AI4b.io Symposium 2025
"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 15 & 16, 2025. We are organizing this meeting for 140 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 see 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 15-16 (Tuesday and Wednesday), 2025

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 Day2, 10:15 – 10:35) to help draw attention to the posters. Presenters have the opportunity to 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 17:30 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 inquiries during the event, you can contact us by phone: +31620947971.
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 see you there!

Best regards,
AI4b.io Symposium Organizing Committee

Organizers
Paul van Lent
Chengyao Peng
Mahdi Naderibeni

Steering Committee
Marcel Reinders
Henk Noorman
Hans Roubos
Wouter van Winden
Jana Weber
Renger Jellema

Timetable – Day 1, 15th 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**: Adding Insight to Artificial intelligence	Gabriel Weymouth Delft University of Technology
11:00-11:15	Break (15 minutes)	
11:15-11:40	Compartment modelling meets Deep Learning: Towards an efficient modelling approach for bioreactor gradients	Hector Maldonado Delft University of Technology
11:40-12:05	The flow must go on: dynamic scheduling algorithms for biomanufacturing	Kim van den Houten Delft University of Technology
12:05-12:30	Use of ensemble modelling for bubble size measurements with an optical fibre probe	Rik Volger Delft University of Technology
12:30-13:45	Lunch	
13:45-14:30	**Keynote**: On the challenges of predicting gut microbial community dynamic	Karoline Faust Katholieke Universiteit Leuven
14:30-14:45	Break	
14:45-15:10	ABaCo: Correcting Technical Heterogeneity for Metagenomic Data integration using Adverserial Generative Models	Edir Vidal Novo Nordisk Foundation Center for Biosustainability
15:10-15:35	Prediction and reengineering of MADS-box proteins using structure-based machine learning	Alejandro Sanchez University of Amsterdam
15:35-15:50	Break	
15:50-16:15	ENFORCE: Exact Nonlinear Constrained Learning with Adaptive-depth Neural Projection	Giacomo Lastrucci Delft University of Technology
16:15-16:40	Inferring Discrete Dynamical Models of Biological Systems with Program Synthesis	Reuben Gardos Reid Delft University of Technology
16:40-17:00	Closing remarks	
17:00-20:00	Borrel + Dinner at Firma van Buiten	

Timetable – Day 2, 16th April

Time	Schedule	Speaker Name & Affiliation
9:00-9:30	Arrive, register, coffee	
9:30-10:15	**Keynote:** AI-driven decentralised biomanufacturing	Vitor Martin Dos Santos Wageningen University and Research
10:15-12:45	Poster session + Networking	
11:45-12:45	Lunch	
12:45-13:10	Machine Learning-assisted pathway optimization in large combinatorial design spaces: a p-coumaric acid case study	Paul van Lent Delft University of Technology
13:10-13:35	Automatic generation of control structures in industrial process flow diagrams with transformers	Dominique Goldstein Delft University of Technology
13:35-14:00	Text2Model: Generating dynamic chemical reactor models using large language models	Sophia Rupprecht Delft University of Technology
14:00-14:30	Break (30 minutes)	
14:30-14:55	All-Atom Novel Protein Sequence Generation Using Discrete Diffusion	Gijs Admiraal Delft University of Technology
14:55-15:20	Predicting the Solubility of Organic Compounds in Solvent Mixtures with different Machine Learning	Simona Buzzi Katholieke Universiteit Leuven
15:20-15:45	AI-guided design of property-optimized copolymers including stoichiometry and chain architecture	Gabriel Vogel Delft University of Technology
15:45-16:00	Break (15 minutes)	
16:00-16:45	**Keynote:** Extending the Canon of Computational Representations of Molecules	Daniel Probst Wageningen University and Research
16:45-17:00	Closing remarks	
17:00-18:00	Borrel	

Day 1, April 15th, 10:00 – 17:00

9:00 – 10:00 | Arrive, register, coffee
10:00 – 10:15 | Welcome Note

Session 1 Chair: David Tax

1 (Keynote) | 10:15 – 11:00 | Adding Insight to Artificial intelligence

Gabriel Weymouth
Delft University of Technology

Generative AI is seeing ever-increasing use in public life and AI based research is making ever-grander claims of progress. But how much of this is real progress and is there still a place for physical knowledge and insight in education, research, and science in the future? In this seminar, I will review some of the high-level findings on the strengths and limitations of generative AI and use it to argue for the necessity of including physical insights in data-driven models.

I will use fluid dynamic systems as the key examples for this talk. Fluid dynamic simulations are used in an ever-increasing range of applications, from designing ships that reduce fuel and noise pollution to controlling schools of robotic fish. Although AI may unlock extremely high-speed predictions for fluid flows, their accuracy depends on constraining the model with physical information at every opportunity. Integration of physical insights such as scaling laws, symmetries in the governing equations, and the use of physics-based input features see enormous quantitative improvements in state-of-the-art systems such as graph neural networks (GNNs) and enable excellent generalization and extrapolation to out-of-distribution test cases.

