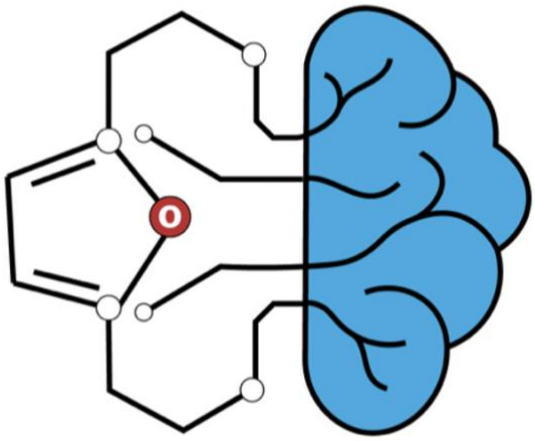
AIChemist



April 28, 2025 - May 2, 2025
CECAM-HQ-EPFL, Lausanne, Switzerland & online

Katya Ahmad
Helmholtz Zentrum München, Germany

Philippe Schwaller
EPFL, Switzerland

Igor Tetko
Helmholtz Munich, Germany

Rodolphe Vuilleumier
Sorbonne Université - ENS-PSL, France

With the support of:

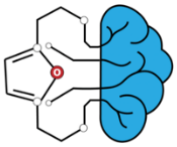
EPFL

EPFL

Università della Svizzera italiana

USI Lugano

Acknowledgement:



AICHEMIST

Funded by the European Union

The European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

SERI

Marie Skłodowska-Curie Actions

Marie Sklodowska Curie Actions

1. Description

The primary goal of the school is to provide young and established scientists (both at industry and academy) working at the intersection of explainable AI and Drug Discovery with a comprehensive overview of the most cutting-edge methods for predicting chemical reactivity, as well as a deep understanding of the remaining hurdles to be overcome in order for the field to advance further. Indeed, in the last few years Machine Learning (ML) has allowed to build upon large databases of chemical reactions to allow predictions of chemical reactions from given reactants, in a way similar to sentence translation. It is then possible to explore the chemical space by following reaction routes, to predict synthesis pathways or retrosynthesis pathways etc. Furthermore, an active field of ML for chemistry is the optimization of single step reaction conditions. The data in that case can be very different whether it comes from High Throughput Experiments (HTE), where 100's of thousands of reactions are performed automatically in a small chemical space and range of reaction conditions, or if it comes from a collection of standard laboratory works, like published works, that can be collected in databases such as the Open Reaction Database. In the latter case, the chemical space can be very large and the reaction conditions very broad and sometimes less documented, for example when using laboratory notebooks as source of information. Choosing the appropriate methods fitting the objective pursued with the data at disposal is then of primary importance and this school will provide tools for students to do so. In parallel, progress has been made as well on the understanding of chemical reactions thanks to ML techniques. ML has been used to develop force-fields or potential energy surfaces, MLP's (Machine Learned Potentials) that have been leveraged to perform Molecular Dynamics simulations. This has led to the accurate description of reaction mechanics at the atomic level. At the same time, new methods in Artificial Intelligence (AI) have been developed to provide explanations along predictions. This so-called explainable AI can not only bring new information from the data but also reinforce the liability of ML, in particular when extrapolating from known data.

These new developments are of interest to a broad range of practitioner of chemistry, and we propose to hold a school on this rapidly evolving field. The school is co-organized by the MSCA Doctoral Network “Explainable AI for Molecules - AiChemist” and as such will be geared towards the 14 doctoral candidates that were recruited into the network. As reaction predictions and explainable AI, both in the context of drug discovery and in general chemistry, is a burgeoning topic, we anticipate that this school will be of great interest to a large number of young scientists as well as academic and industrial researchers.

To start, different topics of ML for chemical reactions will be covered: prediction of chemical reactions and space exploration, single step models and optimization, planning of synthesis and optimization of reactions, machine learning potentials and description of reaction mechanisms, as well as extensions of these methods to biological systems. The participants will then be introduced to advanced ML methods (e.g. graph neural networks, Transformers) that are currently being applied to chemical reactivity problems, as well as methods that could potentially gain traction in this field in the coming years. Finally, a session will be dedicated to explainable AI and Large Language Models.

The participants will benefit from the know-how of world-renowned experts in reaction prediction and AI (explainable AI in particular) from Europe and the United States, and will gain insight into methodologies currently being pursued in the computing, pharmaceutical and life-science industries (including AstraZeneca, IBM and Reaxys), which will be invaluable for students that plan on a building a career in industry but as well as established scientists looking to use automated synthesis prediction in their work. Three hands-on sessions will allow for the participants to get familiar to common ML libraries.

2. Program

Day 1 - Monday April 28th 2025

- 08:00 to 08:45 - Registration
- 08:45 to 09:00 - Welcome & Introduction

Session Chair: Dina Khasanova

- 09:00 to 09:50 - **Jean-Louis Reymond**
  Expanding chemical reaction space for synthesizing the GDB

- 09:50 to 10:40 - **Heather Kulik**
  Leveraging experimental data in machine learning and screening to get from computational model to real world materials fast

- 10:40 to 11:45 - Coffee break

- 11:50 to 12:40 - **Alexandre Varnek**
  Chemography approach to (ultra)big chemical space analysis

- 12:40 to 14:00 - Lunch

Session Chair: Andi Hunklinger

- 14:00 to 14:50 - **Fernanda Duarte**
  Modelling chemical reactions in solution with machine learning potentials – balancing efficiency and accuracy

- 14:50 to 15:40 - **Ursula Roethlisberger**
  Computational chemistry joins forces with artificial intelligence in the design of biomimetic (metallo)enzymes

- 15:40 to 16:00 - Coffee break

- 16:00 to 16:50 - **Anne-Florence Bitbol**
  Capturing homology in protein language models

- 16:50 to 19:30 - Poster session & aperitif

Day 2 - Tuesday April 29th 2025

Session Chair: Paula Torren-Peraire

- 09:00 to 09:50 - **Guillaume Godin**
  How to select your molecular features?

- 09:50 to 10:40 - **Amol Thakkar**
  A mental toolbox for chemical synthesis: insights into representation, algorithms, and their constraints

- 10:40 to 11:00 - Coffee break

Session Chair: Amol Thakkar

- 11:00 to 11:50 - **Esther Heid**
  How (not) to produce data and code for reaction machine learning

- 11:50 to 12:40 - **Paula Torren-Peraire**
  Improving route development using convergent retrosynthesis planning

- 12:40 to 14:00 - Lunch

- 14:00 to 14:50 - **Mike Preuss**
  Monte Carlo Tree Search: from single-objective to multi-objective

- 14:50 to 15:10 - Coffee break

- 15:10 to 17:00 - **Samuel Genheden**
  Hands-on tutorial: retrosynthesis with AiZynthFinder

- 18:30 to 20:30 - Social dinner

Day 3 - Wednesday April 30th 2025

Session Chair: Gilles Marcou

- 09:00 to 09:50 - **Adil Kabylda**
  Introduction to (bio)molecular simulations with ML force fields

- 09:50 to 10:40 - **Damien Laage**
  Exploring chemical reactivity in aqueous solutions and at interfaces with deep potential molecular simulations

- 10:40 to 11:00 - Coffee break

- 11:00 to 11:50 - **Marwin Segler**
  Concepts in synthesis planning and generative molecular design

- 11:50 to 12:40 - **Timur Madzhidov (remote)**
  Navigating the future of reaction informatics: issues solved and challenges ahead

- 12:40 to 14:00 - Lunch

- 14:00 to 14:50 - **Janani Durairaj**
  Deep learning for protein and protein-ligand structure prediction

- 14:50 to 15:10 - Coffee break
- 15:10 to 17:00 - **Gilles Marcou**
  Hands-on tutorial: Multi-task learning: compound profiling using QSAR modeling

