Utrecht Bioinformatics Center

Bioinformatics in the age of AI

The 11th UBC Symposium

7 November 2025

6th Floor
Princess Máxima Center for Pediatric Oncology
Heidelberglaan 25 | 3584 CS Utrecht

Utrecht University · UMC Utrecht · Princess Máxima Center pediatric oncology · Hubrecht Institute · Westerdijk Fungal Biodiversity Institute · Hogeschool Utrecht

Programme

08:30 – 09:00	**Registration**	
09:00 – 09:15	**Welcome by the UBC Executive Committee**	
09:15 – 10:00	**Keynote lecture by Prof. Alexandre Bonvin** 'Solving 3D puzzles of biomolecular interactions by physics- and AI-based integrative modelling'	
10:00 – 10:30	**Coffee break and** Poster presentations	
10:30 – 10:45	Dr. Chris van Oevelen (Hogeschool Utrecht) 'The University of Applied Sciences Utrecht (HU): Integrating Bioinformatics in Education and Research'	
10:45 – 12:00	**Pitch Talks "Machine Learning & Pipelines"**	
	Jan van Eck (UU)	*PLM-eXplain: Divide and Conquer the Protein Embedding Space*
	Charlotte van Dijk (UMC)	*Leveraging single-nucleus long-read sequencing to characterize RNA disruptions and refine genetic association testing in ALS*
	Dr. Michiel Thiecke (UMC)	*Pericode – predicting the regulatory role of the non-coding genome using enhancer MPRAs*
	Dr. Sam Nooij (UU)	*The Campylobacter jejuni and C. coli CRISPRscape*
	Kaylin Palm (UMC)	*Prediction of treatment response in atherosclerosis through integration of drug induced and human plaque transcriptomic profiles*
12:00 – 13:30	**Walking lunch and** Poster presentations	
13:30 – 14:15	**Keynote lecture by dr. Sebastian Lobentanzer** 'Agentic Systems are Upon Us'	
14:15 – 15:45	**Pitch talks "Bioinformatics of complex traits"**	
	Anastasia Spinou (PMC)	*A pathway-informed mutual exclusivity framework to detect genetic interactions in pediatric cancer*
	Joana Marques (Hubrecht)	*On the role of TP53 mutations in cancer aneuploidy*
	Dr. Mao Peng (Westerdijk Institute)	*Bioinformatic exploration of molecular mechanisms of fungal plant biomass conversion*
	Dr. Bart Schimmel (UU)	*Dissecting complex plant traits using pixels and pangenomes*
	Dr. Bastiaan von Meijenfeldt (NIOZ)	*Unlocking an ancient cell membrane mystery*
	Dr. Joana Dopp (UU)	*Single cell RNASeq of insect brains to understand circadian behaviours*
15:45 – 16:15	**Break**	
16:15 – 17:00	**Keynote lecture by Prof. Yiliang Ding** 'Decoding the Language of DNA and RNA in Nature'	
17:00 – 18:00	**Closing ceremony followed by drinks**	

Keynote Presentations

Solving 3D puzzles of biomolecular interactions by physics- and AI-based integrative modelling

Understanding the structure, interactions, and dynamics of biomolecular macromolecules is key to unraveling cellular processes and advancing drug discovery. Accurate modelling of these complexes benefits greatly from incorporating diverse sources of experimental or predictive information. To this end, we have developed HADDOCK (https://www.bonvinlab.org/software), a versatile integrative modelling platform available as a web service (https://wenmr.science.uu.nl). HADDOCK can seamlessly combine data from biochemical, biophysical, and bioinformatics approaches to improve both sampling and scoring of biomolecular assemblies.
Over more than two decades of continuous development, we have also witnessed the transformative rise of AI in structure prediction. While AI has made remarkable progress, physics-based modelling remains essential, with many challenges still requiring their complementary strengths. In this talk, I will present recent advances in HADDOCK and showcase its applications, including examples where AI-generated predictions are combined with physics-based integrative modelling. In particular, I will highlight studies on antibody and nanobody–antigen complexes, demonstrating the synergy between AI and physics-based approaches.

Professor Alexandre Bonvin

Alexandre Bonvin (1964) studied Chemistry at Lausanne University, Switzerland and obtained his PhD at Utrecht University in the Netherlands (1993). After two post-doc periods at Yale University (USA) and the ETHZ (CH) he joined Utrecht University in 1998 where he was appointed full professor of computational structural biology in 2009. In 2006, he received a prestigious VICI grant from the Dutch Research Council. He was director of chemical education from February 2009 until February 2012, vice head of the Chemistry Department (2010-2012 and 2019-2022) and Scientific Director of the Bijvoet Centre for Biomolecular Research from Sept. 2019 until Sept. 2023. He has and is participating to several EU projects including the BioExcel Center of Excellence in Biomolecular Simulations and the EGI-ACE projects. His work has resulted in over 275 peer-reviewed publications.
Research within the computational structural biology group focuses on the development of reliable bioinformatics and computational approaches to predict, model and dissect biomolecular interactions at atomic level. For this, bioinformatics data, structural information and available biochemical or biophysical experimental data are combined to drive the modelling process.

Back to programme

Agentic Systems Are Upon Us

Recent breakthroughs in large language models and multimodal architectures have given rise to new “agentic” AI systems capable of performing complex, context-sensitive tasks with minimal human supervision. Their broad appeal spans domains as diverse as healthcare, finance, and scientific research, yet their rapid proliferation also highlights critical challenges in robustness, reproducibility, and real-world reliability. Proprietary silos of development often prevent thorough auditing and restrain the scientific community’s capacity to independently assess or refine these systems—introducing risks of error propagation, bias, and unaccountable decision-making.

In response, open-source methodologies offer a principled path forward. By designing transparent, domain-relevant frameworks for knowledge management and AI-agent support, we can ensure that researchers and practitioners alike retain the agency to drive innovation. Collaboration and shared governance within global networks help democratise access and federate resources, lowering barriers to entry and expanding the collective expertise that fuels method validation and improvement. These bottom-up efforts invite a broader range of stakeholders to participate in iterative design, stress-testing, and validation, ultimately enhancing the systems’ trustworthiness.

