AI METHODS AND MODELS FOR (BIO)CATALYSIS AND SYNTHETIC BIOLOGY

JUNE 5, 2024

Amphitheater Charles Flahault,
Jardin des Plantes,
University of Montpellier, France

TABLE OF CONTENT

PROGRAM

08:30-09:00 – Welcoming with breakfast

09:00-10:50 – AI Methods and Models in Chemistry and (Bio)Catalysis

- 09:00-09:40 – Invited keynote – **Esther Heid** (MIT, USA and TU Vienna, AT)
  Machine Learning and Data Curation for Bioretrosynthesis
- 09:40-09:55 – Short talk – **Delphine Dessaux** (TBI, FR)
  Design of Symmetrical Multi-Component Proteins using Artificial Intelligence
- 09:55-10:35 – Invited keynote – **Wilhelm Huck** (Radboud U., NL)
  Information Processing in Chemical Reaction Networks
- 10:35-10:50 – Short talk – **Mehdi D. Davari** (IPB, DE)
  Empowering Data-Driven Protein Engineering with Machine Learning

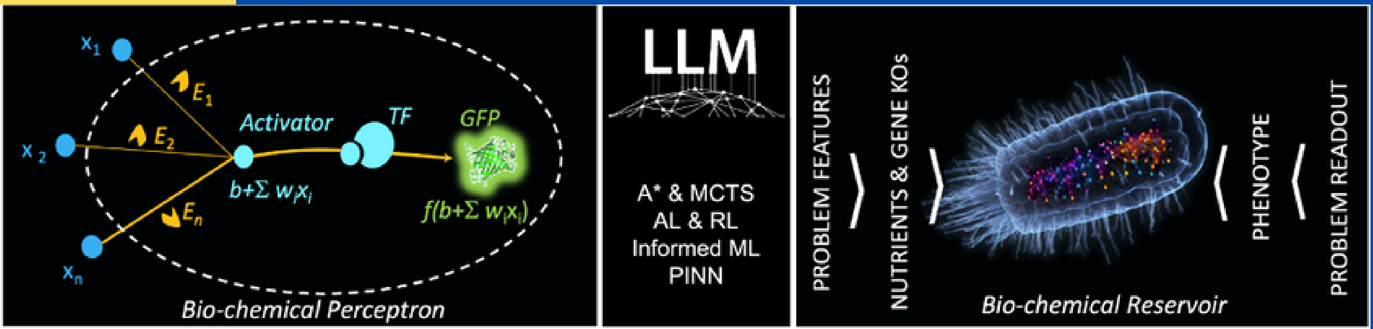
10:50-11:10 – Coffee break and poster session

11:10-13:00 – AI Methods and Models in (Bio)Catalysis and Synthetic Biology

- 11:10-11:50 – Invited keynote – **Ljubisa Miskovic** (EPFL, CH)
  Towards High-Throughput Dynamic Studies of Metabolism: Generative Machine Learning Approaches
- 11:50-12:05 – Short talk – **Amir Pandi** (INSERM, FR)
  A Multispecies Codon Optimizer Using Transformers
- 12:05-12:20 – Short talk – **Soutrick Das** (UCL, UK)
  Designing Neural Network Computation across Engineered Bacterial Communities
- 12:20-13:00 – Invited keynote – **Diego A. Oyarzún** (U. Edinburgh, UK)
  Machine Learning Approaches to Cell Factory Design and Optimization

10:50-11:10 – Coffee break and poster session

- **Yannick Bernard-Lapeyre** (LAAS, FR)
  Building Synthetic Cells through Active Learning and Automation
- **Sudarshan GC** (IIT, IT)
  Machine Learning-Driven Engineering of Shear Stress Sensors in T-Cells to Mitigate Exhaustion
- **Guillaume Gricourt** (INRAE, FR)
  AI Methods and Models for Retro-Biosynthesis
- **Bastien Mollet, Paul Ahavi** (INRAE, FR)
  Escherichia coli-based physical reservoir computing: potential and applications
- **Arnav Upadhyay** (IBE, DE)
  Mathematical Modeling of TX-TL Dynamics in Pseudomonas Putida KT2440