Day 4 - Thursday May 1st 2025

Session Chair: Marcel Hiltscher

- 09:00 to 09:50 - **Michele Ceriotti**
  Integrated machine-learning models for atomic-scale simulations
- 09:50 to 10:40 - **Julia Maria Westermayr**
  Machine learning potentials for excited states
- 10:40 to 11:00 - Coffee break
- 11:00 to 11:50 - **Luc Patiny & Igor Tetko**
  Consensus modelling of reaction pathways: interactive prediction and visualisation of millions of pathways using a java-enabled platform
- 11:50 to 13:15 - Poster Session with Coffee

Day 5 - Friday May 2nd 2025

Session Chair: Mateusz Iwan

- 09:00 to 09:50 - **Charlotte Bunne**
  Virtual cells and digital twins: AI in personalized oncology
- 09:50 to 10:40 - **Philippe Schwaller**
  Teaching language models to speak chemistry
- 10:40 to 11:00 - Coffee break
- 11:00 to 11:40 - **Kevin Jablonka**
  Encoding and decoding chemistry with language models
- 11:40 to 13:10 - **Kevin Jablonka**
  Hands-on tutorial: Transformers for chemistry and materials science
- 13:10 to 14:30 – Lunch
- 14:30 to 15:30 - Roundtable Discussion - Women in Science
- 15:30 to 15:50 - Closing Word

3. Abstracts

A mental toolbox for chemical synthesis: insights into representation, algorithms, and their constraints

Amol Thakkar
Former IBM Research, Switzerland

Over the past six decades, diverse approaches to synthesis prediction have emerged, ranging from rule-based expert systems to modern large language models (LLMs). This work centers on the challenges of synthesis prediction, emphasizing the interplay between representations, algorithms, and their inherent constraints. By examining the practical challenges faced by synthetic chemists, we explore how computational techniques can address these issues effectively. Through case studies and illustrative examples, we provide actionable insights into designing robust solutions and expanding the conceptual toolkit for tackling synthesis prediction problems. This discussion aims to equip researchers with a deeper understanding of the synergies between computational methods and the needs of synthetic chemists.

Capturing homology in protein language models

Anne-Florence Bitbol
EPFL, Switzerland

Protein language models trained on multiple sequence alignments of homologous proteins successfully capture coevolution between amino acids in structural contact. This is one of the ingredients of the success of AlphaFold. We have used such models, especially MSA Transformer, to generate new protein sequences from given protein families, and to predict which proteins interact among the members of two protein families. Despite their successes, a drawback of models based on multiple sequence alignments is that sequence alignment can be imperfect. Thus, we are developing ProtMamba, a homology-aware but alignment-free protein language model. ProtMamba starts from concatenated homologous sequences, is based on the Mamba state-space model architecture, and combines the strengths of autoregressive generation and masked language modeling.

Chemography approach to (ultra)big chemical space analysis

Alexandre Varnek
University of Strasbourg, France

Chemography deals with visualization and analysis of chemical data using 2D maps resulted from application of a dimensionality reduction method to N-dimensional vectorial space encoding molecular structures. Similar to geography, such maps provide with detailed view of studied chemical space and help chemist to discover novel chemical objects - molecules or reactions, integrating human expertise directly into the discovery process. As a dimensionality reduction technique, we use Generative Topographic Mapping (GTM) well suited to treat large arrays of data: it determines not only data points location on the map but also data density distribution. A GTM model (manifold) can be built on relatively small data set, followed by projection of remaining objects on the constructed map. GTM enables transformation of a complex chemical space into a human-interpretable 2D map displaying objects distribution (density landscapes), their properties (e.g., activity and class landscapes), and facilitates the construction of the GTM-based classification and regression models predicting properties of new compounds projected onto the map.

So far, GTM has successfully been used in various tasks including chemical space analysis and visualization, library design, pharmacological profiling and virtual screening. Being coupled with the modern generative AI models, chemography helps to design novel molecules or reactions possessing desirable properties. The GTM-driven chemography can particularly be useful for comparison and design of large chemical libraries: it helps to identify chemotypes (scaffolds, frameworks, MCS) unique

for a given library or to choose a library most similar to the reference data set. This feature was successfully applied to design focused DNA-encoded libraries (DEL) and to build Chemical Libraries Space encompassing several thousand DELs. In combination with deep learning networks, the GTM enables ultra-fast enumeration-free analysis and visualization of large combinatorial libraries.

Computational chemistry joins forces with artificial intelligence in the design of biomimetic (metallo)enzymes

Ursula Roethlisberger
EPFL, Switzerland

Through millions of years of evolution, Nature has accomplished the development of highly efficient and sustainable processes and the idea to understand and copy natural strategies is therefore very appealing. However, in spite of intense experimental and computational research, it is still a difficult task to design efficient functional biomimetic systems. Here we discuss a strategy for the computational design of small single-domain biomimetic compounds that consists of 1) selection of the natural target; 2) selection/design of an appropriate biomimetic scaffold; 3) atomistic and electronic characterization of the wildtype system and the biomimetic compounds; 4) identification of key descriptors and 5) efficient search of chemical and sequence space for optimization of the biomimetic system. To this end we make use of both traditional computational chemistry tools (such as classical and mixed quantum mechanical (QM/MM) molecular dynamics) with approaches from artificial intelligence (diffusion models, convolutional neural networks, feature selection models and genetic algorithm optimization). As a proof-of-principles study, this general approach is illustrated for the computational design of a 'green' catalyst mimicking the action of the zinc metalloenzyme Human Carbonic Anhydrase (HCA) that is a natural model for CO2 fixation since the enzyme is able to convert CO2 into bicarbonate.

[1] S.L. Dürr, A. Levy, U. Rothlisberger. Nat. Commun. 14 (1) 2713 (2023).
[2] E. Bozkurt, M. A. S. Perez, R. Hovius, N. J. Browning, U. Rothlisberger. J. Am. Chem. Soc. 140 (13) 4517 (2018).
[3] E. Brunk, M. A. S. Perez, P. Athri, U. Rothlisberger. ChemPhysChem 17 (23) 3831 (2016).
[4] E. Bozkurt, N. Ashari, N. Browning, E. Brunk, P. Campomanesa, M. A. S. Perez, U. Rothlisberger. Chimia 68 (9) 642 (2014).

Concepts in synthesis planning and generative molecular design

Marwin Segler
Microsoft, United Kingdom

In this talk I will give an overview over important concepts for validation of AI methods in chemistry, modern, generative models for molecular design and predictive synthesis using AI, and discuss limitations and future opportunities.

Consensus modelling of reaction pathways: interactive prediction and visualisation of millions of pathways using a java-enabled platform and Javascript

Luc Patiny & Igor Tetko
EPFL/Zakodium, Switzerland

The development of multistep chemical reactions is advancing rapidly, with new methods frequently emerging and being published, each offering unique advantages. We have implemented a consensus modeling approach that leverages both direct synthesis and retrosynthesis prediction models to forecast multistep synthesis pathways. Using the parallelization capabilities of the OCHEM platform, computations are executed concurrently on multiple GPU and CPU servers, enabling the prediction and enumeration of pathways interactively. This approach typically resolves millions of retrosynthesis pathways within 30–60 minutes. However, prioritizing and selecting the most promising pathways remains a significant challenge.

To address this, we utilize robust interactive JavaScript libraries that enable visualization and prioritization of pathways based on various criteria, such as pathway scores, industrial scalability, or literature support. The challenges associated with further platform development and the indispensable role of human expertise in complementing computational predictions will be discussed.

Deep learning for protein and protein-ligand structure prediction

Janani Durairaj
University of Basel, Switzerland

Predicting three-dimensional protein structures, particularly those involving small molecule ligands, is crucial for understanding biological function and enabling drug discovery. This lecture surveys recent advancements in deep learning-driven protein structure prediction, with a focus on extending these methods to model protein-ligand complexes. We will examine the architecture and validation of AlphaFold2 through extensive benchmarking, and subsequently explore current methodologies for predicting ligand-bound protein structures. We highlight key progress and remaining obstacles in accurately predicting protein-ligand complexes with a focus on the critical need for diverse curated datasets and robust evaluation metrics to assess this progress.

