Expectations, achievements and lessons of AIDD and AIChemist Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate projects

Igor V. Tetko

Helmholtz Munich and BigChem GmbH

ELLIS workshop, HYBRID, Berlin, December 6, 2024

HELMHOLTZ MUNICH

AiChemist

Marie Skłodowska-Curie Actions 2021-2027

Developing talents, advancing research

Under Pillar I of Horizon Europe, the MSCA are the European Union's reference programme for doctoral education and postdoctoral training. They support researchers from all over the world, at all stages of their careers, with a focus on their training, skills and career development.

Under Horizon 2020 (2014-2020), the MSCA:

Funded **1080 doctoral programmes**, of which 156 industrial doctoral programmes and 76 joint doctorates

Involved **4 700 companies**, of which **2 200 SMEs**

Involved **37% of researchers** from non-EU countries and around **1300 organisations** from non-EU and non-associated countries

Since 1996

budget
14 billion €

researchers
140 000
(39 000 PhDs)

Horizon Europe (2021-2027)

budget
6.6 billion €

researchers
65 000
(25 000 PhDs)

Under Horizon Europe, the MSCA will:

Strengthen organisations
The MSCA support excellent doctoral and postdoctoral programmes and collaborative projects worldwide, promoting structuring impact on organisations

Foster research and innovation beyond academia
The MSCA boost ties between academia and other non-academic organisations with various incentives, increasing fellows' exposure to other sectors

Build international links
The MSCA are key in attracting talent to Europe, building international, strategic partnerships, and promoting global research mobility and science cooperation

The MSCA have 5 main actions

Doctoral Networks

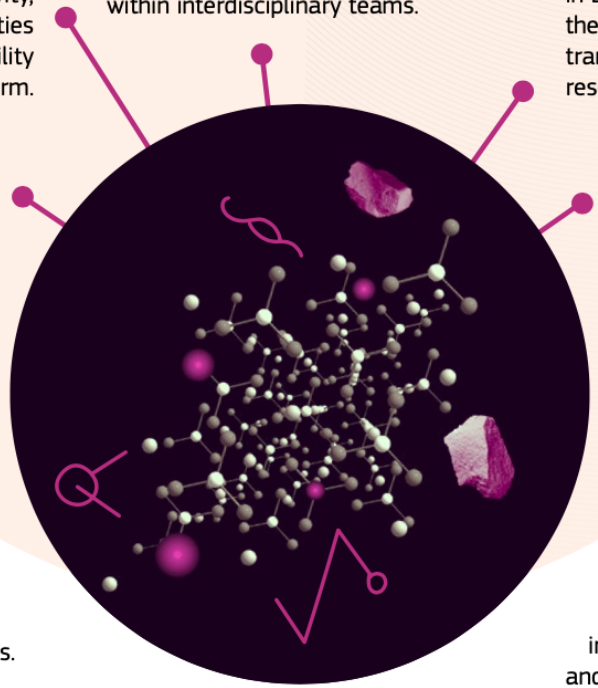
implement doctoral programmes (including joint doctorates and industrial doctorates) **by international partnerships** of organisations from different sectors. They train highly-skilled doctoral candidates, stimulate their creativity, enhance their innovation capacities and boost their employability in the long-term.

Postdoctoral Fellowships

support researchers' careers and foster excellence in research and innovation. Researchers holding a PhD can carry out their research activities, acquire new skills and develop their careers abroad, whilst developing competences in non-academic sectors and working within interdisciplinary teams.

Staff Exchanges

encourage short-term international and inter-sectoral exchanges of research and innovation staff through sustainable, collaborative projects in Europe and beyond. By doing so, they enhance knowledge and skills transfer and increase organisations' research and innovation capacities.



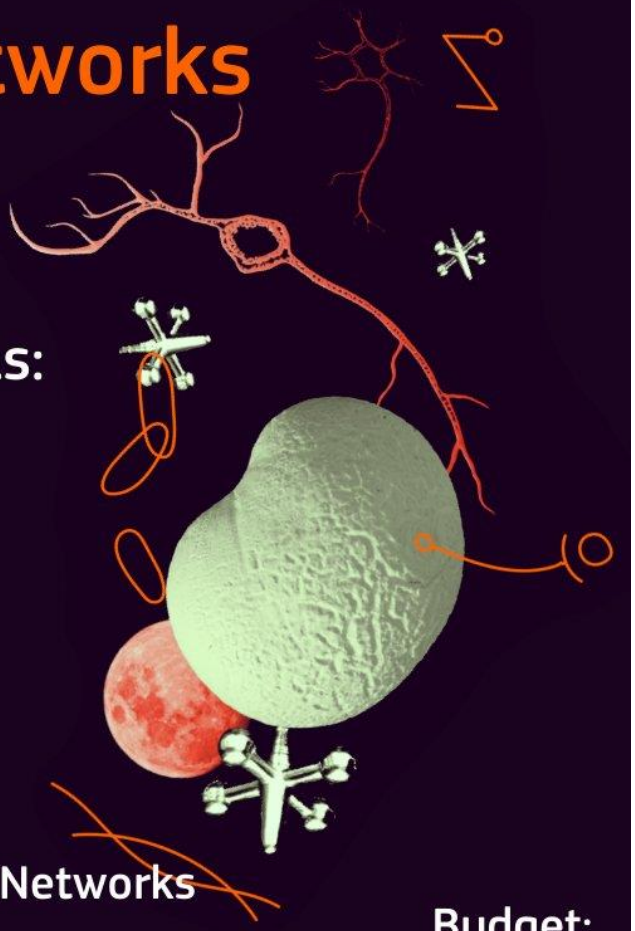
MSCA and Citizens

brings research and researchers closer to children, families and the public at large through the European Researchers' Night - the annual research communication and promotion event taking place at the end of September across EU Member States and Horizon Europe Associated Countries.

COFUND

co-finances regional, national and international doctoral and postdoctoral programmes for researchers' training and career development. The COFUND action spreads MSCA's best practices by setting high standards and excellent working conditions, and boosts training and international, interdisciplinary and inter-sectoral mobility.

Doctoral Networks
Call 2024



Submitted proposals:

1417

of which:

80 Industrial Doctorates

88 Joint Doctorates

1249 Standard Doctoral Networks

Budget:

€608.6 million

MSCA

Marie Skłodowska-Curie **Actions**

Developing talents, advancing research

European Commission

Deadline:
28.11.24

Next one:
25.11.25

https://ai-dd.eu

Advanced machine learning for Innovative Drug Discovery (AIDD)

This project is funded by Horizon2020 research and innovation programme under the Marie Skłodowska-Curie actions

About

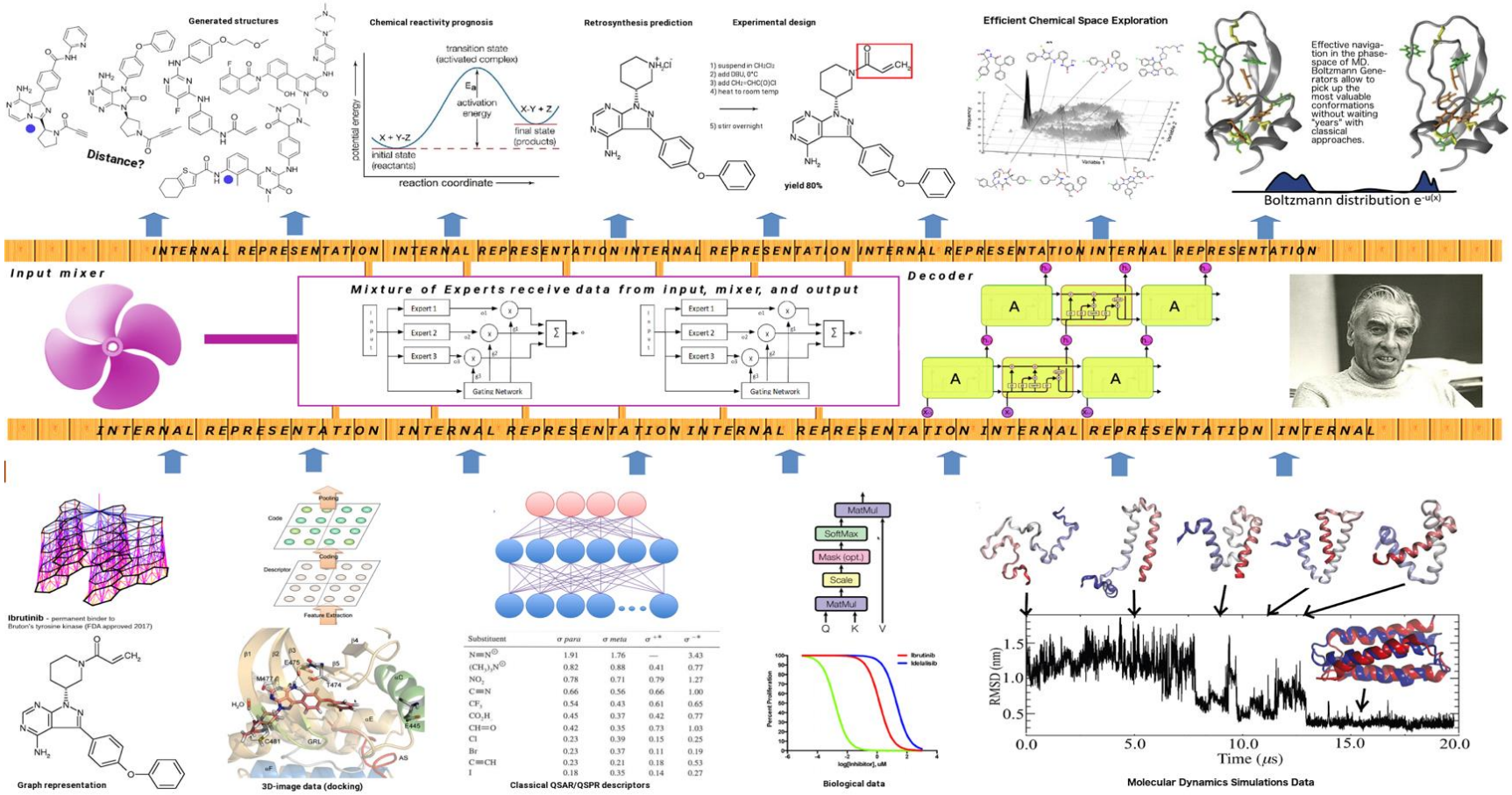
The dramatic increase in using of Artificial Intelligence (AI) and traditional machine learning methods in different fields of science becomes an essential asset in the development of the chemical industry, including pharmaceutical, agro biotech, and other chemical companies. However, the application of AI in these fields is not straightforward and requires excellent knowledge of chemistry. Thus, there is a strong need to train and prepare a new generation of scientists who have skills both in machine learning and in chemistry and can advance medicinal chemistry, which is the prime goal of the AIDD proposal. Research WPs include sixteen topics selected to cover the key innovative directions in machine learning in chemistry. Fellows employed will be supervised by academics who have excellent complementary expertise and contributed some of the fundamental AI algorithms which are used billions of times per day in the world, and leading EU Pharma companies who are in charge of new medicine and public health. All developed methods can be used individually but will also contribute to an integrated "One Chemistry" model that can predict outcomes ranging from different properties to molecule generation and synthesis. Training on various modalities allows the model to understand how to intertwine chemistry and biology to develop a new drug making its design robust. All partners agreed to make the software developed as part of the AIDD project open source. It will boost the field and will provide the broadest possible dissemination of the results both to the academy and industry, including SMEs. The network will offer comprehensive, structured training through a well-elaborated Curriculum, online courses, and six Schools. The IP policy and commercial exploitation of the project results have the highest priority supported by intellectual property asset management organizations. Comprehensive public engagement activities will complement the dissemination of results to the scientific community.