FOREWORDS

Over the past few years, there has been a notable convergence of AI methods in the fields of chemistry and biology. Techniques as varied as active learning, reinforcement learning, informed machine learning, physics-informed neural networks, reservoir computing, generative models, and large language models have gained prominence in both domains.
The primary objective of the workshop will be to facilitate knowledge exchange among communities that typically operate independently fostering a collaborative environment to explore shared experiences, differences, and challenges through the lens of AI techniques. The workshop is financially supported by the PEPR B-BEST, part of the France 2030 initiative, and by the annual symposium of the CNRS International Research Network in Synthetic Biology (IRN SYNSYSBIO).

ABSTRACTS - Invited keynotes

Esther Heid
MIT (USA) & TU Vienna (AT)

Machine Learning and Data Curation for Bioretrosynthesis

Computer-aided retrosynthesis has transformed organic synthesis planning, enabling a quick and interactive planning of reaction pathways. Yet an equally successful enzymatic counterpart has remained elusive until recently, where a simple retraining of popular retrosynthesis frameworks on enzymatic reaction databases obtained a too low accuracy to be applicable to a multi-step pathway search. This talk elucidates new advances on this forefront including the curation of a large high-quality dataset of enzymatic reactions, the development of successful bioretrosynthesis algorithms based on neural networks and transformers, as well as other machine learning models predicting reaction outcomes or regioselectivity.

Wilhelm Huck
Radboud U. (NL)

Information Processing in Chemical Reaction Networks

The flow of information is as crucial to life as the flow of energy. Living cells constantly probe their environment, and processing this information enables cells to adapt their behavior in response to changes in internal and external environmental conditions. Chemical reaction networks such as those found in metabolism and signalling pathways enable cells to sense physical properties of their environment, to search for food, or maintain homeostasis. Current approaches to molecular information processing and computation typically pursue digital computation paradigms and require extensive molecular-level engineering. Despite significant advances, these approaches have not reached the level of information processing capabilities seen in living systems.

In this talk, I will discuss our results on implementing concepts of reservoir computing in molecular systems. I will demonstrate how chemical/enzymatic reaction network can perform multiple non-linear classification tasks in parallel, predict the dynamics of other complex systems, and can be used to time-series forecasting. This in chemico information processing paradigm provides proof-of-principle for the emergent computational capabilities of complex chemical reaction networks, paving the way for a new class of biomimetic information processing systems.

ABSTRACTS - Invited keynotes

Ljubisa Miskovic
EPFL (CH)

Towards High-Throughput Dynamic Studies of Metabolism: Generative Machine Learning Approaches

Metabolism plays a crucial role in various physiological functions, from growth and reproduction to immune responses. Engineering metabolism in microorganisms, plants, and animals allows us to produce chemicals, pharmaceuticals, and fuels, facilitating a transition to a more sustainable, bio-based society. Understanding metabolism, predicting its normal functioning under changing environments, and manipulating genetic interventions' effects are pivotal in advancing biotechnology and medicine.

Nonlinear dynamic models can help us in these tasks because they comprehensively portray metabolic processes system-wide. Their unique capability to integrate omics and physicochemical data within a single mathematical framework makes them particularly effective for data integration. However, despite their potential for studying metabolism, large-scale dynamic models are still rare due to challenges in determining unknown kinetic parameters requiring specialized expertise and advanced computational methods. Consequently, researchers have yet to adopt these models widely in academic and industrial circles.

We introduced a novel generative machine learning framework, REKINDLE (REconstruction of KINetic models using Deep Learning), to democratize access to large-scale nonlinear dynamic models and facilitate dynamic studies of multiple phenotypes and large cohorts [1]. By harnessing the predictive capabilities of neural networks, this framework significantly reduces the substantial computational requirements of conventional kinetic modeling techniques. Its adaptability to large-scale studies is further enhanced through transfer learning, enabling the retraining of neural networks that parameterize kinetic models for new studies using only a small set of data points.