Encoding and decoding chemistry with language models

Kevin Jablonka
Helmholtz Institute for Polymers in Energy Applications, Germany

The field of chemical sciences has seen significant advancements with the use of data-driven techniques, particularly with large datasets structured in tabular form. However, collecting data in this format is often challenging in practical chemistry, and text-based records are more commonly used [1]. Using text data in traditional machine-learning approaches is also difficult. Recent developments in applying large language models (LLMs) to chemistry have shown promise in overcoming this challenge. LLMs can convert unstructured text data into structured form and can even directly solve predictive tasks in chemistry. [2, 3] In my talk, I will present the impressive results of using LLMs, showcasing how they can autonomously utilize tools and leverage structured data and “fuzzy” inductive biases. To enable the training of a chemical-specific large language model, we have curated a new dataset along with a comprehensive toolset to utilize datasets from knowledge graphs, preprints, and unlabeled molecules. To evaluate frontier models trained on such a dataset, we specifically designed a benchmark to evaluate the chemical knowledge and reasoning abilities. I will present the latest results, demonstrating the potential of LLMs in advancing chemical research. [4]

[1] K. M. Jablonka, L. Patiny, B. Smit. Nat. Chem. 14 (4) 365–376 (2022).
[2] K. M. Jablonka et al. Digital Discovery 2 (5) 1233–1250 (2023).
[3] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit. Nat. Mach. Int. 6 161–169 (2022).
[4] A. Mirza et al. arXiv 2404.01475 (2024).

Expanding chemical reaction space for synthesizing the GDB

Jean-Louis Reymond
University of Bern, Switzerland

Designing new drugs molecules is both fun and critically important for addressing unmet medical needs. Our aim is to identify novel but simple and easily synthesizable molecules by enumerating chemical space from first principles in the form of the GDBs (Generated DataBases)[1] and comparing its contents with known molecules.[2] We then move on to predict the synthetic accessibility of GDB molecules,[3,4] and to devise synthetic routes using retrosynthesis planning, preferentially with our triple transformer loop (TTL) method.[5,6]However, GDB molecules often all outside of the applicability domain of retrosynthesis prediction tools trained with known reactions.

To circumvent this limitation, we are investigating data augmentation methods to generate and validate fictive reactions involving GDB molecules.[7]
For more info, visit https://gdb.unibe.ch

[1] J.-L. Reymond, L. Ruddigkeit, L. Blum, R. van Deursen. WIREs Comput. Mol. Sci. 2 (5) 717 (2012).
[2] Y. Buehler, J.-L. Reymond. J. Chem. Inf. Model. 63(20) 6239 (2023).
[3] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, E.J. Bjerrum. J. Cheminformatics. 12(1) 70 (2020).
[4] A. Thakkar, V. Chadimová, E.J. Bjerrum, O. Engkvist, J.-L. Reymond. Chem. Sci. 12(9) 3339 (2021).
[5] D. Kreutter, J.-L. Reymond, Chem. Sci. 14(36), 9959 (2023).
[6] D. Kreutter, J.-L. Reymond, Chem. Sci. 15(43) 18031 (2024).
[7] Y. Grandjean, D. Kreutter, J.-L. Reymond. ChemRxiv (2024).

Exploring chemical reactivity in aqueous solutions and at interfaces with deep potential molecular simulations

Damien Laage
Ecole Normale Supérieure, France

Numerical simulations are essential for gaining a molecular-level understanding of chemical processes, offering insights that complement experimental techniques. For simulations to provide a realistic and predictive description, they must combine an advanced electronic structure representation of the reacting system with an explicit inclusion of surrounding environmental molecules. Additionally, they should be scalable to large systems and capable of capturing dynamics over long timescales. However, these requirements present significant computational challenges, often forcing compromises that reduce the accuracy and predictive power of conventional methods.
Recent advancements in machine learning offer transformative solutions to these challenges. By training neural network potentials on high-quality quantum mechanical data, it is now possible to simulate chemical reactions in condensed phases with remarkable efficiency and accuracy. This presentation will first explore the unique challenges posed by reactive systems for the training such potentials[1]. We will then illustrate the power and limitations of these approaches using examples from our recent work, including Grotthuss proton transport in water[2], the acidity of the air-water interface, and peptide bond formation in aqueous droplets[3]. Finally, we will discuss how these techniques can extend beyond reaction dynamics to applications in spectroscopy, with a particular emphasis on nonlinear vibrational sum-frequency generation[4]. These developments illustrate the potential of machine learning to transform the field of reactive molecular simulations, bridging the gap between accuracy and practical applicability.

[1] A. Gomez, M. de la Puente, R. David, D. Laage. C. R. Chimie in press (2024).
[2] A. Gomez, W. H. Thompson, D. Laage. Nature Chemistry 16 1838 (2024).
[3] R. David, I. Tuñón, D. Laage. J. Am. Chem. Soc. 146 14213 (2024).
[4] M. de la Puente, A. Gomez, D. Laage. Phys. Chem. Lett. 15 3096 (2024).

Hands-on tutorial: retrosynthesis with AiZynthFinder

Samuel Genheden
AstraZeneca, Sweden

This hands-on tutorial provides participants with practical experience in synthesis planning, with a particular focus on retrosynthetic analysis, complementing the concepts introduced in previous lectures. In the session we will use AiZynthFinder, an open-source Python-based tool developed by AstraZeneca. Participants will gain insight into configuring key parameters, including single-step retrosynthesis models, stock specifications, and search algorithms. Attendees are encouraged to bring their own molecular targets for analysis; alternatively, compounds from publicly available sources such as ChEMBL will be provided for exploration.

How (not) to produce data and code for reaction machine learning

Esther Heid
TU Wien, Austria

Machine learning models for retrosynthesis, product prediction, selectivity prediction, and reaction property prediction are increasingly used to assist in synthesis planning. However, effectively integrating chemical priors into the design of model representations and architectures remains a significant challenge. Additionally, the limited availability of open-source datasets poses obstacles to model development and benchmarking efforts. This talk addresses these two critical issues. First, I will introduce EnzymeMap, the largest curated dataset of atom-mapped enzymatic reactions, enriched with sequence information. EnzymeMap is already deployed in several enzymatic synthesis planners, enabling accurate product and precursor predictions both with and without sequence data. I will outline its development, key considerations, and diverse applications. Further, a new benchmarking suite for barrier height prediction will be introduced. Second, I will explore the challenges and recent advancements in machine learning architectures tailored to retrosynthesis and barrier height prediction. A central focus will be the introduction of a new software package for chemical reaction encoding that achieves state-of-the-art performance in barrier height prediction and other reaction property tasks. The package incorporates several innovative reaction representations and model architectures, offering new open-source tools and methodologies to advance the field.

How to select your molecular features?

Guillaume Godin
Osmo Labs, Switzerland

Molecular features play a fundamental role in predictive modeling, with early cheminformatic descriptors serving as the foundation for many approaches. However, classical descriptors often have limitations in capturing the complexity of molecular behavior. In this presentation, we will explore the evolution of molecular representations, from traditional descriptors to modern embeddings, and their impact on predictive performance. Through case studies on **boiling point, flash point, and solubility**, we will highlight scenarios where advanced embeddings outperform classical approaches, demonstrating their potential under specific conditions.

Improving route development using convergent retrosynthesis planning

Paula Torren-Peraire
Former J&J, Switzerland

Computer-aided synthesis planning approaches have allowed a greater exploration of potential synthesis routes; however, these methods are generally developed to produce individual routes from a singular product to a set of proposed building blocks and are not designed to leverage potential shared paths between targets. These methods do not necessarily encompass real-world use cases in medicinal chemistry, where one seeks to synthesize sets of target compounds in a library mode, looking for maximal convergence into a shared retrosynthetic path. Using a graph-based processing pipeline, we explore Johnson & Johnson Electronic Laboratory Notebooks (J&J ELN) and publicly available datasets to identify complex routes with multiple target molecules sharing common intermediates, producing convergent synthesis routes. We find that over 70% of all reactions are involved in convergent synthesis, covering over 80% of all projects in the case of J&J ELN data. We introduce a novel planning approach to develop convergent synthesis routes, which can search multiple products and intermediates simultaneously, enhancing the overall efficiency and practical applicability of retrosynthetic planning. We evaluate the multi-step synthesis planning approach using the extracted convergent routes and observe that solvability is generally high across those routes, being able to identify a convergent route for over 80% of the test routes and showing an individual compound solvability of over 90%.

