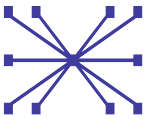
XIX Symposium
of Polish Bioinformatics Society

BOOK of ABSTRACTS

Poznań, September 16–18, 2026

Conference Sponsors:

FACULTY OF COMPUTING
AND TELECOMMUNICATIONS



INGENIX

Symposium Schedule

Wednesday, 16 September					
Hours	Session	Presentation Title	Speaker	Afiliation	Chairperson
11:00 - 12:00	Registration				
12:00 - 12:10	Opening		Maciej Antczak, Tomasz Żok	Poznan University of Technology	
12:10 - 13:00	**Keynote 1**	**How Ancient DNA Is Transforming Our Understanding of the Past**	**Maciej Chyleński**	**Adam Mickiewicz University**	Aleksandra Gruca
13:00 - 14:20	Algorithms for bioinformatics and computational biology	Benchmarking Framework for Taxonomic Classifiers	Patryk Gryz	Warsaw University of Technology	Michał Dąbrowski
		Species-Level Microbiome Dynamics Distinguish Bacterial from Non-Bacterial Sinonasal Infections: Toward Precision Diagnostics	Sylwia Bożek	Sano Centre for Computational Medicine	
		UNet based Approximation of Proton Radiation Dose Distribution in the Human Body for Proton Therapy	Adam Malinowski	Poznan University of Technology	
		Investigating the Impact of Sex-Specific Reference Genomes on RNA-Seq Read Mapping	Julia Rochatka	Poznan University of Technology	
14:20: 15:10	Lunch				
15:10 - 16:30	Modelling of protein and RNA structures	When RNA Sequence Reveals Hidden Geometry	Paulina Hladki	Poznan University of Technology	Tomasz Żok
		Large-scale analysis and clustering of RNA 3D structures	Jan Pielesiak	Poznan University of Technology	
		Analyzing Recurrent Local Protein Packing Motifs in Stable Domains and Dynamic Interfaces	Krzysztof Pysz	Wroclaw University of Science and Technology	
		MD Simulations Reveal How a Tunnel Entrance Mutation Disrupts Transport in an ABCG Transporter	Aleksandra Bigos	Adam Mickiewicz University	
16:30 - 16:50	Coffee break				
16:50 - 17:40	**Keynote 2**	**Integrative large-scale metagenomics for strain-level microbiome analysis**	**Edoardo Pasolli**	**University of Naples Federico II**	Aleksandra Świercz
17:40 - 18:30	Flash poster	NATSEQ: A Nextflow pipeline for identification of natural antisense transcripts from long-read RNA-seq data	Artemi Aleshkevich	Adam Mickiewicz University	Sylwia Bożek
		Statistics for non-canonical interactions	Jan Badura	Poznan University of Technology	
		Functional clustering of low-complexity regions across RNA-binding proteins	Abdeldjalil Bouziane	Institute of Biochemistry and Biophysics PAS	
		Identification of potentially amyloidogenic peptides in gut microbiota bacterial proteins using the updated AmyLoad database	Julia Cieszko	Wroclaw University of Science and Technology	
		G4Composer: A Web-Based Framework for Rapid Assembly of Complex G-Quadruplex Architectures	Dariusz Dziel	Poznan University of Technology	
		Benchmarking deep learning models for lncRNA subcellular localization prediction	Alicja Dzik	Poznan University of Technology	
		Iscores: proper scoring rules for evaluating missing-value imputation methods	Krystyna Grzesiak	University of Wroclaw	
		THE AMYLOGRAPH AMYLOID-ANTIBODY DATABASE: DATA ACQUISITION, CURATION, EXTRACTION FOR AMYLOID-ANTIBODY INTERACTIONS	Valentín Iglesias	Medical University of Bialystok	
		SCSM: Prior-Informed Genetic Demultiplexing for Multiplexed Single-Cell Sequencing of Clinical Cohorts	Katarzyna Jurkowska	AGH University of Krakow, Leiden University Medical Centre	
		Development of confomational ensemble embedding method for flexible target diffusion based generative models	Mikołaj Kikolski	Wroclaw University of Science and Technology	
		Somatic piRNA expression and their putative regulatory roles in two insect models	Krystian Komenda	Jagiellonian University	
		Investigating the Impact of Protonation-State on the Conformational Dynamics of the Glucose Facilitated Transporter (GLF)	Nitheeshkumar Kumar	Adam Mickiewicz University	
		Interface Usability Barriers for Older Adults and Their Implications for Health-Related Applications	Julia Anna Manikowska	Poznan University of Technology	
		LCR-Dataset: A Benchmark for Language Models to Analyze Functions of Low-Complexity Regions in Scientific Literature	Sylwia Morawiec	Silesian University of Technology	
		Computational analysis of sugar–Glf interactions to improve sugar transport efficiency	Abraham Muñiz-Chicharro	Adam Mickiewicz University	
		A Comparative Genomic Framework for Identifying Candidate Functional Sites in the European Bison Genome	Kamil Patan	Adam Mickiewicz University	
		FROM HEALTH TO ADENOMA TO COLORECTAL CANCER: EXPLORING GUT MICROBIOME SIGNATURES	Maja Piątek	Jagiellonian University	
		Automated and autonomous experimental setup (AAES) for long-term and multi-site behavioural experiments	Kacper Roszczak	Adam Mickiewicz University	
		Identification of double-stranded RNA molecules formed by 5'-overlapping protein-coding gene transcripts	Antoni Sawa	Adam Mickiewicz University	
		Democratising structure-based protein function prediction with metagenomic-deepFRI	Filip Schymik	AGH University of Krakow	
		LaRNAI: a web platform for RNA structure analysis	Maksim Serdakov	International Institute of Molecular Mechanisms and Machines PAS	
		TickPacket	Mariia Rutkowska	Medical University of Bialystok	
		Explainable Artificial Intelligence methods for DeepVariant	Piotr Suszyński	Warsaw University of Technology	
		Secretome: finding and characterizing secreted proteins in the human gut microbiome	Alicja Wojciechowska	Alicja Wojciechowska	
		Beyond Binary ORA: Preliminary Results of Unsupervised Clustering on LLM Embeddings for Omics Data	Joanna Żyła	Silesian University of Technology	
18:30 - 20:00	Poster session + welcome reception				