While this framework enhances the efficiency of model generation, it relies on preexisting kinetic modeling methods to produce the necessary data for training the neural networks, which could constrain its wide adoption. To overcome this limitation without sacrificing the efficiency of model construction, we developed a high-throughput generative machine learning framework, RENAISSANCE (REconstruction of dyNAmIc models through Stratified Sampling using Artificial Neural networks and Concepts of Evolution strategies)[2]. RENAISSANCE uses natural evolution strategies to efficiently parameterize dynamic models of metabolism without requiring training data.

We will illustrate the application of these frameworks through several case studies, showcasing their value for studies that analyze fluctuations in metabolism, encompassing variations in metabolite and enzyme levels as well as enzyme activities in health-related and biotechnological contexts. This work could significantly advance the research community's ability to conduct high-throughput, dynamic metabolism studies.

References

1. Subham Choudhury, Michael Moret, Pierre Salvy, Daniel Weilandt, Vassily Hatzimanikatis & Ljubisa Miskovic. Reconstructing Kinetic Models for Dynamical Studies of Metabolism using Generative Adversarial Networks. Nature Machine Intelligence 4, 710–719 (2022)
2. Subham Choudhury, Bharath Narayanan, Michael Moret, Vassily Hatzimanikatis & Ljubisa Miskovic. Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states. under review, https://doi.org/10.1101/2023.02.21.529387 (2023)

Diego A. Oyarzún
U. Edinburgh (UK)

Machine Learning Approaches to Cell Factory Design and Optimization

Machine learning has emerged as a promising paradigm for the optimization of cellular systems engineered for chemical production. In this talk I will describe our recent work at the interface of machine learning and cell factory design. We will first discuss the use of deep learning to predict protein expression from regulatory sequences commonly employed to control transcriptional and translational efficiency. In metabolic engineering, we have developed approaches to predict metabolite production dynamics from the integration of genome-scale metabolic models and kinetic pathway models, as well as robust algorithms for mixed-integer optimization of genetic control circuits for production pathways. We will conclude with recent results on active learning for improving yield of a complex lipopeptide using metabolomics data. The results showcase the many opportunities offered by the combination of data-driven and mechanistic models to improve production of heterologous proteins, secondary metabolites, and more complex molecules.

D. Dessaux
S. Buchet
M. Defresne
L. Barthe
G. Cioci
S. de Givry
L. F. Garcia-Alles
T. Schiex
S. Barbe
TBI (FR)

Design of Symmetrical Multi-Component Proteins using Artificial Intelligence

Natural metabolic pathways have been heralded as a viable route to green synthesis of biofuels and biochemicals. However, these bioprocesses can be difficult to engineer into chassis microorganisms, and, when successful, are often hampered by the limitations imposed by complex cellular metabolism, such as toxicity of products and intermediates, slow growth rates, and maintaining cell viability. An appealing solution to bypass these issues consists in isolating the metabolic production pathways from the organism's cytoplasm. Bacterial microcompartments (BMC) are proteinaceous entities that are composed of a shell encapsulating the enzymes involved in specific pathways. These BMCs spontaneously self-assemble in the cytoplasm of various bacteria and could therefore be repurposed for the optimization of in vivo bioproduction processes. To exploit the properties of the BMCs for the optimization of synthetic pathways, the organization of each components needs to be controlled, starting with the shell components. The most abundant components of BMCs shells are hexameric proteins, called BMC-H, which constitute the most suitable target for engineering new synthetic compartments. Engineering the monomers to achieve specific interactions could help control their spatial organization and, ultimately, that of enzymes in the synthetic microcompartments by covalent links to these monomers. Computational protein design (CPD) methods, more precisely negative multi-state design approaches that can consider favorable (positive) and unfavorable (negative) states, are crucial for designing diverse and specific protein interfaces. To address this negative multi-state design problem, we developed a hybrid generative AI approach, combining a deep-learned coarse-grain scoring function, called Effie, with a multi-state automated reasoning design tool. This approach was applied to RMM, a BMC-H protein, to predict sequence pairs, A and B, that can self-assemble in heterohexamers ABABAB yet fail to form homohexamers. Eventually, interaction between a few of designed AB proteins was experimentally verified using copurification and tripartite GFP techniques.