This project is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956832, and it is Horizon 2020 Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate.

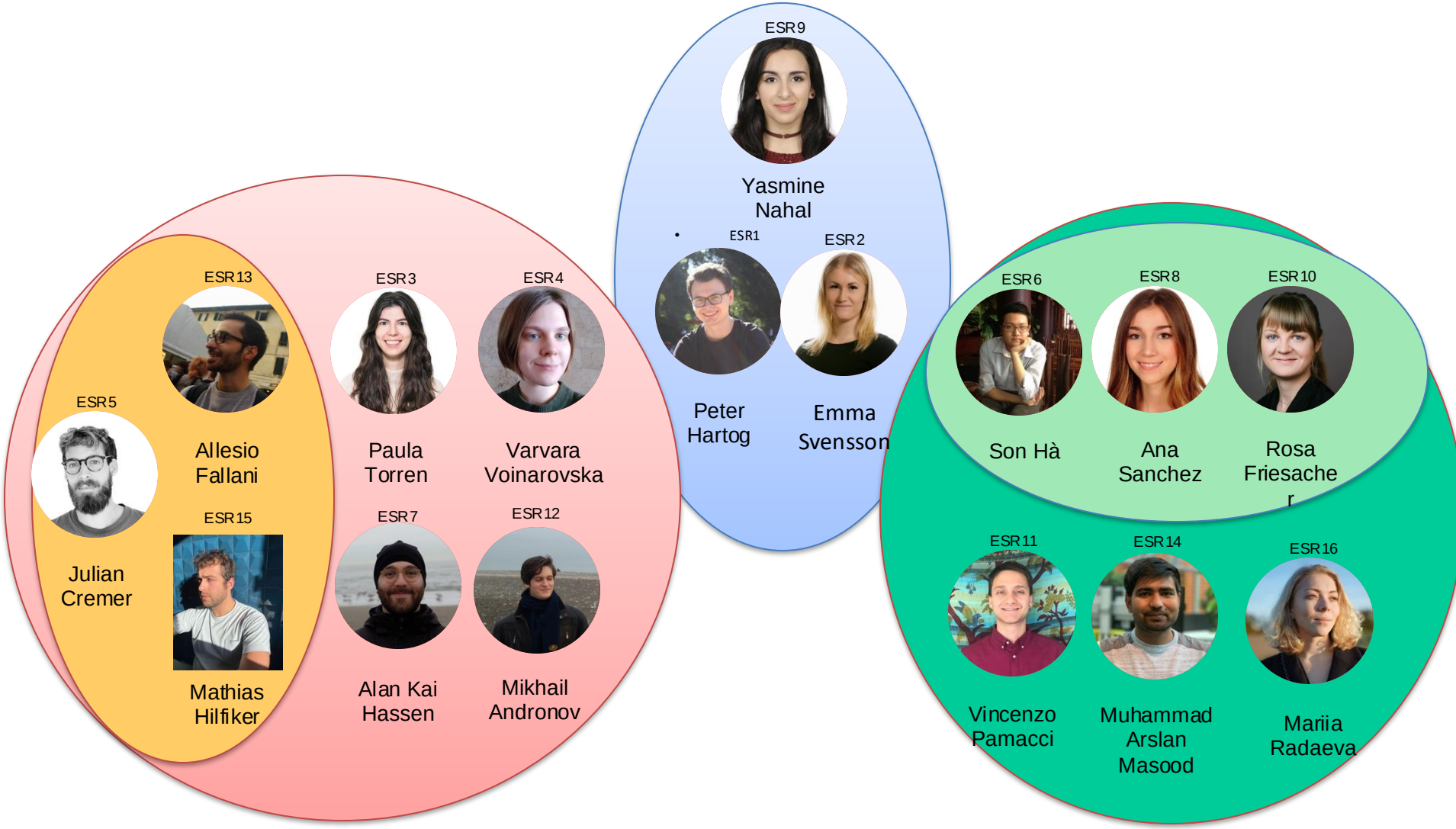
We are on social networks: Twitter, LinkedIn, Facebook. See also project publications at Google Scholar. One project article is currently listed as "highly cited" according to the Web of Knowledge.

Beneficiaries

AIDD overview



The AIDD Fellows



12 nationalities (or 14!)

Julian Cremer, Defended PhD 28.11.2024, postdoc at Pfizer (Berlin)

Chemical Research in Toxicology > Vol 36/Issue 10 > Article

Open Access Editors' Choice

Cite Share Jump to Expand

ARTICLE | September 10, 2023

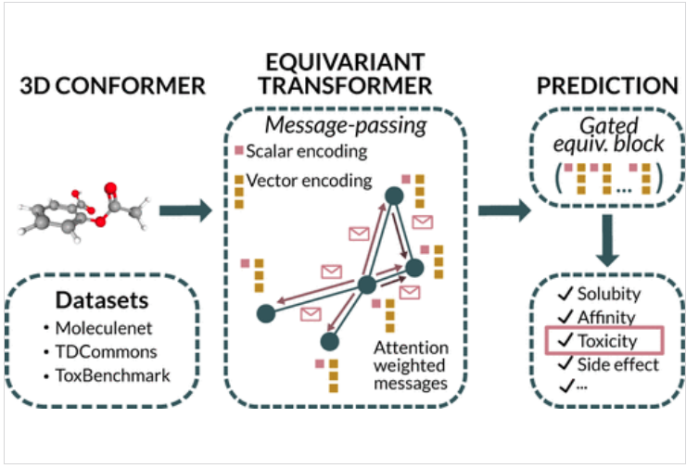
Equivariant Graph Neural Networks for Toxicity Prediction

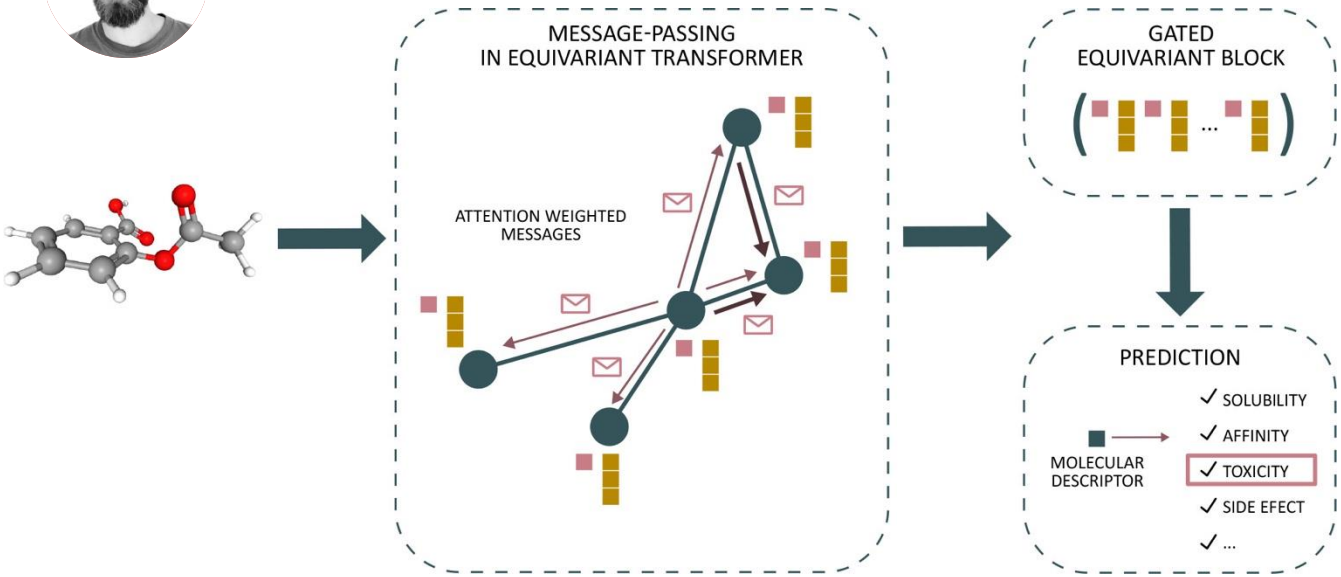
Julian Cremer*, Leonardo Medrano Sandonas*, Alexandre Tkatchenko, Djork-Arné Clevert, and Gianni De Fabritiis

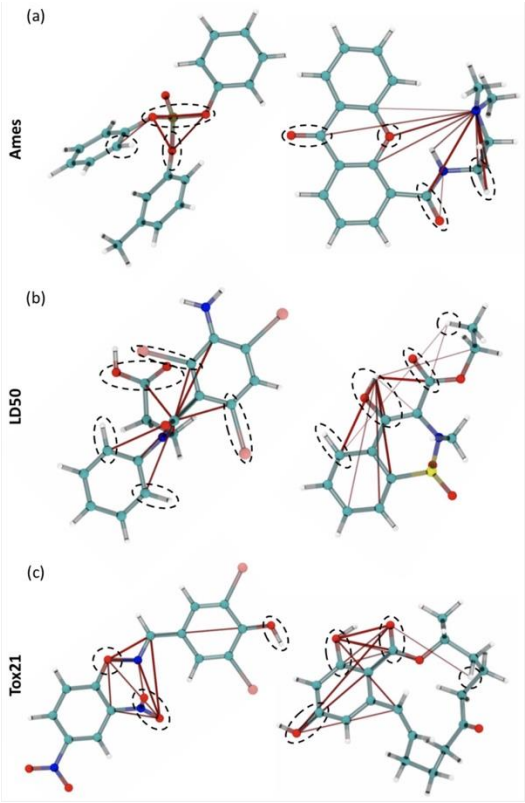
Open PDF | Supporting Information (1)