11:00 – 11:15 | Break

Session 2 Chair: Henk Noorman

2 | 11:15 – 11:40 | Compartment modelling meets Deep Learning: Towards an efficient modelling approach for bioreactor gradients

Hector Maldonado
Delft University of Technology

Reducing the uncertainty when scaling up bioprocesses is crucial for bringing them into the market. A key aspect of the analysis is the assessment of the effect of heterogeneities in the environment that the microorganisms experience. For this, Computational Fluid Dynamics (CFD) simulations have been proven to be a powerful tool. Nevertheless, its inherent computational demand hampers a fast and reliable evaluation of reactor operation schemes. As an alternative, compartment models (CM) have shown the capability to capture hydrodynamic features in stirred tank bioreactors at an affordable computational cost]. This is true for both steady conditions and dynamic ones, as in fed-batch fermentations. In this work, we compare two inferring methods to predict fermentation heterogeneities upon varying operating conditions. The first one implements a self- attention mechanism which is applied to grid points obtained from CFD. The second is inspired by recent

advancements and perspectives in implementing machine learning for fluid dynamics, hence we introduce a transformer-based strategy applied to a compartmentalization technique.

3 | 11:40– 12:05 | The flow must go on: dynamic scheduling algorithms for biomanufacturing

Kim van den Houten
Delft University of Technology

This study investigates scheduling strategies for the stochastic resource-constrained project scheduling problem with maximal time lags (SRCPSP/max)). This problem is recognized in biomanufacturing, in which process durations are stochastic due to the biological nature of unit operations such as fermentation, and the max time-lags due to unstabilized intermediate products that should flow to the next tank. The flow must go on!

Recent advances in Constraint Programming (CP) and Temporal Networks have re-invoked interest in evaluating the advantages and drawbacks of various proactive and reactive scheduling methods. First, we present a new, CP-based fully proactive method. Second, we show how a reactive approach can be constructed using an online rescheduling procedure. A third contribution is based on partial order schedules and uses Simple Temporal Networks with Uncertainty (STNUs). Our analysis shows that the STNU-based algorithm performs best in terms of solution quality, while also showing good relative computation time. We introduce three new methods for SRCPSP/max. The first method is a CP-based version of a proactive model. Then, we present a novel, fully reactive scheduling approach employing the deterministic model for RCPSP/ max. Finally, we propose an STNU-based approach using CP and POS.

4 | 12:05– 12:30 | Use of ensemble modelling for bubble size measurements with an optical fibre probe

Rik Volger
Delft University of Technology

Scale up of bioprocesses is often complicated by uncertainty about the gas-liquid mass transfer at large scale. Bubble size plays a key role in the maximum attainable mass transfer rates and thus advances in bubble size prediction will help realize novel large-scale bioprocesses1. Fiber probes can be used to measure bubble sizes in systems that are visually inaccessible due to e.g. broth opacity or high holdup. Recent advances in single-fibre probes allow for measurement of small bubbles by measurement of speed and fibre residence time from a single fibre.
Unfortunately, these approaches suffer from a low detection rate: only 30 % of bubbles pierced by the fibre can be attributed an accurate speed and size. Here we explore the use of neural networks to increase the validation rate. The central research question is: Can bubble speed be determined for more bubbles through neural network approaches, without sacrificing accuracy of the measurement?

12:30 – 13:45 | Lunch

Session 3 Chair: Ali May

5 (Keynote) | 13:45– 14:30 | On the challenges of predicting gut microbial community dynamic

Karoline Faust
Katholieke Universiteit Leuven

Rationale design of microbiome-based therapeutic interventions requires an in-depth understanding of the human gut microbiome and its interactions with the host, encoded in mathematical models. To gain such an understanding, we follow a bottom-up approach by studying the community dynamics of selected human gut bacteria in controlled conditions.

Model parameterization requires accurate data. We cultivate gut bacteria in anaerobic conditions and combine 16S rRNA gene sequencing with flow cytometry to obtain absolute abundances for community members. We also routinely quantify key metabolites, including sugars and short-chain fatty acids. We saw that a synthetic human gut bacterial community reaches a reproducible steady state in chemostat. However, when exploring a two-species interaction in-depth, we found it to be complex and context-dependent. We also observed that heterogeneity can occur at the population level.

To rationally design microbiome-based therapeutics, we need to systematically measure phenotypic traits of gut microorganisms and explore their survival strategies. Frequently used community models such as the generalized Lotka-Volterra, community extensions of flux balance analysis or standard consumer-resource models do not sufficiently account for metabolic flexibility of gut microorganisms and need to be adapted or replaced with more flexible approaches.

14:30 – 14:45 | Break

Session 4 Chair: Ali May

6 | 14:45– 15:10| ABaCo: Correcting Technical Heterogeneity for Metagenomic Data integration using Adversarial Generative Models

Edir Vidal
Novo Nordisk Foundation Center for Biosustainability

Integration of metagenomic data from multiple studies and experimental conditions is essential to understand the interactions between microbial communities in complex biological systems. The nature of metagenomic data and the increased diversity and biological complexity of samples introduce methodological challenges that require more refined strategies for atlas-level integration from diverse sources. Generative adversarial models have been used in various fields and in particular in single-cell transcriptomics to correct technical heterogeneity and improve data integration, creating an opportunity for a similar concept application into other omics.