Soutrick Das
Chris Barnes
UCL (UK)

Designing Neural Network Computation across Engineered Bacterial Communities

Drawing inspiration from the intricate networks of biological neurons found in the brain, artificial neural networks (ANNs) have emerged as powerful tools in the field of artificial intelligence. In our study we explore the integration of ANN computation within bacterial biofilms. Our novel approach involves spatially arranged, engineered bacterial populations connected by intercellular signals. In our setup, each neuron is represented by a colony of engineered bacteria grown on an agar plate. Communication between colonies is facilitated by diffusible quorum sensing molecules. Signal reception and processing between colonies are influenced by individual location of colonies mimicking the spatial organisation of natural bacteria to perform specialization tasks. The activation function of each neuron is encoded in the bacterial response to diffusible signals, resembling sigmoid functions commonly used in deep learning. This allows for the creation of high-pass, low-pass, or non-monotonic activation functions. By leveraging the additive nature of signal concentration, arbitrarily complex functions are programmed simply by positioning colonies on the agar plate. In-silico, we have designed and optimised desired networks of interconnected bacterial colonies to execute various logic gate functions and achieved desired fold changes in the output based on different input types. This innovative approach presents a promising avenue for creating bio-inspired computational systems and molecular classifiers, merging principles of neural network computation with bacterial behaviour.

ABSTRACTS - Short talks

Mehdi D. Davari
IPB (DE)

Empowering Data-Driven Protein Engineering with Machine Learning

Protein engineering has become an indispensable tool across various domains such as biotechnology, biomedicine, and life sciences[1]. Despite significant technological advances, the full potential of protein engineering remains constrained by limited screening throughput, hindering efficient exploration of the vast protein sequence space [2]. Predicting beneficial amino acid substitutions, their combinations, and their impact on functional properties remains a formidable challenge [3]. In recent years, data-driven models have emerged as a promising avenue in protein engineering, capitalizing on advancements in large experimental databanks, next-generation sequencing (NGS), high-throughput screening (HTS) methods, and artificial intelligence algorithms [2a, 4]. Particularly, machine learning (ML) has garnered attention for its ability to navigate the large libraries of protein variants and uncover underlying rules and effects within the sequence space [4]. ML models optimize protein fitness by discerning relationships between sequences and their corresponding fitness values within the landscape [2a, 4].

To leverage ML methods for optimizing protein properties, we introduce PyPEF (Pythonic Protein Engineering Framework)[5], a versatile ML framework tailored for data-driven protein engineering and ML-assisted directed evolution. PyPEF combines ML techniques with signal processing and statistical physics methods, facilitating the identification and selection of beneficial proteins within a given sequence space through systematic or random exploration of variant fitness and random evolution pathways. We evaluated PyPEF's predictive accuracy and throughput performance using common regression models on four public protein and enzyme datasets. Nonetheless, effectively applying ML methods often demands a substantial amount of experimental data, which can be challenging to obtain within a reasonable timeframe. To tackle this issue and elucidate epistasis and residue coevolution patterns[3], PyPEF integrates ML techniques with evolutionary information (called MERGE method[6]). MERGE combines statistical modelling with an ML model trained on labelled sequence representations to adapt general protein knowledge to specific proteins of interest. By leveraging wet-lab data and insights derived from a protein's evolutionary history, MERGE facilitates data-driven strategies even with limited datasets, typically obtainable from experimentalists. This approach has significantly accelerated protein engineering experiments with data scarcity[7], encompassing both directed evolution and rational design approaches. In essence, MERGE offers a robust solution to current sequence exploration and combinatorial challenges in protein engineering through comprehensive in silico screening of the protein sequence space.

