The Explainable AI for Molecules (AiChemist) Project

Newsletter #1 July 2024

The Explainable AI for Molecules (AiChemist) project is a Marie-Skłodowska-Curie Doctoral Network (DN) funded by the European Commission under the Horizon Europe Programme, Horizon-MSCA-2022 grant agreement number 101120466 started on 1st September 2023. The network brings together sixteen academic and industry partners from 8 European countries in addition to the Korean Institute of Toxicology (South Korea) to train fourteen doctoral candidates (DCs) in close collaboration with associated partners based in Europe and the USA.

This newsletter covers the first stages of the project's development, including the recruitment process, and introduces the main project partners (the beneficiaries).

Project Development

The AiChemist project started on 01.09.2023 and on 14.09.2023 the online kick-off meeting was held, during which the partners formally introduced themselves to one another. The job advertisements for the DC positions had already been published on the AiChemist website in early August, before the official project start date, and by mid-August a large-scale recruitment campaign had begun.

The job adverts were circulated via X (reaching 34k views), LinkedIn (initial post was reposted >100 times), EURAXESS and the "Women In Machine Learning" Google Group. Following the initial official application deadline on 10.09.2023, an online interview process was set up and by the date of the third recruitment meeting on 21.11.2023, 9 DCs had been selected. As of the time of writing, just two DC positions (DC2 and DC14) remain unfilled.

Overview of Recruitment

Up until June 2024, AiChemist DN received 358 applications from all over the world (~34.33% from Southeast Asia, 18% from the Middle East and North Africa, ~12.33% from East Asia, ~10.33% from Eastern Europe/Central Asia, 10% from Northwest Europe, 9% from Southern Europe, 4% from West/Central/East Africa, and 2% from the Americas).

20% of the applications came from female candidates. 123 candidates (~34%) that met the eligibility requirements and basic selection criteria were identified and short-listed.

Of the short-listed candidates, 35% were female. The names of the short-listed applicants, along with the positions that the respective candidate applied for (in order of preference), were listed in a private Google Sheet assessable to all PIs. All PIs had access to the applications mailbox and therefore were able to view the candidates' applications documents and contact them for interviews over Zoom or Microsoft Teams. The interview dates and links were added to a shared Google Calendar, so that other PIs considering the same applicants for their positions could join the interview. With the consent of the interviewee, some interviews were recorded to aid with decision-making.

Over the entire recruitment period, **56 candidates (46% female) were interviewed**. In the Google Sheet, each PI added their comments on the candidates' interview performance, domain-specific knowledge, work experience and mobility experience, and gave the candidates a score between 1 (very poor) and 10 (excellent). Those with the highest scores were offered positions. The consortium made sure to follow the European Charter and Code of Conduct for the Recruitment of Researchers (ECCCRR).

DCs selected between October 2023 and July 2024:

DC1: Fabian Krüger (m), German citizen, accepted offer, hired March 2024

DC2: [redacted] (f), Swiss citizen, female, rejected offer.

DC3: Karoline Schjelde (f), Danish citizen, accepted offer, hired July 2024.

DC4: Matthew Ball (m), British citizen, accepted offer, to be hired August/September 2024.

DC5: Bob van Schendel (m), Dutch Citizen, accepted offer, hired March 2024.

DC6: Mateusz Iwan (m), Polish Citizen, accepted offer, hired February 2024.

DC7: Andrea Hunklinger (f), German citizen, accepted offer, hired March 2024.

DC8: Ghaith Mqawass (m), Syrian citizen, accepted offer, hired June 2024.

DC9: Marcel Hiltscher (m), German citizen, accepted offer, to be hired July/August 2024.

DC10: Subashini Kennedy (f), Indian citizen, accepted offer, hired June 2024.

DC11: Vasilii Fastovski (m), Russian citizen, accepted offer, to be hired August/September 2024.

DC12: Eric Alcaide Aldeano (m), Spanish citizen, accepted offer, hired July 2024.

DC13: Sacha Raffaud (m), French citizen, accepted offer, hired January 2024.