In this research we propose ABaCo, a family of generative models based on Variational Autoencoders (VAEs) combined with the use of an adversarial discriminator. The objective is to achieve the integration of metagenomic data from different studies by minimizing technical heterogeneity without altering biological variability. The VAE learns to encode the data in a latent space and the discriminator is trained to detect the provenance of the data by eliminating the variability associated with this origin. To ensure biological conservation, the data is modeled based on a zero-inflated negative binomial (ZINB) distribution, while the latent space follows a mixture of Gaussian (MoG) distribution.

7 | 15:10– 15:35 | Prediction and reengineering of MADS-box proteins using structure-based machine learning

Alejandro Sanchez
University of Amsterdam

MADS-box proteins are a large family of transcription factors involved in the regulation of every major aspect of plant development. Their functional diversity is determined by their homo/heterodimeric protein-protein interactions. However, proteins with highly similar sequences exhibit vastly different interaction patterns, hinting at structural analysis for protein-protein interaction prediction. Previous studies have pointed at specific regions that seem to contribute to the interaction specificity, but there is currently no clear insight into the molecular determinants that enable or disrupt the interactions.
A large non-redundant dataset of 5,000 experimental interactions was compiled from databases, literature, and unpublished work. Since the protein sequence carries little direct information, we predicted the structure of each dimer using ESMFold and condensed the 3D relationships in distance and contact maps. Assuming that interacting complexes would yield a more stable structure with specific patterns showing in the maps, we built a convolutional neural network model to predict the probability of interaction of two proteins given their modelled structure. Furthermore, we leveraged the information of the gradients with DeepLift to highlight residue-residue contacts important for the model's prediction. This allowed to synthetically reengineer protein-protein interactions by introducing mutations.

15:35 – 15:50 | Break

Session 5 Chair: Kim van den Houten

8 | 15:50– 16:15 | ENFORCE: Exact Nonlinear Constrained Learning with Adaptive-depth Neural Projection

Giacomo Lastrucci
Delft University of Technology

Ensuring neural networks adhere to domain-specific constraints is crucial for addressing safety and ethical concerns while also enhancing prediction accuracy. Despite the nonlinear nature of most real-world tasks, existing methods are predominantly limited to affine or convex constraints. We introduce ENFORCE, a neural network architecture that guarantees predictions to satisfy nonlinear constraints exactly. ENFORCE is trained with standard unconstrained gradient-based optimizers (e.g., Adam) and leverages auto-differentiation and local neural projections to enforce any $C1$ constraint to arbitrary tolerance ϵ. We build an adaptive-depth neural projection (AdaNP) module that dynamically adjusts its complexity to suit the specific problem and the required tolerance levels. ENFORCE guarantees satisfaction of equality constraints that are nonlinear in both inputs and outputs of the neural network with minimal (and adjustable) computational cost.

9 | 16:15– 16:40 | Inferring Discrete Dynamical Models of Biological Systems with Program Synthesis

Reuben Gardos Reid
Delft University of Technology

We frame the problem of discovering qualitative networks as a program synthesis problem, where a qualitative network is seen as a concurrent program. Program synthesis offers several advantages that ensure we are discovering regulatory mechanisms grounded in limited available data. First, through program synthesis, we can discover models in the language of qualitative networks and still leverage existing techniques. Second, the language of programs allows us to incorporate any prior biological knowledge, which can also be invalidated by experiments. Third, we can easily incorporate verification techniques from computer science to ensure that the discovered models stabilize and reproduce only the behaviour observed experimentally.

16:40 – 17:00 | Closing remarks

17:00 – 20:00 | Dinner and Borrel at Firma van Buiten

Day 2, April 16th, 9:00 – 17:00

9:00– 9:30 | Arrive, register, coffee

Session 6 Chair: Mathijs de Weerdt

10 (Keynote) | 9:30– 10:15 | AI-driven decentralised biomanufacturing

Vitor Martin Dos Santos
Wageningen University and Research

Decentralised biomanufacturing offers the possibility of using wide variety of available biological feedstocks and avoids transport of unused fractions, e.g. water or chemicals of minor importance. Despite its great potential, however, the implementation of decentralised biomanufacturing is severely hampered by higher baseline costs due to lack of economies of scale, capital investment, and relatively high costs of specialised personnel to operate the bioreactors. Therefore, smart manufacturing approaches are essential to keep investment and operating costs low enough to create commercially viable business cases. Simple and robust technology is required that can generate consistent output (manufactured compounds) from a variable, complex and low-cost input (waste feedstocks). This requires both the microbes carrying out the conversions to be robust to variable feedstocks and often harsh industrially relevant conditions (pH, temperature, solvents, etc.), and the processes to be operated in an essentially autonomous manner. None of those two conditions are currently sufficiently in place in most industrial biobased processes.