Integrated machine-learning models for atomic-scale simulations

Michele Ceriotti
EPFL, Switzerland

Machine-learning techniques are often applied to perform "end-to-end" predictions, that is to make a black-box estimate of a property of interest using only a coarse description of the corresponding inputs. In contrast, atomic-scale modeling of matter is most useful when it allows one to gather a mechanistic insight into the microscopic processes that underlie the behavior of molecules and materials. In this talk I will provide an overview of the progress that has been made combining these two philosophies, using data-driven techniques to build surrogate models of the quantum mechanical behavior of atoms, enabling "bottom-up" simulations that reveal the behavior of matter in realistic conditions with uncompromising accuracy. I will discuss two ways by which physical-chemical ideas can be integrated into a machine-learning framework.
One way involves using physical priors, such as smoothness or symmetry of the structure-property relations, to inform the mathematical structure of a generic ML approximation. The other entails a deeper level of integration, in which explicit physics-based models and approximations are built into the model architecture. I will discuss several examples of the application of these ideas, from the calculation of electronic excitations to the design of solid-state electrolyte materials for batteries and high-entropy alloys for catalysis, emphasizing both the accuracy and the interpretability that can be achieved with a hybrid modeling approach, and providing an overview of the exciting research directions that are made available by these new modeling tools.

Introduction to (bio)molecular simulations with ML force fields

Adil Kabylda
University of Luxembourg, Luxembourg

Abstract: Molecular simulations help us understand and predict the behavior of complex systems in physics, chemistry, and biology. Traditional methods, such as classical force fields and ab-initio quantum chemistry, have greatly advanced the field but struggle to balance efficiency, accuracy, scalability, and transferability. Machine Learning Force Fields (MLFFs) offer a compelling alternative by combining quantum mechanical accuracy with the speed of classical models [1]. This talk will cover the fundamentals of MLFFs, including the construction of comprehensive training datasets, model design, and ways of capturing long-range interactions. Case studies will illustrate how MLFFs enable simulations of increasingly complex (bio)molecular systems with near ab-initio accuracy [2]. Finally, we will discuss current limitations and future directions.

[1] O.T. Unke et al. Chem. Rev. 121(16) 10142 (2021).
[2] A. Kabylda et al., ChemRxiv (2024).

Leveraging experimental data in machine learning and screening to get from computational model to real world materials fast

Heather Kulik
MIT, United States

Machine learning in transition metal chemistry has lagged behind other areas of chemistry due to the combination of diversity of chemical bonding and limitations in high quality data sets, experimental or computational. I will describe our efforts to overcome these limitations to accelerate the discovery of novel transition metal containing materials using machine learning. I will discuss how we have leveraged experimental data sets through both text mining and semantic embedding to uncover relationships between structure and function, disseminating high quality datasets of transition metal complexes with known function. I will describe how we've used these data sets to build machine learning models that predict the structure of transition metal complexes. Then I will describe how we have leveraged large datasets of synthesized materials to uncover those with novel function in polymer networks. I will demonstrate the success of our design strategy through macroscopically visible changes in network scale properties of polymers once our transition metal complexes are incorporated.

Machine learning potentials for excited states

Julia Maria Westermayr
Leipzig University, Germany

Machine learning (ML) has made significant strides in various fields including foundational models that can be used for the simulation of almost any system in the ground state. However, the integration of ML into photochemistry, which is important for many processes relevant to nature and life as we know it, remains a challenge. In this talk, we will explore the potential of ML models developed specifically for excited states to accelerate the discovery of chemical reactions and improve the prediction of molecular electronic properties [1,2]. In addition, we showcase first steps towards foundational excited-state models using the MACE architecture [3,4] and investigate the transferability from ground to excited states.

[1] J. Westermayr, M. Gastegger, D. Vörös, L. Panzenboeck, F. Joerg, L. González, and P. Marquetand, Nat. Chem. 14(8) 914 (2022).
[2] S. Mausenberger, C. Müller, A. Tkatchenko, P. Marquetand, L. González, J. Westermayr, Chem. Sci. 5(38) 15880(2024).
[3] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, G. Csányi, Advances in Neural Information Processing Systems 35 11423 (2022).
[4] T. S. Gutleb, R. Barrett, J. Westermayr, C. Ortner. arXiv (2024).

Modelling chemical reactions in solution with Machine learning potentials – balancing efficiency and accuracy

Fernanda Duarte
University of Oxford, United Kingdom

While computational characterization of reaction energy pathways has become routine, accurately accounting for solvent and dynamic effects remains a significant challenge. Machine learning potentials (MLPs) offer a promising alternative to traditional ab initio methods by enabling high efficiency and accuracy. This talk introduces efficient strategies for training reactive MLPs with minimal computational cost. Additionally, enhanced sampling techniques are employed to train MLPs where little knowledge of the PES is available. We examine the impact of descriptors, model architectures, training methodologies, and evaluation metrics on the performance and accuracy of MLPs across various systems.

Monte Carlo tree search: from single-objective to multi-objective

Mike Preuss
Universiteit Leiden, Netherlands

Monte Carlo Tree Search (MCTS) is a universal search method for action sequences, especially well-suited to huge search spaces that cannot be covered by deterministic methods. As such, MCTS variants are not only part of well-known AI success stories as AlphaGo but they are applied in various ways in Chemistry as well, e.g. for retrosynthesis.
However, MCTS it is much more a framework than an algorithm and applying it with the right setup is usually not trivial. Additionally, we now have some multi-objective variants that follow several goals at the same time (in different ways). The talk will explain how these methods work, how they can be configured meaningfully and what kind of adjustments are suggested if it does not work out of the box.

Navigating the future of reaction informatics: issues solved and challenges ahead

Timur Madzhidov
Elsevier, United Kingdom

In the rapidly evolving field of reaction informatics, the quality of data and datasets is paramount for developing reliable predictive solutions. This presentation will explore the current landscape of reaction

informatics from the perspective of a leading provider of data and predictive solution. We will discuss our experience in reaction data curation, challenges that we faced in structure curation, atom-to-atom mapping, results of it and our recommendations. We will delve into the intricacies of reaction property modeling, discussing the importance of robust descriptors, validation practices, and benchmarking methodologies to ensure the accuracy of predictive tools. Despite synthesis planning using retrosynthesis became routine, it is still do not fully address the demands of chemists. We want to give a critical analysis of current performance of existing solutions, discuss challenges that still wait for the resolution and features that are demanded from the modern retrosynthesis planners. By addressing the limitations and prospects for improvement in these areas, we aim to foster a deeper understanding of how high-quality data and innovative modeling techniques can drive the future of chemistry.

Teaching language models to speak chemistry

Philippe Schwaller
EPFL, Switzerland

Artificial Intelligence is transforming how we approach chemical research and synthesis. By teaching language models to understand and generate the language of chemistry, we have developed complementary AI systems that bridge the gap between computational design and experimental reality. Our large language model system, ChemCrow, represents one of the first demonstrations of an AI system directly controlling robotic synthesis platforms, successfully executing the synthesis of compounds including organocatalysts and chromophores. Complementing this, our small language model system, Saturn, currently the most sample-efficient molecular design algorithm, enables precise molecular generation with built-in synthesizability constraints. Saturn's innovations include direct optimization against retrosynthetic predictions and integration of building block availability, ensuring that generated molecules are practically accessible. Our work demonstrates how different scales of language models can work together to transform chemical research, from initial molecular design through to physical synthesis, potentially revolutionizing drug discovery, catalysis, and materials development.