Open-source agentic systems are needed to preserve academic freedom and integrity within rapidly advancing AI fields. By learning from early adopters, highlighting the shortcomings of “black-box” solutions through rigorous benchmarks, and discussing strategies to build robust community- and domain-driven ecosystems, we aim to facilitate the creation of frameworks that advance scientific discovery rather than obscuring it. In doing so, we not only promote reproducible science but also uphold the principle that AI’s future—particularly in mission-critical applications—must remain accountable, interdisciplinary, and accessible to all.

Dr. Sebastian Lobentanzer

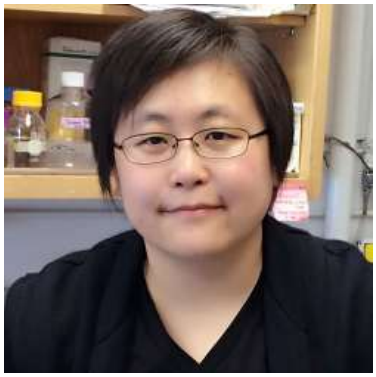
Dr. Sebastian Lobentanzer is a seasoned biomedical researcher and research software engineer with extensive expertise in systems pharmacology. Having completed a PhD in pharmacology and toxicology, he is particularly interested in disentangling causal relationships in molecular biology. He has pursued this focus as a postdoctoral researcher in the lab of Julio Saez-Rodriguez at Heidelberg University Hospital from 2021 to 2025 and, since 2024, in affiliation with the Open Targets group at the European Bioinformatics Institute (EMBL-EBI). Since 2025, he is a Principal Investigator at Helmholtz Munich and leads the Computational Biology Unit at the German Centre for Diabetes Research, where he does Accessible Biomedical AI Research (https://slolab.ai).

Back to programme

Decoding the Language of DNA and RNA in Nature

The structural dynamics of DNA and RNA play critical roles in regulating gene expression. In our work, we have developed a novel in vivo chemical profiling method, enabling the discovery of tertiary RNA G-quadruplex structures in eukaryotes. We have also uncovered that RNA G-quadruplex structure serves as a molecular marker to facilitate plant adaptation to the cold during evolution. Recently, we established two novel DNA and RNA foundation models, PlantDNA-FM and PlantRNA-FM, that facilitates the explorations of functional DNA/RNA structure motifs across genomes and transcriptomes. Our PlantDNA-FM was pretrained on the genome sequences from 156 plant species, while our PlantRNA-FM leveraged a large dataset integrating RNA sequences and structures from 1,124 plant species. Both models demonstrate state-of-the-art performance in plant-specific downstream tasks. Through an interpretable framework, PlantDNA-FM enables genome-wide identification of regulatory motifs, particularly in promoter regions, and PlantRNA-FM identifies functional RNA sequence and structural motifs including secondary and tertiary structures across transcriptomes. Experimental validations confirmed the biological relevance of these FM-guided predictions. Taken together, our PlantDNA-FM and PlantRNA-FM facilitate the exploration of functional DNA/RNA motifs across the complexity of transcriptomes, offering transformative potential for nucleic acid code programming in vivo.

Professor Yiliang Ding



Professor Yiliang Ding FRSB has been a group leader with tenure at the John Innes Centre since 2014, focusing on RNA structure functionality in living cells. She also holds positions as an Honorary Group Leader at the Babraham Institute and an Honorary Professor at the University of East Anglia. Ding completed her bachelor's degree at Shanghai Jiao Tong University in 2005 and her PhD at the John Innes Centre in 2009. Professor Ding was a postdoc researcher at Penn State University from 2010 to 2013. Dr Ding's innovative methods for profiling RNA structures in living cells have delivered new insights into the functional roles of RNA structures in gene regulation. She has received several prestigious awards and grants, such as BBSRC David Phillips Fellowship, three ERC grants and Royal Society Faraday Fellowship. Professor Ding is one of the nine recipients of the 2024 Blavatnik Awards for Young Scientists in Life Sciences in the UK.

Back to programme

Pitch Talks

The University of Applied Sciences Utrecht (HU): Integrating Bioinformatics in Education and Research

Dr. Chris van Oevelen (ILC, University of Applied Sciences Utrecht)

Machine Learning & Pipelines

Session chair: Dr. Sietske van Bentum
Speakers:

Jan van Eck (UU)	PLM-eXplain: Divide and Conquer the Protein Embedding Space
Charlotte van Dijk (UMC)	Leveraging single-nucleus long-read sequencing to characterize RNA disruptions and refine genetic association testing in ALS
Dr. Michiel Thiecke (UMC)	Pericode – predicting the regulatory role of the non-coding genome using enhancer MPRAs
Dr. Sam Nooij (UU)	The *Campylobacter jejuni* and *C. coli* CRISPRscape
Kaylin Palm (UMC)	Prediction of treatment response in atherosclerosis through integration of drug induced and human plaque transcriptomic profiles

Bioinformatics of Complex Traits

Session chair: Dr. Mark Bakker
Speakers:

Anastasia Spinou (PMC)	A pathway-informed mutual exclusivity framework to detect genetic interactions in pediatric cancer
Joana Marques (Hubrecht)	On the role of TP53 mutations in cancer aneuploidy
Dr. Mao Peng (Westerdijk Institute)	Bioinformatic exploration of molecular mechanisms of fungal plant biomass conversion
Dr. Bart Schimmel (UU)	Dissecting complex plant traits using pixels and pangenomes
Dr. Bastiaan von Meijenfeldt (NIOZ)	Unlocking an ancient cell membrane mystery
Dr. Joana Dopp	Single cell RNASeq of insect brains to understand circadian behaviours

Back to programme

The University of Applied Sciences Utrecht (HU): Integrating Bioinformatics in Education and Research

Dr. Chris van Oevelen

Institute for Life Sciences and Chemistry (ILC), University of Applied Sciences Utrecht

The University of Applied Sciences Utrecht (HU) is proud to join the UBC community. Our bioinformatics activities span both education and research within the Institutes of Life Science and Chemistry (ILC) and ICT.

From the first year, ILC students develop strong foundations in reproducible data analysis using R, preparing them for project-based learning in later years. These projects, carried out in collaboration with external partners, address diverse topics such as antimicrobial resistance (AMR) gene detection in sequencing data, single-cell transcriptomics, metagenomics, and R Shiny dashboard development.