		Thursday, 17.09			
9:00 - 9:50	**Keynote 3**	**Ultrafast and accurate sequence alignment and clustering of viral genomes**	**Sebastian Deorowicz**	**Silesian University of Technology**	Bartosz Wilczyński
9:50 - 11:20	Modelling & proteomics	Glycine molecule radical: Predicted properties and dipeptide formation	Jarosław Synak	Poznan University of Technology	Agnieszka Rybarczyk
		Relative quantification of proteins and of post-translational modifications with shared peptides: a weight-based approach	Mateusz Staniak	University of Wroclaw	
		Interpretable protein function predictions with deepFRI2	Paweł Szczerbiak	Sano Centre for Computational Medicine	
		Enhanced Sampling of Free-Energy Surfaces in Biomolecular Systems: Conformational and Reactive Pathways	Juan Francisco Carrascoza Mayen	Poznan University of Technology	
11:10 - 11:30	Coffee break				
11:30 - 12:50	Transcriptomics & proteomics	RRBS-Based Epigenomic Profiling of Type 1 Diabetes: Preliminary Data Quality Assessment and Preprocessing	Joanna Tobiasz	Silesian University of Technology	Piotr Łukasiak
		Predicting Genomic Determinants of Chromatin Compaction Using Machine Learning	Teresa Szczepińska	Warsaw University of Technology	
		Reasoning Across Expression Profiles: A Modality Fusion Architecture for Omics Interpretation	Tomasz Jetka	Ingenix.ai	
		Evolutionary-scale interpretation of protein functions in the human gut microbiome	Jakub Wojciechowski	Sano Centre for Computational Medicine	
12:50 - 13:50	Lunch				
13:50 - 14:40	**Keynote 4**	**CASP and the Long Winding Road to Successful Computation of Protein Structure**	**John Moult**	**University of Maryland**	Krzysztof Fidelis
14:40 - 16:00	Applications of bioinformatics in biology and medicine	Gene-level rare variant scoring in muscular dystrophy: comparative evaluation of GORDIAN and GenRisk	Aleksandra Suwalska	Silesian University of Technology	Tomasz Kościółek
		iFlow - Clinical NGS and PGx Data Analysis Platform	Paweł Sztromwasser	Intelliseq	
		European AI Act and biomedical data: more paperwork or more governance?	Michał Burdukiewicz	Medical University of Bialystok, Vilnius University	
		Discussion about ELIXIR	Izabela Makałowska	Adam Mickiewicz University	
16:00 - 16:20	Coffee break				
16:20 - 19:00	General Assembly PTBI				
20:00	Conference dinner				
		Friday, 18.09			
9:00 - 9:50	**Keynote Honorary**	**Boruta and Beyond: Machine Learning and Information Theory for Marker Discovery**	**Witold Rudnicki**	**University of Bialystok**	Jacek Błażewicz
9:50 - 11:10	Systems biology (Cluster of Excellence in Sustainable Biotechnology)	Complexity Flow Graphs: measuring adaptive complexity in coevolving systems	Mateusz Twardawa	Poznan University of Technology, Poznan Supercomputing and Networking Center	Maciej Antczak
		A Systems Approach to the Study of Complex Biological Systems Based on the Analysis of Biofilm Formation Processes in Escherichia coli	Agnieszka Rybarczyk	Poznan University of Technology	
		Recognizing Ligands in CryoEM using Deep Learning	Dariusz Brzezinski	Poznan University of Technology	
		Deep Neurosymbolic Architectures for Interpretation of Biomedical Imaging	Krzysztof Krawiec	Poznan University of Technology	
11:10 - 11:30	Coffee break				
11:30 - 13:30	Laureates	**Best PhD Dissertation:** Computational methods for leading-edge mass spectrometry techniques	Piotr Radziński	University of Warsaw	Joanna Żyła
		Best MSc Dissertation: Large-scale analysis of 3D chromatin structure changes affecting enhancer-promoter spatial distances in different human tissues	Nikita Kozlov	Warsaw University of Technology	
		Best MSc Dissertation: Sequencing budget allocation for inferring gene regulatory networks from single-cell data	Aleksander Jankowski	University of Warsaw	
		Best BSc Dissertation: Design and Implementation of a Machine Learning-Based System for Cytology Slide Analysis	Aleksandra Walczybok	Wroclaw University of Science and Technology	
		Best BSc Dissertation: Automated Detection of Ki67-Positive and -Negative Cells in Immunohistochemically Stained Tissue Samples	Marcin Kuźniar	Wroclaw University of Science and Technology, CancerCenter.ai	
		HYperbolic Promoter Programming Engine: An iGEM 2026 Warsaw Project on Promoter Optimisation and de novo Design in Trichoderma atroviride	Michał Stanowski, Jakub Giezgała	University of Warsaw	
13:30	Closing				
13:30 - 14:20	Lunch				

Contents

KEYNOTE

How Ancient DNA Is Transforming Our Understanding of the Past

Maciej Chyleński[1]

[1] *Adam Mickiewicz University*

The study of ancient DNA (aDNA) has become a well-recognized field of research. Its results, while maintaining immense scientific value, often resonate with audiences beyond the scientific community. This research helps us understand human evolution, mobility, and aspects of the social structure and economy of past societies. By examining changes in population structures and the ranges of various species, it also provides insight into past environments and ecosystem responses to environmental change. In this presentation, I will provide a brief overview of the field, focusing on the specific nature of aDNA, its characteristics, and the resulting limitations of the genomic data that can be obtained. I will also briefly discuss the methodological toolkit that had to be developed to enable research in this area. Finally, I will demonstrate the wide range of research fields that benefit from aDNA studies by presenting several projects in which my team and I have been involved. I will show how aDNA data have helped us understand the social structure and mobility of prehistoric societies, using Anatolian Neolithic and European Bronze and Iron Age populations as examples. I will then briefly discuss our participation in projects tracking changes in Arctic and Antarctic ice sheets over the last centuries, as well as research on the demographic history and genomic erosion of the European bison.

Integrative large-scale metagenomics for strain-level microbiome analysis

Edoardo Pasolli[1]

[1] *University of Naples Federico II*

Strain-level metagenomics is reshaping the understanding of microbial diversity and dynamics across human hosts and food environments, with significant implications for health, safety, and ecosystem monitoring. This talk presents recent methodological advances and emerging resources that support robust, large-scale, strain-resolved microbiome analysis. Key developments include machine-learning approaches tailored to metagenomic data, alongside the growing integration of foundation models within microbiome research. The session will highlight harmonized reference databases, standardized metadata frameworks, and end-to-end pipelines that encompass both read-based strain profiling and metagenomic assembly with genome-resolved reconstruction. Particular emphasis is placed on initiatives aimed at detecting, annotating, and comparing under-studied microbial species and strains across studies. Examples linking human and food microbiomes will illustrate shared microbial reservoirs, opportunities for strain tracking, and diet-associated microbial flows. The talk concludes by outlining open challenges and priorities for establishing standardized, scalable, and generalizable frameworks within a One Health context.

Ultrafast and accurate sequence alignment and clustering of viral genomes

Sebastian Deorowicz[1]

[1] *Silesian University of Technology*

Viromics produces millions of viral genomes and fragments annually, overwhelming traditional sequence comparison methods. We introduce Vclust, an approach that determines average nucleotide identity by Lempel-Ziv parsing and clusters viral genomes with thresholds endorsed by authoritative viral genomics and taxonomy consortia. Vclust demonstrates superior accuracy and efficiency compared to existing tools, clustering millions of genomes in a few hours on a mid-range workstation.

The talk will describe the tool, which has already been published (A. Zieleziński et al., Nature Methods, 2025), and give a brief look at Vclut extensions we are currently working on.

CASP and the Long Winding Road to Successful Computation of Protein Structure

John Moult[1]

[1] *University of Maryland*

In 1972, Christian Anfinsen received a Nobel Prize for showing that, in principle, protein three-dimensional structure should be calculable from amino acid sequence. In 2024, three investigators received Nobel Prizes for their contributions to an amazingly effective way of doing so. In this talk, I'll try to describe what happened in the half-century between those events. It's a story of many different algorithms, many blind alleys, and the small set of key ideas and algorithms that led to ever more effective and useful methods. The emphasis in the talk will be on the role that community critical assessment, particularly in CASP, played in allowing us to establish what was promising and what was not. But I'll also highlight situations where we did not apply critical assessment principles and as a result, in at least two instances, wasted decades. The moral of that part of the story is that most of us, individually and collectively, fail to adequately critically assess our ideas and results, and we would have more productive and impactful careers if we paid more attention to this process.

Boruta and Beyond: Machine Learning and Information Theory for Marker Discovery

Witold Rudnicki[1]

[1] *University of Bialystok*

All-relevant feature selection is an important tool for marker discovery in high-dimensional complex data. In this lecture, I will revisit our work on all-relevant feature selection, starting from the machine-learning-based Boruta algorithm through the information-theoretic approaches developed within the MDFS framework to our take on the unbiased minimal-optimal approach.

In particular, I will concentrate on the problem of identifying the variables that are relevant due to synergistic interactions with other variables.

Finally, I will discuss what we have learned about identifying relevant variables and the opportunities and limitations of feature selection as a tool for marker discovery.

TALK

Benchmarking Framework for Taxonomic Classifiers

Patryk Filip Gryz[1], Robert Marek Nowak[1]

[1] *Institute of Computer Science, Warsaw University of Technology*

Taxonomic classification is a central step in metagenomic analysis, yet published tool comparisons rarely share identical inputs, ground truth, and execution conditions. In this work, we designed and implemented a unified framework for fair evaluation of read-based and profile-based classifiers on controlled DNA sequence datasets.

Each classifier runs through an identical wrapper contract (input adapter, model adapter, output adapter), standardizing data preparation, recording of parameters, and mapping of native outputs to a shared prediction schema (taxon, rank, abundance). An orchestrator executes the full dataset $\times$ classifier matrix; a metrics engine compares normalized predictions to the gold standard at selected taxonomic ranks, reporting precision, recall, micro-F1, macro-F1, and abundance error. We additionally introduce cross-metric analysis: for each anchor tool we define read-level error slices (FN, FP) and measure how often peer classifiers "rescue" those errors - quantifying complementarity without additional classifier invocations.

