized molecular cohorts become available, scaling this approach with advanced AI/ML analytics will further refine aging-associated pathways and support the systematic development of evidence-based healthy aging strategies.

References

1. Zhu, C., Wang, Y., Yang, X. et al. Multi-dimensional evidence from the UK Biobank shows the impact of diet and macronutrient intake on aging. Commun Med 5, 36 (2025). https://doi.org/10.1038/s43856-025-00754-5
2. Tessier, AJ., Wang, F., Korat, A.A. et al. Optimal dietary patterns for healthy aging. Nat Med 31, 1644–1652 (2025). https://doi.org/10.1038/s41591-025-03570-5

16:50 – 17:00 | Closing remarks

Hans Roubos, dsm-firmenich

17:00 – 18:00 | Borrel

Day 1 - 14:50 – 16:30 | Poster presentations

Multi-agent system for separation process design

Qinghe Gao, Sophia Rupprecht and Artur M. Schweidtmann, Process Intelligence Research Team, Department of Chemical Engineering, Delft University of Technology

Graphical abstract

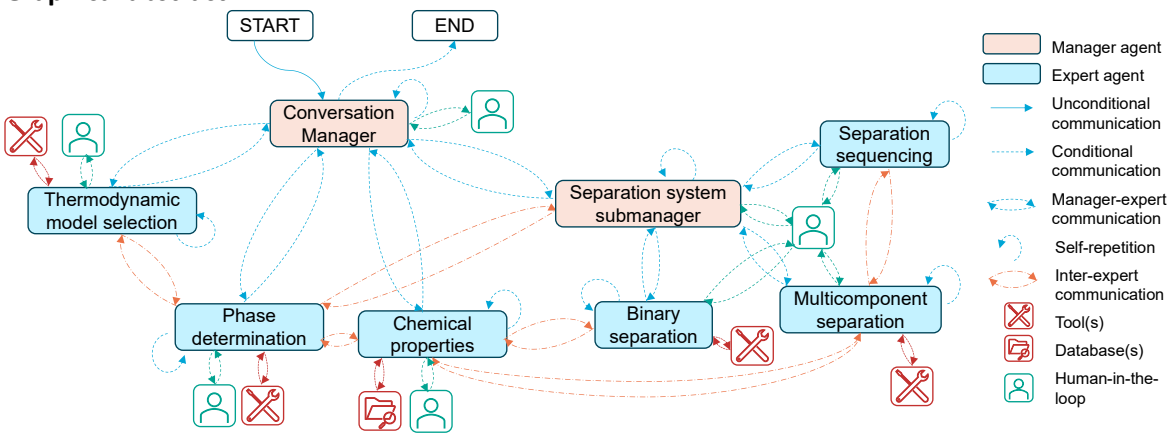


Figure 1: Multi-agent system architecture. A supervisor agent manages LLM-based expert agents dynamically. The expert agents have access to tool(s) or database(s) and can request advice from a human-in-the-loop. The MAS allows communication between a manager and its subordinate expert agents, but also inter-expert communication, depending on required information

Abstract

Motivation: Large language model (LLM)-based multi-agent systems (MASs) are a rapidly evolving and promising concept in the field of process systems engineering (PSE). MASs integrate specialized agents and aim to address complex tasks by a collaborative group effort. Larger tasks are subdivided into smaller subtasks and assigned to agents by expertise. Previous work has shown remarkable progress in using MASs in chemical and process systems engineering and adjacent domains [1,2]. In the PSE domain, MAS designs address tasks such as P&ID synthesis [3], process optimization [4], and operational decision-making [5]. Tool-enabled systems in experimental and computational chemistry, such as ChemCrow [6] and El Agente [7], have especially underlined the potential of MASs for an efficient user-MAS symbiosis.

Methods: We present a MAS that comprises the capabilities of thermodynamic model selection, phase determination, chemical property retrieval, and separation system exploration. The MAS aims at facilitating multi-step problem solving through dynamic agent interaction. Therefore, the agents are arranged in a semi-hierarchical architecture as displayed in Figure 1. LLM-based expert agents cover a specific field of expertise each and are equipped with access to tailored tools and databases. The conversation between the expert agents is managed by a supervisor agent. The supervisor agent receives the user's initial question, initiates and manages the expert conversation, and returns a final answer to the user. Each expert agent can additionally interact with the fellow expert agents, bypassing the supervisor agent. Each expert agent can consult a human-in-the-loop for advice or error-handling. The MAS is implemented using the LangGraph framework [8].

Results: We plan to assess the MAS shown in Figure 1 on a selection of case studies with increasing level of complexity. We hypothesize that the complexity of a task depends on the number of steps and agent interactions required to achieve a suitable response to a user question. Single-step tasks include property retrieval, or thermodynamic model selection for a given system of chemicals. Multi-step tasks include phase calculations without a known thermodynamic model, or separation method selection requiring prior phase calculations. We intend to specifically investigate the ability of the MAS to request and integrate human-in-the-loop feedback.

Impact: Our contribution investigates the potential and suitability of MASs for separation process design. The current design of the MAS targets separation method screening but will be expanded to separation process design based on the findings of this study. The results of this work are constrained to the scope of the case studies, thus limiting the ability to generalize from the conclusions drawn based on this study. Furthermore, the expertise of the individual agents is limited by their tool and database access. For instance, phase calculations currently neglect the presence of solids.

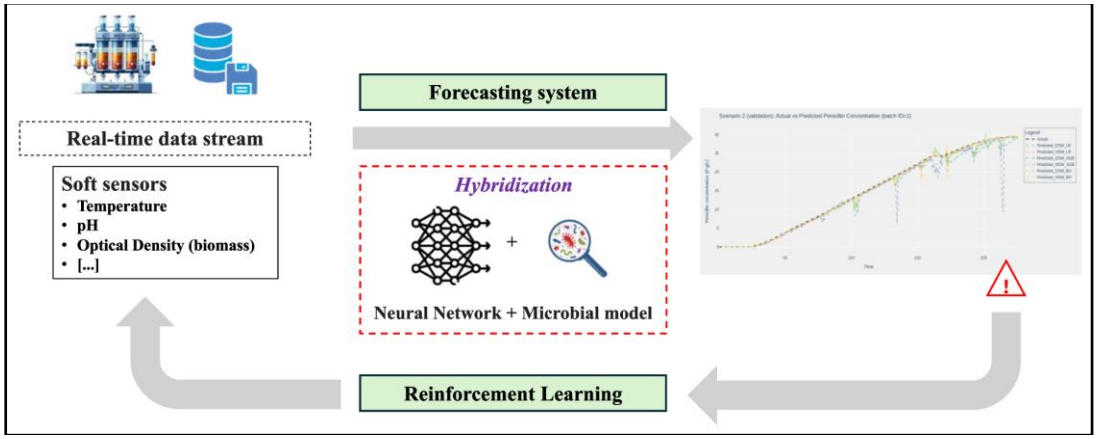
References

1. T. Woo et al., "Leveraging Generative AI and Large Language Model for Process Systems Engineering: A State-of-the-Art Review," Korean Journal of Chemical Engineering, Jul. 2025, doi: 10.1007/s11814-025-00524-y.
2. S. Rupprecht et al., "Multi-agent systems for chemical engineering: a review and perspective," Current Opinion in Chemical Engineering, vol. 51, p. 101209, Mar. 2026, doi: 10.1016/j.coche.2025.101209.
3. S. Gowaikar et al., "An Agentic Approach to Automatic Creation of P&ID Diagrams from Natural Language Descriptions," Dec. 2024, doi: 10.48550/ARXIV.2412.12898.
4. T. Zeng et al., "LLM-guided chemical process optimization with a multi-agent approach," Machine Learning: Science and Technology, vol. 6, no. 4, p. 45067, Dec. 2025, doi: 10.1088/2632-2153/ae2382.
5. E. Pajak et al., "Multi-Agent LLMs for Automating Sustainable Operational Decision-Making," in Proceedings of the 35th European Symposium on Computer Aided Process Engineering (ESCAPE 35), in ESCAPE 35, vol. 4. PSE Press, Jul. 2025, pp. 1824–1829. doi: 10.69997/sct.156776.
6. A. M. Bran et al., "Augmenting large language models with chemistry tools," Nature Machine Intelligence, vol. 6, no. 5, pp. 525–535, May 2024, doi: 10.1038/s42256-024-00832-8.
7. Y. Zou et al., "El Agente: An autonomous agent for quantum chemistry," Matter, vol. 8, no. 7, p. 102263, Jul. 2025, doi: 10.1016/j.matt.2025.102263.
8. "LangGraph." LangChain, Inc. [Online]. Available: https://www.langchain.com/langgraph.

Digital Twin for bioprocess control

Gommeren T[1], Martín-Martín I[1,2], Koehorst JJ[1], Metcalfe B[1], Martins dos Santos VAP[2], Suarez Diez M[1]
[1]Laboratory of Systems & Synthetic Biology, Wageningen University & Research
[2]Laboratory of Bioprocess Engineering, Wageningen University & Research

Graphical abstract



Abstract

Context: Biomanufacturing leverages microbial potential to substitute fossil fuel-based compounds with biobased and sustainable products. However, bioprocesses are complex requiring knowledge of, and controlling for, biotic and abiotic factors. Here, a software prototype is presented for bioprocess monitoring, prediction, and control (i.e. a Digital Twin), combining time-series forecasting and microbial mechanistic knowledge (Figure 1). This proof-of-concept Digital Twin will incorporate Reinforcement and Hybrid Learning to help integrate physical bioprocess modelling with knowledge on microbial genetics and strain physiology.