Four of the thirteen selected candidates were female (~31%). Given that 30% of the short-listed candidates were female, this was the expected result, as we did our best to achieve a gender balance by ensuring that roughly equal numbers of male and female candidates were interviewed, while also treating candidates equally, as specified in the ECCCRR. Since the selected DC2 candidate rejected the offer, the supervising PIs have continued their search, and have recently extended an informal offer to another female candidate, which will be made formal once the HMGU human resources department completes the screening process and approves the appointment. In case approval is not granted, the supervising PIs will continue searching for candidates. The supervising PI of DC14 has extended an offer to a candidate and expects to receive a final decision from the candidate by mid-July.

Introducing the Beneficiaries

The Project Coordinator, Helmholtz Munich - German Research Centre for Environmental Health (HMGU) is a member of the Hermann von Helmholtz Association of German National Research Centres and is managed as a limited non-profit company. Investigations are carried out in complex systems sustaining human life and health at the interface of environmental influences and genetic predisposition. The HMGU runs 52 scientific institutes divided into seven departments. 2475 employees from 85 nations, including 650 PhDs, work in close linkage to the other Helmholtz Research Centres as well as external research partners. HMGU was a participant in >200 FP7/Horizon2020/Horizon Europe projects, including 48 ERC grants, and has coordinated six MSCA DNs (four of which were coordinated by Dr. Igor Tetko). HMGU is in charge of the Helmholtz Artificial Intelligence Cooperation Unit (Helmholtz-AI). The DCs employed at HMGU will gain full access to this impressive AI environment through seminars, courses, tutorials, and personal interactions with its members. Examples of recent studies include state-of-the-art methods for retrosynthesis, innovative representation learning Transformer CNN, as well as openOCHEM https://ochem.eu. The DCs will not only receive training in the development of new algorithms but will also have access to experimental validation on a number of targets being developed in the centre.

AstraZeneca has extensive expertise in drug discovery and development, from early-stage hit identification to worldwide production of registered drugs. Computational chemists at AZ have made seminal discoveries in machine learning for drug discovery (e.g. developing new methods for chemical structure generation using deep neural networks in the BIGCHEM http://bigchem.eu and AIDD https://ai-dd.eu projects). The employed DCs will learn cutting-edge computational methods for drug discovery and the theoretical foundations of novel algorithms for molecular generation, molecular dynamics, reaction prediction, and quality of focused library assessment, in a real-world drug discovery industry environment. The Molecular AI section that will host PhD students is co-located with the chemistry automation laboratory, giving students ample opportunity to explore the synergy between AI and automation.

Bayer AG has established the Machine Learning Research group to develop the tools needed to impact drug discovery in the age of digitalization. Molecular embeddings are developed using input sequences, graphs and 3D shapes of molecules. Such representations are used downstream to optimise drug candidates, predict their ADMETox properties, or prioritise synthesis to name a few. Models are based on high-quality proprietary data and put in the hands of the end users via our deployment platform. Explainability has been in the spotlight recently, with new methods to improve the explainability of graph neural networks, explain QSAR model outcomes or make sense of unsupervised embeddings. The DCs employed at Bayer will learn how to apply and develop explainability methods in relevant use cases for the pharmaceutical and agrochemical industry.

Pfizer was established more than 150 years ago and is a world-leading biopharmaceutical company with an established heritage in drug discovery and development. Pfizer has recently established a new Machine Learning Research department, which will bring its strong scientific expertise to this project. The new group consists of highly respected scientists in the field of machine learning whose scientific career includes publications in the most prestigious forums of machine learning, such as ICLR, ICML and NeurIPS. The group has an excellent track record in learning on graphs, e.g., for equivariant message passing for the prediction of tensorial properties and molecular spectra or for learning unsupervised permutation-invariant graph-level representations. The DCs will learn how to apply machine learning techniques in the scope of drug discovery projects and how to develop molecular descriptors that can be used in the context of optimization of chemical reactions.

Sanofi is a global life sciences company committed to improving access to healthcare and supporting the people we serve throughout the continuum of care. The In Silico Design group, within the Integrated Drug Discovery unit, develops and applies Machine Learning algorithms to drug discovery programs. Group members have recently published, e.g. on molecular generation under scaffold constraints and on the predictability of reaction yields. The DCs will be integrated within a dense and international network of inhouse experts in drug design as well as in computational methods development and deployment.