Transformers for chemistry and materials science

Kevin Jablonka
Helmholtz Institute for Polymers in Energy Applications, Germany

During this hands-on session, we will build some key pieces for recent applications of transformers, in particular LLMs, in chemistry and materials science from scratch. We will build a GPT-style model on molecules from scratch using simple Python code, and will use simple Python code and the OpenAI API to let a model use external tools such as RDKit or Web APIs.
Materials: https://github.com/lamalab-org/llm-tutorial.

Tutorial on multi-task learning: compound profiling using QSAR modeling

Gilles Marcou, Dragos Horvath, Alexandre Varnek
University of Strasbourg, France

This tutorial is dedicated to prediction of pharmacological profiling of compounds. The pharmacological profile of a given compound corresponds to ensemble of its biological activities (or any other properties). A “classical” strategy to build the compound’s profile implies obtaining individual QSAR models for each activity. There are two main drawbacks of this approach: for each individual model one should (i) tune the model’s parameters, and (ii) setup validation procedure. This problem could be efficiently resolved using multi-task learning algorithms. In this tutorial, several algorithms are considered. Their efficiency is compared to the corresponding single task (classical) approach. The tutorial describes different ways to assess the models’ performance, parameters selection for multitasks algorithms and interference of individual modeling tasks.

Virtual cells and digital twins: ai in personalized oncology

Charlotte Bunne
EPFL, Switzerland

The complexity of cancer demands understanding biological processes across scales, from molecular interactions to tissue architecture. This talk explores how artificial intelligence enables the creation of digital twins at both cellular and tissue levels, with the aim to predict cellular phenotypes, function and their responses to perturbations such as cancer therapies. Concretely, I will introduce the Virtual Tissues (VirTues) platform, a foundation model framework that transforms how we analyze multiplexed tissue data and seamlessly integrates molecular, cellular, and tissue-scale information to increase diagnostic precision and biological understanding in personalized oncology. VirTues employs a multi-modal vision transformer architecture designed to learn from heterogeneous, high-dimensional datasets spanning different biological markers, measurement characteristics, and variable clinical annotations. While existing approaches often focus on H&E-stained slides, our framework incorporates highly multiplexed imaging techniques that capture hundreds of proteins within single tissue sections. Through unsupervised learning and a multi-scale neural network architecture, VirTues unifies these diverse data sources into a coherent virtual tissue space. As a result, new patient biopsy samples can be automatically mapped into this common representation. This enables integrative analyses of morphological, molecular and spatial complexity while facilitating clinically relevant predictions. To bridge insights from the analysis of patient samples with personalized treatment, we employ generative models trained on large biomedical datasets. These models predict treatment responses of biopsied cells from metastatic melanoma patients by revealing patterns of signaling pathway modulation associated with driver mutations and metastasis sites. Together, these approaches enable a multi-scale understanding of cancer biology and treatment response, advancing the development of personalized therapies guided by comprehensive digital twins of patient biology.

4. Posters

A full assessment of the template generative approach for computer-aided retrosynthetic planning

Nguyen Xuan Vu Nguyen, Zlatko Jončev, Daniel Armstrong, Philippe Schwaller
EPFL, Switzerland

Most approaches in data-driven computer-aided synthesis planning (CASP) rely either on *de-novo* generation of reactants, which are highly entropic, or reaction template selection, which is limited to a fixed template library. Combining the strengths of these two approaches, in this work, we formulate retrosynthesis as template generation without using any additional input besides the target molecule. We extensively study the utility and performance of this approach across three scenarios: Single-step, search-based multi-step, and direct multi-step. On USPTO-PaRoutes benchmark, we show that template generative models achieve up to 92% top-10 accuracy on single-step retrosynthesis, and 53% top-10 accuracy on search-based multi-step retrosynthesis on a subset of 1000 reactions from PaRoutes' n5 set, outperforming both traditional SMILES generative and template classification (AiZynthFinder) models. These preliminary results highlight the potential of template generation for retrosynthesis. The final version of this work will include a fully functional template-based direct multi-step model, along with comprehensive benchmarking of the multi-step models on the complete PaRoutes' n1 and n5 sets.

A heme location prediction tool with 3dCNN/docking procedure

Qihao Zhang[1], Simon Dürr[1], Ursula Röthlisberger[1], Ute F. Röhrig[2]
[1]EPFL, Switzerland
[2]Swiss Institute of Bioinformatics, Switzerland

Heme-containing proteins are involved in many important metabolic processes like oxygen transport in blood or gas sensing. However, most methods for heme prediction either only predict binding residues or require the number of hemes as the input. In this work, we develop a 3dCNN-based method Heme3d combining with force-field based docking by Swissdock for blind heme prediction. Given a single protein structure as input, Heme3d can generate an accurate full-atom heme location prediction. Now, in our test set of 1140 proteins, RMSD values in general are either <2.5 A (good vinyl positions) or 5.5-7 A (flip by 180 deg).

A web-based multi-target cytotoxicity prediction for multi-component nanoparticles: nano-QSAR model with extended applicability domain

Jaehyeon Park, Hyun Kil Shin
Korea Institute of Toxicology(KIT), Republic of Korea

As nanotechnology advances, increasingly complex nanoparticles are being developed for various applications, raising critical concerns about their potential toxicity. Not only Nano-QSAR models have been developed to predict their toxicity by cell lines separately, but also their applicability domain (AD) has been limited to specific nanoparticle types (i.e., bare metal oxide, coated metal, or carbon-based nanomaterials). This research introduced multi-target nano-QSAR model, being developed with improved AD by training the model on multi-component nanoparticles (MC NPs) to use size-dependent electron configuration fingerprint (SDEC FP) and with one-hot encoded cell features to predict cytotoxicity of MC NPs over 110 cell lines. The CatBoost regression model showed good performance (R2 test = 0.877) and is now accessible through user friendly web interface (https://www.kitox.re.kr/nanotoxradar). NanoToxRadar allows users to input nanoparticle specifications-including core, shell, doping, and coating materials, along with particle diameter-and receive predicted pIC50 values across 110 cell lines.

Automatic charge assignment for accurate hydration free energy calculations

Mathias Hilfiker[1], Marco Klähn[1], Ola Engkvist[1], Leonardo Medrano Sandonas[2], Alexandre Tkatchenko[3]
[1]AstraZeneca, Sweden
[2]TU Dresden, Germany
[3]University of Luxembourg, Luxembourg

Assessing the solubility of drug candidates is a critical step in hit optimization, as poor solubility can significantly limit bioavailability. In high-throughput screening, experimental determination of solubility-related properties is often constrained by the number of compounds that can be feasibly analysed. To address this, computational methods have become essential tools. The accuracy of absolute hydration free energy (AHFE) calculations strongly depends on the quality of atomic charge assignment. While density functional theory (DFT)-derived charges offer high accuracy, their computational cost and sensitivity to molecular conformation limit their practical use. Semiempirical methods, such as AM1-BCC, although less accurate, are the most common choice due to their computational efficiency and lower dependence on the input conformation.
In this work, we present a novel approach for generating DFT-quality atomic charges that are representative of the conformational ensemble of a molecule. Our method employs an equivariant, descriptor-based machine learning (ML) model trained to predict electrostatic potential (ESP) charges combined with molecular dynamics (MD) sampling, thereby capturing the conformational dependence of atomic charges without the need for expensive quantum mechanical calculations.
We evaluated the performance of our ML-based charge assignment by computing AHFE values for 22 compounds from the FreeSolv dataset and comparing them to those obtained using AM1-BCC charges. Our model achieved a reduction in root mean squared error (RMSE) by 1.3 kcal/mol and improved the Kendall's τ correlation coefficient from 0.75 to 0.84.
These results demonstrate that our method provides enhanced accuracy and ranking performance without increasing the computational cost of free energy calculations, making it a promising tool for high-throughput drug discovery workflows.