Approach: First, a time-series forecasting system (under development) anticipates the evolution of a variable of interest over time, through real-time readings of process inputs, i.e. 'soft sensors' [1, 2]. Secondly, a Reinforcement Learning implementation will allow the system to adjust input signals over time and suggest corrective actions to avoid deviations in desired trends. Finally, a genome-scale metabolic model (GEM) module will be incorporated to relay microbial mechanistic knowledge in the form of a "hybrid system" [3].

Results: A preliminary framework already in place allows predictions of bioreactor inputs through a real-time data stream (Pioreactor ®). Current work focuses on the design of: (i) a Reinforcement Learning implementation for real-time control, and (ii) alternatives to plug in microbial mechanistic models to build a hybrid system.

Impact: A hybrid mechanistic-genomic digital twin will allow for the genetic characteristics of the microbial workhorse to be considered as part of the process forecasting (Figure 1).This hybridization will reduce the amount of data required to obtain reliable predictions, facilitating the transferability of such framework to different (microbial and process) settings and reactor scales.

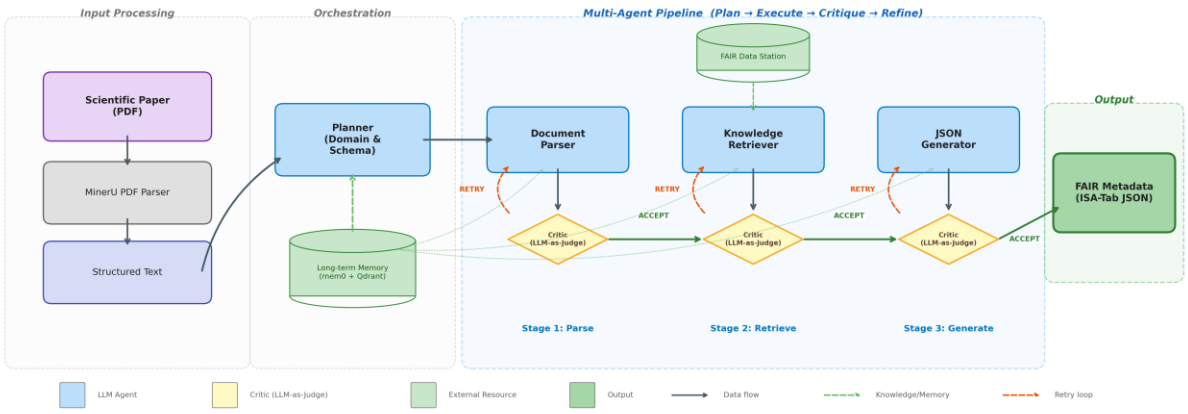
References

1. Acosta-Pavas JC, Robles-Rodriguez CE, Griol D, Daboussi F, Aceves-Lara CA, Corrales DC (2024). Soft sensors based on interpretable learners for industrial-scale fed-batch fermentation: learning from simulations, Computers and Chemical Engineering. doi: https://doi.org/10.1016/j.compchemeng.2024.108736.
2. Metcalfe, B., Acosta-Pavas, J. C., Robles-Rodriguez, C. E., Georgakilas, G. K., Dalamagas, T., Aceves-Lara, C. A., Daboussi, F., Koehorst, J. J., & Corrales, D. C. (2025). Towards MLOps soft sensor for real-time predictions in industrial-scale fed-batch fermentation. Computers & Chemical Engineering, Volume 194, DOI: https://doi.org/10.1016/j.compchemeng.2024.108991.
3. Faulon JL, Dursoniah D, Ahavi P (2025). dAMN: a genome scale neural-mechanistic hybrid model to predict bacterial growth dynamics. HAL-INRAe. https://hal.inrae.fr/hal-05233970v1.

FAIRiAgent: Agentic AI for Automated FAIR Metadata Generation in Biological Research

Changlin Ke, Jasper Koehorst, Michael Schon, and Maria Suarez Diez, Laboratory of Systems and Synthetic Biology, Wageningen University & Research

Graphical abstract



Abstract

Context: Life sciences research generates data far faster than it can be curated. A key bottleneck is transforming the unstructured information scattered across publications into FAIR (Findable, Accessible, Interoperable, Reusable) metadata [1]—without which datasets remain siloed and downstream AI-driven analyses lack essential context. Manual curation of a single paper can take hours and is error-prone at scale. While recent LLM-based extraction tools can parse short texts in a single pass [4], they lack the iterative reasoning needed for full-length scientific documents with complex methods sections, nested tables, and domain-specific terminology. We ask: can a team of collaborating AI agents match a human curator's ability to produce publication-ready metadata from complete papers?

Approach: We developed FAIRiAgent, a multi-agent LLM system built on LangGraph that implements a Plan–Execute–Critique–Refine loop. Given a PDF, FAIRiAgent first converts it to structured text (MinerU), then orchestrates a pipeline of specialized agents—a Planner selects the appropriate MIxS checklist [2], a Document Parser extracts field values, a Knowledge Retriever grounds terms against

FAIR Data Station [3] and ontologies, and a JSON Generator produces ISA-Tab compliant metadata. At every stage, a Critic Agent (LLM-as-Judge) evaluates the output for schema compliance and evidence coverage, deciding to ACCEPT or RETRY with targeted feedback. A vector-database-backed memory layer (mem0 + Qdrant) enables cross-session learning. Unlike single-pass approaches, FAIRiAgent processes complete papers - methods, tables, and supplements—through multiple rounds of self-correction.

Results: As a proof of concept, we evaluated FAIRiAgent on two biological manuscripts (metagenomics, whole-cell biosensor) with eight LLM backends across three provider families (OpenAI, Anthropic, Qwen), totalling over 160 agentic runs plus 30 single-prompt baseline runs for comparison.

- **Agentic self-correction is essential**. Single-prompt baselines—even with GPT-5.1—achieved roughly half the mandatory field coverage and never met the 100% mandatory criterion. The best agentic model (Qwen-Max) achieved an 80% success rate for fully publication-ready metadata, reaching 100% on the metagenomics document ($p < 0.001$, Fisher's exact test) — reducing a task that takes a curator hours to under ten minutes.
- **Practical reliability via pass@k**. Under moderate quality thresholds, top models reach near-certain success within a small number of independent attempts (pass@5 ≈ 1.0), making FAIRiAgent suitable for human-in-the-loop curation where a curator selects the best candidate.
- **Choice of LLM backbone matters**. Success rates range from 80% (cloud API) to 0% (locally-hosted open model), but even lower-performing models correctly identify the right metadata schema — suggesting the workflow is sound and will improve as models do.
- **Domain understanding is robust**. All models correctly selected the appropriate MIxS metadata packages, indicating the bottleneck is comprehensive field-level extraction, not schema identification.

Impact: FAIRiAgent demonstrates that collaborative AI agents with iterative self-correction can automate FAIR metadata curation—a task that currently requires hours of manual expert effort per paper. The key lesson is that single-prompt LLM approaches, even with frontier models, are insufficient; the Plan-Execute-Critique-Refine loop is what bridges the gap to publication-ready quality. By lowering the barrier to FAIR compliance [5], FAIRiAgent enables scalable bio-curation of legacy and incoming literature, providing richer sample-level context for downstream AI analyses such as metabolic modeling and functional annotation.
Next steps: expand evaluation to a multi-domain benchmark corpus (50+ documents), integrate user feedback loops, and release FAIRiAgent as an open-source tool for the community.

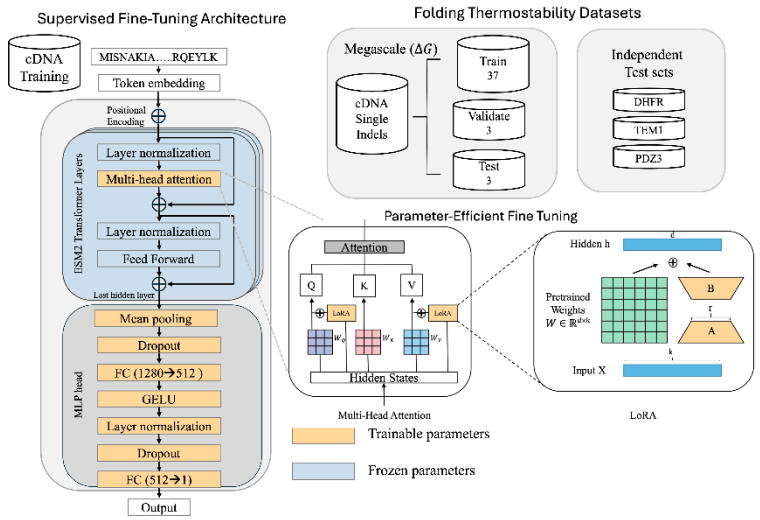
References

1. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016).
2. Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415–420 (2011).
3. Nijsse, B., Schaap, P. J. & Koehorst, J. J. FAIR data station for lightweight metadata management and validation of omics studies. GigaScience 12, giad014 (2023).
4. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. Nat. Commun. 15, 1418 (2024).
5. Hughes, L. D. et al. Addressing barriers in FAIR data practices for biomedical data. Sci. Data 10, 98 (2023).