References

1. S. Pramanik, F. Contreras, M. D. Davari, U. Schwaneberg, Protein Engineering: Tools and Applications 2021, 153-176.
2. aJ. Yang, F.-Z. Li, F. H. Arnold, ACS Central Science 2024; bN. E. Siedhoff, U. Schwaneberg, M. D. Davari, Methods in Enzymology 2020, 643, 281-315.
3. M. Wittmund, F. Cadet, M. D. Davari, ACS Catalysis 2022, 12, 14243-14263.
4. P. Kouba, P. Kohout, F. Haddadi, A. Bushuiev, R. Samusevich, J. Sedlar, J. Damborsky, T. Pluskal, J. Sivic, S. Mazurenko, ACS Catalysis 2023, 13, 13863-13895.
5. N. E. Siedhoff, A.-M. Illig, U. Schwaneberg, M. D. Davari, Journal of Chemical Information and Modeling 2021, 61, 3463-3476.
6. A.-M. Illig, N. E. Siedhoff, U. Schwaneberg, M. D. Davari, bioRxiv 2022, 2022.2006. 2007.495081.
7. S. Hemmer, N. E. Siedhoff, S. Werner, G. Ölçücü, U. Schwaneberg, K.-E. Jaeger, M. D. Davari, U. Krauss, JACS Au 2023, 3, 3311-3323.

Amir Pandi
INSERM (FR)

A Multispecies Codon Optimizer Using Transformers

The genetic code is a universal set of DNA instructions for cellular protein production, composed of 64 three-nucleotide codons for 20 natural amino acids. Codon usage refers to the species-specific preferential selection of synonymous codons. This necessitates optimization of DNA sequences for gene expression in heterologous hosts especially in the era of de novo protein design and ever-decreasing DNA synthesis cost. Here, we introduce CodonTransformer, a novel approach leveraging transformer-based deep learning models trained on diverse organisms with over a million DNA sequences. Our model integrates organism-specific codon usage patterns and amino acid-codon embeddings to efficiently generate host-specific DNA sequences. Evaluation against existing codon optimization tools demonstrates CodonTransformer's ability to generate optimal DNA sequences with natural-like long-range code usage patterns across organisms.

ABSTRACTS - Posters

Y. Bernard-Lapeyre
A. Sakai
M.-J. Huguet
C. Danelon
LAAS (FR)

Building Synthetic Cells through Active Learning and Automation

Building a living cell from separate components faces a major hurdle: the huge number of parameters that must be explored as the system's complexity increases. We address this challenge by combining automation and active learning algorithms to navigate the vast experimental parameter space. Our approach integrates (i) robotics for large-scale exploration of molecular contents (e.g., lipids and PURE system components), (ii) high-throughput screening of gene-expressing vesicles, and (iii) artificial intelligence to accelerate the searching of biochemical compositions that lead to improved or novel vesicle properties.
We developed a workflow for enhancing protein synthesis yield and kinetics using active learning [1] and Echo-assisted dispensing of 20 different PURE constituents. New compositions resulting in higher expression levels in bulk reactions have been discovered. Follow-up experiments aim at encapsulating optimized PURE inside liposomes to boost up the occurrence of phenotypes that are relevant to build a synthetic cell. This integrated approach will be applied to the expression and evolution of larger 'synthetic genomes'. Moreover, first steps towards a closed-loop optimization workflow will be established, whereby all key operational steps will be executed in an autonomous manner.

References
1. Pandi et al. (2022) Nat Commun 13, 3876.

Sudarshan GC
David Dannhausser
IIT (IT)

Machine Learning-Driven Engineering of Shear Stress Sensors in T-Cells to Mitigate Exhaustion

Synthetic biology holds promise for revolutionizing biomedical interventions by engineering living organisms to exhibit novel functionalities. However, the unpredictable and inefficient nature of the engineering process poses challenges. In this study, we present a novel approach leveraging machine learning frameworks to enhance the predictability and efficiency of engineering mechanobiological sensors aimed at combating T-cell exhaustion—a critical hurdle in immunotherapy. Specifically, we focus on engineering shear stress sensors in T-cells, followed by actuators designed to alleviate T-cell exhaustion. Our methodology integrates principles from both biology and engineering, harnessing machine learning algorithms to optimize sensor design and functionality. Through rigorous experimentation and validation, we demonstrate the efficacy of our approach in mitigating T-cell exhaustion, offering a promising avenue for the advancement of synthetic biology in immunotherapeutic applications.