Computational cost (wall time, CPU cycles, peak memory) and scaling survey over a sequence-count ladder are recorded alongside accuracy; we report Δquality/Δcost relative to the fastest tool. We further built a meta-classifier on benchmark calibrated statistics: a base pair votes per read (disagreements resolved by higher accuracy), each read gets error probability p_i, and additional classifiers run only on the highest-risk tail, when expected accuracy gain justifies the extra cost.

Species-Level Microbiome Dynamics Distinguish Bacterial from Non-Bacterial Sinonasal Infections: Toward Precision Diagnostics

Sylwia Bożek[1], Aldona Olechowska-Jarząb[3], Małgorzata Foryś[3], Tomasz Kościółek[1], Joanna Szaleniec[1,2]

[1]*Sano Centre for Computational Medicine, Kraków, Poland*

[2]*Department of Otolaryngology, Faculty of Medicine, Jagiellonian University Medical College, Kraków, Poland*

[3]*Microbiology Department, University Hospital, Krakow, Poland*

The role of the sinonasal microbiome in chronic rhinosinusitis (CRS) and acute exacerbations of CRS (AECRS) remains poorly understood due to limited longitudinal data and low taxonomic resolution of previous studies. While bacterial ARS can be clinically distinguished, there is no clear framework for AECRS, leading to frequent antibiotic use. We aimed to characterize temporal microbiome dynamics and identify clinically relevant microbial patterns. In this study, we collected 293 sinonasal samples from 36 patients and 19 healthy volunteers. Full-length 16S rRNA gene sequencing ONT enabled species-level taxonomic resolution. We defined pathogen-associated microbial blooms as dominance of a single species during ARS and AECRS, classified as appearance blooms (emergence of a previously undetected species) or expansion blooms (increase in abundance relative to the previous sample). To assess their impact, we performed in silico perturbation by removing bloom-associated taxa and evaluated shifts relative to baseline states. The sinonasal microbiome showed high inter-individual variability with no CRS-specific signature. Longitudinal analysis revealed that bacterial infections were characterized by microbial blooms, whereas most viral or post-viral ARS episodes showed profiles similar to baseline. Importantly, expansion of potentially pathogenic species, was also detected during clinically silent periods, indicating that pathogen presence alone does not reflect active infection. Longitudinal microbiome profiling enables detection of dynamic microbiome changes associated with bacterial sinonasal infections and may help distinguish bacterial from non-bacterial CRS exacerbations, supporting more targeted antibiotic use and precision diagnostics.

UNet based Approximation of Proton Radiation Dose Distribution in the Human Body for Proton Therapy

Adam Malinowski[1], Wojciech Kotłowski[1]

[1] *Poznań University of Technology*

Proton therapy is a highly accurate radiation therapy that enables therapeutic doses to be delivered to target tissues whilst sparing the surrounding healthy structures. Accurate prediction of proton dose distributions is therefore essential for treatment planning, as even minor uncertainties may impact both tumour control and the risk of radiation-induced toxicity in normal organs. Currently, Monte Carlo simulations offer the most accurate dose calculation, but their high computational cost limits their regular use in time-limited clinical workflows.

In this work, we explore the application of a 3D U-Net neural network as a surrogate model to approximate the proton dose distributions in heterogeneous anatomical environments. The model is trained on datasets generated with Monte Carlo simulations with TOPAS, learning the relation between tissue composition, spatial geometry and deposited energies.

Preliminary results indicate that dose prediction can be performed within a few to a few tens of milliseconds per inference, while achieving a level of accuracy comparable to Monte Carlo simulations employing on the order of several thousand primary protons. Furthermore, training of the proposed network for datasets with spatial resolutions up to 64^3 voxels can be carried out efficiently on commercially available GPU hardware. Ultimately, faster dose estimation may contribute to more personalized radiation treatments, improving the balance between effective tumor irradiation and preservation of healthy biological structures.

Investigating the Impact of Sex-Specific Reference Genomes on RNA-Seq Read Mapping

Julia Rochatka[1], Aleksandra Świercz[1,2], Weronika Kraczkowska[3]

[1]*Institute of Computing Science, Poznan University of Technology*

[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences,*

[3]*Department of Biochemistry and Molecular Biology,Poznań University of Medical Sciences,*

Homologous regions of the X and Y chromosomes present a challenge in RNA-Seq data analysis, as their high sequence similarity can lead to ambiguous read mapping and inaccurate estimation of gene expression levels. This issue particularly affects genes located within pseudoautosomal regions (PARs) and other homologous regions of the sex chromosomes. The aim of this study is to evaluate the impact of masking selected sex chromosome regions on RNA-Seq read mapping quality and downstream gene expression analyses. RNA-Seq datasets from samples with known biological sex will be analyzed. Reads will be mapped to both a standard reference genome and sex-specific modified reference genomes. For female samples, the Y chromosome will be masked, whereas for male samples, the PAR1 and PAR2 regions of the Y chromosome will be masked. Read alignment will be performed using two widely used mapping algorithms, HISAT2 and STAR. The analysis pipeline includes quality control, read mapping, alignment quality assessment, gene expression quantification, and differential expression analysis. The proposed approach will be evaluated by comparing the number of uniquely mapped reads, gene expression estimates, and differential expression results obtained with standard and modified references. Additionally, the influence of the mapping algorithm on these outcomes will be assessed. The study will provide insight into whether sex-specific reference genomes improve the accuracy of transcriptomic analyses and facilitate more reliable interpretation of expression patterns for genes located on the sex chromosomes.

When RNA Sequence Reveals Hidden Geometry

Paulina Hładki[1], Marta Szachniuk[1,2]

[1]*Institute of Computing Science, Poznan University of Technology, Poznan, Poland*

[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

RNA molecules fold into complex structures that allow them to perform catalytic, regulatory, and structural functions. Computational prediction of RNA structure has traditionally focused on canonical Watson-Crick-Franklin base pairs, which form the main scaffold of secondary structure. Less frequent non-canonical interactions, although essential for tertiary contacts, local geometry, and conserved RNA motifs, remain difficult to predict from sequence alone.

In this work, we investigate whether non-canonical interaction patterns in recurrent RNA motifs can be inferred directly from primary structure. We use a sequence-only pairwise neural model that predicts nucleotide-nucleotide interaction classes, including canonical base pairs and non-canonical Leontis-Westhof classes. The model is evaluated with two encoder variants: a simple transformer trained from nucleotide identity and position, and a RiNALMo-Mega-based model using pretrained RNA language-model representations.

Our results show that sequence-based prediction of non-canonical interactions is highly motif-specific. Most tested motifs show weak or absent prediction, whereas the sarcin-ricin motif exhibits a strong and reproducible sequence-encoded signal. These findings suggest that some RNA motifs carry unusually readable sequence determinants of non-canonical geometry, making sequence-only models useful not only for prediction, but also for probing where RNA structure is locally encoded in sequence.

Large-scale analysis and clustering of RNA 3D structures

Tomasz Żok[1], Marta Szachniuk[1,2], Maciej Antczak[1,2], Jan Pielesiak[1], Wojciech Rączka[1]

[1]*Poznan University of Technology*

[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences*

Multiloop is a motif within an RNA structure in which three or more helices connect. While multiloops have a high impact on the RNA structural topology, they are rare and therefore not thoroughly researched. The main goal of this project is to create datasets dedicated to RNA structures with multiloop motifs in order to create machine learning algorithms for RNA secondary structure classification. With such algorithms, RNAs that are not yet solved experimentally will be utilised to create a database of possible multiloops. This talk will present the current state of the work: the development of a pipeline allowing fast and consistent processing of mmCIF (Macromolecular Crystallographic Information File) files through RNA analysis tools. To achieve this, we store RNA datasets obtained from EMBL-EBI, which are then processed, and the results are gathered in a single CSV file. This allows us to have reliable access to the results and detect potential discrepancies between them with different tools used. The next thing shown is the progress on a clustering tool that would allow us to look for similarities between observed motifs found in the RNA secondary structures. To do this, we are looking at the dot-bracket representations of the secondary structures and dividing them by the distance (differences) between them. In doing so, we aim to obtain information about redundancy, the exclusion of which is necessary to reliably train machine learning models. This work is supported by the National Science Centre, Poland (grant no. 2023/51/D/ST6/01207).