Improving Protein Indel Stability Prediction via Efficient Fine-tuning of a Protein Language Model

Minxing Wang and Amelie Stein, Department of Biology, University of Copenhagen

Graphical abstract



Abstract

Context: Understanding the effect of protein variation on protein folding stability (ΔG) is crucial for protein evolution, disease etiology, and engineering. While the consequences of single amino acid substitutions are well explored, short insertions or deletions (indels) represent a more complex class of variation that could profoundly alter protein backbone structures and possess high-risk, high-reward potential for applications [1]. Computationally predicting the stability effects of indels, however, remains a significant challenge as existing predictors often struggle with indels due to the historical scarcity of experimental data and most traditional variant effect predictors explicitly support only missense variants.

Approach: To address the complex backbone alternations, limited predictor support for indels and data-intensive modelling requirments, this work presents SFT-ESM2, a lightweight, data-efficient framework that incorporates parameter-efficient fine-tuning (PEFT), specifically Low-Rank Adaptation (LoRA), on the pre-trained protein language model (PLM) ESM2 [2, 3]. Updating low-rank weight matrices parallel to the frozen pre-training backbone (weights), LoRA enables an efficient way to fine-tune large language models, with updates equivalent to 0.2% of full-size fine-tuning. This work is specifically tuned to predict and rank the effects of short amino acid indels on folding stability using a subset of approximately 6,000 indel variants from a large-scale in vitro folding thermodynamic stability dataset [4].

Results: The fine-tuned model outperformed the native ESM2 backbone (650M) and demonstrated a competitive performance to state-of-the-art zero-shot PLMs, inverse folding models, structure-based metrics (confidence scores), and Rosetta modelling. Beyond the held-out datasets, the model generalises to three independent deep mutational scanning (DMS) datasets (86.7% increase over the native ESM2 model), which contain unseen proteins and diverse indel types. Further studies confirmed its robustness to data depletion, maintaining high predictive ability even when training data reduced by 60%.

Impact: This work explores the utility of pre-trained language models for predicting amino indel-induced protein stability changes, mitigating an important gap in current protein engineering tools. Adapting the efficient fine-tuning method, it demonstrates a data-efficient approach to mitigating overfitting in data-scarce scenarios. While fine-tuning narrows the gap, models still face challenges in estimating the absolute Gibbs free energy and identifying stabilising variants. Therefore, there is an urgent need to construct symmetric indel effect datasets. Interpretability also remains a key area of study, as understanding how models make decisions about discrimination and prediction will clarify the alignment between PLMs and biophysical functions.

References

1. Gerasimavicius, L., B.J. Livesey, and J.A. Marsh, Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. Protein Sci, 2023. 32(7): p. e4688.
2. Hu, E.J., et al., Lora: Low-rank adaptation of large language models. ICLR, 2022. 1(2): p. 3.
3. Lin, Z., et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023. 379(6637): p. 1123-1130.
4. Tsuboyama, K., et al., Mega-scale experimental analysis of protein folding stability in biology and design. Nature, 2023. 620(7973): p. 434-444.

Multi-Label Node Classification in Biological Graphs

Tianqi Zhao[1], Ngan Dong[2], Alan Hanjalic[1], Megha Khosla[1]

[1]Department of Intelligent Systems, Delft University of Technology

[2]L3S Research Center, Hanover

Graphical abstract

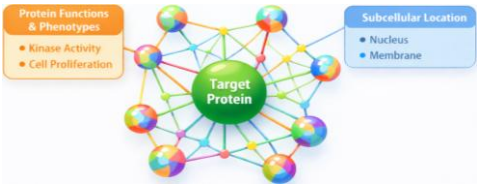


Figure 1: example of biological multi-label graphs

Abstract

Introduction: Graph Neural Networks (GNNs) have set the state of the art for node classification on graphs, yet this progress has largely been measured under a single-label scenario. In many real applications, especially in biological graphs, nodes naturally have multiple labels: for instance, in protein–protein interaction networks (Figure 1), a protein can have several molecular functions or be linked to multiple disease phenotypes.

We start by analysing the existing datasets for multi-label node classification which already pointed to severe data quality issues. As an example, the OGB-Proteins dataset in the Open Graph Benchmark (OGB)[1], has around 90% of the nodes unlabelled in the test set. The lack of labels in the test set combined with the use of the Area Under the ROC Curve (AUROC) metric leads to overly exaggerated performance scores on the leaderboard.

Furthermore, while GNNs typically excel on homophilic graphs, where neighbouring nodes share the exact same label, neighbours in multi-label graph may only overlap on a small subset of labels [2]. As a result, multi-class methods built around homophily/heterophily assumptions do not generalize well to multi-label graphs.

Finally, real-world graphs are often generated in a streaming manner. For example, protein–protein interaction networks evolve as new interactions are discovered and existing proteins receive additional functional or phenotypic annotations over time [3]. This motivates learning paradigms in which a deployed model is updated continuously from sequential graph snapshots while preserving previously acquired knowledge.

Approach We perform a data-centric analysis in both static and continual graph learning setting. Firstly, we analyse the properties of multi-label graph datasets—examining label distributions and label-induced similarities—and quantify how these properties affect model performance. Secondly, we curate three biological graph datasets and introduce a synthetic multi-label graph generator with controllable properties, enabling rigorous and reproducible comparisons. Thirdly, we conduct a large-scale study evaluating eight representative methods across nine datasets, finding that simple baselines (e.g., DeepWalk) can outperform more sophisticated GNNs on several benchmarks. Finally, we develop a generalized CGL evaluation framework applicable to both multi-class and multi-label node classification (and readily extensible to multi-label graph/edge classification).

Results and Discussion: Our results on static multi-label node classification, summarized in Table 1, indicate that existing techniques are still insufficient: simple baselines such as DeepWalk often outperform GNNs on several real-world (and most synthetic) datasets, while methods designed for low-homophily multi-class graphs (e.g., H2GCN) and multi-label-specific approaches (e.g., LANC) yield only modest or inconsistent improvements.

Table 1: Mean performance scores (Average Precision) on real-world datasets.

Method	BlogCat	Yelp	OGB-Proteins	DBLP	PCG	HumLoc	EukLoc
Mlp	0.043	0.096	0.026	0.350	0.148	0.170	0.120
DeepWalk	**0.190**	0.096	0.044	0.585	**0.229**	0.186	0.076
LANC	0.050	OOM	0.045	0.836	0.185	0.132	0.062
Gcn	0.037	0.131	**0.054**	**0.893**	0.210	**0.252**	**0.152**
Gat	0.041	0.150	0.021	0.829	0.168	0.238	0.136
GraphSage	0.045	**0.251**	0.027	0.868	0.185	0.234	0.124
H2Gcn	0.039	0.226	0.036	0.858	0.192	0.172	0.134
GCN-LPA	0.043	0.116	0.023	0.801	0.167	0.150	0.075

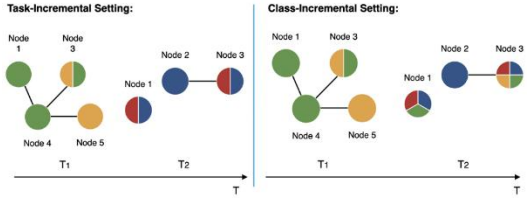


Figure 2 illustrates our two generalized incremental settings for both multi-class and multi-label node classification, characterized by the node overlap across tasks in the task-incremental setting and the expansion of the label set over time in the class-incremental setting

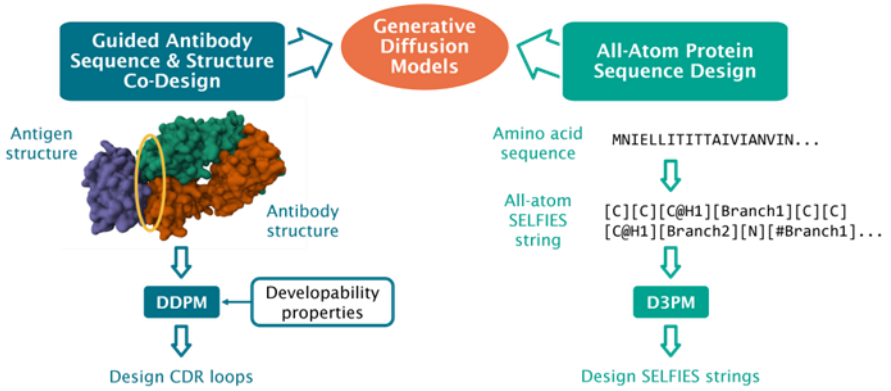
References

1. W.Hu, et al. Open graph benchmark: Datasets for machine learning on graphs. arXiv:2005.00687, 2020.
2. T.Zhao, et al. Multi-label Node Classification on Graph-Structured Data (MLGNC). https://openreview.net/forum?id=EZhkV2BjDP.
3. T.Zhao, et al. AGALE: A Graph-Aware Continual Learning Evaluation Framework. arxiv.org/pdf/2406.01229.

From loops to atoms: Diffusion models for immunoprotein design

Amelia Villegas-Morcillo, Gijs J. Admiraal, Marcel J. T. Reinders, and Jana M. Weber, Dept Intelligent Systems, Delft University of Technology

Graphical abstract



Abstract

Context: Generative protein design has advanced rapidly with diffusion models that can jointly reason over sequence and structure. However, most existing methods either focus on structure design with limited property awareness, or on sequence-level generation restricted to the 20 canonical amino acids. This limits their applicability to immunoproteins, where therapeutic success depends not only on binding and structure, but also on developability properties and atomic-level chemical modifications such as post-translational modifications.