Our research focuses on developing innovative, data-driven solutions in the life sciences. Ongoing projects include alternatives to animal testing (ombion-cpbt.nl), predictive toxicology using AI (vhp4safety.nl), and advanced image analysis pipelines (UMC Utrecht – IMAGINE). Together, these initiatives reflect HU's commitment to integrating bioinformatics into both education and applied research.

Back to programme

Back to Pitch Talks

PLM-eXplain: Divide and Conquer the Protein Embedding Space

Jan van Eck
Utrecht University

Protein language models (PLMs) have revolutionised computational biology through their ability to generate powerful sequence representations for diverse prediction tasks. However, their black-box nature limits biological interpretation and translation to actionable insights. We present an explainable adapter layer - PLM-eXplain (PLM-X), that bridges this gap by factoring PLM embeddings into two components: an interpretable subspace based on established biochemical features, and a residual subspace capturing non-interpretable but predictive information. Using embeddings from ESM2 and ProtBert, our adapters incorporate well-established properties, including secondary structure and hydropathy while maintaining high performance. We demonstrate the effectiveness of our approach across three protein-level classification tasks: prediction of extracellular vesicle association, identification of transmembrane helices, and prediction of aggregation propensity. PLM-X enables biological interpretation of model decisions without sacrificing accuracy, offering a generalisable solution for enhancing PLM interpretability across various downstream applications. This work addresses a critical need in computational biology by providing a bridge between powerful deep learning models and actionable biological insights.

Back to programme

Back to Pitch Talks

Leveraging single-nucleus long-read sequencing to characterize RNA disruptions and refine genetic association testing in ALS

Charlotte van Dijk
University Medical Center Utrecht

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease with an average life expectancy of 3–5 years after diagnosis. About half of the risk for developing ALS is determined by genetics, yet a specific genetic risk factor can be identified in only ~15% of cases. We aim to improve the explained heritability by fine-tuning established genetic tests to the transcriptomes of vulnerable cell types. ALS is characterized by the loss of motor neurons (MNs), large and rare cells located in the motor cortex and spinal cord. Using single-nucleus RNA sequencing (snRNA-seq) and PacBio Kinnex long-read RNA sequencing of postmortem tissue, we characterize the isoform landscape expressed in MNs. Because long-read RNA sequencing from nuclei has only recently become feasible, there are no established methods for analyzing such data. We therefore aim to develop a pipeline that removes technical and biological artifacts while retaining sensitivity. By including data from both patients and controls, we aim to move beyond characterizing the isoform landscape to identify isoform-level disruptions associated with disease.

Back to programme

Back to Pitch Talks

Pericode - predicting the regulatory role of the non-coding genome using enhancer MPRAs

Dr. Michiel Thiecke
University Medical Center Utrecht

Deep learning models that predict transcriptional activity from DNA sequence have enabled a wide range of in-silico analysis directions, including mutational effect prediction and large-scale context-aware study of functional elements. Despite these achievements, there is a lack of models that accurately produce cell-type specific predictions, and the amounts of required training data can be prohibitive. We are developing PARM V2; a streamlined set of wet- and dry-lab techniques that enable building transcription-predictive deep learning models for any cell type. To this end, we first constructed a Massively Parallel Reporter Assay (MPRA) targeted at a diverse selection of distal regulatory elements, as well as at all promoters. The resulting data enable us to train predictive models that learn from distal-elements as well as core-promoters, thereby improving the capacity to predict transcription in a cell-type-specific manner and at noncoding as well as coding loci. This presentation focuses on selecting a set of distal regulatory elements that function in a wide range of cell types, to improve overall model performance as well as cell-type specificity.

Back to programme

Back to Pitch Talks

The Campylobacter jejuni and C. coli CRISPRscape

Dr. Sam Nooij
Utrecht University

Campylobacter jejuni and *C. coli* are pathogens that have several reservoirs. We aim to use the adaptive DNA defence system CRISPR-Cas to type *Campylobacter* strains and trace their source. CRISPR-Cas contains spacers that are chronologically integrated and represent exposure to foreign DNA like phages and plasmids. To assess this as potential typing method, we characterised CRISPR-Cas in all *C. jejuni* and *C. coli* genomes.

CRISPR-Cas alone cannot type all *Campylobacters*: it is present in 67% of *C. jejuni* genomes and 29% of *C. coli*. The size of the CRISPR arrays differs between *C. jejuni* and *C. coli* genomes. Analogous to pangenome terminology, we find that *C. jejuni* and *C. coli* have an open CRISPRscape with many unique spacers.

Based on a subset of genomes, we find that different defence systems are not mutually exclusive. Genomes often have both restriction-modification (RM) systems and CRISPR-Cas. However, linear modelling identified a putative evolutionary trade-off between CRISPR-Cas and RM type IIG.

Mapping spacers to bacteriophages and plasmids indicates that CRISPR-Cas more commonly targets bacteriophages (27% and 2.6% of spacers hit, respectively). For 67% of spacers we detected no targets. Further research will enhance understanding of CRISPRscape and its potential applications in epidemiology.

Back to programme Back to Pitch Talks

Prediction of treatment response in atherosclerosis through integration of drug induced and human plaque transcriptomic profiles

Kaylin Palm
University Medical Center Utrecht

Despite advances in prevention, atherosclerosis remains the leading cause of death worldwide. The limited success of new therapies reflects the reliance on simplified models that fail to capture disease complexity. Patient-specific factors such as sex, genetics, and comorbidities shape the molecular and cellular landscape of atherosclerotic plaques, resulting in distinct disease phenotypes. Many drug-responsive genes show subtype-specific expression, underscoring the need for therapies tailored to plaque biology.

Here, we present an approach to predict treatment response by modeling drug-induced transcriptional changes onto human plaque transcriptomic profiles. We trained a Random Forest classifier to distinguish plaque subtypes using Athero-Express transcriptomic data and then reclassified samples after simulated in-silico Colchicine treatment.

Modeling of Colchicine-treated endothelial cells indicated shifts toward less vulnerable plaque subtypes, with a significant reduction in patients classified with the lipomatous subtype. This effect was not uniform across all subtypes, suggesting that patients with lipomatous plaques may be more responsive to Colchicine.

Overall, the model predicted an increased proportion of less vulnerable plaques following simulated Colchicine exposure. These findings demonstrate that atherosclerosis comprises molecularly distinct plaque phenotypes with variable drug sensitivity and highlight the potential of modeling drug effects on human plaque transcriptomes to guide targeted, personalized therapy.