Analyzing Recurrent Local Protein Packing Motifs in Stable Domains and Dynamic Interfaces

Krzysztof Pysz[1], Krzysztof Fidelis[2], Witold Dyrka[1]

[1]*Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology*

[2]*Genome Center, University of California, Davis*

Proteins, as complex macromolecules, warrant analysis on multiple levels. Here, we represent proteins as sets of local structural descriptors that capture spatial packing arrangements around amino acids, including contacts between structural fragments that may be distant in sequence but close in three-dimensional space. By extracting these descriptors from identical or homologous proteins in different conformational states, we track how local packing arrangements are conserved or altered across conformations. The primary goal of our study is to determine whether descriptor-defined local packing solutions, which are already expected in stable domain interiors, can also be detected at experimentally observed moving interfaces between domains or protein partners. We achieve this by evaluating structural similarity scores between descriptors extracted from these regions. Identifying these recurring conservation patterns will allow us to build a database of protein annotations that captures conserved local packing motifs and their behavior across conformational states. Such a resource could support computational protein design by providing guidance for engineering proteins with specific, dynamic interface transitions.

MD Simulations Reveal How a Tunnel Entrance Mutation Disrupts Transport in an ABCG Transporter

Aleksandra Bigos[1], Wanda Biała-Leonhard[2], Konrad Pakuła[2], Bartłomiej Surpeta[3], Michał Jasiński 2[4], Jan Brezowský[1]

[1]*1 Laboratory of Biomolecular Interactions and Transport, Department of Gene Expression, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznan, ul. Uniwersytetu Poznańskiego 6, 61-614 Poznan*

[2]*2 Departement of Plant Molecular Physiology, Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Z. Noskowskiego 12/14, 61-704 Poznan*

[3]*3 International Institute of Molecular and Cell Biology in Warsaw, ul. Ksiecia Trojdena 4, 01-109 Warszawa*

[4]*4 Department of Biochemistry and Biotechnology, Poznan University of Life Sciences, ul. Dojazd 11, 60-632 Poznan, Poland*

ATP-binding cassette transporters of the G subfamily (ABCG) play essential roles in plant physiology by transporting metabolites across cellular membranes, contributing to homeostasis and environmental responses.[1] ABCG46 from Medicago truncatula mediates export of phenylpropanoids — specialized metabolites essential for pathogen resistance. Previously, we revealed the molecular basis of ABCG46 multispecificity and identified F562, located in the central tunnel region, as a residue critical for substrate transport.[2] In this follow-up study, we investigated how a mutation at the tunnel entrance leads to loss of transport function, motivated by experimental evidence of uncoupling of transport and ATPase activity in the closely related ABCG36 transporter from Arabidopsis thaliana carrying an equivalent substitution.[3] Using molecular dynamics (MD) simulations, we examined the structural and energetic consequences of mutating a residue at the site of initial substrate entry to the transmembrane region of ABCG46. Simulations were performed using an AlphaFold3 model embedded in a membrane environment, enabling detailed analysis of substrate access pathways. Comparative simulations of the wild-type and mutant transporter revealed mutation-induced changes in transmembrane helix conformations, tunnel geometry, and thermodynamically unfavorable substrate transport to the central cavity. These findings provide a mechanistic explanation for the experimentally confirmed loss of transport activity and highlight the critical role of the tunnel entrance in substrate gating and interdomain communication in ABCG transporters. 1. W. Biała-Leonhard et al., Plant Physiol. 198 (2025). 2. K. Pakuła et al., Cell. Mol. Life Sci. 80 (2023). 3. J. Xia et al., Nat. Commun. 16 (2024).

Glycine molecule radical: Predicted properties and dipeptide formation

Jaroslaw Synak[1,2], Jacek Blazewicz[1,2,3]

[1]*Poznan University of Technology, Institute of Computing Science, Piotrowo 2, 60-965 Poznan, Poland*

[2]*European Center for Bioinformatics and Genomics, Piotrowo 2, 60-965 Poznan, Poland*

[3]*Polish Academy of Sciences, Institute of Bioorganic Chemistry, Noskowskiego 12/14, 61-704 Poznan, Poland*

One of the vital direction of research into beginning of life is prebiotic chemistry. Reactions which eventually led to formation of crucial biological molecules are actively searched for, with new interesting pathways being discovered every year. When talking about the primordial soup, one of the main aspects of a desired sequence of chemical reactions is that it cannot require any sophisticated catalysts or conditions, since it is supposed to have taken place spontaneously on young Earth.

The authors analysed one of the possible pathways, while investigating the properties of radicals derived from glycine (the simplest aminoacid). One of the possible interactions leads to formation of a peptide bond with a relatively moderate energy threshold (<20 kcal/mol). While the results are based on theoretical and computational chemistry, the authors hope that their findings will open a new, interesting direction of research.

Enhanced Sampling of Free-Energy Surfaces in Biomolecular Systems: Conformational and Reactive Pathways

Francisco Carrascoza[1,2], Konrad Gorzelańczyk[1], Jacek Błażewicz[1,2,3]

[1]*1 Institute of Computing Sciences, Poznan University of Technology, ul. Piotrowo 2, 61-138 Poznan, Poland*

[2]*2 European Centre of Biology and Informatics (ECBiG), ul. Piotrowo 2, 61-138 Poznan, Poland*

[3]*3 Institute of Chemistry and Biology, Polish Academy of Sciences, Zygmunta Noskowskiego 12/14, 61-704 Poznań, Poland*

Free-energy surfaces (FES) unify the thermodynamics and kinetics of molecular processes, yet mapping them remains hard: the landscapes are rugged and the transitions of interest are rare on accessible timescales. This talk offers a methodological perspective on enhanced sampling for discovering conformational transitions and reactive pathways, built on several years of work with metadynamics, umbrella sampling, and replica exchange simulations across quantum and classical regimes. The approach is illustrated with ab initio molecular dynamics of prebiotic reactivity. The condensed-phase formation of glycine from carbon monoxide, and how to properly model water at ultra-low temperatures in the region 200 – 300 K. Finally, I discuss transferring these strategies to nucleic acids, where conformational complexity grows steeply with size. Schemes already established for RNA, where enhanced sampling methods, —define a natural testbed, and I sketch ongoing exploratory work toward integrating template-based 3D modeling with enhanced sampling for small RNA targets, presented as a forward-looking direction rather than a finished result. This work acknowledges support from the Poznań Supercomputing and Networking Centre (PSNC), and from grants 0311/SBAD/0781 and the Excellence Initiative – Internship Programs, both from Poznan University of Technology.

Keywords: free-energy surfaces; metadynamics; umbrella sampling; machine-learning collective variables; ab initio molecular dynamics; enhanced sampling; RNA conformational landscapes.

Relative quantification of proteins and of post-translational modifications with shared peptides: a weight-based approach

Mateusz Staniak[1], Ting Huang[2], Amanda Figueroa-Navedo[2], Devon Kohler[2], Meena Choi[3], Trent Hinkle[3], Tracy Kleinheinz[3], Robert Blake[3], Christopher Rose[3], Yingrong Xu[4], Pierre Beltran[4], Liang Xue[4], Olga Vitek[2], Małgorzata Bogdan[1]

[1] *University of Wrocław*

[2] *Northeastern University*

[3] *Genentech, Inc.*

[4] *Pfizer Inc.*

Bottom-up mass spectrometry-based proteomics studies changes in protein abundance and structure across conditions. Since the currency of these experiments are peptides, i.e. subsets of protein sequences that carry the quantitative information, conclusions at a different level, e.g., at the level of proteins or of post-translational modifications, must be computationally inferred. The inference is particularly challenging in situations where the peptides are shared by multiple proteins or post-translational modifications. While many approaches infer the underlying abundances from unique peptides, there is a need to distinguish the quantitative patterns when peptides are shared.