Approach: We present two complementary diffusion-based frameworks that expand the design space for immunoproteins. First, we introduce a guided diffusion model strategy [1] for antibody design that jointly generates CDR sequences and backbones, conditioned on antigen structure [2]. During generation, we guide sampling toward improved developability properties, optimizing hydropathy and predicted folding energies. Second, we propose an all-atom discrete diffusion model for full protein sequence generation [3] using the SELFIES chemical representation [4] and the discrete diffusion (D3PM) framework [5]. By modelling proteins at the atomic level, this approach enables the inclusion of non-canonical amino acids and post-translational modifications.

Results: Guided antibody diffusion produces CDR-H3 loops with improved developability profiles, including hydropathy shifts from −0.73 to −1.62 and predicted folding energy improvements from 0.14 to −2.20, while preserving structural fidelity to reference loops. Pareto-optimal solutions demonstrate that diverse sequences can converge to similar functional structures. Our all-atom sequence diffusion model generates highly novel and diverse proteins (up to 99.9% novelty and 99.7% diversity) with structural foldability comparable to amino-acid-based baselines. Our proposed atom-level evaluation pipeline further enables systematic filtering and analysis of canonical and non-canonical generated proteins.

Impact: Together, these contributions demonstrate how diffusion models can integrate structural awareness, property optimization, and atomic-level chemical flexibility for immunoprotein design. This enables the generation of antibodies, immunopeptides, and other therapeutic proteins that go beyond canonical sequence constraints, while remaining compatible with downstream structural and

developability assessment. Our work lays the groundwork for next-generation immunoprotein design by unifying structure, function, and chemistry in generative models.

References

1. A. Villegas-Morcillo, et al. Guiding diffusion models for antibody sequence and structure co-design with developability properties. PRX Life, 2024.
2. S. Luo, et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. NeurIPS, 2022.
3. A. Villegas-Morcillo, et al. All-atom protein sequence design using discrete diffusion models. Journal of Cheminformatics, 2026.
4. M. Krenn, et al. Self-referencing embedded strings (SELFIES). Mach Learn Sci Technol, 2020.
5. J. Austin, et al. Structured denoising diffusion models in discrete state-spaces. NeurIPS, 2021.

From Mechanistic Model to Digital Twin: A Framework for Real-Time Optimization of Ethanol Production in Saccharomyces cerevisiae

Omar Bayomie[1,2,3], Mauditra Matin[3]

[1]The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London

[2]SSPC, The Science Foundation Ireland Research Centre for Pharmaceuticals, School of Chemical & Bioprocess Engineering, University College Dublin

[3]Delft Institute of Applied Mathematics, Delft University of Technology

Abstract

The transition to autonomous bioprocessing requires control strategies capable of managing the nonlinear dynamics and limited observability inherent to industrial fermentation. This work presents an integrated Digital Twin framework for the real-time optimization of fed-batch ethanol production, combining mechanistic modeling, nonlinear state estimation, and predictive control. A global trajectory optimization first demonstrated that variable-volume fed-batch operation yields a significant increase in total ethanol mass compared to batch baselines, identifying a critical physical trade-off where volumetric gain outweighs concentration dilution. To enable real-time implementation, a comparative benchmarking of nonlinear observers including Extended (EKF), Ensemble (EnKF), and Particle Filters (PF) was conducted, with these estimates, a Nonlinear Model Predictive Control (NMPC) strategy was developed to robustly track kinetically near-optimal substrate setpoints while strictly enforcing reactor volume and inhibition constraints. The closed-loop framework demonstrated remarkable performance even in the presence of significant kinetic parameter mismatch and heteroscedastic measurement noise confirming that the proposed Digital Twin can effectively stabilize high-yield production strategies in uncertain biological systems.

Context: Fermentation processes are inherently difficult to control. Biomass the key driver of productivity cannot be measured online. Substrate and product concentrations fluctuate unpredictably. Model parameters drift between batches. Traditional batch operation leaves significant yield on the table: glucose is loaded upfront, inhibition kicks in early, and the process terminates with substrate unconsumed or product underproduced. Fed-batch operation offers a solution, but only if feeding decisions are made intelligently, in real-time, based on the actual state of the fermentation. This requires a digital twin: a computational replica that estimates what cannot be measured and optimizes what can be controlled.

Approach: We developed a closed-loop digital twin framework for Saccharomyces cerevisiae ethanol fermentation integrating three components: a mechanistic kinetic model, an Extended Kalman Filter (EKF) for state estimation, and Nonlinear Model Predictive Control (NMPC) for optimal feeding. The mechanistic model captures Monod kinetics with ethanol inhibition. The EKF fuses Raman spectroscopy measurements of glucose and ethanol with model predictions to infer unmeasured biomass concentration a critical capability since online yeast sensors remain impractical at industrial scale. Covariance matrices were calibrated from six laboratory experiments, not tuned to optimize fit, measurement noise R derived from model-data residuals, process noise Q fixed at $(0.1^2)\cdot I$ across all batches. The NMPC maximizes total ethanol mass while respecting reactor volume constraints and feed system limits. Robustness was evaluated through closed-loop simulations with parametric mismatch between the virtual plant and the model used by the controller, heteroscedastic measurement noise and partial observability where biomass was inferred from glucose and ethanol dynamics.

Results: The framework works. Under ideal conditions, the digital twin produced 475 g of ethanol in 20 hours. Under 15-20% model-plant mismatch, it produced 483 g, actually more, because the real plant grew faster than the model predicted, and the feedback loop captured this benefit automatically. State estimation remained accurate: MSE of 1.6 $(g/L)^2$ for glucose, 0.4 $(g/L)^2$ for ethanol, and 0.08 $(g/L)^2$ for biomass in the ideal case. With mismatch, biomass MSE increased to 0.22 $(g/L)^2$ yet production efficiency stayed above 99%. The NMPC discovered a three-phase feeding strategy without being told: moderate feeding (0.18 L/h) to build substrate, near-zero feeding during exponential growth, then aggressive feeding (up to 0.4 L/h) during peak production. Glucose was maintained around 30 g/L, avoiding both starvation and accumulation. Results were consistent across five independent noise realizations (coefficient of variation < 1%), confirming that performance does not depend on lucky measurement sequences.

Impact: This work demonstrates that model-based state estimator digital twins can deliver robust, near-optimal control of fed-batch fermentation even when the model is wrong and biomass is unmeasured. The framework requires no online parameter adaptation, no neural network training, no experiment-specific tuning just a calibrated mechanistic model and principled covariance specification. Estimation errors do not translate proportionally to production losses; feedback compensates. We found 807% improvement over batch operation quantifies the economic value of real-time optimization. Next steps: experimental validation on a laboratory-scale bioreactor with actual Raman spectroscopy, implementation of the unified aerobic-anaerobic model for processes with oxygen transitions, and extension to multi-objective optimization balancing yield, productivity, and probably energy consumption.

References

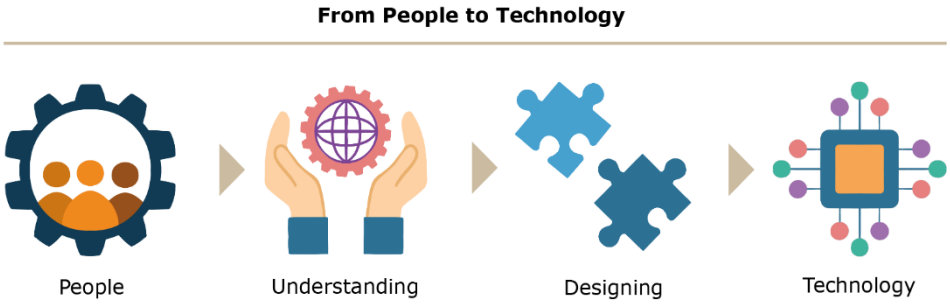
1. Agarwal, P., McCready, C., Ng, S. K., Ng, J. C., van de Laar, J., Pennings, M., & Zijlstra, G. (2025). Hybrid modeling for in silico optimization of a dynamic perfusion cell culture process. Biotechnology Progress, 41, e3503.
2. Gargalo, C. L., Udugama, I., Pontius, K., Lopez, P. C., Nielsen, R. F., Hasanzadeh, A., Mansouri, S. S., Bayer, C., Junicke, H., & Gernaey, K. V. (2020). Towards smart biomanufacturing: a perspective on recent developments in industrial measurement and monitoring technologies for bio-based production processes. Journal of Industrial Microbiology & Biotechnology: Official Journal of the Society for Industrial Microbiology and Biotechnology,47, 947–964.
3. Kramer, D., & King, R. (2016). On-line monitoring of substrates and biomass using near-infrared spectroscopy and model-based state estimation for enzyme production by s. cerevisiae. IFAC-

PapersOnLine, 49, 609–614. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB 2016.