In the scope of a series of research programmes, we advantage of recent advances in artificial intelligence, combined with modern synthetic biology and microbial ecology-inspired bioprocessing approaches, work on implementing and demonstrate a decentralised biomanufacturing process based on an AI-driven “digital twin” workflow, that is to be capable of autonomously matching product needs to local waste availability in the most economically competitive manner.

10:15 – 10:30 | Poster pitches

10:30 – 12:45 | Poster session + networking (see below for abstracts)

11:45– 12:45 | Lunch

Session 7 Chair: T.B.A.

11 | 12:45 –13:10 | Machine Learning-assisted pathway optimization in large combinatorial design spaces: a p-coumaric acid case study.

Paul van Lent
Delft University of Technology

Combinatorial pathway optimization is an important tool for industrial metabolic engineering to improve titre/yield/productivity of strains. The Design-Build-Test-Learn (DBTL) cycle, an engineering framework that aims to navigate through the large landscape of combinatorial designs using an iterative approach, has been increasingly augmented with machine learning approaches on all its

aspects. However, in terms of combinatorial pathway optimization, approaches have so far been limited to small design spaces with few targeted pathway elements, therefore limiting the advantages that machine learning assisted recommendation strategies may offer.

In this work, two DBTL cycles are performed on Saccharomyces cerevisiae for p-coumaric acid production. We first perform a library transformation on 19 genes with twenty promoters, which expands the size of the combinatorial design space significantly (>500 million configurations) compared to common combinatorial pathway optimization experiments. The machine learning method XGBoost, with a customized loss function, was trained to predict p-coumaric acid production outcomes. We propose a model-guided gradient bandit recommendation algorithm to choose strain designs for a next round of the DBTL cycle. This uses an exploration/exploitation parameter that we propose to balance by using a gene set diversity measure for the proposed strains versus predicted average production trade-off.

12 | 13:10 –13:35 | Automatic generation of control structures in industrial process flow diagrams with transformers

Dominique Goldstein
Delft University of Technology

An immense potential to assist in the design of a process plant lies in novel, generative artificial intelligence methods, which currently transform workflows across industries. In particular, transformer models demonstrate great generative power as the underlying technology in large language models, such as ChatGPT. Recent works show that transformers can learn patterns of complex PFDs and P&IDs. However, due to a lack of openly available PFDs for training, current models fail to predict complex control structure in real PFDs.
Recognizing this challenge, we recently initiated a cooperative research project with Linde Engineering and Siemens called P&ID CoPilot. Within the project scope, we explore the industrial applicability of generative AI for the generation of P&IDs by leveraging large quantities of real, P&ID data.

13| 13:35 –14:00 | Text2Model: Generating dynamic chemical reactor models using large language models

Sophia Rupprecht
Delft University of Technology

We fine-tune Llama 3.1 8B Instruct to synthesize a full, simulatable Modelica model based on a textual description of a reactor scenario in the user's prompt. The fine-tuning dataset is generated synthetically from a set of templated dynamic reactor scenarios. The supervised fine-tuning procedure is conducted using the parameter-efficient fine-tuning technique low-rank adaptation. We compare the performance of our fine-tuned model to the baseline Llama 3.1 8B Instruct model as well as GPT-4o as a benchmark. We assess the models' predictions manually with regards to syntactic and semantic accuracy of the generated dynamic models.

14:00 – 14:30 | Break

Session 8 Chair: Chengyao Peng

14| 14:30 –14:55 | All-Atom Novel Protein Sequence Generation Using Discrete Diffusion

Gijs Admiraal
Delft University of Technology

Traditional approaches for representing protein sequences rely on the 20 canonical amino acids, limiting their ability to incorporate non-canonical amino acids and post-translational modifications. Expanding the representations to include these could enhance the diversity and functionality of designed proteins.

Our research explores an all-atom protein sequence representation that captures the atomic composition of each amino acid, unlocking the design of unnatural or synthetic amino acids. Using a discrete diffusion model, we investigate how this representation influences the quality, novelty, diversity, and structural foldability of generated Protein sequences. Additionally, we examine the impact of different noise schedules on model performance. We employ a sequence-to-sequence convolutional neural network within the Discrete Denoising Diffusion Probabilistic Models (D3PM) framework, which supports categorical data and adaptable noise schedules. Our findings highlight the promising potential of all-atom representations in generating novel protein sequences. While challenges remain in consistently generating fully valid proteins, it is remarkable that successful sequences exhibit improved novelty and diversity.

15| 14:55 – 15:20 | Predicting the Solubility of Organic Compounds in Solvent Mixtures with different Machine Learning Approache

Simoni Buzzi
Katholieke Universiteit Leuven

Predicting solubility limits of organic compounds in solvent mixtures is crucial in many areas including environmental science, Chemical engineering, pharmaceuticals, and biomedical research. Advancements in machine learning (ML) have enabled fast and accurate predictions, offering a valid alternative to time-consuming experimental solubility measurements. Over the years, many ML models have been developed to determine physiochemical properties, including the solubility, of pure compounds. However, studies regarding mixtures remain limited. In this work, we propose hybrid models to estimate the solubility of organic molecules in mixtures of solvents.
Building on our previous work, SolProp, we developed SolProp-mix, an extension designed for solvent mixtures. Our novel model predicts the solubility of organic molecules in solvent mixtures across different temperatures (T< 350 K). To handle multiple solvents, we introduced MolPool, a permutation-invariant pooling function. The SolProp-mix implementation effectively predicts solubility trends in various solvent mixtures and temperatures, making it a valuable tool for industrial and pharmaceutical screening of organic compounds solubility.