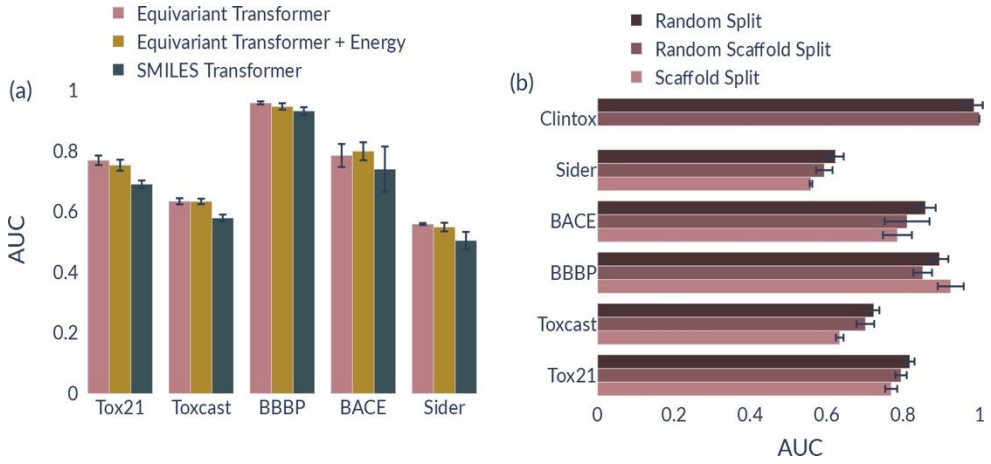
Abstract

Predictive modeling of toxicity is a crucial step in the drug discovery pipeline. It can help filter out molecules with a high probability of failing in the early stages of de novo drug design. Thus, several machine learning (ML) models have been developed to predict the toxicity of molecules by combining classical ML techniques or deep neural networks with well-known molecular representations such as fingerprints or 2D graphs. But the more natural, accurate representation of molecules is expected to be defined in physical 3D space like in ab initio methods. Recent studies successfully used equivariant graph neural networks (EGNNs) for representation learning based on 3D structures to predict quantum-mechanical properties of molecules. Inspired by this, we investigated the performance of EGNNs to construct reliable ML models for toxicity prediction. We used the equivariant transformer (ET) model in TorchMD-NET for this. Eleven




MESSAGE-PASSING
IN EQUIVARIANT TRANSFORMER
ATTENTION WEIGHTED
MESSAGES
GATED
EQUIVARIANT BLOCK
PREDICTION
MOLECULAR
DESCRIPTOR
SOLUBILITY
AFFINITY
TOXICITY
SIDE EFECT
...


(a)
Ames
(b)
LD50
(c)
Tox21


Equivariant Transformer
Equivariant Transformer + Energy
SMILES Transformer
(a)
AUC
1
0.8
0.6
0.4
0.2
0
Tox21
Toxcast
BBBP
BACE
Sider
Random Split
Random Scaffold Split
Scaffold Split
(b)
Clintox
Sider
BACE
BBBP
Toxcast
Tox21
0
0.2
0.4
0.6
0.8
1
AUC

Rosa Friesacher, finishing PhD during 4th year at Katholieke Universiteit Leuven

Why do we need Uncertainty Quantification (UQ)?

UQ can provide valuable information:

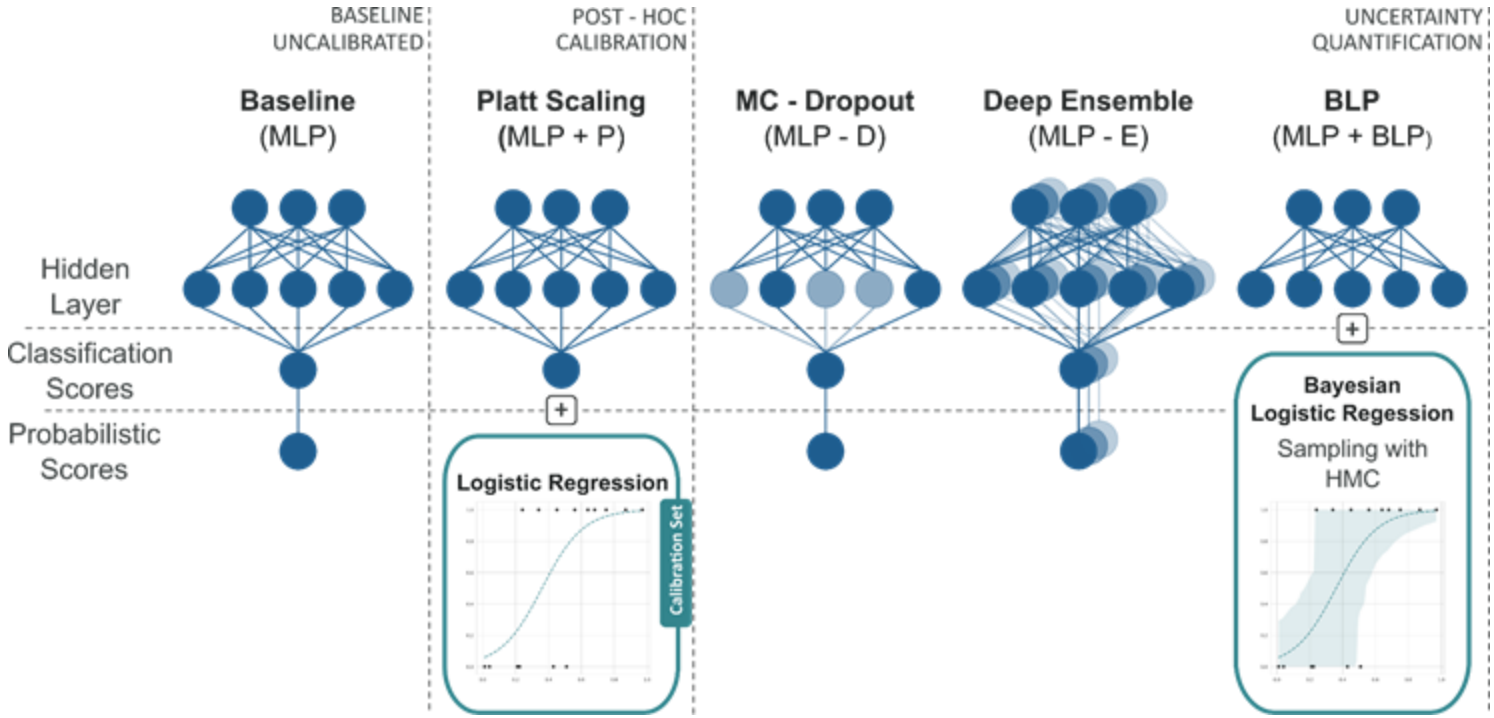
- for which compounds can the model **confidently** make predictions?
- about which compounds is the model **uncertain**?

UQ can help with:

- assessment of **risks, costs and benefits**
- **prioritization** of test compounds for further analysis

→ **Models are often poorly calibrated**

Uncertainty Quantification Approaches

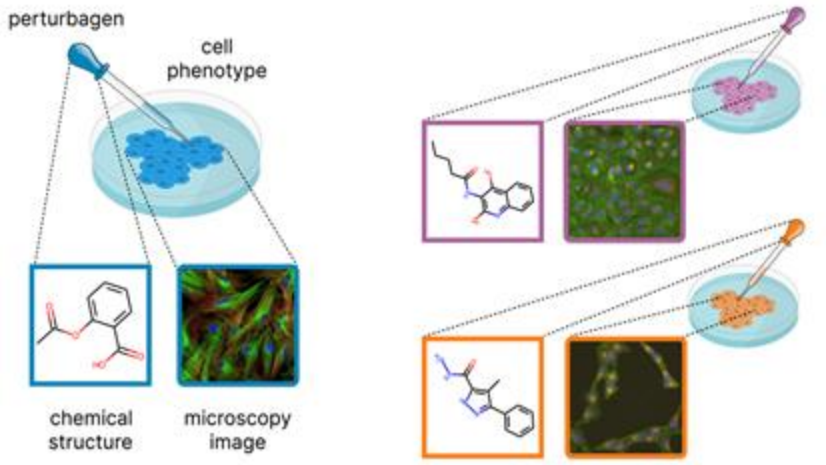


ICML 2024 AI for Science Workshop, https://ai4sciencecommunity.github.io/icml24.html

Ana Sánchez Fernández, finishing PhD during 4th year at Johannes Kepler Universität Linz

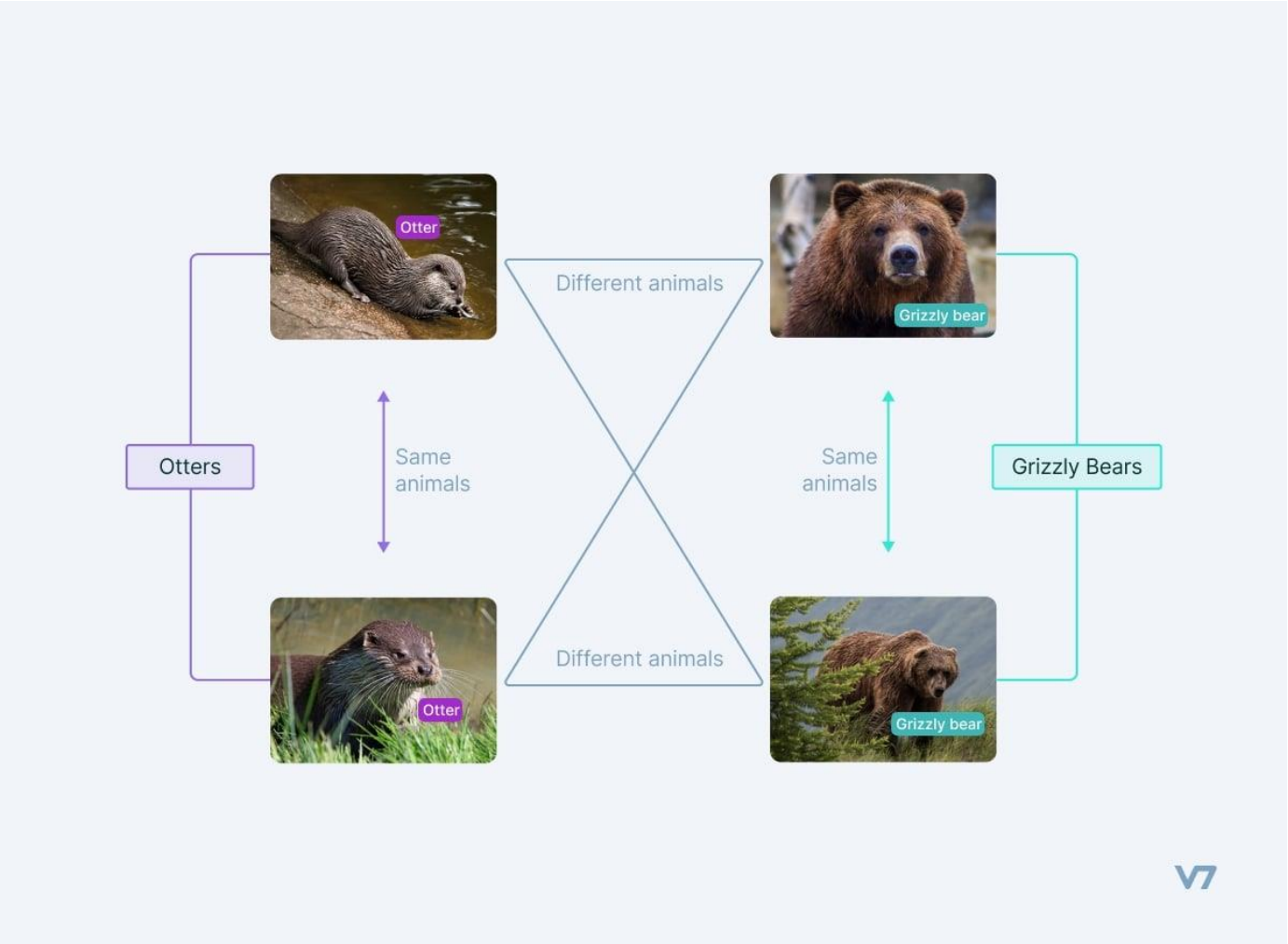
CLOOME. Contrastive Leave-One-Out boost for Molecule Encoder