DesAIgning Bioscience: Strategic Innovation Begins with People

Giulia Bacchi, dsm-firmenich

Graphical abstract



Abstract

The biotech industry is experiencing a rapid expansion in data volume, experimental complexity, and cross-functional dependencies. While AI promises unprecedented acceleration of scientific processes—from knowledge extraction to hypothesis generation—the real bottleneck often lies elsewhere: human collaboration. Misaligned mental models, fragmented communication, and heterogeneous workflows limit the impact of even the most advanced tools. As emerging studies on human–AI teaming highlights, effective innovation requires understanding how people work, think, decide, and collaborate before layering technology on top (Fragiadakis et al., 2024).

I present a human-centered approach to design for Biotech which positions people, not technology, as the primary design material. Qualitative inquiry, collaboration-mapping, and future-state scenario workshops have been used to identify the frictions experienced by scientists, data experts, and decision-makers across modern DBTL cycles. I have coupled these human-insight methods with AI-enabled sense-making, automated clustering of interview transcripts, extraction of recurrent patterns, and synthesis of shared needs. This dual lens aligns with recent methodological frameworks that emphasize symbiotic models of human–AI collaboration rather than AI–centric automation (Song et al., 2024; Stenhouse et al., 2025).

Results consist in three core findings: 1. Human collaboration gaps are the real burden, not technological limitations. Teams struggle more with alignment, context-sharing, and joint decision-making than with the adequacy of digital tools. 2. AI can accelerate understanding of the human system. AI consistently surface latent patterns identifying conflicting mental models of the "future lab" inconsistent definitions of expertise, governance and invisible dependencies across interviews and project documentation. 3. AI as collaborator, not replacer. This analysis supports the view that AI can meaningfully assist scientists with data curation, cognitive offloading, and hypothesis-level reasoning, only if its integration is guided by human needs and trust-enabling design. (Chetouani et al., 2025).

If biotech organizations want AI innovations that scale, they must start from the human dimension. The future is not a choice between "technology" or "people," but a deliberate sequencing: people → understanding → design → technology. By grounding AI deployment in human reality, collaboration patterns, cognitive load, incentives, and aspirations, we can pave the way for more adaptive, reliable, and trustworthy human–machine systems. My next step of my work is strategically operationalizing this approach: embedding human-centered discovery as a prerequisite to AI adoption, developing shared collaboration protocols, and cultivating hybrid human centered intelligence where humans and machines learn together.

References

1. Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M. (2024). Evaluating Human-AI Collaboration: A Review and Methodological Framework. arXiv:2407.19098.
2. Stenhouse, A., et al. (2025). A vision of human–AI collaboration for enhanced biological collection curation and research. BioScience, 75(6), 457–471.
3. Song, B., Zhu, Q., & Luo, J. (2024). Human-AI Collaboration by Design. Proceedings of the Design Society, 4, 2247–2256.
4. IEEE (2025). A Multifaceted Vision of Human-AI Collaboration. IEEE Xplore. 5. Chetouani, M., Nowak, A., & Lukowicz, P. (2025). Handbook of Human-AI Collaboration. Springer Nature.

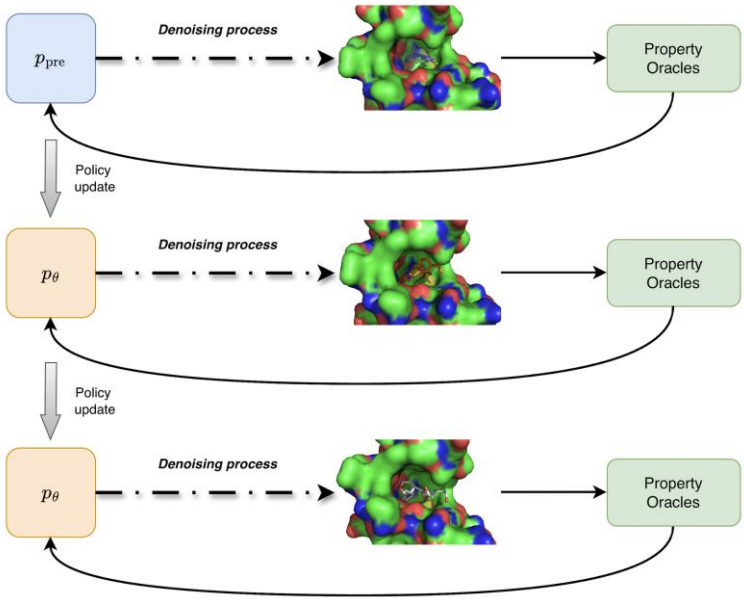
Fine-tuning Pocket-Aware Diffusion Models via Denoising Policy Optimization

Yuan Xue[1], Daniel Kudenko[1] and Megha Khosla[2]

[1]L3S Research Centre, Leibniz University Hannover

[2]Dept. Intelligent Systems, Institute of Electrical Engineering, Mathematics, Delft University of Technology

Graphical abstract



Abstract

Context: Structure-based drug design (SBDD) involves identifying 3D molecules with high binding affinity to the target protein pocket. Most of the existing SBDD methods, from autoregressive [1] to

diffusion-based methods [2], mainly focus on resembling the training distribution via likelihood maximization, leading to degraded performance when the training data does not align well with the desired molecular properties required by the tasks at hand. We seek to establish a flexible and efficient framework for structure-based molecule optimization (SBMO) that enhances binding affinity while supporting task-specific multi-property optimization.

Approach: In our study, we propose to employ online reinforcement learning for fine-tuning a pretrained pocket-aware diffusion model. Specifically, we formulate the denoising process of the diffusion model as a multi-step Markov Decision Process and view each denoising step as an action taken by the denoising policy. In the realm of SBMO with diffusion models, recent research has explored strategies including gradient-based guidance [3] and direct preference optimization [4], whereas the systematic adoption of online reinforcement learning for finetuning pocket-conditional diffusion models remains underexplored, which is a gap filled by our study.

Results: Our evaluation on the CrossDocked2020 benchmark demonstrates substantial improvements over all state-of-the-art baselines in binding affinity, while also achieving the best performance in drug-likeness and molecular diversity, along with competitive results in synthesizability. To the best of our knowledge, our approach is the first to achieve Vina Score of -8.5 kcal/mol, representing a 33.7% improvement over the reference molecules in the dataset. In addition, we also demonstrate strong conformation stability of the generated molecules.

Impact: We show that applying online reinforcement learning for the alignments of pre-trained pocket-conditional diffusion models can be a promising paradigm for SMBO. The next step is to expand the application of our framework to a broader range of diffusion models including flow-matching based methods.

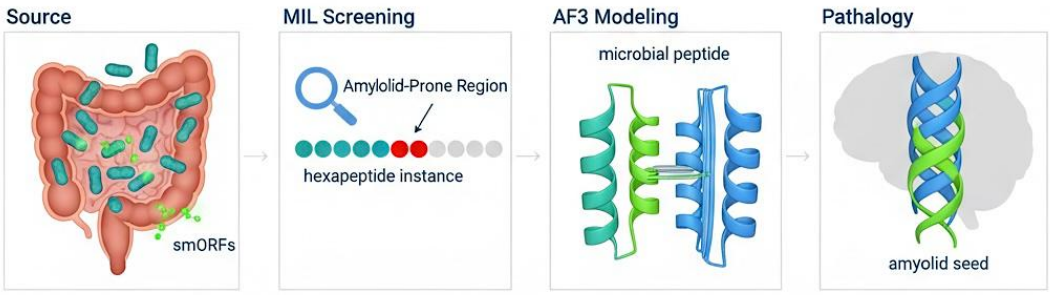
References

1. Peng, Xingang, et al. "Pocket2mol: Efficient molecular sampling based on 3d protein pockets." International conference on machine learning. PMLR, 2022.
2. Guan, Jiaqi, et al. "3d equivariant diffusion for target-aware molecule generation and affinity prediction." arXiv preprint arXiv:2303.03543 (2023).
3. Dorna, Vineeth, et al. "Tagmol: Target-aware gradient-guided molecule generation." arXiv preprint arXiv:2406.01650 (2024).
4. Gu, Siyi, et al. "Aligning target-aware molecule diffusion models with exact energy optimization." Advances in Neural Information Processing Systems 37 (2024): 44040-44063.

Microbiome-derived proteins as seeds for protein aggregation in neurodegenerative diseases

Abdul Rafay Pirzada, Patrick May and Paul Wilmes, Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Graphical abstract



Abstract

Context: Microbiome dysbiosis is a recognized hallmark of neurodegenerative diseases [1], yet the molecular mechanisms linking gut bacteria to brain pathology remain poorly defined. We hypothesize that certain microbial small open reading frame products (smORFs) act as "amyloid seeds" capable of crossing the gut-brain axis [2]. These proteins may initiate or accelerate the aggregation of human proteins like alpha-synuclein through cross-species nucleation [3]. Characterizing these interactions is vital for moving beyond epidemiological correlations toward mechanistic interventions.

Approach: To address this, we are developing a multi-stage computational and experimental pipeline. Traditional protein classifiers often fail because amyloidogenic signal, 'Amyloid-Prone Regions (APRs)', are short sequences buried within longer chains. We overcome this by utilizing a Multi-Instance Learning (MIL) framework trained on experimentally validated amyloidogenic proteins, where full sequences act as "bags" and constituent hexapeptides as "instances. [4]" This architecture allows the model to learn the local features driving aggregation, which we then apply to screen microbial smORF for high confidence amyloidogenic signatures. Top candidates from the MIL screen undergo structural interaction modeling. We employ AlphaFold3 to predict the binding interfaces between microbial smORFs and human amyloid proteins. By simulating various stoichiometries (e.g., monomeric vs. oligomeric templates), we analyze the predicted complexes for beta-sheet alignment and hydrophobic packing, structural prerequisites for cross-species nucleation.