Back to programme

Back to Pitch Talks

A pathway-informed mutual exclusivity framework to detect genetic interactions in pediatric cancer

Anastasia Spinou
Princess Màxima Center

The exponential increase of sequenced cancer genomes has enabled the in-silico study of genetic interactions (GIs) - particularly synthetic lethality, where two gene alterations lead to cell death - and identify new candidate therapeutic targets. This rise is primarily present in adult cancer, while in-silico GI studies remain challenging in pediatric cancer. To advance our understanding of GIs in pediatric oncology, we developed the pathway-informed genetic interaction framework (PIGI) that employs mutual exclusivity and co-occurrence testing and leverages biological pathways to infer candidate GIs. Pathways facilitate the detection of hidden biology by grouping genes in functional units, and alleviate key confounders of these analyses: pathway epistasis and cancer subtypes - thereby highlighting genes of greater interest. PIGI detected 35 mutually exclusive and 2 co-occurring mutated gene pairs in two primary pediatric cancer datasets, DKFZ and TARGET. Over half of the identified gene pairs have not been previously described in the literature and only TP53-DROSHA gene pair in Wilms tumors has been reported before. Four of them could be promising candidates for synthetic lethal GIs. These findings highlight the benefits of GIs inference by exploring pediatric cancer data through pathways and propose new gene pairs for follow-up synthetic lethality experimentation

Back to programme

Back to Pitch Talks

On the role of *TP53* mutations in cancer aneuploidy

Joana F. Marques
Hubrecht Institute | University Medical Center Utrecht | Oncode Institute

Aneuploidy and *TP53* mutations are among the most frequent genetic alterations in cancer, and *TP53* inactivation is considered an important contributor to the emergence of cancer aneuploidy. It is unclear, however, if p53 protects against particular aneuploidy features and whether it does so universally. By analyzing *TP53* mutations and aneuploidy features in 31 cancer types of TCGA data set, we verify that on a pan-cancer level *TP53* mutant cancers tend to have a higher degree of aneuploidy. However, for many cancer types, the average degree of aneuploidy is similar in *TP53* wild type and mutant samples, and a substantial degree of aneuploidy can accumulate with intact *TP53* in most cancer types. Neither arm-level nor whole chromosome aneuploidy but rather chromosome loss events distinguish *TP53* mutant from wild type cancers. We conclude that *TP53* inactivation is neither sufficient nor necessary for the emergence of cancer aneuploidy, but it is associated with the degree of aneuploidy and chromosome loss in a subset of cancer types. Our findings underscore the strong and poorly understood cancer type specificity for the association between *TP53* mutations and aneuploidy in cancer.

Back to programme Back to Pitch Talks

Bioinformatic exploration of molecular mechanisms of fungal plant biomass conversion

Dr. Mao Peng
Westerdijk Institute

Fungal plant biomass conversion (FPBC) is of great importance to the global carbon cycle and has been increasingly applied to produce biofuel, enzymes and biochemicals. During the past 8 years we have developed a series of bioinformatics frameworks that combined evolutionary, statistical and machine learning methods for integrative analysis of multiple omics data. These computational tools greatly facilitated the discovery of novel genes, enzymes and pathways associated with FPBC, and accelerated their experimental validation.
In this talk, I will highlight two of our recent works: (1) discovering novel transcription factors controlling expression of crucial FPBC relevant genes; (2) elucidating evolutionary diversity of plant biomass utilization approaches across different fungal species. These findings not only enhanced our understanding of the complex molecular mechanisms underlying FPBC across diverse species, but also provided novel insights that can guide future genetic engineering of fungi for valorization of agriculture waste into value-added bioproducts.

Back to programme Back to Pitch Talks

Dissecting complex plant traits using pixels and pangenomes

Dr. Bernardus C. J. Schimmel
Utrecht University

When plant defenses are activated, growth is often inhibited. This growth–immunity trade-off is a complex, poorly understood trait, which involves the reallocation of resources and numerous interactions between signaling pathways. Here, we have leveraged advances in high-throughput automated digital phenotyping and pangenomics to gain novel genetic insights into the growth–immunity trade-off in cultivated lettuce. We transiently exposed 180 lettuce genotypes to two defense-eliciting (growth-inhibiting) chemicals, and monitored these plants before, during and after the treatments with four top-view imaging systems in an NPEC facility. This produced 12 TB of RGB, chlorophyll fluorescence, thermal, and hyperspectral images. Using custom AI-powered bioinformatics pipelines, we extracted over 32,000 time-resolved digital phenotyping traits from these images and conducted genome-wide association studies (GWAS) to identify genomic loci that mediate growth and immune responses. These loci provide a foundation for developing lettuce varieties with robust growth under biotic stress, while the paired genome–phenome data enables machine learning-based predictive tools for genomic breeding and precision agriculture.

Back to programme Back to Pitch Talks

Unlocking an ancient cell membrane mystery

Dr. Bastiaan von Meijenfeldt
NIOZ

Cellular membranes are composed of lipids and embedded proteins. Although all living cells are surrounded by a membrane, the main lipid components remarkably differ between the domains of life; archaea incorporate isoprenoids ether-linked to glycerol-1-phosphate into their membrane, whereas bacteria and eukaryotes incorporate fatty acids ester-bound to glycerol-3-phosphate. Why this 'lipid divide' exists is still unknown. To unbox the mystery, I will zoom in on the proteins that are embedded inside the membrane. Investigating sequence and structure of integrated membrane proteins and genomic presence of lipid biosynthesis genes throughout the tree of life, I will create the first large-scale comparison of archaeal and bacterial membranes. Using phylogenetics and ancient protein reconstruction, I will investigate key ancestors in the tree of life that can shed light on the lipid divide; the membrane lipids of the Last Universal Common Ancestor (LUCA) are central to understanding the origin of the divide, and the membrane of the Last Asgard archaea and Eukaryotes Common Ancestor (LAECA) can provide insights into a putative membrane lipid transition during eukaryogenesis. My findings will be important for understanding early cellularity, eukaryogenesis, and the detailed functioning of cellular membranes.