16 | 15:20 – 15:45 | AI-guided design of property-optimized copolymers including stoichiometry and chain architecture

Gabriel Vogel
Delft University of Technology

This study proposes a generative model for inverse design of copolymers, extending previous works to consider also monomer stoichiometry and chain architecture as opposed to only the monomer chemistry. We developed a semi-supervised variational autoencoder (VAE) that integrates graph and string-based polymer representations. The model encodes stochastic polymer graphs[a] using a weighted directed edge message-passing neural network[a] and decodes them into string representations. We further include a property prediction network on top of the latent space of the VAE, enabling the model to generate polymers with specific properties. While the framework generates novel candidates, the results indicate that more diverse monomer structures and a larger number of stoichiometries and chain architectures could enhance the generative capabilities.

15:45– 16:00 | Break

Session 9 Chair: Wouter van Winden

17 (Keynote) | 16:00– 16:45 | Extending the Canon of Computational Representations of Molecules

Daniel Probst
Wageningen University and Research

Computational representation of molecules can take many forms, including graphs, string encodings of graphs—known as SMILES, binary vectors, or real-valued vectors comprised of numerical molecular descriptors or learned embeddings. These representations are then used in downstream classification, regression, or generative tasks, forming the basis of a wide range of machine learning models. Here, we present an overview of new approaches to the canon of computational representation of molecules that we have introduced over the last few years.
In our recent work, we have extended the canon of computational representations of molecules. We have introduced a framework of set-based representations for molecules, protein-ligand complexes, and reactions; the representation of chemical reactions as the symmetric difference of two sets; the use of vector-valued functions to learn representations of molecules, as well as protein-ligand and macromolecular complexes; the representation of protein structure and dynamics as heterogeneous graphs; and finally, the representation of small-molecule mass spectra as path graphs.
Various types of computational representations of molecules, from graphs to physicochemical descriptors, form the basis for a rich ecosystem of machine learning methodologies in biology and chemistry research. By introducing new approaches and methods, we believe we can lay the basis for further developments of machine learning techniques in biology and chemistry research that extend and diversify the currently available approaches.

16:45 – 17:00 | Closing remarks
17:00 – 18:00 | Borrel

Poster presenters – Day 2, 16th April

Title: T.B.A.

Stefan Loonen
Delft University of Technology

Production of valuable compounds, such as biofuels, vitamins and antibiotics, relies on heterologous overexpression of biosynthetic pathways. The effective yield of such production often is limited by the effectiveness of electron transfer between small-molecule electron carrier pools and redox-dependent enzymes such as iron-sulphur (Fe-S) proteins. Protein electron carriers (PEC) mediate the electron transfer and are widely conserved across the domains of life. Currently, it is not possible to predict whether PEC-enzyme pairs are compatible, which poses a wide-spread limitation in bioengineering.

Recent breakthroughs in computational protein structure prediction and protein complex assembly methods allow us to explore the features that determine electron flow efficiency in silico. We aim to develop a predictive computational model for activity of a particular Fe-S enzyme, IspG, in E. coli. To this end we combine molecular dynamics and machine learning methods.

Program Synthesis for Chemical Reaction Networks

Richard Wijers
Delft University of Technology

Many advancements in chemistry rely on accurate models of chemical networks to ensure stability, reliability, and effectiveness. For instance, in the development of medicines, an expert might be interested in maintaining concentrations of active ingredients, often requiring stabilizing agents. These chemical processes are often represented using Chemical Reaction Networks (CRNs). CRNs are a framework describing the interactions between species through reactions, making them a powerful tool to model, simulate, and analyze reactions. However, constructing CRN is a challenging task. Even when, for example, the products of a chemical network are known, the input compounds and the necessary reactions might be difficult to determine. Similarly, in chemical systems with incomplete data, inferring unknown reactions and states presents a significant challenge.

The goal of this work is to develop a tool for constructing CRNs through program synthesis.
This study focuses on two key problem scenarios: 1) Designing Stable Medicine Formulations; selecting appropriate substances and their quantities to maintain consistent concentrations of active ingredients, and 2) Inferring Unknown Chemical Reactions; deducing hidden reactions and states from partial experimental data. The challenge lies in the vast search space that the solver will have to traverse.

Self-supervised pretraining on polymer graphs with Joint Embedding Predictive Architecture

Gabriel Vogel
Delft University of Technology

Machine learning (ML) has shown promise in accelerating polymer discovery by enabling rapid property prediction for new materials. However, the
accuracy of these models is often constrained by the scarcity of labelled data. Self-supervised learning (SSL) has been successfully applied to small molecules and text-based polymer representations, allowing models to extract meaningful features from large, unlabelled datasets and improve performance on small labelled datasets. Despite the known scarcity of labelled polymer data, graph-based SSL remains underexplored for polymers with only one recently published work. In this study, we adapt the recently developed Joint Embedding Predictive Architecture (JEPA) for stochastic polymer graphs, introducing a polymer-JEPA model for self-supervised pretraining. We then investigate for two different datasets whether our pretraining approach enhances downstream prediction performance in label-scarce scenarios.