Results: In a preliminary pilot study, we utilized an existing ensemble tool to screen several hundred thousand smORFs. From this screening, a number of selected candidates were experimentally confirmed to exhibit amyloidogenic properties in vitro. However, this pilot also exposed significant limitations: existing tools produce high false-positive rates because they do not account for the structural context of APRs. Consequently, our current work focuses on training the more robust MIL-based model to reduce these inaccuracies and employing structural features for better outcome.

Impact: By establishing a mechanistic link between specific microbial products and human protein aggregation, this research moves the field from correlation to causation. The development of a structure-aware discovery tool allows for the systematic mapping of the "microbial amyloidome." Ultimately, this work provides a blueprint for early diagnostic biomarkers and potential therapeutic

interventions, such as small-molecule inhibitors or microbiome "edits," designed to stop neurodegeneration at its enteric source before it reaches the central nervous system.

References
1. Intili, G. et al. (2023). From Dysbiosis to Neurodegenerative Diseases through the Gut–Brain Axis. Life, 13(2), 346.
2. Wilmes, P., Davin, M., et al. (2025). Expanding the human metaproteome. Research Square [Preprint].
3. Fernández-Calvet, A. et al. (2024). Microbial amyloids seed α-synuclein. Nat. Commun., 15, 6062.
4. Carbonneau, M.-A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognition, 77, 329–353.

Hyperbolic Molecule-Text Contrastive Learning for Hierarchy-Aware (Bio)Molecular Representations

Lorenzo Di Fruscia, Amelia Villegas-Morcillo and Jana M. Weber, Department of Intelligent Systems, Delft University of Technology

Abstract

Context: Large-scale molecular machine learning increasingly relies on joint representations that connect (bio)chemical structures to natural language descriptions (e.g., functional annotations) to support retrieval, discovery, and generalization. Both molecule structures and their semantics are inherently multi-level and hierarchical: atoms form functional groups and scaffolds, which then form molecules; natural language captions range from broad chemical classes to highly specific descriptions. However, most molecule–text alignment methods learn embeddings in Euclidean spaces, which can struggle when representing hierarchy. We aim to improve cross-modal molecular representations by leveraging geometric structure that is better suited to hierarchical data.

Approach: We improve molecular representation learning by explicitly accounting for hierarchies in chemical annotations. Particularly, we introduce a hyperbolic variant of molecule-text contrastive pretraining to capture multi-scale chemical semantics. A common way to learn these joint representations is molecule-text retrieval, where a model is trained so that, given a molecule, it retrieves the most relevant description, and vice versa. Building on an established molecule-text alignment framework [1], we replace the shared Euclidean embedding space of the two data modalities with a hyperbolic manifold.

Notably, hyperbolic embeddings are known to represent taxonomy-like relationships with lower distortion [2], providing a principled inductive bias for hierarchical organization. The resulting model aims to learn embeddings that are simultaneously retrieval-effective across modalities and structurally organized with respect to the hierarchy of chemical semantics.

Results: Qualitatively, the hyperbolic alignment yields a more interpretable latent space: broad, generic captions and higher-level chemical concepts tend to map closer to the origin, while more specific molecule/caption pairs occupy larger radii, consistent with a learned notion of specificity as observed in prior hyperbolic multimodal work [3]. Quantitatively, hyperbolic embeddings achieve retrieval performance competitive with Euclidean baselines on molecule-text retrieval benchmarks, while additionally inducing a clearer hierarchical organization of the joint space.

Impact: By moving molecule-text alignment from a flat space (Euclidean) to a negatively curved geometry (hyperbolic), we take a step toward multimodal representations that better reflect the hierarchical structure of chemical knowledge. This can benefit semantic search in molecule databases, zero-shot annotation, and dataset curation via cross-modal retrieval. Next steps include incorporating structural constraints (e.g., fragments/motifs) to further improve interpretability.

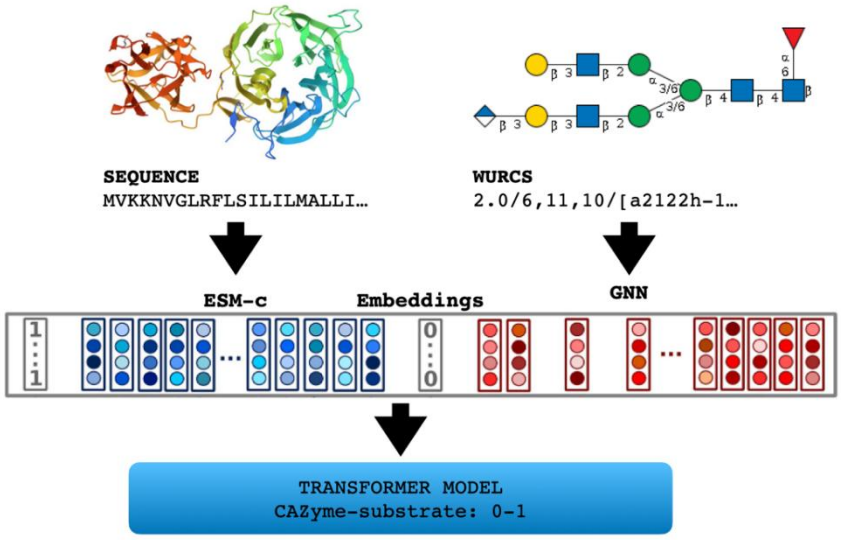
References

1. S. Liu et al., "Multi-modal molecule structure–text model for text-based retrieval and editing," Nat Mach Intell, vol. 5, no. 12, Art. no. 12, Dec. 2023, doi: 10.1038/s42256-023-00759-6.
2. R. Sarkar, "Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane," in Graph Drawing, M. van Kreveld and B. Speckmann, Eds., Berlin, Heidelberg: Springer, 2012, pp. 355–366. doi: 10.1007/978-3-642-25878-7_34.
3. K. Desai, M. Nickel, T. Rajpurohit, J. Johnson, and R. Vedantam, "Hyperbolic Image-Text Representations," Jan. 18, 2024, arXiv: arXiv:2304.09172. doi: 10.48550/arXiv.2304.09172.

Predicting CAZyme-substrate specificities using AI

Maarten Boneschansker, Wagening University & Research

Graphical abstract



Abstract

Context: Carbohydrates are the most abundant biomolecule on Earth by mass; some such as cellulose and chitin serve a structural purpose, others such as starch or glycogen are used for energy storage. Carbohydrates are natural graph structures consisting of linked sugar monomers, where combinations of different monomers, linkage types, and branching yield a vast and complex chemical space.

To break down such a diverse set of molecules, many microorganisms employ specialized carbohydrate-active enzymes (CAZymes), many of which show activity only toward specific carbohydrate substructures. Although the number of experimentally characterized CAZymes is steadily growing, they still account for only ~0.01% of the millions of CAZymes currently listed in the CAZy database (Drula et al., 2022). The current state of the art in CAZyme annotation relies heavily on

sequence similarity-based methods, which work well when a highly similar sequence is available, but for the vast majority of CAZymes this is not the case (Zheng et al., 2023).

Approach: We aim to employ machine learning methods to model the rules governing CAZyme–substrate specificity, providing more accurate annotations especially where sequence similarity falls short. Methods developed for general enzyme function prediction show promise but do not transfer directly to CAZymes, whose substrates are polysaccharides — large, modular structures that differ fundamentally from the small molecules these methods were designed for (Kroll et al., 2023).

As a first step, we employ a protein language model (pLM) to generate dense embeddings of CAZyme sequences. We analyze these embeddings to assess what substrate-specificity information is encoded and compare this to what is recoverable from sequence similarity alone. Next, we aim to connect pLM embeddings to graph neural network (GNN) representations of carbohydrate structures — a natural fit given their inherent graph topology — in a multimodal cross-attention framework like the recently published EZspecificity (Cui et al., 2025).

Results: Preliminary classification results indicate that pLM embeddings capture substrate-specificity signal beyond what sequence similarity provides. Encouragingly, the embedding space also reflects CAZyme (sub)family membership and taxonomic origin. These results validate the pLM-based representation as a meaningful starting point for substrate prediction and motivate the planned multimodal extension. Full results from the multimodal model are forthcoming.

Impact: The vast majority of known CAZymes lack functional annotation, limiting our understanding of microbial carbohydrate metabolism in ecosystems ranging from the human gut to industrial fermentation. By moving beyond sequence similarity, our approach aims to unlock substrate-specificity predictions for the many millions of unannotated CAZymes. Connecting protein and carbohydrate representations in a joint multimodal model would enable reasoning over both enzyme and substrate structure, opening a path toward annotations beyond the reach of current methods.