Learn molecular **representations** with **contrastive** learning using **microscopy** images and molecular **structures**



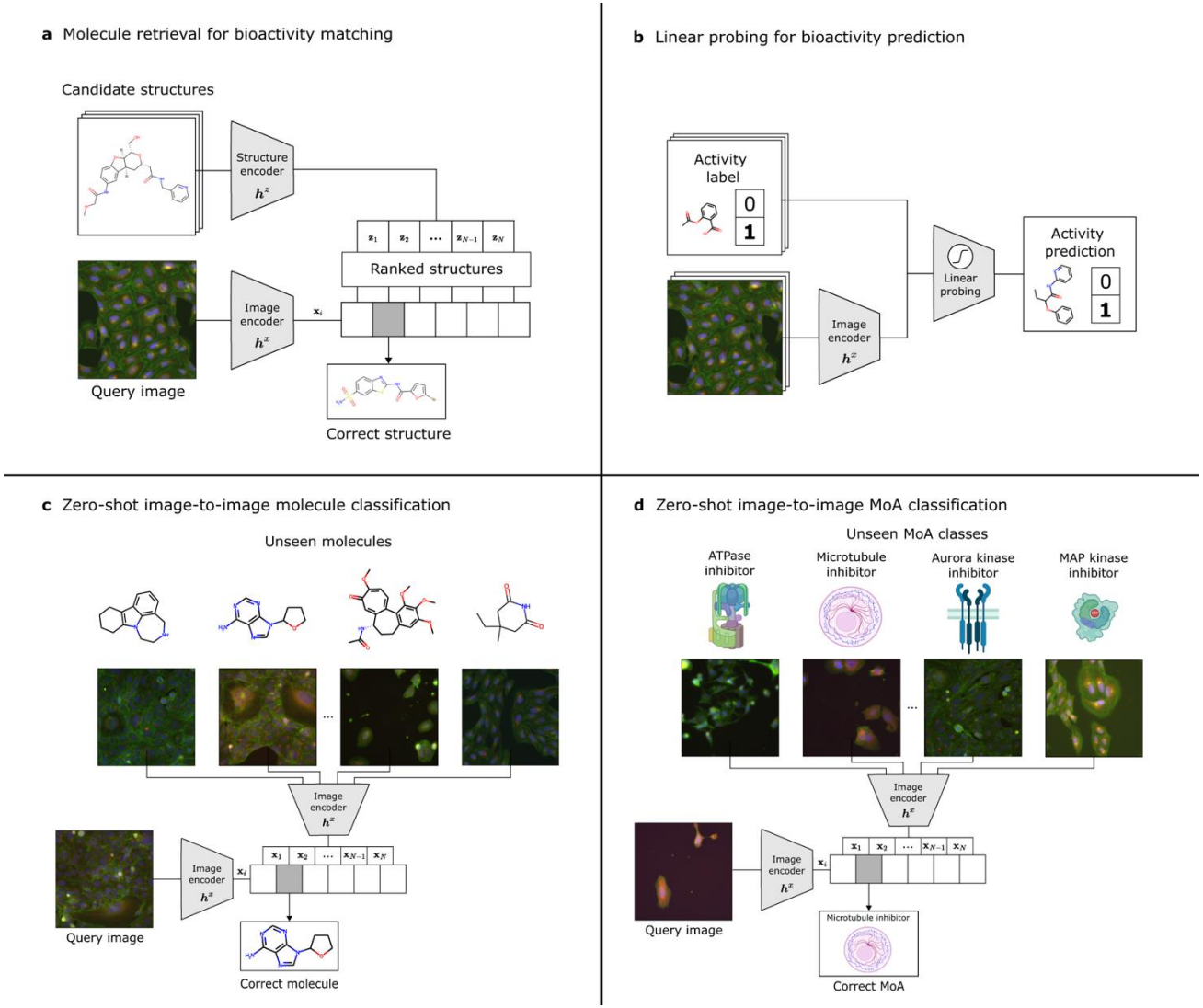
Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. & Klambauer, G. (2023). **CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures.** *Nature Communications*, 14, 7339. https://doi.org/10.1038/s41467-023-42328-w

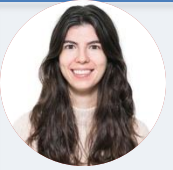
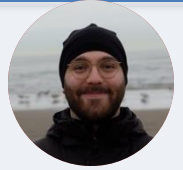
Principles of Contrastive Learning



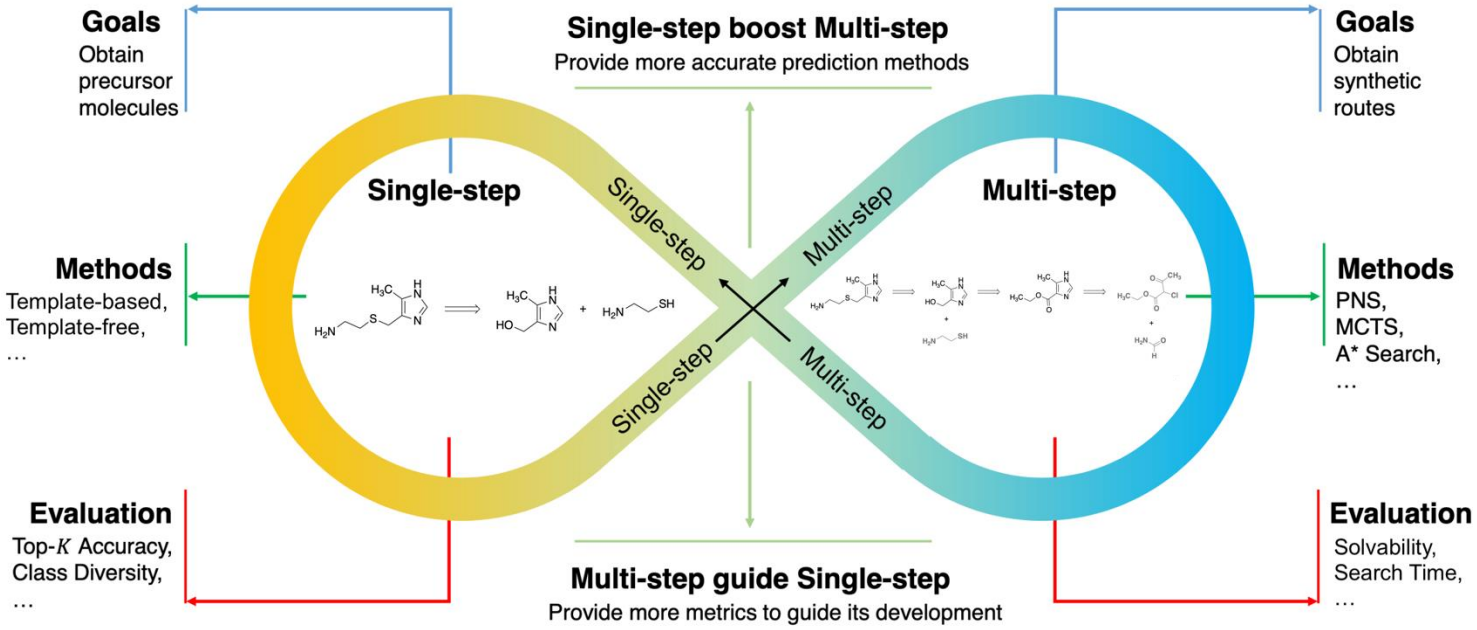
https://www.v7labs.com/blog/contrastive-learning-guide

Use cases of CLOOME

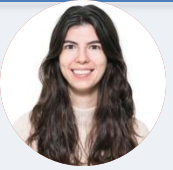
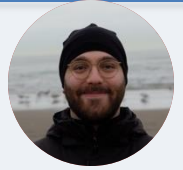


Retrosynthesis Prediction



Adapted from: Fig. 1, DOI: arXiv:2301.05864

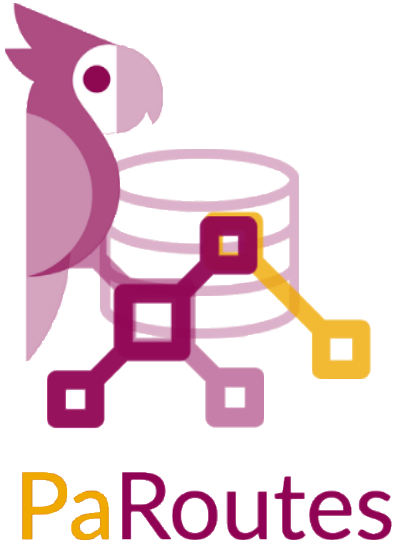
Multi-Step: PaRoutes

10,000 compounds from USPTO

Each compound has one recorded retrosynthetic route (n-1 set)

Assess success rate and accuracy

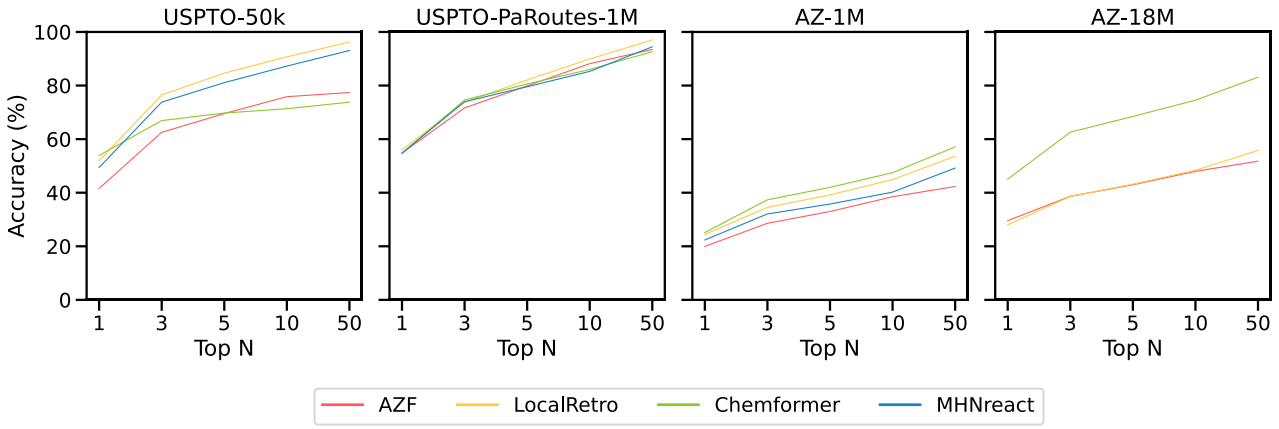
Building blocks: specialised PaRoutes set