Back to programme Back to Pitch Talks

Single cell RNASeq of insect brains to understand circadian behaviours

Dr. Joana Dopp
Utrecht University

The ant *Camponotus floridanus* turns into a zombie once infected by *Ophiocordyceps*, a fungal parasite. The infection timeline is stereotypic and correlates with the light-dark cycle. Our prior bulk transcriptomic work has further highlighted that the biological clock may drive the abnormal rhythmic behaviour of infected ants. However, what remains unclear are the detailed molecular mechanisms of this clock-controlled extended phenotype, such as the cell populations most affected by fungal manipulation. Utilising a published ant brain single-cell transcriptomic atlas, we show that our bulk candidate genes are expressed heterogeneously across the ant brain and that their expression clusters specifically in glial cell subtypes. This suggests that the fungal parasite targets cells selectively rather than attacking its host's brain as a whole. The finding also highlights the need for investigating molecular correlates of this behaviour at single-cell resolution. Here, we propose to perform 10x single-cell RNA-sequencing around the clock to reveal detailed underpinnings of behaviour-manipulation by leveraging the annotated *C. floridanus* genome, cross-species predictions of cell annotations based on extensive single-cell transcriptomic work in the insect model organism *Drosophila melanogaster* and a wide range of available bioinformatic tools for single-cell analysis.

Back to programme Back to Pitch Talks

Poster Presentations

# Poster	Presenter	Title
1	Daphne van Ginneken (UMCU)	Deciphering antibody repertoire evolution using protein language models and B cell lineage inference with AntibodyForests
2	Gianpaolo Cardellini (PMC)	Expanding the M&M Pan-Cancer Classifier with Batch Effect Correction Using Transfer Learning and Autoencoders
3	Irene Gonzales Ortega Villena (UU)	Exploring rare event detection techniques to monitor reproduction and health in dairy cows
4	Dr. Joanna von Berg (PMC)	The Dutch childhood cancer genome project: Identifying driver gene candidates from multiple mutation types
5	Dr. Kristof Van Avondt (UMCU)	Spatial Biology Accelerator at UMC Utrecht: Enabling High-Resolution Insights Through Spatial Omics Technologies
6	Eveline Ilcken (UMCU)	High-throughput pegRNA screen for prime editing of monogenic disorders
7	Viktorija Vodilovska (UU)	Mutational Signature Analysis of MACROD2 and PRKN deletions in Colorectal Cancer
8	Tristan Schadron (RIVM)	Source attribution of Shiga-toxin-producing Escherichia coli (STEC): a multinational study in Europe using virulence profiles and core-genome multi locus-sequence typing
9	Roula Farag (PMC)	Uncovering the Landscape of Chromosomal Instability in Pediatric Solid Tumors
10	Dr. Julian A Paganini (IRAS)	Comparing Machine Learning Algorithms and Genome Representations for Source Attribution of Salmonella Typhimurium
11	Dr. Kaixin Hu (UU)	Metagenomic deep learning to support risk assessment of antibiotic resistant genes
12	Dr. Flip Mulder (UBEC, CMM, UU)	UBEC – Assisting and connecting in bioinformatics
13	Dr. Victor Reys (UU)	On the road to predict the human PDZ domains interactome by combining both structure based and machine learning methods

Back to programme

Deciphering antibody repertoire evolution using protein language models and B cell lineage inference with AntibodyForests

Daphne van Ginneken

University Medical Center Utrecht

B cell selection and evolution are key processes in regulating successful adaptive immune responses. Recent advances in single-cell sequencing and deep learning strategies have unlocked new potential to study affinity maturation of B cells at unprecedented scale and resolution. To unravel the complex dynamics of B cell repertoire evolution during immune responses and to facilitate Protein Language Model (PLM)-guided antibody engineering, we created the R package AntibodyForests (van Ginneken, Tromp, et al. bioRxiv, 2025). AntibodyForests encompasses pipelines to infer B cell lineages, quantify inter- and intra-antibody repertoire evolution, and analyze somatic hypermutation (SHM) using PLMs and protein structure. Using AntibodyForests, we explore how general and antibody-specific PLM-generated likelihoods relate to features of in vivo B cell selection, evolution, antigen specificity and binding affinity (van Ginneken, Samant, et al. Briefings in Bioinformatics, 2025). We find that PLM likelihoods correlate with biologically relevant features including isotype and V-gene usage, mutational load, and SHM patterns. Additionally, we observed that mutating residues along evolutionary trajectories tend to have lower PLM likelihoods than conserved residues. These results indicate that PLMs could predict to what amino acid SHM will most likely mutate and at which position. Interestingly, our findings challenge in vitro observations (Hie et al. Nature Biotechnology 2023) by revealing a negative correlation between PLM likelihoods and antigen binding affinity in in vivo repertoires. In our exploitation of these discoveries using six different PLMs and varying sequence regions, we uncovered that the region of antibody sequence (Complementarity-Determining Region (CDR3) or full-length VDJ) provided to the PLM, as well as the type of PLM used, influences the resulting likelihoods. These comparisons emphasize the importance of PLM long-range interaction, potential training data biases, and pairing heavy and light chains. Together, these studies highlight the power of combining repertoire-wide phylogenetic inference with PLMs to better understand the principles governing antibody evolution and selection, and offer new tools for therapeutic antibody discovery and engineering.

Back to programme

Back to Poster presentations

Expanding the M&M Pan-Cancer Classifier with Batch Effect Correction Using Transfer Learning and Autoencoders

Gianpaolo Cardellini

Prinses Màxima Centrum

Accurate molecular classification of pediatric cancers using RNA-seq is essential for diagnosis and research, yet integrating data from multiple sequencing centers remains challenging due to batch effects - non-biological expression differences introduced by factors such as library preparation, sequencing platform, and tissue preservation. The M&M classifier, a state-of-the-art RNA-seq–based diagnostic model, performs robustly within a single center dataset but exhibits sensitivity to cross-center variability (Wallis et al. 2025). Current batch effect correction methods (e.g., ComBat-seq) reduce technical variance for small subsets of genes but alter the reference dataset, rendering them unsuitable for clinical workflows where new data must be mapped to a fixed model without retraining.

This project aims to develop a biologically informed Conditional Variational Autoencoder (CVAE) that corrects batch effects while preserving tumor-specific biology. Inspired by single-cell integration frameworks such as scArches (Lotfollahi et al. 2022), the model employs transfer learning to map new datasets onto a stable reference space through architecture surgery; freezing reference decoder weights and fine-tuning only dataset-specific encoders. Biological priors, including pathway activity scores derived via Gene Set Variation Analysis (GSVA), are integrated as conditional inputs to guide the latent representation toward biologically meaningful structure.