We introduce a statistical approach for estimating protein abundances, as well as site occupancies of post-translational modifications, based on quantitative information that includes shared peptides. This method extends the existing MSstats and MSstatsTMT frameworks for summarization and differential analysis of MS data. by treating the quantitative patterns of shared peptides as convex combinations of abundances of individual proteins or modification sites, and estimating the abundance of each source in a sample together with the weights of the combination. Relative abundances of proteins obtained this way take into account the information contained in shared peptides, and can be used in downstream statistical analysis. We demonstrate the utility of this model by using computer simulations and examples based on data from experiments with diverse biological objectives, including protein degradation and changes in protein post-translational modifications.

Interpretable protein function predictions with deepFRI2

Paweł Szczerbiak[1,2], Rund Tawfiq[1], Adam Nowak[1], Witold Wydmański[3,4], Jakub Wojciechowski[1], Jakub Adamczyk[2], Valentyn Bezshapkin[5], Lukasz Szydlowski[1,6], Tomasz Kosciolek[1,2]

[1]*Sano Centre for Computational Medicine, Kraków, 30-054, Poland*

[2]*Faculty of Computer Science, AGH University of Kraków, Kraków, 30-059, Poland*

[3]*Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, 30-387, Poland*

[4]*Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, 30-348, Poland*

[5]*Institute of Microbiology, ETH Zürich, 8093 Zürich, Switzerland*

[6]*International Research Agenda 3P – Medicine Laboratory, Medical University of Gdańsk, Gdańsk, 80-211, Poland*

Protein function prediction remains a key challenge in biology due to the rapid growth of sequence and structure data enabled by high-throughput technologies and recent advances in protein structure prediction. While deep learning methods and protein language models have improved predictive performance, effectively integrating sequence and structural information for large-scale function prediction remains an open problem. Here, we introduce deepFRI2, an upgraded framework of deepFRI (Deep Functional Residue Identification) for predicting protein function using Gene Ontology terms and Enzyme Commission numbers. Like its predecessor, deepFRI2 operates in two complementary modes: sequence-based and sequence+structure-based. Architecturally, the model consists of two main components: a structural prober, which leverages shallow convolutions over distograms for improved interpretability, and a sequence analyzer, powered by ESM with lightweight attention, to capture evolutionary and sequence-derived signals. This dual design enables robust functional inference in metagenomic settings, where protein structures are unavailable, as well as high-precision annotation when structural information is known. Moreover, its modular architecture allows disentangling structure- and sequence-driven signals while maintaining interpretability. Guided by principles of simplicity and robustness, deepFRI2 uses only a few million parameters yet achieves strong performance across all evaluated benchmarks, improving upon deepFRI and reaching state-of-the-art results. At the same time, it preserves a high level of interpretability and scalability — two often overlooked, but crucial features.

RRBS-Based Epigenomic Profiling of Type 1 Diabetes: Preliminary Data Quality Assessment and Preprocessing

Joanna Tobiasz[1], Walter Wolfsberger[2], Khrystyna Shchubelka[2,3], Taras Oleksyk[2,3], Joanna Polańska[1]

[1]*Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*

[2]*Department of Biological Sciences, Oakland University, Rochester, MI 48309, USA*

[3]*Department of Biology, Uzhhorod National University, Uzhhorod 88000, Ukraine*

Type 1 diabetes (T1D) is a chronic autoimmune disease that may arise due to genetic and environmental factors. DNA methylation is also investigated as a potential contributor to the development of T1D. In this study, we aim to explore the DNA methylation profiles of patients with T1D. The paired-end reduced-representation bisulfite sequencing (RRBS) of 43 samples was conducted to measure DNA methylation levels. The cohort consisted of both males and females aged 18-65 years, with the age at onset ranging from 3 to 49 years. We used FastQC and MultiQC to assess sequencing quality prior to and after trimming. Trim Galore served for the elimination of adapter contamination and for the removal of the cytosines introduced experimentally during the end-repair reaction. We examined the first three bases of each sequence to adjust Trim Galore parameters and select the appropriate RRBS mode. Among all reads, CGA was the most common sequence-start triplet, which indicated sequencing of complementary PCR-derived strands and using methylated cytosines for the end-repair reaction. The other most common first triplets were CGG and TGG, typical for the original DNA strands. After trimming, the median sequence length varied from 50 to 63 bp, while all the original reads were 151 bp long. Trimming decreased the GC content from 49-55Detailed inspection of read composition allowed the selection of trimming parameters and removal of technical artifacts to prepare the reads for alignment and methylation calling. Our results highlight the importance of RRBS-specific preprocessing approaches and will serve for further analysis of DNA methylation patterns in T1D. The work was supported by SUT's grant for Support and Development of Research Potential (02/070/BKM_26/0082).

Predicting Genomic Determinants of Chromatin Compaction Using Machine Learning

Ryan Burke[1], Ewelina Holm-Bidstrup[1], Shuting Liu[2], Ilaria Lupi[1], Monika Staniszewska[1], Andrew Belmont[2], Teresa Szczepińska[1]

[1] *1. CEZAMAT, Warsaw University of Technology, Warsaw, Poland*

[2] *2. University of Illinois at Urbana-Champaign, Urbana, United States*

The hierarchical 3D organization of chromatin in eukaryotic cells plays a crucial role in gene regulation and adapts during development, in response to stimuli, and in disease. While large-scale chromatin domains at scales larger than TADs (up to hundreds of nanometers) have been observed using microscopy, recent advances such as PCC-seq enable high-throughput measurement of chromatin compaction at kilobase resolution, revealing local variability, including those around transcription start sites (TSS). Here, we applied machine learning approaches, including Logistic Regression, Random Forest, and Histogram-based Gradient Boosting, to classify genomic regions according to chromatin compaction levels. Models were trained for regions of TSS (±1 kb, ±5 kb), genome-wide (1 kb, 100 kb), considering 2, 3, and 5 compaction classes. Predictions were based on 519 genomic features, including proteins and histone modifications ChIP-seq data, chromatin accessibility (ATAC-seq, DNA-seq), and transcription (BRU-seq, RNA-seq). As a baseline comparison, an analogous classification was performed for ATAC-seq signal in TSS ±1 kb regions. Using permutation feature importance and single-feature ROC AUC scores, we ranked genomic marks associated with chromatin compaction. In TSS ±1 kb regions, EP400, ELF4, POLR2H, H2AFZ, and ATAC-seq signals were associated with decompaction, while H3K27me3 and STAG1 correlated with compaction. Moreover, in TSS ±5 kb regions, H3K27ac and PCBP1 were linked to decompaction, whereas MCM3 correlated with compaction. At whole genome scales, in 100 kb bins THRAP3 and MBD1 were associated with decompaction, while ZBTB33 and H3K9me3 correlated with compaction. At 1 kb resolution, SETDB1, ZC3H4, and ZNF263 were associated with increased compaction.

Reasoning Across Expression Profiles: A Modality Fusion Architecture for Omics Interpretation

Maciej Sypetkowski[1], Joanna Krawczyk[1], Łukasz Smoliński[1], Remigiusz Kinas[1], Przemysław Pietrzak[1], Tomasz Jetka[1], Rafał Powalski[1]

[1]*Ingenix.ai, Warsaw, Poland*

Interpreting transcriptomic data requires both access to quantitative expression measurements and biological reasoning. Existing omics foundation models typically analyse profiles without generating explanations, whereas biomedical language models reason over text but lack direct access to expression values.

We present OmicsLM, a multimodal large language model that integrates single-cell and bulk transcriptomic profiles with natural-language instructions. Each profile is encoded as a compact continuous representation within the LLM context, preserving quantitative expression signal while allowing natural-language instructions, explicit gene mentions, and multiple interleaved biological samples to be processed together

OmicsLM was instruction-tuned on >5.5 million examples spanning >70 task types, including cell and tissue annotation, disease and clinical prediction, perturbation response, pathway reasoning and open-ended biological question answering. We also introduce GEO-OmicsQA, a publication-disjoint benchmark of 3,000 questions grounded in real GEO expression profiles and designed to evaluate multi-sample omics reasoning.