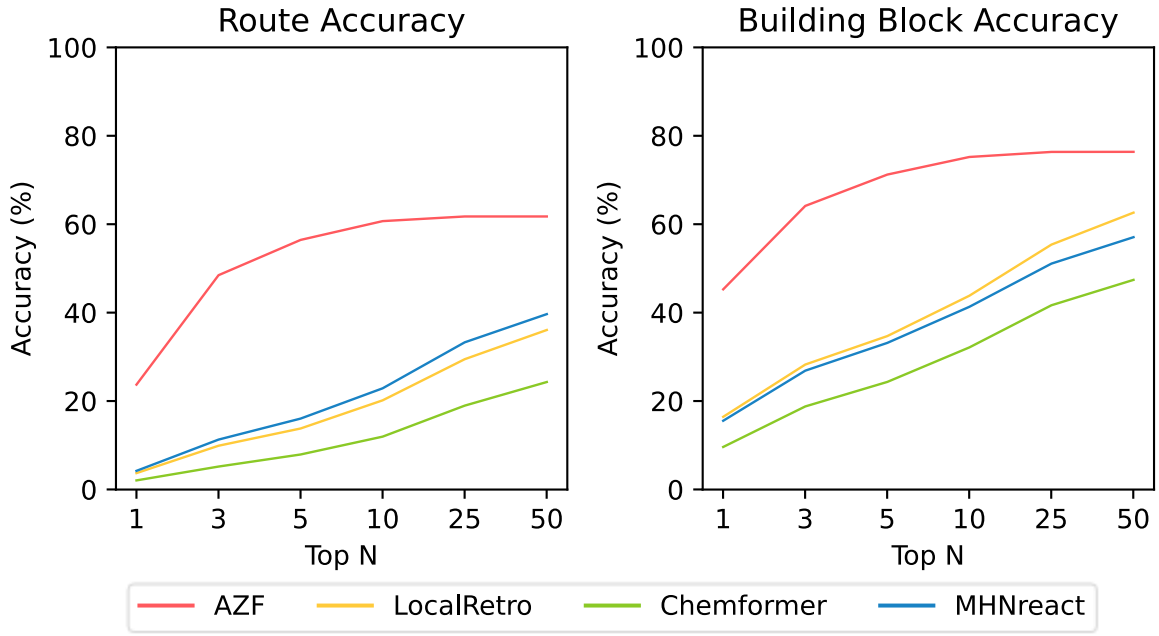
Genheden, S. & Bjerrum, E. PaRoutes: towards a framework for benchmarking retrosynthesis route predictions. Digital Discovery 1, 527–539 (2022).


Accuracy PaRoutes

Single step



Multi step



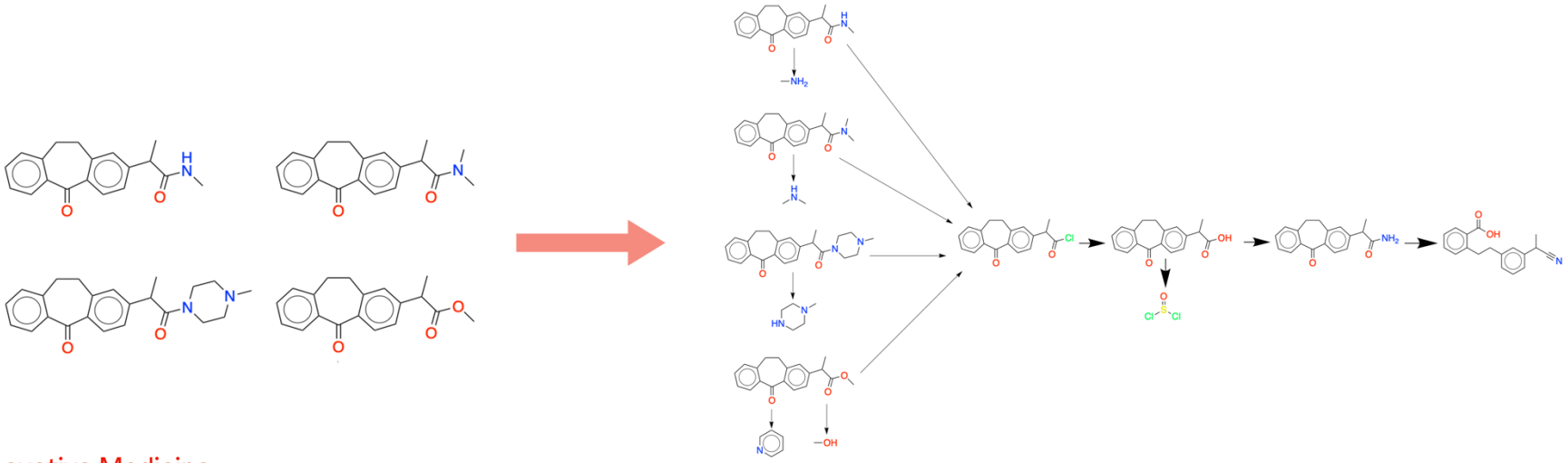
P. Torren-Peraire, et al Models Matter: the impact of single-step retrosynthesis on synthesis planning *Digital Discovery*, **2024**.

Paula Torren-Peraire, postdoc at Novartis starting 15.12.2024 (preparing PhD thesis)

Convergent Routes

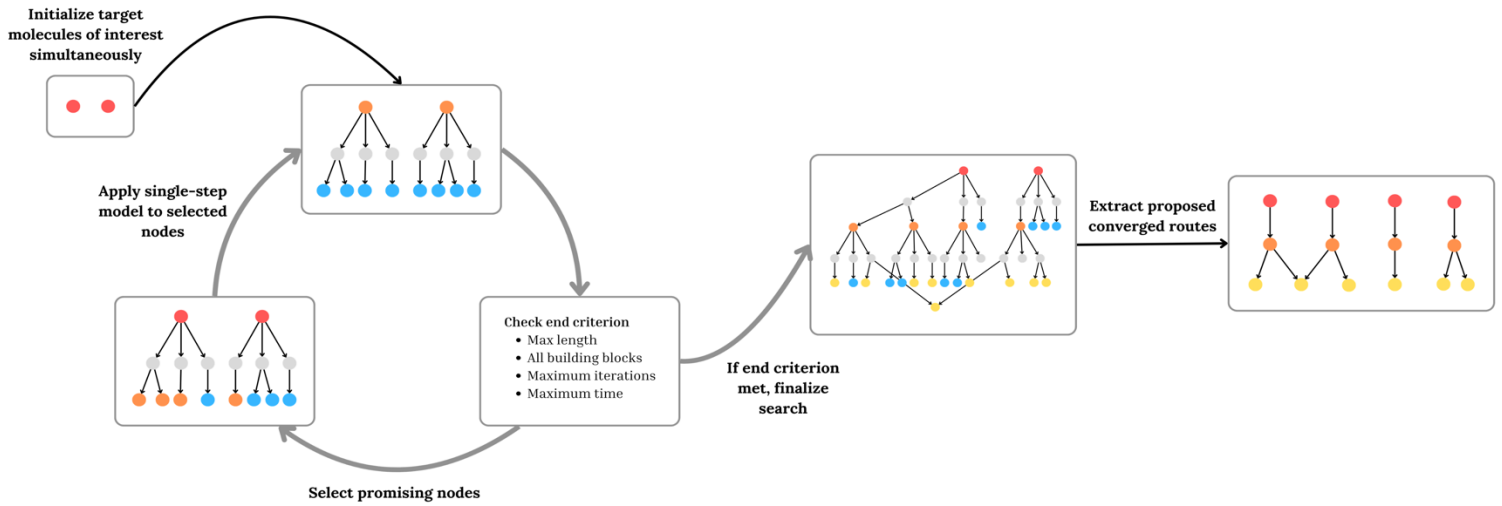
Computer-Aided Synthesis Planning (CASP) proposes retrosynthetic routes for a compound of interest, however, medicinal chemists commonly work in compound libraries

Convergent routes allow the synthesis of multiple target molecules while reducing the time/cost of synthesis by using a joint common path



Convergent Search Approach

Multi-step retrosynthesis planning search allows for the use of multiple compounds and can identify convergent routes

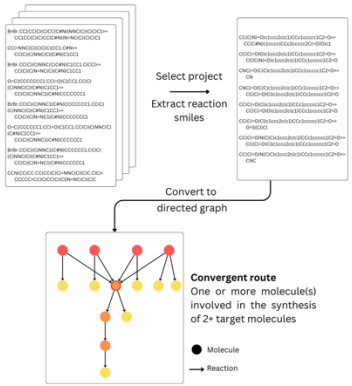


Convergent Routes Dataset

We create a novel convergent routes dataset, extracting common routes across multiple target molecules, identifying common intermediates

J&J ELN	USPTO
Collection of proprietary reaction data from Johnson & Johnson Electronic Laboratory notebooks (ELN)	Publicly available data source of 3.7 million reactions based on 40 years of patent applications and grants

We establish the convergent routes dataset to quantify the prevalence and characteristics of convergent routes in public and J&J data



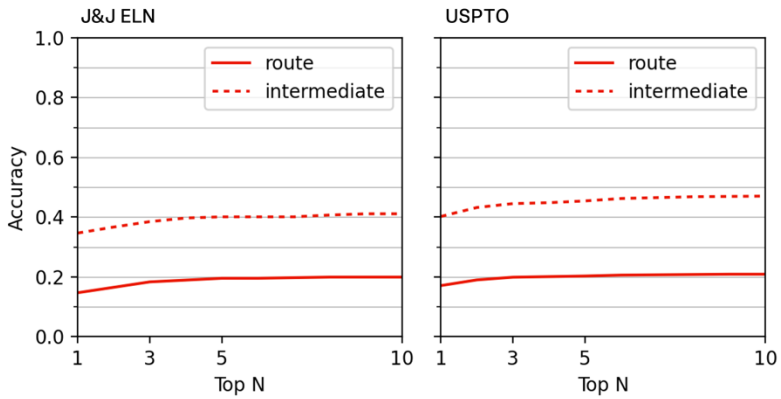
D. M. Lowe, "Extraction of chemical structures and reactions from the literature," Thesis, 2012
P. Neves et al., J. Cheminformatics, vol. 15, no. 1, p. 20, 2023