Preliminary analyses across Princess Máxima Center and St. Jude pediatric RNA-seq cohorts reveal subtype-dependent batch effects, while removal of batch-sensitive DEGs does not significantly impact M&M performance, indicating classifier robustness to moderate technical variation. Pathway-level analyses highlight distinct biological programs that will serve as conditioning features in the CVAE.

Back to programme

Back to Poster presentations

Exploring rare event detection techniques to monitor reproduction and health in dairy cows

Irene Gonzalez Ortega Villena

Utrecht University

There is a lack of visualization or decision support tools in the livestock sector, which makes difficult for farmers to interpret their own performance data. At the same time, prediction models such as logistic regression or random forest have been previously used to study pregnancy outcomes in dairy cattle. In this project, we propose a combination of those models with Statistical Process Control (SPC) methods and Tableau, a visualization software, to help dairy farm producers monitor cow health in real time and improve the smart decision-making level. To do so, we used datasets containing inseminations records, performed data processing using R, and developed the prediction model in Python, integrating it with Tableau. We built a random forest model with the metrics: 0.859 of Area Under Curve (AUC), 0.872 of sensitivity and 0.686 of specificity. Moreover, we linked a feature importances function to that model, which is able to describe which predictors contribute most to the model's variance and predictions. Lastly, we developed an interactive method to test the feature impact across all the inseminations, showing whether individual outcomes are influenced by specific features or if improvements or declines in farm performance are more likely due to farm-specific management actions. As global result, we created a Tableau dashboard that include these analyses in a single, intuitive view, allowing farmers to easily interpret the results and make informed decisions.

Back to programme

Back to Poster presentations

The Dutch childhood cancer genome project: Identifying driver gene candidates from multiple mutation types

Dr. Joanna von Berg

Prinses Màxima Centrum

Despite improvements in cure rates, cancer is still the leading cause of disease-related deaths among children in high-income countries. As childhood cancers are rare, concerted efforts are needed to identify relevant genomic alterations. The centralized position of the Princess Máxima Center for Pediatric Oncology enables uniform data generation from patients in The Netherlands. Whole-genome sequencing and RNA-sequencing are routinely performed. Currently, the dataset consists of 1,478 primary tumor samples.

Previously, candidate tumor driver genes have been identified, predominantly based on somatic single nucleotide variants (sSNVs). However, certain pediatric tumor types show an enrichment of larger somatic variants; copy number alteration (sCNAs) and structural variants (sSVs). Therefore, we aimed to identify candidate drivers by first testing variant-type-specific enrichment per gene. sSNVs were analysed with OncodriveFML, where variants with a higher predicted functional effect get a larger weight. sCNAs were analysed with GISTIC2.0, detecting genomic segments with significantly increased or decreased copy number. sSV breakpoints were analysed with our own approach; we used a Poisson-binomial test to calculate enrichment, correcting for gene length and number of sSV breakpoints per patient. Finally, we use the harmonic mean p-value (HMP) to integrate the variant-type-specific p-values to identify variant-type-agnostic candidate driver genes.

We identified 2,863 genes that are significantly affected (HMP empirical q-value < 0.05) by any variant type and are altered in at least 5 patients. If we had only considered individual variant types, we would have identified (variant-specific q-value < 0.05): 43 genes affected by sSNVs, 2,437 by sCNVs, and 829 by sSVs (these sets are not mutually exclusive). 2,680 of the candidate tumor drivers have not previously been identified in the Cancer Gene Census. These potential drivers will be further assessed through differential expression analysis. Eventually, this study will contribute to our understanding of tumor biology and may support future clinical decision-making.

Back to programme

Back to Poster presentations

Spatial Biology Accelerator at UMC Utrecht: Enabling High-Resolution Insights Through Spatial Omics Technologies

Dr. Kristof Van Avondt

Center for Molecular Medicine, University Medical Center Utrecht

Spatial biology offers a transformative approach to studying the molecular landscape of tissues by preserving spatial context during analysis. To facilitate the integration of these powerful techniques across biomedical research, UMC Utrecht has launched the Spatial Biology Accelerator — a collaborative platform that provides access to advanced spatial omics technologies and expert support.

The Accelerator supports a range of cutting-edge platforms, including spatial transcriptomics (Bruker's GeoMx and CosMx), spatial proteomics and multiplexed tissue imaging (Standard BioTools' Hyperion system for Imaging Mass Cytometry), and Bruker's timsTOF fleX for MALDI mass spectrometry imaging. These technologies allow researchers to visualize gene and protein expression patterns within intact tissue architecture, enabling insights into cell–cell interactions, microenvironmental changes, and disease mechanisms.

Beyond technology access, the Accelerator offers end-to-end support, including project design, sample preparation, data analysis, and integration with existing molecular or imaging datasets. Whether you are investigating cancer heterogeneity, tissue regeneration, immunology, or neuroscience, our team is ready to collaborate and help tailor spatial strategies to your research needs.

This poster presents an overview of the technologies and services provided by the Spatial Biology Accelerator and highlights opportunities for collaboration across disciplines. We invite researchers, clinicians, and data scientists at UMC Utrecht and beyond to engage with us and explore how spatial biology can accelerate discovery in their fields.

Back to programme

Back to Poster presentations

High-throughput pegRNA screen for prime editing of monogenic disorders

Eveline Ilcken

University Medical Center Utrecht

Genetic diseases affect over 10,000 families in the Netherlands, with significant child morbidity and mortality. Gene-editing technologies offer potential cures to these diverse diseases by correcting disease-causing mutations. Although gene-editing technologies such as prime editing can, in theory, correct any (smaller) mutation, editing efficiency is limited by the prime-editing guide RNAs (pegRNAs). Several aspects of the pegRNA that can influence editing also depend on the genomic context of the mutation, such as the PBS or RTT length and the folding of the pegRNA. Finding an optimal pegRNA, therefore, requires extensive experimental screening, as these parameters must be optimised for each mutation individually. To aid in the design of pegRNAs for efficient editing, we have set up a genomic screen to learn the design rules for efficient prime editing. For this, we designed a pegRNA library consisting of 37.894 pegRNAs targeting 34 clinically relevant mutations, with each mutation represented by 200-2000 unique pegRNAs varying in PBS and RTT lengths. The editing efficiency of these pegRNAs will be assessed with six different prime editors in both a mismatch repair proficient (K562) and deficient (HEK293T) cell lines. This creates 454,728 unique prime editing conditions from which important features for efficient prime editing can be learned in the context of different editors and DNA damage repair backgrounds. The screening data will be used in conjunction with large public screening data from the PRIDICT and DeepPrime studies to build a model for predicting prime editing efficiency. Using active learning cycles, this model will be iteratively adapted to new scenarios, such as primary patient-derived cell lines with minimal experimental data, aiding in a more efficient translation from in vitro to in-vivo experiments and, ultimately, patient implementation.