Across four evaluation settings, OmicsLM retained near-ceiling GTEx tissue-classification performance (99.4%) and achieved strong zero-shot annotation. On GEO-OmicsQA, it outperformed general-purpose and biology-specialized language models (binary QA accuracy 0.752; free-text score 0.623). On an unseen CRISPRi perturbation screen, it achieved Recall@50 of 0.633. OmicsLM provides a unified conversational interface for quantitative transcriptomic data and evidence-grounded biological interpretation.

Evolutionary-scale interpretation of protein functions in the human gut microbiome

Jakub W. Wojciechowski[1], Filip Schymik[1], Jacopo Pasqualini[2,3], Paweł Szczerbiak[1], Alicja W. Wojciechowska[1], Rund Tawfiq[1], Wei Wei[2,3], Torsten Schwede[2,3], Joana Pereira[4], Tomasz Kosciolek[1]

[1]*1 Sano Centre for Computaional Medicine*

[2]*2 .Biozentrum, University of Basel, Basel, Switzerland,*

[3]*3 SIB Swiss Institute of Bioinformatics, Basel, Switzerland*

[4]*4 VIB Center for AI and Computational Biology, Leuven, Belgium*

In recent years, the gut microbiome has emerged as one of the most important modulators of human health. Its composition has been linked to a number of conditions, including inflammatory bowel disease, cancer, as well as some neurodegenerative diseases. The vast majority of research in this area focuses on differences in microbial composition between patients and healthy individuals. While this approach has provided important insights into the role of the microbiome, it tends to generalize poorly between different populations due to the large variation in the structure of microbiomes across cohorts. To overcome this limitation and gain mechanistic insight into the role of these microorganisms in our health, we need to understand their functional potential. To address this problem, we are developing Gutlas, a unified catalogue of the sequences, structures, and functions of proteins produced by the human gut microbiome. Leveraging millions of high-quality structures from publicly available databases, as well as predicted de novo with AlphaFold and ESM, we assigned structures to more than 12 million proteins from the Unified Human Gastrointestinal Proteome. Using these structures, we were able to functionally annotate them with our newly developed tool, metagenomic-deepFRI, which enables de novo protein function prediction at the scale of metagenomes.

Gene-level rare variant scoring in muscular dystrophy: comparative evaluation of GORDIAN and GenRisk

Aleksandra Suwalska[1], Sara Pini[2], Karolina Kulis[1], Ana Töpf[3], Marco Savarese[4], Volker Straub[3], Filippo M. Santorelli[5], Jocelyn Laporte[6], Rossella Tupler[2], Joanna Polanska[1]

[1]*Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*

[2]*Department of Biological, Metabolic and Neural Science, University of Modena and Reggio Emilia, Modena, Italy*

[3]*John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle Hospitals NHS Foundation Trust and Newcastle University, Newcastle Upon Tyne, UK*

[4]*Folkhälsan Research Center, University of Helsinki, Helsinki, Finland*

[5]*Department of Molecular Medicine, IRCCS Fondazione Stella Maris, Pisa, Italy*

[6]*Department of Translational Medicine and Neurogenetics, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch-Graffenstaden, France*

Rare variant analysis in neuromuscular disorders (NMD) is complex because variants are often individually weak and not directly comparable between genes of different lengths. This study compares the proposed GORDIAN gene-level scoring method with GenRisk by assessing numerical differences and biological consistency of results. The analysis included rare variants from 620 patients with muscular dystrophy and 665 reference samples. GORDIAN scores were calculated using multiple variant-level predictors and annotations, with GenRisk used as a comparator. Score distributions were compared with the Mann-Whitney U test, then Gene Set Enrichment Analysis was conducted. GORDIAN and GenRisk were compared for score differences, patient-level correlations, gene rankings, and the biological coherence of enriched pathways. 46Overall, GORDIAN and GenRisk highlighted different gene and pathway signals, with GORDIAN providing a more biologically coherent gene ranking. The results support the use of GORDIAN as a numerical ranking tool before downstream pathway analysis and patient stratification in independent neuromuscular disease cohorts. The COMPASS-NMD project is financially supported by the European Union funding scheme HORIZON-RIA Research and Innovation Actions, programme HORIZON.2.1 Health, grant no 101080874.

iFlow - Clinical NGS and PGx Data Analysis Platform

Paweł Sztromwasser[1], Mateusz Dawidziuk[1], Mateusz Marynowski[1], Monika Krzyżanowska[1], Jakub Otręba[1], Maria Paleczny[1], Dżesika Hoinkis[1], Sławomir Gołda[1], Katarzyna Tomala[1], Klaudia Pacewicz[1], Klaudia Szklarczyk-Smolana[1], Michał Korostyński[1], Marcin Piechota[1]

[1] *1 Intelliseq SA, Kraków*

Background: NGS drives modern molecular diagnostics, yet translating genomic data into clinical reports introduces considerable infrastructure and regulatory overhead that burdens in-house bioinformatics teams. Rather than focusing on research innovation, teams are consumed by maintaining fragmented infrastructure stacks, container orchestration, and continuous pipeline updates to keep annotation sources current. These challenges are exacerbated by regulatory overhead—strict data security and access tracing, and constantly evolving compliance frameworks. Even with a high-performing team, these factors limit the scalability of lab operations. Solution: To address this, we developed iFlow—an integrated, cloud-based platform that lifts the burden of infrastructure and pipeline management. iFlow provides production-ready, pre-validated pipelines for germline, somatic, and PGx analyses, delivering native scalability and built-in compliance. Workflows, run via a web GUI or CLI/API, provide rich clinical annotation and automated reporting. The platform features VariantMiner—an interactive browser and reclassification module for custom report generation. Importantly, engineering teams can use the CLI/API to plug analyses directly into existing in-house systems and workflows. Conclusion: Supporting the infrastructure and regulatory requirements of a clinical diagnostics lab while catering to research can be overwhelming. iFlow empowers bioinformatics teams to reclaim their time and shift focus from routine stack maintenance to high-impact genomic research. This work was supported by PGx Plus project No. POIR.01.02.00-00-0089/18-00.

European AI Act and biomedical data: more paperwork or more governance?

Michał Burdukiewicz[1,2], Valentín Iglesias[1], Mariia Rutkowska[1]

[1]*1 Medical University of Białystok*

[2]*2 Vilnius University*

The European AI Act (EAA) introduces a new framework for regulating artificial intelligence, with direct implications for bioinformatics, especially studies made on biomedical data. At first glance, its requirements may appear to be another layer of seemingly pointless administrative burden.

Importantly, many of the activities that appear novel in the context of the EAA are already well established by methodological and reporting guidelines for machine learning models and clinical decision support systems. Frameworks such as TRIPOD+AI (10.1136/bmj-2023-078378), FAIR guidelines for software development (10.1038/s41597-022-01710-x) and similar ones (10.1038/s41592-026-03037-6) already require transparent reporting of prediction model development and validation. Therefore, the EAA should not be approached as yet another regulatory checklist, but as one component of a broader system of requirements. By systematically identifying the parallels between those components, we can produce outputs such as model cards, dataset descriptions or validation reports that simultaneously meet the requirements of multiple standards.

In this presentation, I will provide an in-depth comparative analysis of the EEA and selected methodological, reporting, and software development frameworks relevant to bioinformatics, especially in the context of biomedical data processing. Based on this comparison, I will propose a consolidated checklist to guide the development of biomedical AI descriptors in a way that is compatible with multiple requirements. Thanks to that, we could minimize the effort necessary to meet all these requirements while maximizing the governance over the newly produced models.

Complexity Flow Graphs: measuring adaptive complexity in coevolving systems

Mateusz Twardawa[1,2,3], Piotr Formanowicz[1,2]

[1] *1) Division of Algorithm Theory and Programming Systems, Institute of Computing Science, Poznan University of Technology*

[2] *2) Sustainable Biotechnology Cluster, Poznan University of Technology*

[3] *3) Cybersecurity and Data Protection Department, Poznan Supercomputing and Networking Center affiliated to the Institute of Bioorganic Chemistry, PAS*

Measuring adaptive complexity in coevolving biological systems remains an open problem. Classical information-theoretic and algorithmic complexity measures treat complexity as an intrinsic property of isolated systems, making them ill-suited for settings where system and environment co-evolve and mutually reshape one another. We introduce the Complexity Flow Graph (CFG), a relational framework that grounds complexity measurement in coevolutionary dynamics rather than system-internal properties.