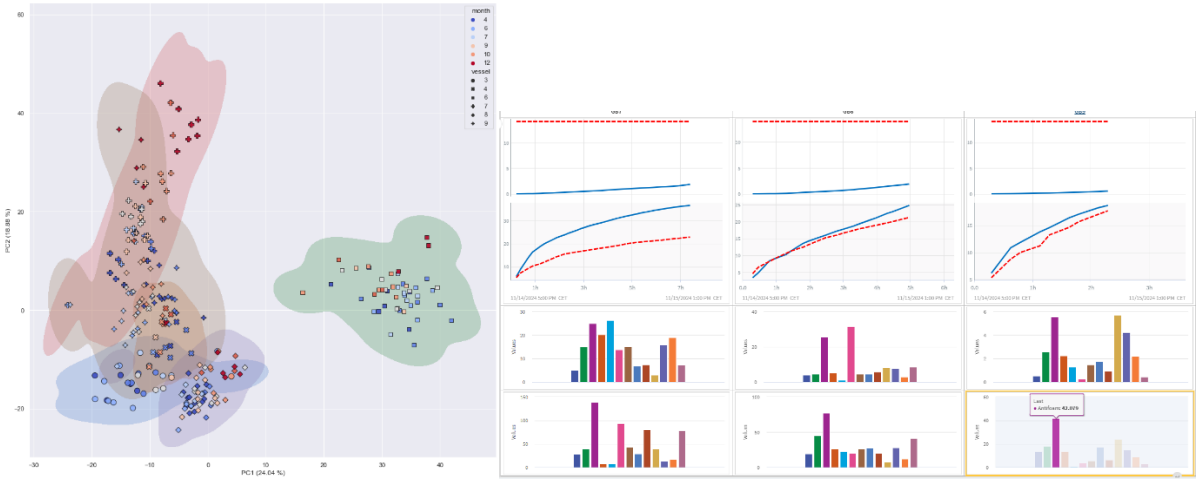
References

1. Cui, H., Su, Y., Dean, T. J., Yu, T., Zhang, Z., Peng, J., Shukla, D., & Zhao, H. (2025). Enzyme specificity prediction using cross attention graph neural networks. Nature. https://doi.org/10.1038/s41586-025-09697-2.
2. Drula, E., Garron, M.-L., Dogan, S., Lombard, V., Henrissat, B., & Terrapon, N. (2022). The carbohydrate-active enzyme database: Functions and literature. Nucleic Acids Research, 50(D1), D571–D577. https://doi.org/10.1093/nar/gkab1045.
3. Kroll, A., Ranjan, S., Engqvist, M. K. M., & Lercher, M. J. (2023). A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nature Communications, 14(1), 2787. https://doi.org/10.1038/s41467-023-38347-2.
4. Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., & Yin, Y. (2023). dbCAN3: Automated carbohydrate-active enzyme and substrate annotation. Nucleic Acids Research, 51(W1), W115–W121. https://doi.org/10.1093/nar/gkad328.

Real-time statistical process evaluation at a fermentation plant

Tom Ploeger, dsm-firmenich

Graphical abstract



Abstract

Context: Process monitoring in a factory is usually still done by checking many separate process variables. Especially in a factory with multiple vessels running simultaneously, this is a time-consuming task with the risk of not noticing deviations. Advanced techniques that perform data interpretation are ubiquitous, but care should be taken that the results are conveyed to the operator in a clear way and allow for corrective action.

Approach: Monitoring of the fermentation process in six vessels simultaneously was made possible by making a PCA model based on historical batches. Scaling and modelling all the data as one set showed that the fermentation vessels behaved differently. Separate scaling of the vessels into one model was subsequently chosen as method. The model was implemented via the existing IT/OT systems of the factory.

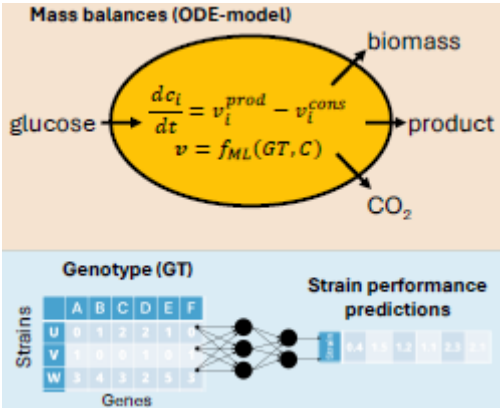
Results: A dashboard was made that shows whether each running batch is within normal boundaries (Hotellings T2 and Squared prediction error), along with the loadings of each variable to show the impact on the scores. A PCA model can easily be implemented in a SCADA system as a linear combination of existing variables.

Impact: Interpreting the status of a running fermentation batch from a single parameter saves a lot of time for operations personnel looking at (mostly in-control) data. It allows faster and better focus on the areas that are not running well.

Neural-ODE assisted metabolic engineering of p-coumaric acid production

Rik van Rosmalen, Zheng Zhao, Sara Moreno Paz, Wenjun Tang, Wouter Touw, Liang Wu, Hans Roubos, Ben Smith and Joep Schmitz, dsm-firmenich

Graphical abstract



Abstract

Context: Developing high-performing microbial cell factories remains slow due to empirical, hit and miss Design–Build–Test–Learn cycles. More predictive methods are needed to guide targeted genetic modifications and accelerate optimization of complex biosynthetic pathways, such as p-coumaric acid production in S. cerevisiae.

Approach: We implemented a Neural Ordinary Differential Equation (Neural ODE) framework using the jaxkineticmodel toolkit [1] to model combinatorial pathway designs. Automated strain engineering, high throughput screening, and whole genome sequencing generated a rich dataset for training this hybrid kinetic–ML model. We benchmarked the Neural ODE approach against established ML approaches to assess its performance in accuracy.

Results: The Neural ODE framework successfully captured pathway dynamics and performed on par compared to established ML methods on test datasets.

Impact: By strengthening predictive power within the DBTL cycle, the Neural ODE approach offers a scalable strategy for accelerating metabolic engineering campaigns. In the future, we will assess if Neural ODE provides improved capability of extrapolation and interpretation as compared to traditional ML methods.

References

1. van Lent P, Bunkova O, Magyar B, Planken L, Schmitz J, Abeel T. Jaxkineticmodel: Neural ordinary differential equations inspired parameterization of kinetic models. PLoS Comput Biol. 2025 Jul 7;21(7):e1012733. doi: 10.1371/journal.pcbi.1012733.

A Closer Look at Antidepressant–Microbiome Interactions: Is There Evidence for Bacterial Biotransformation?

Stefanie Malan-Müller[1], Fernanda Parraguez Contreras[2], Nto Johnson Nto[3], Christopher A Lowry[4], Albert Garcia Valiente[5], Mireia Valles-Colomer[5], Sian MJ Hemmings[6], Sahar El Aidy[7], Anja Lok[8], Halima Mouhib[2], Walter Pirovano[2]

[1] Universidad Complutense de Madrid (UCM), [2] Vrije Universiteit Amsterdam, [3] University of Nigeria, [4] University of Colorado Boulder, [5] Universitat Pompeu Fabra,[6] Stellenbosch University, [7] University of Amsterdam, [8] Amsterdam UMC (University of Amsterdam)

Graphical abstract

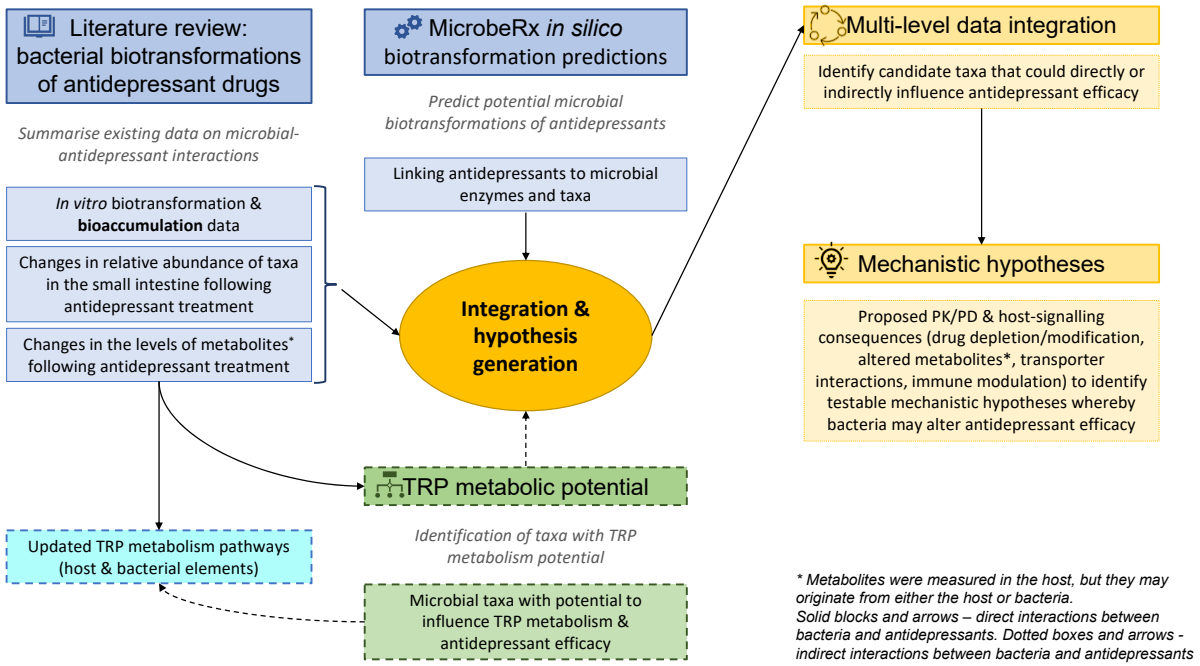


Figure 1. Integrated workflow combining literature review and in silico analysis to study gut microbiome-antidepressant interactions

Abstract

Context: Major depressive disorder remains one of the most disabling psychiatric conditions globally, affecting over 280 million individuals and ranking among the top contributors to disability-adjusted life years worldwide [1]. Emerging evidence implicates the gut microbiota as a key mediator in the development and progression of mood and anxiety disorders [2]. At the same time, growing data suggests a bidirectional interaction between antidepressants and the gut microbiome [3,4], however the mechanistic nature of these interactions remains poorly understood. Consequently, current models of antidepressant action largely ignore microbial contributions, despite growing evidence that microbial metabolism, efflux systems, and neurotransmitter pathways may influence drug efficacy, side effects, and interindividual response variability.