Interpretable models for quality control

Sjoerd de Haan
Data Driven Lab Consulting, Amsterdam

Nano particles are widely used in biomedical research and applications. Their size, shape and material can critically alter outcomes and should thus be monitored and controlled. It is challenging to observe nano particles directly. The go-to method for monitoring, nano tracking analysis, avoids direct observation by capturing the Brownian motion of nano particles suspended in a fluid. From the distribution of displacements, particle masses are derived, but not shapes. Fukuda et. al demonstrated that black-box deep learning can distinguish nano particles of rod and spherical shape. Black box models do not generalize to unseen particles and cannot be understood easily. In this work we build up towards a physics-based model.

We make use of latent variable models that we build out gradually. In order to build, test and critique quickly, we generate hypotheses, and we create simple models that we fit with probabilistic methods. We visualize the results, and we refine our hypotheses and models incrementally in the spirit of Blei et al.

LAST-PAIN: Learning Adaptive Spike Thresholds for Low Back Pain Biosignals Classification

Mahyar Shahsavari
Radboud University, Nijmegen

Spiking neural networks (SNNs) present the potential for ultra-low-power computation, especially when implemented on dedicated neuromorphic hardware. However, a significant challenge is the efficient conversion of continuous real-world data into the discrete spike trains required by SNNs. In this paper, we introduce Learning Adaptive Spike Thresholds (LAST), a novel, trainable encoding strategy designed to address this challenge. The LAST encoder learns adaptive thresholds to transform continuous signals of varying dimensionality–ranging from time series data to high

dimensional tensors–into sparse spike trains. Our proposed encoder effectively preserves temporal dynamics and adapts to the characteristics of the input. We validate the LAST approach in a demanding healthcare application using the EmoPain dataset. This dataset contains multimodal biosignal analysis for assessing chronic lower back pain (CLBP). Despite the dataset's small sample size and class imbalance, our LAST-driven SNN framework achieves a competitive Matthews Correlation CoeDicient of 0.44 and an accuracy of 80.43% in CLBP classification. The experimental results also indicate that the same framework can achieve an F1-score of 0.65 in detecting protective behaviour. Furthermore, the LAST encoder outperforms conventional rate and latency-based encodings while maintaining sparse spike representations. This achievement shows promises for energy-efficient and real-time biosignal processing in resource-limited environment.

Trust-Region Twisted Sequential Monte Carlo (TRT-SMC)

Joery de Vries
Delft University of Technology

Monte-Carlo tree search (MCTS) has driven many recent breakthroughs in deep reinforcement learning (RL). However, scaling MCTS to parallel compute has proven challenging in practice which has motivated alternative planners like sequential Monte-Carlo (SMC). Many of these SMC methods adopt particle filters for smoothing through a reformulation of RL as a policy inference problem. Yet, persisting design choices of these particle filters often conflict with the aim of online planning in RL, which is to obtain a policy improvement at the start of planning. Drawing inspiration from MCTS, we tailor SMC planners specifically for RL by improving data generation within the planner through constrained action sampling and explicit terminal state handling, as well as improving policy and value target estimation. This leads to our Trust-Region Twisted SMC (TRT-SMC), which shows improved runtime and sample-efficiency over baseline MCTS and SMC methods in both discrete and continuous domains.

BioLizard's Bio|Verse® framework for AI-powered visual data analytics

Christian Rausch
Biolizard B.V.

Bio|Verse is an AI-powered framework developed by BioLizard to facilitate data-driven research and development in the life sciences sector. Designed for computational scientists and biologists, it offers user-friendly software solutions tailored to the needs of biotech and pharmaceutical researchers. By streamlining data exploration and interpretation workflows, Bio|Verse aims to reduce time-to-insight for scientists and accelerate scientific discovery.

Using Bio|Verse, researchers can accelerate literature reviews, analyze microbiome compositions, identify biomarkers, and derive insights from multi-omics datasets. The platform's interactive visual tools enhance the interpretability of complex data, while its AI-powered analytics enable faster hypothesis generation and validation.

Kolmogorov Arnold Networks (KANs) as surrogate models for global optimization

Tanuj Karia
Delft University of Technology

Multi-layer perceptrons (MLPs) is a popular surrogate model to address the issue of tractability in global optimization of (bio)-chemical processes due to its ability to universally approximate any input/output relationship. Recently, a new class of machine learning models called Kolmogorov Arnold Networks (KANs) have been proposed. KANs are fully connected networks similar to MLPs, yet based on the Kolmogorov representation theorem instead of the universal approximation theorem. In this work, we investigate whether we can leverage the parametric efficiency offered by KANs relative to MLPs when using them as surrogate models for global optimization of (bio)-chemical processes. To embed KANs as surrogate models into optimization formulation, we propose a Mixed-Integer Nonlinear Programming (MINLP) formulation of a KAN and tested the proposed formulation on a case study considering the optimization of the auto-thermal reforming process. We also compared the performance of KANs with MLPs.