In the CFG, complexity is defined relative to selective pressure: the complexity of a biological system is the proportion of contextually relevant information it has accumulated with respect to its environment. This framing draws a formal parallel to the underfitting and overfitting regimes in machine learning, and yields a testable upper bound — system complexity cannot exceed the complexity of the environment that shaped it.

The CFG represents coevolving entities as paired nodes, with directed weighted arcs encoding survival under foreign selective pressures and host-driven environmental transitions. Arc weights capture transition probability, and complexity change in both system and environment. The circularity inherent in mutually defined complexity measures is resolved by treating the host-pathogen interaction matrix as an eigenvector centrality problem.

We instantiate the framework in adaptive immune system evolution, where immune repertoires are modelled as sequence-based systems competing against dynamic pathogen antigenic landscapes. The CFG recovers known coevolutionary signatures including arms race trajectories and Red Queen cycling, and provides a complexity metric that distinguishes genuine adaptive accumulation from repertoire overfitting.

A Systems Approach to the Study of Complex Biological Systems Based on the Analysis of Biofilm Formation Processes in Escherichia coli

Agnieszka Rybarczyk[1,2], Wojciech Smułek[3], Ewa Kaczorek[3], Piotr Formanowicz[1]

[1] *1 Institute of Computing Science, Poznan University of Technology*

[2] *2 Institute of Bioorganic Chemistry, Polish Academy of Sciences*

[3] *3 Institute of Chemical Technology and Engineering, Poznan University of Technology*

Bacterial biofilms are structured communities of microorganisms embedded in a self-produced extracellular matrix composed of polysaccharides, proteins, and nucleic acids. They form at liquid–solid and air–liquid interfaces and are highly resistant to conventional eradication methods, posing major challenges in environmental, clinical, and bioindustrial settings. As complex biological systems, biofilms exhibit properties that arise not only from the behavior of individual cells but also from dense networks of interactions among their components. Thus, systems-level modeling is essential for understanding the mechanisms underlying biofilm development and persistence. In this study, biofilm formation in Escherichia coli was investigated using a Petri net-based model. Petri nets provide an intuitive and mathematically grounded framework for representing biological processes and analyzing their structural properties. The model was examined using t-invariant analysis, MCT set analysis, knockout analysis, and simulation experiments. These analyses provided biologically relevant insights into the mechanisms governing biofilm formation and maintenance in E. coli.

Recognizing Ligands in CryoEM using Deep Learning

Dariusz Brzezinski[1], Adam Mazur[1], Ashley Jurga[1], Jakub Radziejewski[1], Agata Bielska[1], Jacek Karolczak[1], Anna Przybyłowska[1], Witold Taisner[1], John Heumann[2], Michael Stowell[2]

[1]*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland*

[2]*Department of Molecular, Cellular and Developmental Biology, University of Colorado Boulder, Boulder, CO 80309, USA*

Accurate identification of small-molecule ligands is crucial for interpreting protein structures and supporting structure-based drug design. In cryo-electron microscopy (cryoEM), ligands are typically modeled manually from Coulomb potential maps—a process that is time-consuming and prone to cognitive bias, particularly at the lower resolutions. Although automated ligand identification tools exist, they were developed for X-ray crystallography and rely on feature engineering or iterative fitting rather than end-to-end learning. Here, we present a deep learning approach that treats density map fragments as 3D point clouds and predicts ligand identity using sparse 3D convolutions. Trained predominantly on X-ray data supplemented with a smaller set of cryoEM ligands, the model attains accuracy competitive with established crystallography methods and is, to our knowledge, the first ligand identification method applicable to cryoEM maps. To address the lack of standardization in cryoEM map processing, we introduce a custom map thresholding scheme that makes cryoEM blobs comparable to crystallographic ones. The method is distributed as a ChimeraX plugin available through the ChimeraX Toolshed: `https://cxtoolshed.rbvi.ucsf.edu/apps/chimeraxligandrecognizer`. We further describe ongoing work on using probabilistic point sampling and rotation-equivariant networks built with E3NN to improve predictive performance. Our work demonstrates the potential of machine learning to accelerate structural biology research, providing a valuable tool for the broader cryoEM community.

Deep Neurosymbolic Architectures for Interpretation of Biomedical Imaging

Krzysztof Krawiec[1]

[1] *Poznan University of Technology*

Modern deep learning is primarily defined by high-dimensional pattern recognition, yet it is fundamentally limited by a reliance on statistical correlations. These "black-box" models often fail to capture the causal and algorithmic structures inherent in complex domains, leading to extreme demands for training data, poor generalization, and a lack of adherence to physical constraints. This talk is about neurosymbolic synthesis, the integration of symbolic and programmatic priors into neural blueprints, and physics-informed manifolds, embedding of inductive biases to enforce physical consistency directly within the learning process. We demonstrate that these hybrid approaches can effectively overcome the optimization plateaus and data inefficiencies that plague traditional models in interpretation of biomedical imaging.

Computational methods for leading-edge mass spectrometry techniques

Piotr Radziński[1], Anna Gambin[1], Jakub Karasiński[1]

[1] *1 University of Warsaw*

The rewarded dissertation explores three topics related to data modeling and analysis in mass spectrometry, each corresponding to a different measurement technique. The first part concerns the determination of monoisotopic mass, whose signals are often not directly visible in the mass spectra of high molecular mass molecules due to a low signal-to-noise ratio. However, knowing this mass is essential for molecular identification. Two novel algorithms are presented to address this issue, enabling the determination of monoisotopic mass for proteins and oligonucleotides. Next, the dissertation introduces an image compression algorithm using neural networks for mass spectrometry imaging (MSI) data. This method significantly enhances the computational efficiency of data analysis on the compressed images. Finally, the last part of the dissertation focuses on atomic mass spectrometry. After a brief overview of selected measurement techniques using a multicollector inductively coupled plasma mass spectrometer, a new algorithm is proposed for estimating uncertainty in the sample-standard bracketing calibration method for exact isotope ratio measurements.

Large-scale analysis of 3D chromatin structure changes affecting enhancer-promoter spatial distances in different human tissues

Nikita Kozlov[1]

[1] *Warsaw University of Technology*

Background

The 3D structure of chromatin is one of the factors regulating gene expression, primarily through spatial relationships between promoters and enhancers (E-P). Experimental methods like ChIA-PET, Hi-C provide insights into chromatin architecture, but their high costs and resource requirements limit their scalability. This study leverages computational tools to model whole-chromosomal 3D structures and conduct large-scale analysis of chromatin variability across different human tissues.

Results

Across GM12878, H1ESC and HFFC6, genes with a predicted enhancer contact are observed to sit significantly closer to their closest enhancer. The data ingestion stage processed 66 whole-chromosomal and 8,000 regional model ensembles, producing a database of 30M E-P pairs, integrated with EnhancerAtlas, EAGLE functional links, ENCODE RNA-Seq and gene ontology databases. Integrating strand-specific ENCODE RNA-Seq with distance data shows a robust anti-correlation between $\log_2$ fold-change and distance change ($|r| \approx -0.57$, $p < 10^{-11}$ in all three contrasts): genes whose enhancers move two quantile tiers closer tend to be highly expressed, whereas genes whose enhancers drift apart are usually observed to have lower expression. GO enrichment showed that these gene sets are enriched for biologically coherent terms matching the lineage-defining programme of each cell line.

These findings further support the hypothesis that 3D chromatin architecture contributes to cell-type-specific gene regulation.

Sequencing budget allocation for inferring gene regulatory networks from single-cell data

Krzysztof Łukasz[1], Aleksander Jankowski[1]

[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

The inference of gene regulatory networks (GRNs) from single-cell sequencing data is a fundamental challenge in systems biology. Such networks provide insights into cell identity and differentiation mechanisms. However, the high dimensionality and sparsity of single-cell sequencing data, coupled with financial constraints, necessitate optimized experimental designs. This thesis explores how different resource allocation strategies might influence the quality of inferred networks. Specifically, we examined the trade-offs between sequencing depth, cell number, and the inclusion of chromatin accessibility data.