Convergent Search Results

Accuracy

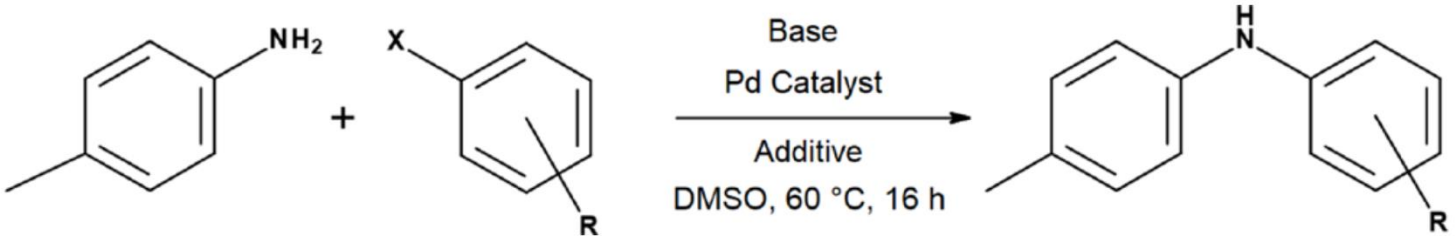
Ability to propose the experimentally validated synthesis route within the top-*N*

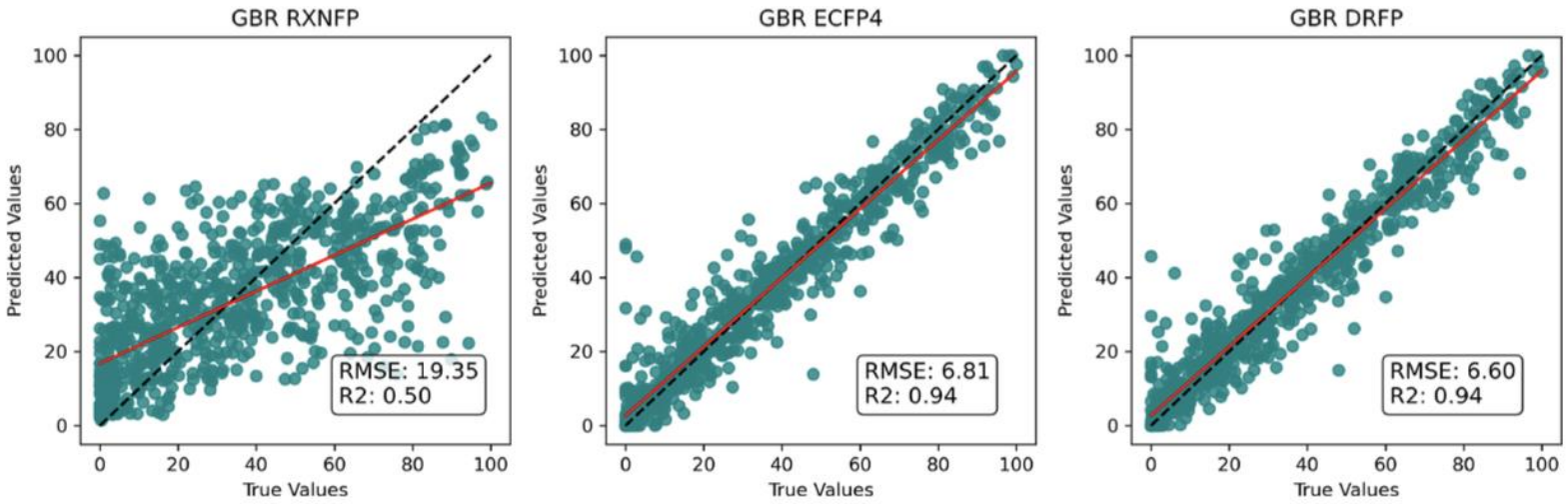
- Route: Match between reactions of experimentally validated and proposed route
- Intermediate: Match between common intermediates of experimentally validated and proposed route



Varvara Voinarovska, Defended PhD 17.10.2024, postdoc at AstraZeneca

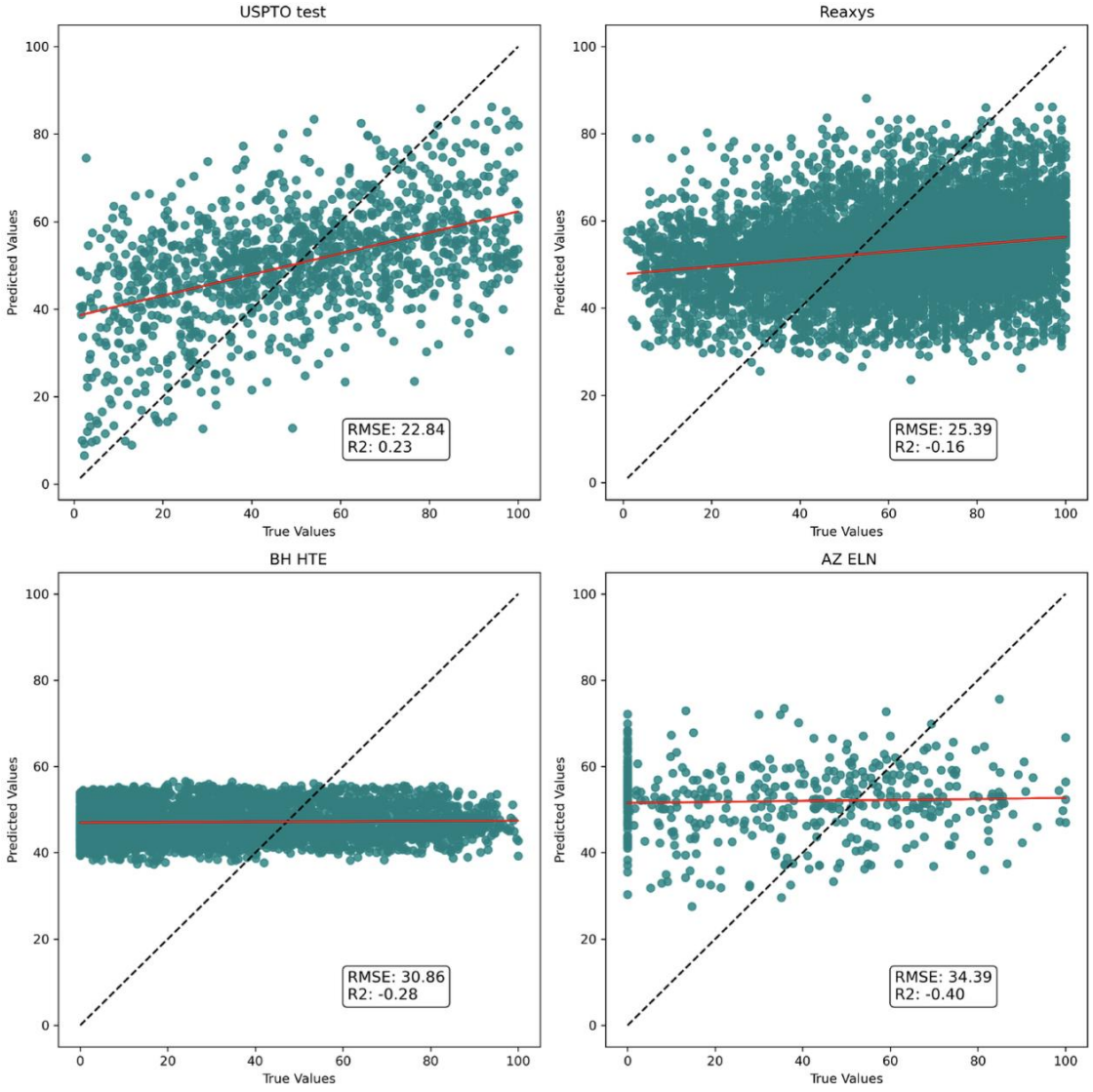
Models for HTS data from Buchwald-Hartwig reaction



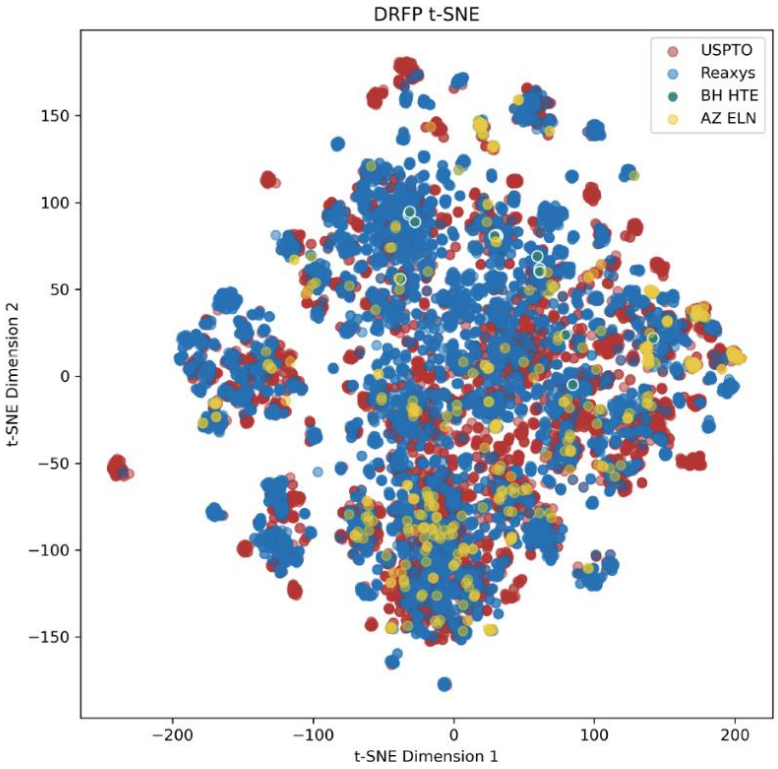


HTS – high quality data run by the same group

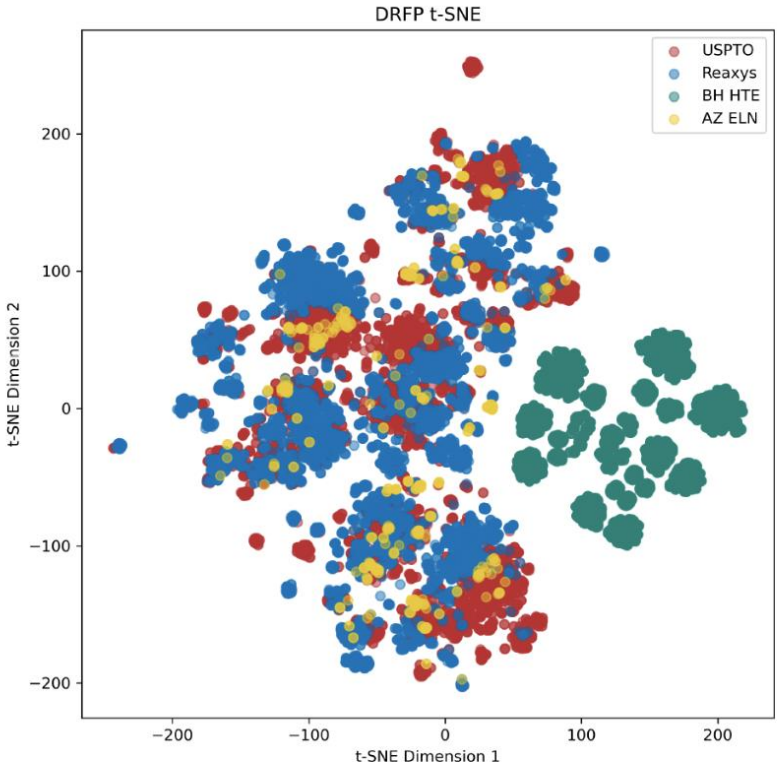
Yield prediction using model based on USPTO