Leveraging Large Language Models for enzymatic reaction prediction and characterization

Lorenzo Di Fruscia
Delft University of Technology

Biochemistry is essential for various applications, including medicine, food production, and sustainable energy. Identifying novel biochemical reactions and biocatalysts is crucial for innovation but remains labor- intensive and costly. Recent advances in Machine Learning (ML), particularly Transformer-based models, have facilitated breakthroughs in molecular sciences, enabling tasks such as reaction prediction, retrosynthesis, and enzyme-substrate interaction modeling. However, traditional ML models often require extensive task-specific training and datasets, limiting their scalability. Large Language Models (LLMs) offer a promising alternative due to their generalization capabilities and transferability across multiple biochemical tasks.

We investigate the capability of LLMs in biochemical reaction prediction. We assess them on three tasks: Enzyme Commission (EC) number prediction, forward synthesis (FS), and retrosynthesis (RS). We use enzyme classification numbers as hierarchical classification (e.g. EC 1.2.3.4) and represent chemical reactions as reaction SMILES (Simplified Molecular Input Line Entry System). We compare two set-ups: a single-task and a multitask fine-tuning approach and assess parameter-efficient fine-tuning (PEFT). Further, we compare fine-tuning performance with zero-shot in-context learning, analyze the effects of different data regimes, and assess model generalization across biochemical tasks.

Hazard and Operability Study (HAZOP) development using Large Language Models (LLMs) and Knowledge Graphs

Achmad Anggawirya Alimin
Delft University of Technology

A Hazard and Operability Study (HAZOP) is critical to identify, evaluate, and minimize risk in chemical and bioprocess facilities. HAZOP systematically reviews the Piping & Instrumentation Diagram (P&IDs) of a facility by breaking them into nodes, analyzing potential deviations and risks, and suggesting recommendations. This review process can be tedious and lengthy, considering the scope and details in the P&ID, which affects the completeness and quality of a HAZOP study. Fortunately, the advancement in flowsheet digitization, data standards, and generative AI offer new opportunities to improve the HAZOP study process.

In this work, we aim to assist the HAZOP study development using knowledge graphs and LLMs. Our method is divided into steps: 1) we improve the flowsheet readability by representing them as knowledge graphs, 2) we integrate the P&ID knowledge graph to LLMs using graph-based retrieval augmented generation (graph-RAG), and we develop HAZOP data model development, which acts as guidelines and validation for LLM to develop HAZOP scenarios. The response from LLM is then evaluated and compared from a direct approach with LLMs.

Accelerated prediction of continuous fluid flow in stirred vessels

Mahdi Naderibeni
Delft University of Technology

Prediction of microbial performance in industrial fermentation can be achieved using models at different levels of fidelity, ranging from reduced-order models and compartment models to Direct Numerical Simulations. These levels of fidelity come with varying degrees of accuracy and computational cost. Recent developments in the Machine Learning field have motivated research on accelerated simulation of fluid flow systems. These attempts range from operator learning and enhanced turbulence modelling to generative modelling and physics-informed neural networks [2, 3]. This study focuses on accelerating the simulation of fluid flow inside stirred vessels. This task is a crucial step towards developing a digital twin for fermentation processes.

We utilize a Computational Fluid Dynamics (CFD) solver to simulate the flow field inside an industrial scale fermentation tank for multiple operating conditions (with stirring rate and liquid height as parameters). While the flow inside stirred vessel is inherently transient, by transforming the governing partial differential equations (PDEs) to a rotating reference frame, the system becomes steady in relation to this new frame of reference. We put our focus on learning implicit representations of the flow field (Velocity and pressure). These representations perform a mapping between operating parameters and velocity and pressure fields and are also open for constraining the model with the governing PDEs leveraging automatic differentiation.

Machine-Learning Guided Protein Engineering

Brakel Rowan
Delft University of Technology

The brain is a highly complex and sophisticated organ. Fully understanding how the brain works is crucial for curing diseases like Parkinson's and Alzheimer's disease. However, the currently available tools for studying brain dynamics are not yet fully optimised to capture the swift signalling that occurs in brain tissue. Therefore, several projects aimed to develop a new genetically encoded voltage indicator (GEVI) are being conducted across the globe. One such project is at the Brinks Lab (TU Delft), where rhodopsin are being utilised as a GEVIs. This family of proteins has already shown promising results, yet there was still room for improvement.

In this study, we explored the potential of a Machine-Learning tool to assist us in finding an enhanced version of our protein of interest.To build such a tool, we first tested two representative Large Language Models specially adapted for proteins: ProSE and ESM. These models can represent proteins numerically as vectors by embedding their sequences into a multi-dimensional protein space based on (structural) similarity. Subsequently, several regression techniques were tested to extrapolate the feature values from new mutants (e.g. brightness, sensitivity or reactivity) from previously screened rhodopsins. To verify the best combinations of parameters, we tested the aforementioned procedure on two big datasets of GFP and β-lactamase mutants.