Digital chemical synthesis at syngenta crop protection

Nataliya Lopanitsyna, Marta Pasquini
Syngenta, Switzerland

The exponential growth of chemical reaction data is accelerating the adoption of Artificial Intelligence (AI) in reactivity prediction. As both the volume and quality of data increase, enabling the training of more sophisticated models, three key needs have emerged:
Making AI models easily accessible for everyday use by scientists,
Generating high-quality, machine-readable data to fuel better model performance, and
Developing advanced tools to explore and navigate complex chemical data.
At Syngenta, we are addressing these needs through the **Data4Synthesis Platform**—a solution designed to close the loop between data generation and model improvement. The platform ensures that reaction data is captured at the source in a machine-ready, fully annotated format, allowing it to feed directly into the continuous training and refinement of AI models. It also offers intuitive access to predictions from multiple reactivity tools, as well as literature-derived data, enabling informed decision-making across the synthesis lifecycle.
This effort is powered by two in-house software tools: **LinChemIn** and **NOCTIS**.
LinChemIn acts as an abstraction layer between retrosynthetic route providers—whether computational tools or human experts—and the platform. It standardizes formats and harmonizes route representations into a unified data model. This unification enables meaningful analysis and comparison of synthetic routes, including metric computation and descriptor extraction, supporting data-driven synthesis planning.
NOCTIS builds chemical knowledge graphs by linking reaction datapoints across diverse sources, constructing comprehensive reaction networks. It offers tools to explore these networks, mine synthetic routes from experimental data, and integrate them with AI model outputs. NOCTIS also serves as the foundation for a forthcoming language model-powered recommendation system, driving data-informed synthesis design through advanced network analytics.

Explainable random forest predictions of polyester biodegradability using high-throughput biodegradation data

Madeleine Parker, Philippa Jacob, Jonathan Hirst
[1]University of Nottingham, United Kingdom

Biodegradable plastics offer a promising solution to environmental challenges posed by traditional plastics. However, conventional methods for assessing their biodegradability are time-consuming and resource-intensive. With limited data available, predicting polymer biodegradability remains a significant challenge. In this study, we present an integrated experimental and computational framework for accelerating the identification of biodegradable polyesters. We report the development of a high-throughput enzymatic biodegradation assay, and use the experimental data generated to train machine learning models for biodegradability classification. Among the configurations tested, a random forest classifier using RDKit fingerprints achieved 78% accuracy, balancing predictive performance with interpretability. SHAP analysis revealed key structural features influencing model predictions, supporting chemically intuitive interpretation. Notably, the model only requires the chemical structure of a polymer as input, enabling in-silico screening of candidate materials before synthesis and reducing experimental overhead.

Attempts to enhance performance using transfer learning and chained modelling with external literature data did not yield improvement, highlighting the challenge of integrating datasets across differing chemical and experimental contexts. These findings suggest future efforts should focus on dataset expansion and domain adaptation strategies to enable broader generalisation and improved predictive capability. This work establishes a scalable and interpretable platform for the early-stage discovery of biodegradable polyesters, streamlining the path toward environmentally responsible polymer design.

Guide the training of enzyme-substrate prediction models with explainable AI

Andrea Hunklinger[1], Noelia Ferruz[1], Santiago Villalba[2]
[1]Centre for genomic regulation, Spain
[2]Bayer AG, Germany

Advancements in deep learning are revolutionizing protein research. Transformer-based models have demonstrated success across a range of tasks, from predicting protein structures to designing functional enzymes. However, these models often operate as black boxes, leveraging mechanisms that are not directly human-interpretable. Explainable AI (XAI) has been used to evaluate whether such models capture underlying scientific principles, yet it remains challenging to determine if this knowledge is utilized during prediction. We propose a shift in the use of XAI - from relying solely on post-hoc analysis to incorporating interpretability insights directly into the training process. Starting from a pretrained enzyme-substrate prediction model, we introduce a novel explanation loss that integrates gradient-based feature attribution results, evaluated against prior knowledge of substrate-interacting residues and functional groups - to coach model behavior in a more biologically-informed manner.

Improving multi-task learning for drug discovery: descriptors, data, and evaluation

Ekaterina Sosnina
Unaffiliated, Austria

Multi-task learning (MTL) is a machine learning approach in which a single model is trained to solve multiple related tasks simultaneously. Unlike traditional single-task learning (STL), where models are trained separately for each task, MTL uses shared information across tasks, which helps capture patterns that improve generalization and prediction accuracy. To fully benefit from MTL, several key aspects of model development must be considered.
First is the choice of input representation. STL usually depends on predefined descriptors, while MTL can also work with raw interaction data between compounds and targets. I found that including

molecular and target descriptors remains important in MTL, especially for predicting interactions involving new compounds or targets not seen during training.
The quality and diversity of the training data also play a critical role. I explored how expanding the training set with more diverse and informative examples affects model performance. This led to clear improvements, although I observed that the effect depends on both the amount and the variability of the added data.
Another key factor is how the model is evaluated. I recommend using both warm-start (previously seen compounds) and cold-start (new compounds) test sets to better reflect real-world prediction challenges and provide a more complete evaluation of model performance.
This study provides practical recommendations for improving MTL models in drug discovery by focusing on descriptor use, data enrichment, and evaluation strategies.

In silico design of enzymes to degrade short-chain per- and polyfluoroalkyl substances (PFAS)

Salomé Guilbert, Ursula Röthlisberger, Philippe Schwaller
[1]EPFL, Switzerland

Increasing concern over the persistence of per- and polyfluoroalkyl substances (PFAS), often referred to as “forever chemicals”, has pushed for the need to design technologies for their remediation. Exposure to these molecules has been associated with increased health defects and to date, the only commercially PFAS degradation method involves costly and unsustainable thermal deconstruction. Bioremediation stands as an attractive, cost-effective and sustainable solution for degrading such pollutants, although few bacteria and no enzymes have yet been discovered to achieve the full degradation of PFAS.

The in silico rational design of novel enzymes for PFAS degradation first involved the selection of a suitable biological target that had the potential to be optimised for better substrate binding, catalytic activity and thermal stability. A theozyme was created with the catalytic residues of fluoroacetate dehalogenase, a well-studied enzyme capable of breaking the strong carbon-fluorine bond (~460 kJ/mol) in fluoroacetate. First, RFDiffusionAA helped generate a library of suitable scaffolds around the theozyme. After sequence design with ProteinMPNN, AlphaFold was used to filter the sequences that folded in the desired structures.

To assess the catalytic activity and binding energies of these enzyme designs, QM/MM simulations will be performed, such that C-F bond dissociation energies can be computed. Further, selected biomimetic systems will be optimised with genetic algorithms, classical molecular simulations and machine learning tools such as AI.Zymes. This process will be applied to short-chain PFAS molecules, often overlooked as degradation products of longer-chain PFAS.

Modern XAI methods: hype or help?

Dina Khasanova, Igor Tetko
HMGU, Germany

The results of neural networks can be questionable, especially for experimental scientists, and the architecture itself often seems like a black box. That is the reason Explainable AI (XAI) is gaining more and more popularity each year, but the reliability of these methods remains an open question. Previous work has already shown inconsistencies between methods, models, and human intuition. However, the accuracy of the models for which explanations were retrieved is not high enough, which could be the cause of the ambiguous results. Additionally, not all methods can be compared straightforwardly, especially when transformers are part of the overall architecture and attention weights are used for explanation. In our work, we compare several XAI methods on highly accurate models, taking into account the specificity of each method, and show that modern techniques can be reliable and may eventually be used to convert learned rules into structural alerts, CSRML, or SMARTS.