Analysis of chemical space for Buchwald-Hartwig reaction



(a) Conditions excluded

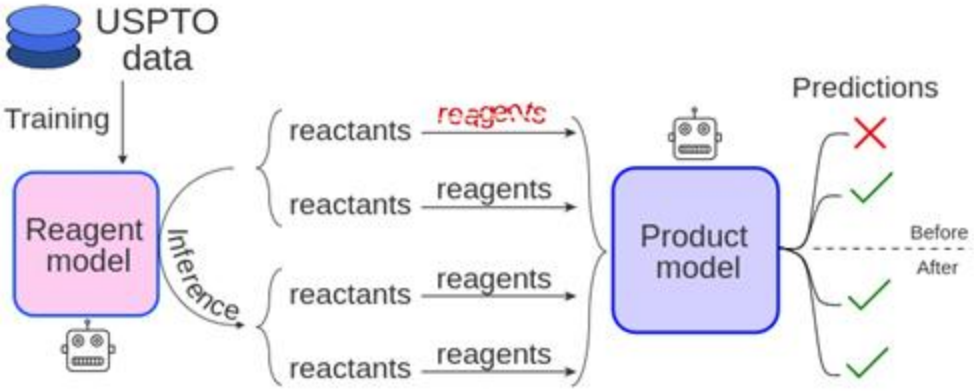


(b) Conditions included

V. Voinarovska, et al When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges *JCIM*, **2024**.


Reagent prediction with the transformer

Prediction of missing reagents in reactions with a SMILES-to-SMILES transformer, recovering missing reagents to improve the data for reaction prediction.



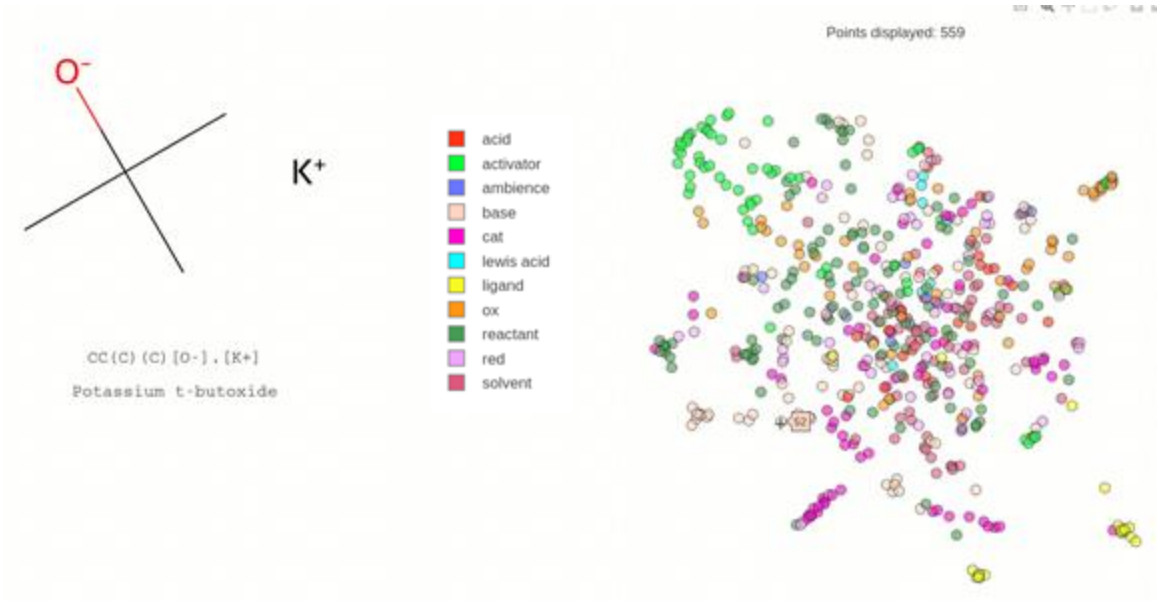
https://github.com/Academich/reagents

Andronov et al. (2023). Reagent prediction in a molecular transformer improves reaction data quality. Chemical Science, 14, 3235-3246.

Mikhail Andronov, finishing PhD during 4th year at Scuola Universitaria Professionale della Svizzera Italiana

Reagent data curation in an interactive app

A word2vec-inspired algorithm for grouping reagents by roles based on their co-occurrences. Helps reaction data curation and reagent labeling.

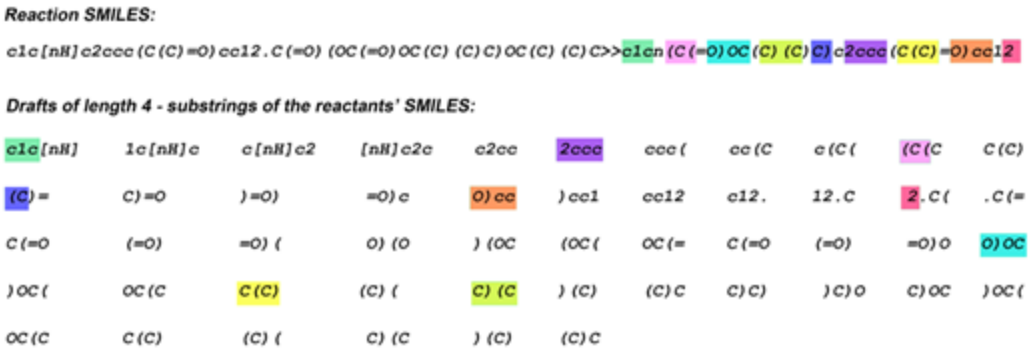


https://github.com/Academich/reagent_emb_vis

Andronov et al. (2024). Curating Reagents in Chemical Reaction Data with an Interactive Reagent Space Map. In: Clevert, DA., Wand, M., Malinovská, K., Schmidhuber, J., Tetko, I.V. (eds) AI in Drug Discovery. AIDD 2024. Lecture Notes in Computer Science, vol 14894. Springer, Cham.

Fast SMILES-to-SMILES with speculative decoding

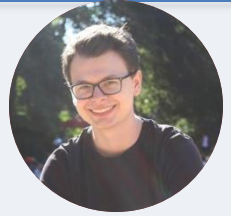
Generating several tokens per forward pass in conditional SMILES generation for 3X faster inference with greedy and beam search decoding without losing accuracy.



The target SMILES can be assembled using the patches of the source SMILES

https://github.com/Academich/translation-transformer

Andronov et al. (2024). Accelerating the inference of string generation-based chemical reaction models for industrial applications. *arXiv preprint arXiv:2407.09685*.



Peter Hartog, preparing PhD thesis, has an offer for postdoc at Switzerland

Are XAI interpretations consistent?

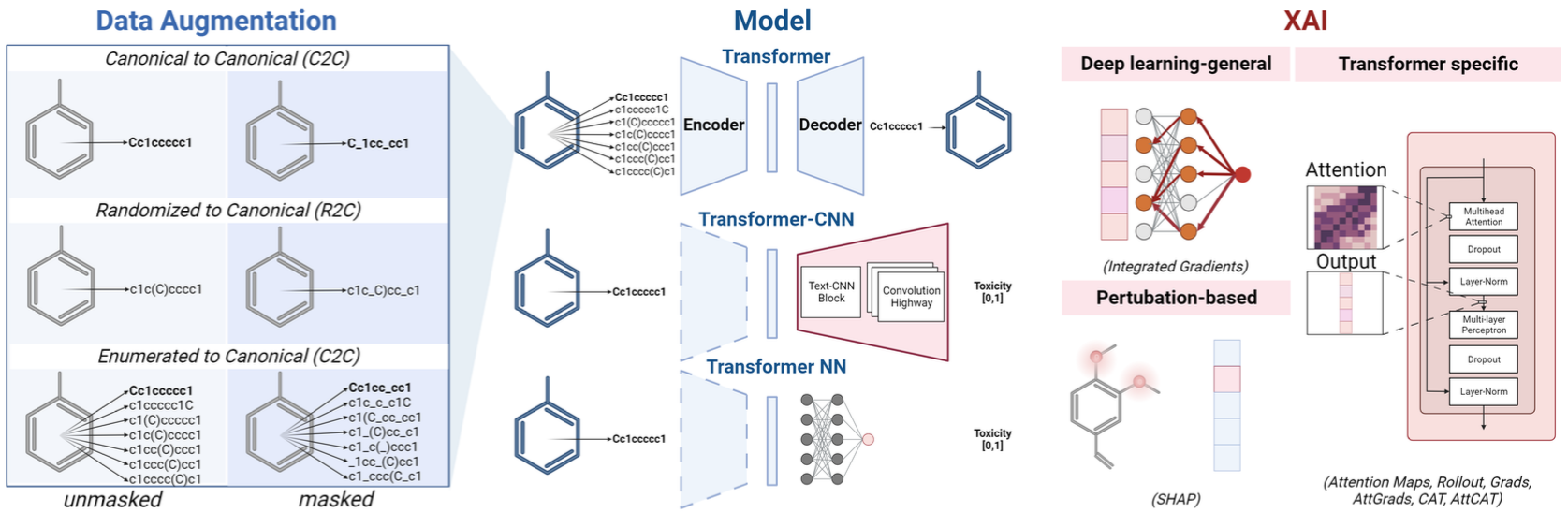
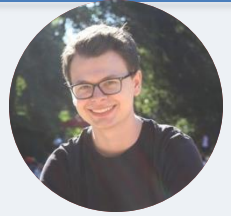


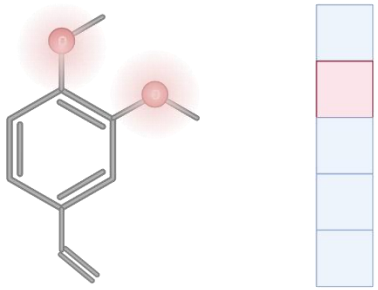
Figure 1: **Overview of the methods used throughout the research.** Data augmentation is used during pre-training. Transfer learning uses the pre-trained transformer encoder together with a small neural network or CNN. Thereafter, eight XAI methods subdivided into three groups were used for interpretation.