Back to programme

Back to Poster presentations

Mutational Signature Analysis of MACROD2 and PRKN deletions in Colorectal Cancer

Viktorija Vodilovska

Utrecht University

Structural variants (SVs) are large genomic mutations that are highly prevalent in colorectal cancer. A recent study observed that deletions or defects of two genes in particular, MACROD2 and PRKN, had significant system-wide impact measured using the CIBRA score. This work aims to identify mutational signatures associated with deficient MACROD2 and PRKN, with the aim of exploring their role in genome instability and tumor progression. For this analysis we utilize three whole-genome sequencing datasets, covering primary and metastatic CRC, as well as a pan-cancer dataset for technique validation. To enrich the exploration of SV mutational signatures we explore three different approaches for mutational signature analysis, (1) using cataloged signatures from COSMIC, (2) extracting de-novo signatures using SigProfiler, as well as (3) our proposed deep learning approach for SV signature modeling. Additionally, we validate these three approaches on different downstream tasks such as cancer type prediction. Through comparative analysis of defined patient subgroups and using different analysis techniques for we provide insights into the functional impact of these alterations. Findings indicate consistent significance of DNA damage repair deficiencies, chromothripsis, and whole-genome duplication patterns across MACROD2 and PRKN-altered samples, supporting their relevance for SV-based mutational trajectories in colorectal cancer. An interesting finding from this study is the prevalence of inversion related SV-signatures which have largely been understudied in this context, offering a novel direction for research.

Back to programme

Back to Poster presentations

Source attribution of Shiga-toxin-producing Escherichia coli (STEC): a multinational study in Europe using virulence profiles and core-genome multi locus-sequence typing

Tristan Schadron

Rijksinstituut voor Volksgezondheid en Milieu (RIVM)

Shiga toxin producing Escherichia coli (STEC) is a pathogenic E. coli bacterium capable of producing a Shiga toxin (Stx) (Yano et al., 2023). The public health relevance of STEC is highlighted by its potential virulence, subsequent severe symptoms and its foodborne outbreak potential.
The main reservoirs of STEC are animals capable of maintaining STEC colonization in absence of continuous exposure to STEC from other sources, mainly are ruminants like cattle, sheep and goat. Additionally, a wide range of other animals can serve as spill-over hosts, like other livestock animals, birds and wildlife (Mughini-Gras et al., 2017). However, the relative importance of these reservoirs and spill-over hosts for human infections is not completely understood.
Here we use a large dataset (n=3418), encompassing whole genome sequences from STEC isolates from human patients and potential sources, from 11 European countries, to determine the relative importance of putative sources for human infections. The wide range of countries and sources sampled allow for an accurate and generalizable source attribution of STEC. For this various machine models were trained with the best performing used for further analysis.
With the metadata containing disease symptoms associated with the human isolates, we were able to relate disease severity to particular sources. Based on these results all disease severities had the highest propensities for being originated from cattle, with HUS standing out with the largest fraction. Wild animal were also predicted as a relevant sources for isolates with symptomatic disease. This may be informative for the (re-)design of One Health STEC surveillance activities.
Additionally, permutation importances were retrieved to assess contributions of genetic features to the model's performance. A few genetic features were important for the model performance, however the effects are small, which might mean there is information overlap, or combinations of genes are important for the model's performing.

Back to programme

Back to Poster presentations

Uncovering the Landscape of Chromosomal Instability in Pediatric Solid Tumors

Roula Farag

Princess Màxima Centrum

Chromosomal instability (CIN) is a hallmark of many cancers and is characterized by ongoing structural and numerical chromosomal alterations. Recently, 17 recurrent copy number signatures were defined capturing distinct CIN processes in adult cancers. However, the prevalence and characteristics of CIN in pediatric cancers remain largely unexplored, limiting our understanding of CIN etiology and the development of targeted therapies for pediatric tumors.

We applied a copy number signature framework to 312 high quality pediatric solid tumor samples to investigate the prevalence and etiology of CIN. CIN was detected in ~60% of the tumors. Out of the 17 adult-derived CIN signatures, nine were present in our pediatric cohort. The signatures with the highest activities have been associated with impaired homologous recombination (IHR), replication stress, and high-level amplifications. These results highlight a distinct CIN landscape enriched for these processes.

We next explored cohort-wide covariates, including TP53 mutation status and whole genome duplication (WGD). Tumors harbouring both TP53 mutations and WGD exhibited significantly higher replication stress-associated CIN ($p<0.04$) than those with either TP53 mutations or WGD only, suggesting increased tolerance to replication stress in these contexts. Whereas tumor type-specific analyses showed that neuroblastomas had the highest CIN signature activity associated with mitotic error and telomere shortening ($p<x10\text{-}11$) compared to other tumor types, likely driven by MYCN amplifications that lead to telomeric loss.

Overall, CIN is prevalent in pediatric solid tumors and is predominantly mediated by IHR deficiency, replication stress, and high-level amplifications. WGD and TP53 alterations enhance tolerance to CIN, while tumor types such as neuroblastoma display distinct etiologies likely involving MYCN-driven telomere shortening. By resolving the underlying mechanisms of CIN in pediatric tumors, this study provides the foundation for CIN-directed therapeutic strategies tailored to pediatric cancer subtypes.