Molecular Networks (MN) is a software and knowledge development company specialised in the areas of chemoinformatics and computational toxicology. The international team of scientists is headed by Prof. Chihae Yang, a recognized and leading domain expert in these areas. Established in 1997, the company has also a long tradition and strong expertise in the application of machine learning and artificial intelligence methods and approaches to chemical and pharmaceutical R&D challenges including toxicity and metabolism prediction and profiling, calculation of physical, chemical, physicochemical, and biological properties,QSA/PR modelling as well as the modelling of chemical reactivity.

Københavns Universitet (UCPH) is the leading university in Denmark and one of the top universities in Europe. The Department of Chemistry has 50 academic staff members, with more than 10 working within theoretical chemistry. This strong focus on theory provides a strong basis for top research in different areas of theoretical chemistry, chemical reactivity being one of these. Among others, state-of-the-art methods in reaction predictions, genetic algorithms in application to molecular design, metadynamics at semi- empirical level were developed.

University of Strasbourg (UNISTRA) has a long experience of Chemoinformatics, with a 5-year study program dedicated to Chemoinformatics started in 2001 and a laboratory founded in 2004. The pedagogical program is backed by an internationally recognized research team. For instance, the team of Strasbourg is developing a toolkit in Chemoinformatics named ISIDA. This toolkit combines management of chemical information, analysis of data, machine learning, artificial intelligence and model publication tools. These tools are highly original, such as the Generative Topographic Maps used as a chemography tool to draw maps of the chemical universe in an analogous way to a geographic map. The maps can be colour coded according to a chemical property and combined with artificial intelligence algorithms to generate new chemical structures for engineering applications such as drug design, materials, etc. The team also contributed interpretable QSAR.It is also developing the concept of Condensed Graph of Reaction that was recently used to generate new reactions.

Leiden University (ULEI) is a public research university in Leiden, Netherlands. The university has seven academic faculties and over fifty subject departments, housing more than forty national and international research institutes, and it has produced twenty-six Spinoza Prize Laureates and sixteen Nobel Laureates. Mike Preuss is an associate professor at LIACS, the Computer Science department of ULEI. He is a renowned scientist in the fields of multi-modal and multi- objective optimization, search and reinforcement learning algorithms, and machine learning methods that are first explored in the area of computer games but then transferred to real-world problems. His algorithms show great potential for solving real chemical synthetic tasks where all reagents, reactants, catalyst, solvent, and temperature needs to be taken into account to optimise the outcome and the yield of reactions and reaction planning. The employed DC will learn and implement the algorithms to select appropriate conditions for the reactions and also score alternative retrosynthetic pathways suggested by the models. Capabilities in multi-objective optimization in this area are critical for the success of the AiChemist project.

The Mario Negri Institute (IRFMN) is a non-profit pharmacological and biomedical foundation for training and research activities. The Laboratory of Environmental Chemistry and Toxicology has been active in the field of predictive toxicology for more than 20 years. In most recent years, the group research was focused on the development and use of machine learning (ML) based models and their use as NAMs to inform on toxicological properties and fulfil regulatory needs.

The Centre for Genomic Regulation (CRG) is a biomedical and genomics research center based in Barcelona. Most of its facilities and laboratories are located in the Barcelona Biomedical Research Park, in front of Somorrostro Beach. It provides an excellent scientific and technological framework at the forefront of life sciences. Dr. Noelia Ferruz is a group leader and expert in unsupervised generative models for protein design and will provide access to the incoming DC to an impressive experimental and computational array of platforms.

The Technical University of Munich (TUM) has a unique profile in Germany with its core domains in natural science, engineering, life science and medicine, and is regularly among the high performers in international rankings, with the latest Times Higher Education World University Ranking TUM among the three best universities in the EU. Prof. Theis is a pioneer in biomedical artificial intelligence and machine learning, particularly in the context of single-cell genomics. His group develops novel machine learning algorithms to solve complex biological and medical questions. The DC will benefit and learn from the recent work from the Theis lab in the scope of modelling drug perturbation responses at the single-cell level as well as from ongoing work from Prof. Theis's ERC advanced grant focusing on systematically modelling single cell behaviour under external perturbations, particularly drug-induced perturbations with single-cell readouts.