Approach: We integrated microbial genetics, pharmacology, host-metabolite interactions, and targeted in silico enzyme predictions to construct a multi-level mechanistic framework for antidepressant–microbiome interplay. More specifically, we: 1) synthesized current evidence on bacterial interactions with SSRIs and SNRIs; 2) summarized changes in the relative abundance of

common small intestine bacteria following antidepressant exposure; 3) applied an in silico enzymatic prediction approach to model potential gut bacterial biotransformation pathways for representative SSRIs and SNRIs; 4) examined secondary literature for changes in TRP metabolism and LPS-related signaling after antidepressant exposure; 5) assessed the bacterial potential to produce TRP metabolites.

Results: Our analyses revealed convergent microbial responses to SSRIs and SNRIs – including efflux pump induction, increased plasmid transfer, transporter interference, neurotransmitter modulation, and broad-spectrum duloxetine bioaccumulation. Moreover, we identified a previously unrecognized candidate enzyme, a putative digallate acylhydrolase, with the capacity to biotransform fluoxetine; offering the first plausible molecular entry point for SSRI metabolism by gut bacteria. Additionally, we identified bacterial taxa with functional potential across the indole, serotonin, and kynurenine pathways. We thus generated a preliminary catalogue of candidate bacterial enzymes that may directly or indirectly interact with these drugs.

Impact: Our study provides a comprehensive assessment of potential bacterial mechanisms involved in antidepressant biotransformation, and proposes testable hypotheses for how microbial processes could modulate antidepressant pharmacodynamics and psychiatric outcomes. From a translational and clinical perspective, these findings underscore that the gut microbiome is an active determinant of antidepressant response. Incorporating bacterial, metabolic, and genetic profiling into psychiatric research and treatment could revolutionize how antidepressants are prescribed, monitored, and optimized. We believe this mechanistic synthesis fills a critical conceptual gap and provides a foundation for the development of microbiome-informed antidepressant strategies.

References

1. World Health Organization (2023). Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression.
2. Cryan, J.F., O'Riordan, K.J., Cowan, C.S.M., Sandhu, K.V., Bastiaanssen, T.F.S., Boehme, M., Codagnone, M.G., Cussotto, S., Fulling, C., Golubeva, A.V., et al. (2019). The Microbiota-Gut-Brain Axis. Physiol. Rev. 99, 1877–2013. https://doi.org/10.1152/physrev.00018.2018.
3. Munoz-Bellido, J.L., Munoz-Criado, S., and Garc.a-Rodr.guez, J.A. (2000). Antimicrobial activity of psychotropic drugs: selective serotonin reuptake inhibitors. Int. J. Antimicrob. Agents 14, 177–180. https://doi.org/10.1016/s0924-8579(99)00154-5.
4. Klünemann, M., Andrejev, S., Blasche, S., Mateus, A., Phapale, P., Devendran, S., Vappiani, J., Simon, B., Scott, T.A., Kafkia, E., et al. (2021). Bioaccumulation of therapeutic drugs by human 1031 gut bacteria. Nature 597, 533–538. https://doi.org/10.1038/s41586-021-03891-8.

High-Throughput Platform for AI-Derived Antibody Drug Identification and Developability Assessment

Alexander Jansma, Sino Biological

Antibodies and engineered minibinders play central roles in modern drug discovery by enabling highly specific targeting of disease-associated molecules. Monoclonal antibodies have become a dominant therapeutic modality for oncology, immunology, and infectious diseases due to their high affinity, selectivity, and adaptable effector functions. Minibinders—small, engineered binding proteins—complement antibodies by offering advantages such as reduced molecular size, improved tissue

penetration, and simplified manufacturing. Together, these binding modalities expand the therapeutic landscape and create new opportunities for precision medicine.

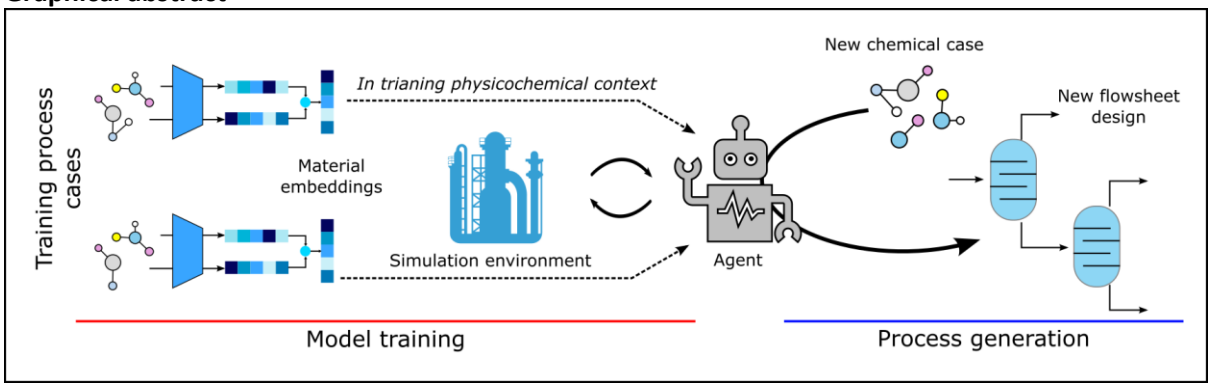
Artificial intelligence is transforming biologics discovery by enabling rapid in silico design, optimization, and screening of vast sequence spaces that would be impractical to explore experimentally. However, the true potential of AI depends on the availability of large, high-quality experimental datasets and efficient validation pipelines to iteratively refine predictive models.

Sino Biological has established an integrated, high-throughput technology platform specifically designed to accelerate AI-driven lead discovery and optimization. The platform combines automated antibody production systems—including high-throughput mammalian cell expression and cell-free protein synthesis—with comprehensive developability assessment tools such as stability, aggregation, and functionality profiling. This infrastructure enables the rapid generation and characterization of thousands of AI-derived antibody and minibinder variants, producing robust datasets essential for AI model training and performance improvement. By seamlessly translating AI-designed sequences into experimentally validated therapeutic candidates, the platform effectively bridges the gap between computational design and real-world drug development, accelerating the path from concept to clinic.

Context-aware agent for process flowsheet synthesis

Ulderico Di Caprio and Artur M. Schweidtmann, Department of Chemical Engineering, Delft University of Technology

Graphical abstract



Abstract

Context: Modern pharmaceutical industry relies on trains of unit operations to perform reaction, transformation, and separation tasks. Process design is a crucial phase of drug development that needs to be accelerated because of business and patients needs [1]. Creating optimal process systems is a complex task, particularly when dealing with feeds of varying composition and chemical nature of the components. In recent years reinforcement learning (RL) has emerged as a promising tool supporting such design effort [2,3], however existing approaches requires agent retraining to deal with new mixture, reducing employment scalability and slows down practical deployment. Thus, there is a need for RL-based methods that maintain design performance while generalising to unseen mixtures without additional training.

Approach: This work proposes a physics-informed context-aware RL framework based on Proximal Policy Optimisation (PPO) for automatic process generation, using purification train as case study. The flowsheet is represented as a process graph [3], where the RL agent selects the next unit operation, its placement within the graph, and associated design variables such as light/heavy keys, reflux ratio, and process targets. To overcome the retraining requirement in existing models, physical and thermodynamic information of the present components within the mixture is embedded directly into the graph representation. This enriched, mixture-invariant state space enables the trained agent to operate in inference mode when encountering new mixtures, without modifying model parameters.

Results: The proposed method successfully generates feasible multi-column distillation trains for mixtures of up to four components, meeting purity specifications and achieving competitive energy performance without any retraining. When tested on mixtures unseen during the training, the agent demonstrated strong generalisation capabilities and produced valid solutions without retraining. Compared to the non-context inform baseline RL methods, the novel approach improved robustness and reduced computational effort

Impact: In this work a context-aware RL agent for process generation was proposed, showing its potential to serve as a scalable and reusable tool for flowsheet synthesis in pharmaceutical processing. By enabling inference-only usage of the agent for new mixtures, the approach reduces engineering time, computational cost, and barriers to industrial adoption. Future directions include extending the

action space to additional unit operations, incorporating heat-integration decisions, and exploring transfer learning across broader separation tasks.

References

1. J. Kim, K. Okamura, M.R. Gaddem, Y. Hayashi, S. Badr, H. Sugiyama, Curr. Opin. Chem. Eng. 47 (2025) 101093.
2. Q. Göttl, J. Pirnay, J. Burger, D.G. Grimm, Comput. Chem. Eng. 194 (2025) 108975.

L. Stops, R. Leenhouts, Q. Gao, A.M. Schweidtmann, AIChE J. 69 (2023) e17938.