Back to programme

Back to Poster presentations

Comparing Machine Learning Algorithms and Genome Representations for Source Attribution of Salmonella Typhimurium

Dr. Julian A. Paganini

Institute for Risk Assessment Sciences (IRAS), Utrecht University

Background
Salmonella Typhimurium is a major cause of foodborne illness, with infections arising from multiple animal reservoirs. Whole-genome sequencing and machine learning (ML) enable high-resolution source attribution, offering new insights into pathogen ecology. Here, we trained and compared ML classifiers using diverse genomic representations to predict isolate sources and identify genomic features driving attribution.

Methods
We compiled over 11,000 draft genomes of S. Typhimurium with source metadata from EnteroBase and AllTheBacteria. Data were split into training (80%) and test (20%) sets, stratified by population structure and source. To prevent data leakage, closely related isolates (<5 allele differences) were grouped within the same subset.
Genomes were represented as (1) one-hot encoded cgMLST profiles, (2) unitig presence/absence matrices derived from a colored compacted de Bruijn graph, or (3) a combination of both. Dimensionality reduction was performed using Multivariate Unbiased Variable Selection (MUVR). Reduced features were used to train Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB) classifiers. To address class imbalance, we evaluated SMOTE and random upsampling. Hyperparameters were optimized for balanced accuracy using 10-fold cross-validation, and model performance was assessed on the independent test set. Feature importance was evaluated using permutation importance (SVM) and SHAP values (RF, XGB). All workflows were implemented in a new Python package, maGeneLearn, enabling reproducible ML-based genomic analyses.

Results
RF models trained on combined representations (cgMLST + unitigs) achieved the best test-set performance (accuracy = 0.792–0.796). The upsampled RF model reached high F1-scores for poultry (0.90) and swine (0.82), while bovine isolates were more challenging to classify (F1 = 0.61) due to frequent confusion with swine (114/362).
Feature interpretation revealed biologically relevant markers: the top poultry-associated feature was a 36 bp unitig aligning to fucR, an activator of the fuc operon involved in L-fucose metabolism, while the most predictive swine-associated unitigs mapped to fliC, encoding flagellin.

Conclusions
Machine learning applied to genomic data can accurately attribute S. Typhimurium sources, particularly for poultry and swine. Lower accuracy for bovine isolates underscores challenges posed by genomically similar sources. Feature importance analysis highlights candidate genomic regions potentially linked to host adaptation or transmission routes.

Back to programme

Back to Poster presentations

Metagenomic deep learning to support risk assessment of antibiotic resistant genes

Dr. Kaixin Hu

Utrecht University

The increased adoption of genomic surveillance of antimicrobial resistance (AMR) has proven to be a valuable tool in rapidly identifying genetic markers of resistance, which, in general, are highly predictive of susceptibility to antibiotics. However, leveraging metagenomics data to understand the genomic drivers of resistance spread in the One Health spectrum remains underdeveloped. While specific antimicrobial resistance gene (ARG) contexts, particularly mobile genetic elements, have been explored, comprehensive utilization of full genomic contexts in metagenomic samples to characterize the risk of AMR spread is still needed.

We annotated 778 metagenome samples of both ARGs and other coding regions, identifying 440 ARGs with a total of 33,385 occurrences. A pseudo-pan-genome graph network was constructed to connect all samples, where each node represented a coding region. We learned aggregator functions of the network nodes through unsupervised learning using the Graph Neural Networks GraphSAGE framework. Based on these learned aggregator functions, we extracted an embedding for each subgraph centered on an ARG. Furthermore, we quantified the dissemination risk for each ARG across four aspects: the number of samples containing the ARG, its frequency across samples, the number of species associated with contigs containing the ARG, and the number of countries from which samples containing the ARG were collected. Finally, these embeddings were used to train regression-based machine learning models—multilayer perceptron and support vector machine—to predict the four-dimensional ARG dissemination risk score.

Our work develops an aggregative scoring scheme that provides a per-sample risk metric of AMR threat. Given a new metagenomics sample, our pipeline, equipped with trained machine learning models, annotates ARGs and subsequently generates a score reflecting the ARG dissemination risk.

Back to programme

Back to Poster presentations

Poster Number: 12

UBEC – Assisting and connecting in bioinformatics

Dr. Flip Mulder

UMCU BIOINFORMATICS EXPERTISE CORE (UBEC)

University Medical Center Utrecht | Center for Molecular Medicine

The UBEC (Utrecht Bio-informatics Expertise Core) aspires to supply a centralized repository for the most commonly used and standardized pipelines and tools. This in the form of hosting, maintaining, developing and also supplying guides tutorials on and user-friendly ways of using them. But also, where possible and relevant for a wider audience, assist in developing new solutions.

But the goals is broader than this, it is to also connect bio-informaticians which might all work on more solitary islands. To assist with issues related to storing data, aspiring to provide a standardized way and guide to follow.

All this with researchers in mind and with the goal to make the possibilities available in a user friendly and approachable way to a broader audience.

Back to programme

Back to Poster presentations

On the road to predict the human PDZ domains interactome by combining both structure based and machine learning methods

Dr. Victor Reys

Computation Structural Biology, Utrecht University

PDZ domains are involved in major cellular functions, such as the localization of cellular elements and the regulation of pathways. They do this by interacting with partner proteins PDZ binding motifs (PBM), which are C-terminus linear motifs. With 266 PDZ domains and up-to 5000 potential PBMs that can be found in the human proteome, the corresponding interactome could involve more than 1 million putative interactions. Not only endogenous proteins are interacting with PDZ domain, but also pathogens are hijacking this system to better navigate through the cell, making the understanding of the underlaying interacting network of great interest.

While experimental data requires specialized setup and tedious work, in silico predictions together with the training of machine-learning models can bridge the gap between available binding affinities and the complete description of the interactome involved with the human PDZome. In this work, we investigate various machine-learning methods for the prediction of the PDZ-PBM interaction specificities and binding affinities; from sequence-based Bayesian models, to more complex deep-learning architectures using protein language models embeddings together with convolutional attention network for a global training on the human PDZome. Finally, structure-based information, first requiring to model tens of thousands of complexes, used with graph convolutional neural network through the DeepRank2 framework, are bringing additional contact-based features and therefore describing the system in its most complete form.

The accurate prediction of the interactions involving the human PDZ domains will not only complement our understanding of the complexity to maintain homeostasis in the human cell but also allow the deciphering of pathways used by pathogens to infect the cell and exploit its function, and therefore envision potential new therapeutic solutions.

Back to programme

Back to Poster presentations