The Technical University of Eindhoven (TUE) is a research-driven university and one of the leading technical universities in Europe. TU/e is the Dutch member of the EuroTech Universities Alliance, a strategic partnership of universities of science & technology in Europe. TU/e hosts the Institute for Complex Molecular Systems (ICMS), in which various departments are clustering their strengths to understand complex molecular systems at the fundamental level. The Eindhoven AI Systems Institute (EAISI) brings together all AI activities of the TU/e, with top researchers from various research groups working together to create new and exciting AI applications with a direct impact on the real world. The Molecular Machine Learning team led by Grisoni works at the interface between method development and experimental validation, in the field of de novo design,active learning, molecular property prediction, and XAI. Their research is supported by the Centre for Living Technologies and the The Dutch National Supercomputer facilities.

The École normale supérieure (ENS) is one of the constituent members of Paris Sciences et Lettres University (PSL). Due to its special historical role, large endowment, and influence within French society, the ENS is generally considered the most prestigious of the *grandes écoles.* The ENS benefits from a French government funded program (CMA-IA) to establish a new curriculum on AI for physics, chemistry, and biology, building upon the world leading mathematics departments of the ENS and Paris-Dauphine. Prof. Rodolphe Vuilleumier of the ENS Chemistry Department is a renowned scientist on the description of chemical reactions in complex environments from quantum and statistical mechanics approaches. Recently, he has made contributions, in collaboration with Sanofi, in AI for drug design and for the optimization of reaction conditions to improve reaction yields using a database extracted from literature.

The Università della Svizzera italiana (USI) sometimes referred to as the University of Lugano is a public Swiss university established in 1995, with campuses in Lugano, Mendrisio and Bellinzona (Canton Ticino, Switzerland). The Dalle Molle Institute for Artificial Intelligence Research (IDSIA), a joint institute split between USI and Scuola universitaria professionale della Svizzera italiana (SUPSI) has been co-directed by Prof. Jürgen Schmidhuber, a highly acclaimed computer scientists considered one of the "Fathers of Deep Learning" since 1995. He has co- invented LSTM for recurrent neural networks and has a clear vision on how machine learning will be revolutionised soon. His research group focuses on fundamental research in machine learning and AI, but also covers innovative application areas, e.g., in the framework of medical technology. The DC will be working in a vibrant creative environment in one of the best machine learning specialist groups in the world.

The École Polytechnique Fédérale de Lausanne (EPFL) is a public research university in Lausanne, Switzerland with a very strong international standing in science, technology, engineering, mathematics and natural sciences. Prof. Philippe Schwaller has made an outstanding contribution in the fields of ML for chemical reactions, more specifically he developed seminal works on the use of chemical language models for reaction prediction, synthesis planning, and data-driven reaction representations, for which he received multiple prizes. The DC will work in an excellent environment in one of the best Universities in the world in the field of chemistry.

The Korea Institute of Toxicology (KIT) is a global toxicity research institute for the nation's health and to build a safer society. KIT is affiliated with National Research Council of Science and Technology (NST) which is under the Ministry of Science and ICT (MSIT) in South Korea. It was founded to contribute to the public health and welfare enhancement by safety assessments of chemical and biological materials by research developments of related technologies, and their GLP system has been certified by Korean and international regulatory authorities based on OECD and US FDA GLP criteria. KIT was the first organization in Asia accredited by AAALAC International for humane laboratory animal treatment. Given that importance of alternative testing method gains momentum, the Department of Predictive Toxicology at KIT aims to develop advanced predictive toxicology technology based on in vitro and in silico approaches. KIT plays a leading role in developing alternative toxicology technology, and toxicity prediction AI model development is currently a key research project. Examples of AI models from KIT are drug-induced liver injury prediction model (http://toxstar.kitox.re.kr/), blood-brain barrier penetration prediction model (https://toxbbb.kitox.re.kr/), nanotoxicity prediction models (OpenRiskNet services: https://openrisknet.org/e-infrastructure/services/ and
NanoToxRadar: http://nanotoxradar.kitox.re.kr/). The University of Science and Technology (UST) and KIT have founded a joint campus (called KIT campus in UST) to develop an education program focused on human and environmental toxicology.

Important events

- AiChemist is co-organizing the ICANN2024 conference https://icann2024.org and co-running the Tox24 Challenge https://e-nns.org/icann2024/challenge (deadline 31.08.2024) – the results of which will be announced during ICANN2024.

Coming soon

The next issue (Newsletter #2) will cover the first AiChemist School.