Pharmacovigilance meets demographics: towards personalized cardiotoxicity predictions

Mateusz Iwan[1], Marina Garcia de Lomana[2], Francesca Grisoni[3], Alessandra Roncaglioni[1]
[1]Mario Negri Institute of Pharmacological Research IRCCS, Italy
[2]Bayer AG, Berlin, Germany
[3]Eindhoven University of Technology, Eindhoven, Netherlands

Drug-induced cardiotoxicity is a significant concern in drug development, often leading to late-stage clinical trial failures and post-market withdrawals. Adverse effects, such as arrhythmias, cardiomiopathies, and QT prolongation may arise unexpectedly, posing significant risks to patient safety. Moreover, cardiotoxic responses can vary across populations due to biological, genetic, and demographic differences, underscoring the need for a more personalize approach to risk assessment. To address this, we developed a novel pharmacovigilance-based dataset incorporating sex, age, and weight and used it to train machine learning models for personalized cardiotoxicity prediction.
Patient reports from the FDA Adverse Event Reporting System (FAERS), spanning from 2012 Q4 to 2024 Q3, were retrived. The drug descriptions were cleaned and mapped to active ingredients using multiple external databases. Cardiotoxicity events were identified through MedDRA Preferred Terms (PT) within selected High-Level Group Terms from the Cardiac Disorders System Organ Class. Disproportionality Analysis (DPA) metrics and their Confidence Intervals were used to detected cardiotoxicity signals and assign labels to compounds. We systematically evaluated the impact of different DPA metrics, PT terms selected to denote cardiotoxicity, and the inclusion of drugs labeled as "Secondary Suspect" in our analyses.
The developed dataset demonstrated a reasonable agreement with DICTRank, a recently released Cardiotoxicity dataset based on FDA labelling documents, with a tendency to be more conservative in labeling cardiotoxic drugs. Moving forward, we will explore different model-descriptor combinations and optimize the integration of demographic variables to improve prediction accuracy.

Predicting radical addition reaction feasibility via qm-ml integration in sparse data regions

Bob van Schendel[1], Mike Preuss[2], Mikhail Kabeshov[3]
[1]AstraZeneca & Leiden University, Sweden
[2]Leiden University, Netherlands
[3]AstraZeneca, Sweden

There are many opportunities within the reaction data space, especially for drug discovery. These opportunities are often hampered by regions of sparse data. This research focuses on these regions, which are crucial for ML models, used for prediction and generation within this space. Our overall aim is integrating quantum mechanical (QM) simulations with machine learning (ML), to predict feasibility of candidates for a wide range of reactions. Our first project focuses on predicting the feasibility of radical addition reactions. We have developed a workflow involving simulating Gieselike reactions by conducting conformational sampling, performing density functional theory (DFT) optimization, and calculating different energies of the components in question. Derived energies serve as predictors for the ML models. Model reliability has been validated over multiple rounds, forming a solid foundation for active learning in ML models. This approach has the potential to significantly enhance reactivity prediction across a wide range of reactions, specifically in sparse data regimes. It will therefore contribute to the advancement of computational synthesis planning as well as guiding experimental setups by providing a fundamental prediction of reactivity.

Publishing neural networks in drug discovery might compromise training data privacy

Fabian Krüger[1], Johan Östman[2], Lewis Mervin[1], Igor Tetko[3], Ola Engkvist[1]
[1]AstraZeneca, Sweden

[2]AI Sweden, Sweden
[3]Helmholtz Munich, Germany

Neural networks have become central to modern drug discovery, enabling accurate predictions of molecular properties. While sharing these models promotes open science and collaboration, it can also expose proprietary chemical data to privacy risks. This study presents the first systematic assessment of such risks using membership inference attacks in a black-box setting. We evaluate various neural network architectures and molecular representations, revealing significant vulnerabilities—particularly for minority-class molecules, which are often the most valuable in drug discovery. Interestingly, graph-based molecular representations and message-passing neural networks appear to offer some privacy resilience. Our findings underscore the need for balancing model openness with data confidentiality in cheminformatics.

Quantification of molecular selection by assembly theory

Ramil H. Babazade, Sebastian Pagel, Abhishek Sharma, Leroy Cronin
University of Glasgow, United Kingdom

The observation of complex molecules is the outcome of evolutionary processes that occur within all biological life. However, **quantification of evolutionary selection at the molecular level** to create biologically relevant molecules remains a complicated and open problem. The approach presented here compares natural products (NPs), which can be regarded as pre-validated by nature, with a broader molecular database to provide insight into **how molecular novelty emerges and persists.** It also explores the **systematic reduction of construction constraints**, referred to as contingency loss resulting in generation of novel molecular structures.

Topology-aware GNNs: teaching molecules their quantum features

Subashini Kennedy[1], Rodolphe Vuilleumier[1], Marc Bianciotto[2]
[1]École normale supérieure - PSL, France
[2]Integrated Drug Discovery, Sanofi R&D, Vitry-sur-Seine F-94403, France, France

Understanding the electronic structure of molecules is essential for modeling reactivity, stability, and chemical behavior. This project explores a novel methodology that integrates electron density topology, specifically Electron Localization Function (ELF) descriptors into graph neural network (GNN) architectures. The aim is to evaluate whether features derived from quantum topological analysis can enhance explainability and predictive power in molecular property prediction.
Using Auto-Qchem, an automated DFT workflow tool, twenty-four quantum mechanical (QM) descriptors were generated for each molecule. These included HOMO-LUMO gaps, dipole moments, and thermodynamic energies. To enrich this feature space with quantum topological information, TopChem2 was incorporated into the workflow. This program computes critical points in the electron density—such as nuclear, bond, and ring critical points—along with descriptors like electron localization function (ELF), Laplacian, delocalization indices, and basin populations. Outputs include visualizable files (.cube, .xyz) and text-based reports, enabling detailed topological analysis.
TopChem2's data, particularly the ELF-based descriptors, were parsed and organized into structured JSON files using custom Python scripts. These files capture coordinates, charge distribution, critical point classification, and ELF intensity, creating a comprehensive molecular fingerprint. The resulting descriptors are mapped onto chemical graphs, where nodes represent atoms and topological features, and edges encode chemical connectivity and spatial relations.
The finalized pipeline enables the seamless transformation of quantum chemical outputs into tensor formats compatible with graph neural networks (GNNs). Visualization using VMD has played a key role in verifying the spatial fidelity of electron density features, offering intuitive insights into their topological distribution. Preliminary models trained on ELF-informed molecular graphs exhibit encouraging performance, revealing quantum-level patterns often missed by conventional descriptors. This foundation supports the development of a quantum-topology-aware GNN framework.
To scale this approach, the methodology is being extended to larger and more diverse datasets such as QM9 and OCELOT, with the goal of constructing a comprehensive, centralized database of ELF-

enhanced molecular representations. Concurrently, efforts are underway to replace computationally demanding DFT calculations with faster quantum approximations or machine learning-based surrogates.

In the final stage, these descriptors are encoded as graph-based tensors, enabling GNNs to learn structure-property relationships by combining standard molecular connectivity with topological features derived from electron density. This hybrid strategy provides a novel interpretable alternative for molecular property prediction. While initially demonstrated in the context of drug discovery, the framework is broadly applicable—offering significant potential for modeling catalytic behavior, chemical reactivity, and photophysical properties in a scalable and interpretable manner.