G. Gricourt
P. Meyer
T. Duigou
J.-L. Faulon
INRAE (FR)

AI Methods and Models for Retro-Biosynthesis

Retrosynthesis aims to efficiently plan the synthesis of desirable chemicals by strategically breaking down molecules into readily available building block compounds. Having a long history in chemistry, retro-biosynthesis has also been used in the fields of biocatalysis and synthetic biology. Artificial intelligence (AI) is driving us towards new frontiers in synthesis planning and the exploration of chemical spaces, arriving at an opportune moment for promoting bioproduction that would better align with green chemistry, enhancing environmental practices. In this review, we summarize the recent advancements in the application of AI methods and models for retrosynthetic and retro-biosynthetic pathway design. These techniques can be based either on reaction templates or generative models and require scoring functions and planning strategies to navigate through the retrosynthetic graph of possibilities. We finally discuss limitations and promising research directions in this field.

ABSTRACTS - Posters

B. Mollet [1]
P. Ahavi [2]
A. Cornuejols [1]
A. Le Gouellec [3]
E. Lutton [1]
A. Tonda [1]
J.-L. Faulon [2]

1. UMR MIA, AgroParisTech, INRAE - Paris-Saclay University (FR)
2. Micalis Institute (UMR 1319), INRAE - Paris-Saclay University (FR)
3. Grenoble Alpes University, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, CHU Grenoble Alpes, TIMC (FR)

Keywords: reservoir computing, metabolism, hybrid model, COVID-19

Escherichia coli-based physical reservoir computing: potential and applications

Synthetic circuits have been the cornerstone of bacteria-based computing since synthetic biology's early days. The principle consists in harnessing the unique physical properties of micro-organisms and particularly bacteria to execute specific tasks such as biosensing. Based on gene regulation mechanisms, these systems enabled the implementation of higher-order functions in living cells, but display serious limitations (such as noise, metabolic burden, and orthogonality issues) restricting the complexity of the tasks achievable. Here, we present a conceptually different alternative based on reservoir computing that circumvents some of these limitations.

Bacteria exhibit a variety of non-linear dynamics in response to changes in their growth environment. Such complex dynamics are not fully understood yet but process information about the growth conditions. Thus, bacterial phenotype can be considered as a projection of the growth conditions into an observable space functioning as a physical reservoir.

In this poster, we describe two possible use cases of bacteria-based physical reservoir computing. First, using in-silicomodels of Escherichia coli, we demonstrate that bacteria can be used in a problem-solving framework and we compare it to classic machine learning methods. Secondly, we designed a biosensing system which would be able to predict the severity of COVID-19 based on a blood sample from early infected patients.

Arnav Upadhya
Ralf Takors
IBE (DE)

Mathematical Modeling of TX-TL Dynamics in Pseudomonas Putida KT2440

Transcription-translation (TX-TL) coupling represents a potential rate-limiting factor in the biosynthesis of proteins. Precise and efficient modeling of this process is crucial for enhancing the overall yield of the target protein product.

The study uses a deterministic, dynamic model founded on differential algebraic equations (DAE) designed to simulate the in vivo transcription and translation of individual gene sequences into their respective protein sequences in Pseudomonas putida. This model originates from an in vitro model developed by Arnold *et al.* [1] and was further refined by Nieß et al. [2] to accurately represent Escherichia coli protein biosynthesis under in vivo conditions. Currently, the model is being translated from Matlab to the programming language Julia for application in P. putida.

Julia's open-source advantage and computational efficiency make it the optimal choice over Matlab and Python. The TX-TL model currently under development in Julia closely approximates the results of the Matlab-based model when applying kinetic parameters identified for E. coli, with a low error percentage for protein concentration. Following the assessment of metabolite concentrations and distinctive physiological attributes, the model will undergo further refinement to suit the requirements of the P. putida.

TX-TL modeling can be used for precise simulation of gene-to-protein translation in P. putida, crucial for optimizing protein production in synthetic biology. By employing Julia's computational framework, this approach significantly improves the accuracy and efficiency of synthetic biological systems.

References
1. https://doi.org/10.1007/b136414.
2. https://doi.org/10.1021/acssynbio.7b00117.

ORGANIZATION TEAM

Philippe Meyer
Data Scientist
PhD

Guillaume Gricourt
Engineer
PhD student

Joan Hérisson
Research Engineer
PhD

Jean-Loup Faulon
Group leader
PhD