Hartog P et al, *J. Cheminformatics*, **2024,** 16, 39.



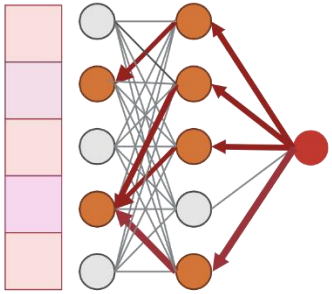
XAI methods

Pertubation-based XAI

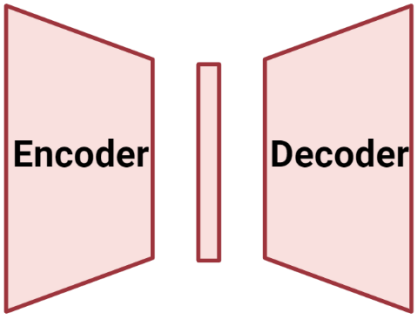


SHAP

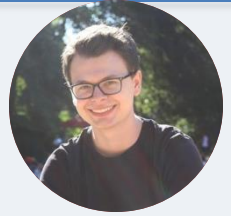
Gradient-based XAI



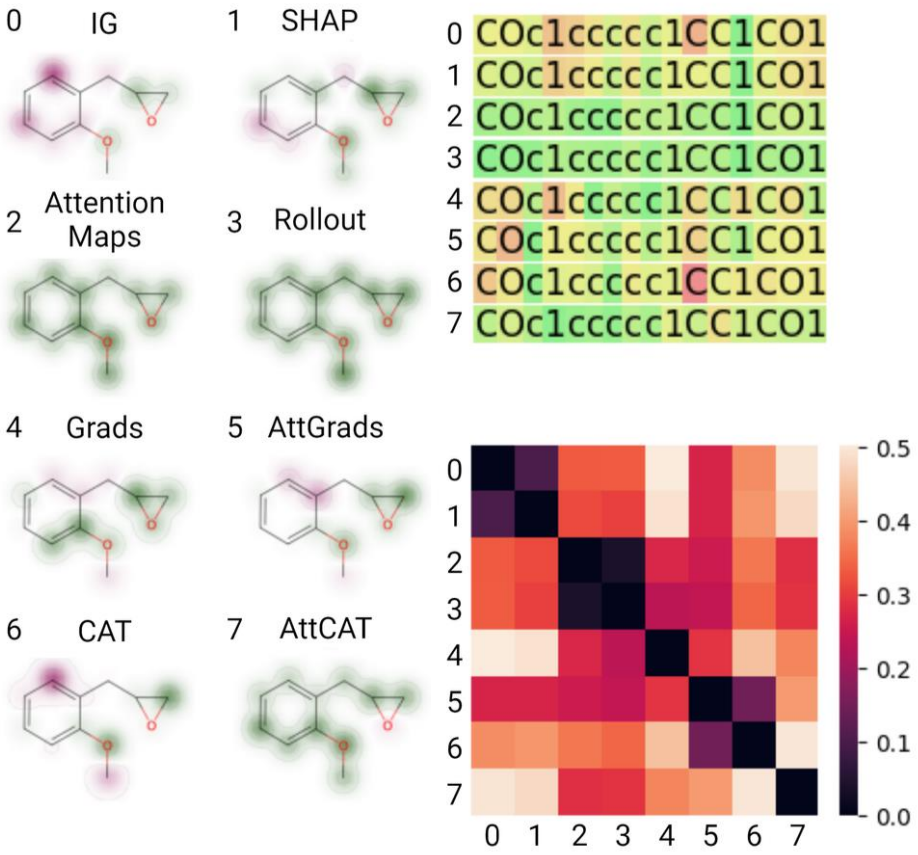
Integrated Gradients (IG)



Attention Maps, Rollout, Grads, AttGrads, CAT and AttCAT



Importance of features across XAI methods



PhD students

Peter Hartog, Helmholz Zentrum München, Germany and AstraZeneca AB, Sweden

Emma Svensson, Johannes Kepler Universität Linz, Austria and AstraZeneca AB, Sweden

Paula Torren Peraire, Helmholz Zentrum München, Germany and Janssen Pharmaceutica NV, Belgium

Dr. Varvara Voinarovska, Helmholz Zentrum München, Germany, AstraZeneca AB, Sweden and Enamine Limited Liability Company, Ukraine; PhD defended on 17th October, 2024 at Technical University of Munich; now Postdoc at AstraZeneca

Dr. Julian Cremer, Universitat Pompeu Fabra, Spain and Pfizer Pharma GmbH, Germany; PhD defended on 28th November, 2024 at Universitat Pompeu Fabra; now postdoc at Pfizer Pharma GmbH

Son Hà, TU Dortmund, Germany/Johannes Gutenberg-Universität Mainz and Janssen Pharmaceutica NV, Belgium.

- Alan Kai Hassen, Universiteit Leiden, Netherlands and Pfizer Pharma GmbH, Germany
- Ana Sánchez Fernández, Johannes Kepler Universität Linz, Austria and Janssen Pharmaceutica NV, Belgium
- Yasmine Nahal, AstraZeneca AB, Sweden and Aalto University, Finland.
- Rosa Friesacher, Katholieke Universiteit Leuven, Belgium and AstraZeneca AB, Sweden.

Vincenzo Palmacci, Bayer Aktiengesellschaft, Germany and University of Vienna, Austria

- Mikhail Andronov, Scuola Universitaria Professionale della Svizzera Italiana, Switzerland and Pfizer Pharma GmbH, Germany

Alessio Fallani, Université du Luxembourg and Janssen Pharmaceutica NV, Belgium

- Muhammad Arslan Masood, Aalto University, Finland and Janssen Pharmaceutica NV, Belgium.
- Mathias Hilfiker, Université du Luxembourg and AstraZeneca AB, Sweden.

Dr. Mariia Radaeva, Vancouver Prostate Center, The University of British Columbia, Canada; PhD defended 23rd August, 2024 at The University of British Columbia; now Board Member of Innovation OnBoard

On average 2-3 articles per fellow as the first author

- finishing PhD during 4th year at the respective University

Schools, conferences, challenges

ICANN24

33rd International Conference on Artificial Neural Networks

Tox24 Challenge: How accurately can we predict binding to transthyretin?

Start:17/05 » Submit:31/08 » Winner:18/09 » Article:31/12

AiChemist

Chemical Research in Toxicology

https://ochem.eu

ICANN24

33rd International Conference on Artificial Neural Networks

MENU

Six schools
Transferable skills
On-line presentations

Explainable AI for Molecules - AiChemist MSCA DN Horizon Europe

AiChemist-DN developing and implementing exlplainable representation learning approaches in drug discovery

Pharmaceutical Manufacturing · Munich · 1K followers · 11-50 employees https://aichemist.eu

Djork-Arné & 5 other connections work here

Message | ✓ Following | ···

Home | My Company | About | Posts | Jobs | People

Overview

AiChemist is funded by the European Union's Horizon Europe under the Marie Skłodowska-Curie grant agreement No 101120466

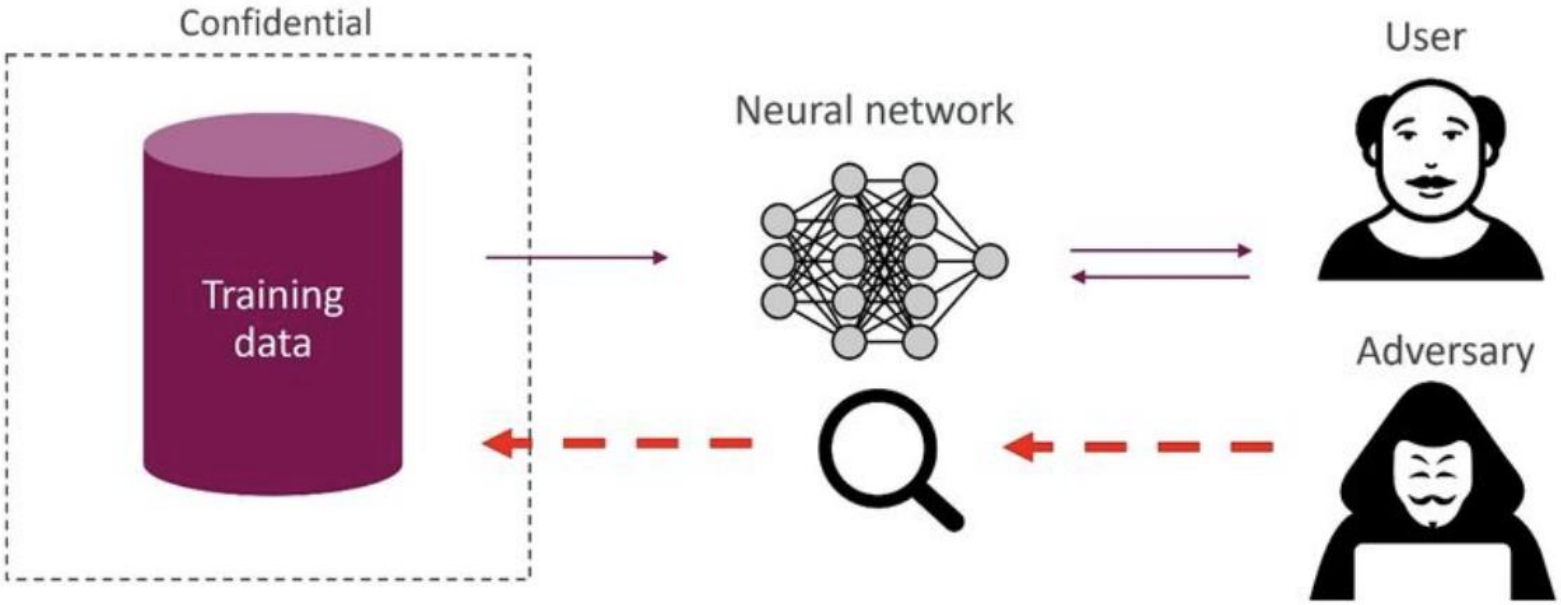
https://aichemist.eu/news

Explainable AI for Molecules - AiChemist MSCA DN Horizon Eu...

1,171 followers

1d •

On Monday the 9th of December at 14:00 CET, AiChemist fellow **Fabian Krüger** will give a talk on the vulnerability of proprietary training data in open-science frameworks, within the drug discovery context. All are welcome to join - a ...more



Dina Khasanova and 25 others

3 reposts

Like Comment Repost Send

Thank you for your attention!

https://aid-dd.eu

https://aichemist.eu

https://github.com/aidd-msca

CECAM Flagship School: 28/04 – 02/05/25, Lausanne (apply soon!)

Follow at X/twitter: @aichemist_dn

https://www.linkedin.com/company/aichemist