Machine Learning-Enhanced Digital Twins: Utilizing Industrial Data for Spray Drying Operation

Maxmilian F. Theisen
Delft University of Technology

Spray drying plays a crucial role in various industries for producing powders from liquids and slurries. Reliable industrial-scale operation of spray dryers can be challenging due to intertwined phenomena. Digital twin frameworks such as soft sensors could help in spray drying operations, e.g. by providing real-time estimates of key powder properties. However, soft sensors need accurate dynamic models of spray dryers. At the same time, current machine learning approaches often lack access to sufficient industrial data and validation, limiting their ability to produce accurate and generalizable models for real-world spray drying processes.

We propose to utilize topology-aware graph neural networks to incorporate information about the spray dryer together with the process data. To capture thespray drying plant, we model major unit operations in a multi-stage spray dryer as nodes, material flows as edges and sensor measurements as graph attributes to build a process graph. Based on the approach, we develop a soft sensor model that is able to predict two key powder properties in spray drying, moisture content and tapped density. We train our ML models on an extensive dataset of multistage spray drying process data from 1.5 years of infant formula production. We evaluate the topology-aware graph neural network approach against five commonly used, black-box ML models.

Bayesian uncertainty quantification of graph neural networks using stochastic gradient Hamiltonian Monte Carlo

Qinghe Gao
Delft University of Technology

Graph neural networks (GNNs) have proven state-of- the-art performance in molecular property prediction tasks. However, a significant challenge with GNNs is the reliability of their predictions, particularly in critical domains where quantifying model confidence is essential. Therefore, assessing uncertainty in GNN predictions is crucial to improving their robustness. Existing uncertainty quantification methods, such as Deep ensembles and Monte Carlo Dropout (MC-dropout), have been applied to GNNs with some success, but these methods are limited to approximate the full posterior distribution.

In this work, we propose a novel approach for scalable uncertainty quantification in molecular property prediction using Stochastic Gradient Hamiltonian Monte Carlo (SGHMC). Additionally, we utilize a cyclical learning rate to facilitate sampling from multiple posterior modes which im-proves posterior exploration within a single training round. Moreover, we compare the proposed methods with MC-dropout and Deep ensembles, focusing on error analysis, calibration, and sharp-ness, considering both epistemic and aleatoric uncertainties.

In-feed antibiotic growth promoter and its natural alternative consistently alter the topology of microbial co-occurrence network in chicken gut

Chengyao Peng
Delft University of Technology

Due to increasing concerns over antibiotic resistance, several countries, including the EU, have banned the routine use of antibiotics for livestock. Therefore, the livestock industry is increasingly shifting away from antibiotic growth promoters (AGPs) to safer alternatives, such as phytogenic feed additives (PFAs). However, developing effective replacements is difficult due to our limited mechanistic understanding of AGPs. Since the dosage of antibiotics used is too low for direct pathogen inhibition, delineating th actual impact on animal gut microbiome is essential to improve such understanding. An overlooked aspect of previous studies characterizing such influence is the community-level gut microbiome dynamics.

In this study, we performed a comparative network analysis to investigate the effects of a common AGP and an alternative PFA on cecum microbial dynamics in broiler chickens. Using metagenomic data from a randomized controlled trial, we constructed microbial co-occurrence networks representing microbiome dynamics under basal, AGP-supplemented and PFA-supplemented diets across key development stages of poultry. Through network-level topological analysis and node-centrality analysis, we profiled the changes associated with the tested feed additives. Additionally, we assessed the ability of important nodes prioritized from the networks to characterize different diets using random forest (RF) model.

Rule-based autocorrection of Piping and Instrumentation Diagrams (P&IDs) on graphs

L. Schulze Balhorn
Delft University of Technology

Undetected errors or suboptimal designs in Piping and Instrumentation Diagrams (P&IDs) can cause increased financial costs, hazardous situations, unnecessary emissions, and inefficient operation. These errors are currently captured in extensive design processes leading to safe, operable, and maintainable facilities. However, grassroots engineering projects can involve tens to thousands of P&ID pages, leading to a significant revision workload. With the advent of digitalization and data exchange standards such as the Data Exchange in the Process Industry (DEXPI), there are new opportunities for algorithmic support of P&ID revision.

We propose a rule-based, automatic correction (i.e., autocorrection) of errors in P&IDs represented by the DEXPI data model. Our method detects potential errors, suggests improvements, and provides explanations for these suggestions. Specifically, our autocorrection method represents a DEXPI P&ID as a graph. Thereby, nodes represent DEXPI classes and directed edges the connectivity between them. The nodes retain all attributes of the DEXPI classes. Additionally, each rule consists of an erroneous P&ID template and the corresponding correct template, represented as a graph. The correct template includes the rule explanation as a graph attribute. Then, we apply the rules at inference time. The autocorrection method searches the erroneous template via subgraph isomorphism and replaces the erroneous with the corresponding correct template in the P&ID graph.