5. Participant list

Organizers

Ahmad, Katya
Helmholtz Zentrum München, Germany
Schwaller, Philippe
EPFL, Switzerland
Tetko, Igor
Helmholtz Munich, Germany
Vuilleumier, Rodolphe
Sorbonne Université - ENS-PSL, France

On-site participants

Alcaide, Eric - Università della Svizzera Italiana, Switzerland
Ball, Matt - University of Strasbourg, France
Bianciotto, Marc - Sanofi, France
Bitbol, Anne-Florence - EPFL, Switzerland
Breznikar, Rok - University of Basel, Switzerland
Bunne, Charlotte - EPFL, Switzerland
Ceriotti, Michele - EPFL, Switzerland
Coletti, Silvano - Universitä degli Studi Guglielmo Marconi, Italy
Criscuolo, Emanuele - Eindhoven University of Technology, Netherlands
Duarte, Fernanda - University of Oxford, United Kingdom
Durairaj, Janani - University of Basel, Switzerland
Fastovskii, Vasilii - Strasbourg University, France
Fenogli, Juliette - Ecole Normale Supérieure, France
Fischer, Markus - Elsevier, Germany
Genheden, Samuel - AstraZeneca, Sweden
Godin, Guillaume - Osmo Labs, Switzerland
Guilbert, Salomé - EPFL, Switzerland
H. Babazade, Ramil - University of Glasgow, United Kingdom
Heid, Esther - TU Wien, Austria
Hilfiker, Mathias - AstraZeneca, Sweden
Hiltscher, Marcel - Eindhoven University of Technology, , Netherlands
Hunklinger, Andrea - Centre for Genomic Regulation, Spain
Iwan, Mateusz - Mario Negri Institute of Pharmacological Research IRCCS, Italy
Jablonka, Kevin - Helmholtz Institute for Polymers In Energy Applications, Germany
Kabylda, Adil - University of Luxembourg, Luxembourg
Kennedy, Subashini - École Normale Supérieure - PSL, France
Khasanova, Dina - HMGU, Germany
Krüger, Fabian - AstraZeneca, Sweden
Kulik, Heather - MIT, United States
Kuster, Tom - University Freiburg, Germany
Laage, Damien - Ecole Normale Supérieure, France
Lopanitsyna, Nataliya - Syngenta, Switzerland
Mansouri, Amin - EPFL, Switzerland
Marcou, Gilles - University of Strasbourg, France
Mezghani, Ines - Ecole Normale Supérieure Paris Saclay, France
Nguyen, Nguyen Xuan Vu - EPFL, Switzerland
Novitskaia, Olga - FU Berlin, Germany

Paiz, Paulina - Stanford, United States
Park, Jaehyeon - Korea Institute of Toxicology (KIT), Republic of Korea
Parker, Madeleine - University of Nottingham, United Kingdom
Pasquini, Marta - Syngenta Crop Protection AG, Switzerland
Patiny, Luc - EPFL/Zakodium, Switzerland
Preuss, Mike - Universiteit Leiden, Netherlands
Randriharimanamizara, Ulrick Fineddie - ICMUB Université Bourgogne Europe, France
Reymond, Jean-Louis - University of Bern, Switzerland
Roethlisberger, Ursula - EPFL, Switzerland
Schjelde, Karoline - AstraZeneca, Sweden
Segler, Marwin - Microsoft, United Kingdom
Smith, Imogen - University College London, United Kingdom
Sosnina, Ekaterina - Unaffiliated, Austria
Thakkar, Amol - Former IBM Research, Switzerland
Torren-Peraire, Paula - Former J&J, Switzerland
Van Schendel, Bob - AstraZeneca & Leiden University, Sweden
Varnek, Alexandre - University of Strasbourg, France
Watanabe, Yuji - EPFL, Switzerland
Westermayr, Julia Maria - Leipzig University, Germany
Zhang, Qihao - EPFL, Switzerland

On-line participants

Aniagyei, Albert - University of Health and Allied Sciences, Ghana
Adeyinka, Adedapo - University of Johannesburg, South Africa
Alibakhshi, Amin - SISSA, Italy, Italy
Azzouzi, Mohammed - Imperial College London, United Kingdom
Basak, Sumon - Bar-Ilan University, Israel
Bayer, Kutand - İzmir Institute of Technology, Turkey
Biton, Dionessa - Carl Von Ossietzky University of Oldenburg, Germany
Boldtumur, Bayar - Forschungszentrum Jülich, Germany
Borel, Alain - EPFL, Switzerland
Borisova, Anna - EPFL, Switzerland
Bournez, Colin - SpiroChem AG, Switzerland
Chen, Shuai - Groningen University, Netherlands
Choo, Hyunkyung - VTT, Finland
Cirqueira, Leonardo - Universidade de Sao Paulo, Brazil
Deshmukh, Pratham - EPFL, Switzerland
Ding, Yuhui - ETH Zurich, Switzerland
Dingiş Birgül, Serap İpek - Bezmialem Vakif University, Turkey
Dos Santos Domingues, Tiago - Yale University, United States
Dost, Katharina - Jožef Stefan Institute, Slovenia
El Moutaoukal, Yassir - Norwegian University of Science and Technology, Norway
Elizaveta, Lozhkina - Université de Montpellier, France
Feng, Qiushi - University of Turin, Italy
Flaibani, Matteo - Università di Trieste, Italy
Galvez Reverte, Jose Pablo - Universidad de Murcia, Spain
Ghasemi, Mehdi - The University of Manchester, United Kingdom
Golbamaki, Azadi - Syngenta Crop. Protection, Switzerland
Grebner, Christoph - Sanofi, Germany
HASHIM, Hani - KU Leuven, Belgium
Jin, Ruoyan - Aalto University, Finland
Jingyuan, Zhu - BUCT, China
Karnati, Konda Reddy - Bowie State University, United States

Kazemi, Mohammad - Slippery Rock University, United States
Kleuker, Felix - Leiden University, Netherlands
König, Patricia - Fritz-Haber-Institut of the Max Planck Society, Germany
Kopf, Sarina - EPFL, Switzerland
Kumar, Prashant - University College London, United Kingdom
Kumar, Abhinav - Elsevier, United Kingdom
Laricheva, Elena - Utah Valley University, United States
Lawal, Abdulfatai - Tu Wien, France
Le Breton, Guillaume - Institut de Chimie Moléculaire de L'Université de Bourgogne, France
Liew, Zheng Jie - University of Cambridge, United Kingdom
Lima, Anderson - Universidade Federal Do Pará, Brazil
Liu, Ji - Tyndall National Institute, Ireland
Lorenzini, Paolo - ASTAR, Singapore
Lv, Xinrui - Beijing University Of Chemical Technology, China
Madzhidov, Timur - Elsevier, United Kingdom
Mojica Jiménez, Paula Andrea - Ghent University, Belgium
Mojica Urrego, Harold Mateo - University of Groningen, Netherlands
Mondal, Soumya - IIT DELHI, India
Mqawass, Ghaith - Pfizer GmbH/Technical University of Munich, Austria
Naghdi Dorabati, Amirhossein - NCBJ, Poland
Nair, Renjith - Madanapalle Institute of Technology & Science, India
Olivieri, Jean-François - Université Paris-Saclay, IUT Orsay, France
Olshanova, Aleksandra - The University of Southern Denmark, Denmark
Osipova, Iuliia - Università degli Studi di Padova, Italy
Ould Mohamed, Lina – Univ. Des Sciences et de la Technologie Houari Boumediene, Algeria
Pahuja, Anushka - Universite Paris Saclay, France
Pan, Xiaolin - New York University, United States
Patra, Alok - A*Star, Singapore
Phung, Cheng Fei - National University of Singapo, Singapore
Qin, Ting - Sygnature Discovery, United Kingdom
Qin, Xueping - Technical University of Denmark, Denmark
Rafik, Amine - Universidad de Huelva & University Mohammed V In Rabat, Spain
Roehrig, Ute - SIB Swiss Institute of Bioinformatics, Switzerland
Rossi, Federico - Norwegian University of Science and Technology, Norway
Sabanza Gil, Victor - EPFL, Switzerland
Sappati, Subrahmanyam - Gdańsk University of Technology, Poland
Sethi, Harshit - Aalto University, Finland
Sharma, Himanshu - Seth G. L. Bihani Sd Pg College, India
Stany Lionel, Tsomene - University of Yaoundé 1, Cameroon
Sugathan Nair, Akhil - Fritz-Haber Institute of the Max-Planck Society, Germany
Sulaiman, Mohammed - Schrodinger, India
Sun, Liuziyue - Ludwig-Maximilians Univerity of Munich, Germany
Tang, Xuechen - University of Innsbruck, Austria
Whalley, Gowan - University College London, United Kingdom
Worakul, Thanapat - EPFL, Switzerland
Yeamin, Md Bin - The Helmholtz-Zentrum Dresden-Rossendorf, Germany
Zaher, Leen - University of Milan, Italy
Zhang, He - Beijing Academy of Artificial Intelligence, China