Using the BEELINE benchmarking framework and functional reference networks derived from the STRING database, we evaluated the performance of multiple inference algorithms across diverse human single-cell datasets. We employed a simulation approach with a fixed total read count, systematically downsampling cells and reads to model realistic resource constraints.

Our results demonstrate that inference accuracy is significantly more sensitive to sequencing depth than to sample size. In low-budget scenarios, the strategy of sequencing fewer cells at high coverage consistently outperformed other approaches. Additionally, the performance of the evaluated methods proved sensitive to data preprocessing, with aggregation techniques yielding mixed results depending on the algorithmic approach. Furthermore, contrary to our expectations, integration of multi-omic data (scATAC-seq) did not yield global performance improvements over expression-only methods. These observations suggest that prioritizing the quality of expression profiles over increasing cell throughput or additional data modality is the most beneficial strategy when designing experiments for GRN inference.

Design and Implementation of a Machine Learning-Based System for Cytology Slide Analysis

Aleksandra Walczybok[1]

[1] *1. Department of Artificial Intelligence, Wrocław University of Science and Technology*

Cytological analysis plays a key role in the early detection of cancers such as cervical cancer. Despite considerable technological progress, this field still does not fully exploit its potential, and the current system for assessing cytological smears faces numerous challenges. This project introduces a diagnostic support tool designed to reduce the subjectivity of assessments, increase the accuracy of analyzes, and accelerate the work of cytologists. The proposed solution enables cancer classification at two levels: individual cells extracted from the slide image and whole microscopic slides. Classification experiments were conducted using both deep neural network models and classical machine learning algorithms, as well as their fusion. The system has been implemented as a dedicated application that ensures data security and archiving. The application allows for reading morphometric features of cells, their classification, and explaining model decisions using explainable artificial intelligence methods, which visualize regions important for the prediction and the contribution of individual features. Thanks to the level of detail and interpretability of the predictions, the project has real potential for use in everyday diagnostic practice. Additionally, the implemented data archiving facilitates the consultation of difficult cases and supports the creation of a diagnostic repository.

Automated Detection of Ki67-Positive and -Negative Cells in Immunohistochemically Stained Tissue Samples

Marcin Kuźniar[1,2,3], Mateusz Maniewski[4,5], Izabela Wasiak[6], Jarosław Kwiecień[2], Piotr Krajewski[2], Piotr Giedziun[2,7], Witold Dyrka[1]

[1]*Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, Poland*

[2]*Technology and R&D Team, Cancer Center Sp. z o. o., Poland*

[3]*Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Poland*

[4]*Department of Tumor Pathology and Pathomorphology, Oncology Centre Prof. Franciszek Łukaszczyk Memorial Hospital, Poland*

[5]*Doctoral School of Medical and Health Sciences, Nicolaus Copernicus University in Toruń, Poland*

[6]*Department of Pathology, 10th Military Research Hospital in Bydgoszcz, Poland*

[7]*Department of Artificial Intelligence, Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Poland*

The manual assessment of the Ki-67 proliferation index in immunohistochemically (IHC) stained tissue samples is a crucial prognostic factor in breast cancer diagnosis. However, this conventional approach is time-consuming, highly subjective, and prone to inter-observer variability. This work presents the design, development, and evaluation of an automated, end-to-end deep learning pipeline for the precise evaluation of the Ki-67 index. To address the severe scarcity of expert-annotated, pixel-level histopathological data, a Teacher-Student semi-supervised learning framework leveraging pseudo-labeling was implemented. The proposed pipeline comprises two main stages: tissue segmentation and nuclei segmentation. In the first stage, an ensemble of two U-Net models combining a high-sensitivity "Specialist" and a robust "Generalist" isolates cancerous regions from background and non-cancerous tissues, effectively defining the region of interest (ROI). In the second stage, a dedicated nuclei segmentation model localizes and classifies individual Ki-67 positive and negative nuclei within these predefined ROIs. By integrating these components, the unified pipeline generates an accurate Ki-67 proliferation score while producing interpretable visual evidence of tissue masks and localized nuclei keypoints, enabling clinical verification. Experimental results demonstrate that the proposed data-centric methodology effectively mitigates confirmation bias in pseudo-labeling and achieves robust performance. The framework provides a transparent and generalizable solution that bridges the gap between deep learning automation and clinical applicability in digital pathology.

HYperbolic Promoter Programming Engine: An iGEM 2026 Warsaw Project on Promoter Optimisation and de novo Design in Trichoderma atroviride

Jakub Giezgała[1], Magdalena Osełka[2], Stanisław Gołębiewski[1], Aleksander Janowiak[1], Agnieszka Prudło[1], Michał Stanowski[1]

[1] *University of Warsaw*

[2] *Karolinska Institutet*

iGEM is the world's largest synthetic biology competition, in which student teams build genetic systems to solve real-world problems. Our project focuses on promoter engineering in Trichoderma atroviride, a fungus that is used in biocontrol and is becoming increasingly important in industrial biotechnology. Promoters determine how strongly a gene is expressed and under which conditions. Currently, achieving a defined strength is mostly trial and error because the mapping from sequence to expression is not obvious and does not follow a simple rule.

HYPPE is a framework, not a single predictor. Promoters are encoded into a hierarchical, discrete latent space. The levels are placed in hyperbolic space so that distance from the origin indicates how general the code is. Near the centre are a handful of codes describing overall promoter architecture. Further out are exponentially more of them, covering local detail. In practice, this means that an edit at one level reorganises the entire promoter, while an edit at another level shifts a single base. Strain and growth conditions are used for conditioning, and we encode the experimental descriptions as text. This means that a new condition can be specified without retraining. Reconstruction is self-supervised and can use any promoter, labelled or not. The expression head, however, sees pairs and determines which of the two is stronger. Labelled promoters are scarce in non-model fungi, and ranking is what promoter design actually requires.

We plan to test this on pks1, the polyketide synthase responsible for 6-pentyl-α-pyrone. Candidates originate from minimal edits to the native promoter, guided motif placement and sampling the latent directly. Each candidate is screened for off-target effects before synthesis, and our wet-lab team measures them over successive design-build-test rounds.

POSTER

NATSEQ: A Nextflow pipeline for identification of natural antisense transcripts from long-read RNA-seq data

Artemi Aleshkevich[1], Izabela Makałowska[1]

[1]*Institute of Human Biology and Evolution, Faculty of Biology, Adam Mickiewicz University, Poznań*

Antisense and overlapping transcripts comprise a large fraction of the human transcriptome and are involved in regulatory mechanisms such as transcriptional interference and RNA duplex formation. However, identification of overlapping transcript pairs remains challenging due to isoform diversity and limitations of gene-level overlap approaches that fail to resolve transcript-specific relationships.

Here we present NATSEQ, a Nextflow pipeline for transcript-level identification and classification of overlapping transcript pairs. NATSEQ integrates splice-aware alignment, quality control, and transcriptome reconstruction optimized for long-read technologies, including ONT direct RNA, cDNA, and PacBio HiFi datasets. A key component is a strand-aware overlap detection module that integrates multiple levels of evidence, enabling classification of antisense relationships, including head-to-head, tail-to-tail, and embedded configurations.

To enable systematic evaluation, we constructed a synthetic benchmark based on NanoSim models trained on SG-NEx error profiles. The benchmark includes positive overlapping transcript pairs, non-overlapping gene pairs, and hard negatives with proximal transcription start and end sites, enabling assessment under realistic sequencing noise.

Application of NATSEQ to SG-NEx datasets demonstrates its ability to reconstruct known antisense pairs and reveal distinct overlap landscapes across cellular contexts. Comparative analysis of H9 and MCF7 indicates tissue-specific patterns, with more unique overlapping pairs in the cancer-derived MCF7 line.

Overall, NATSEQ provides a reproducible framework for transcript-level discovery of overlapping antisense transcripts, enabling comparative analysis beyond gene-centric annotation frameworks.

Statistics for non-canonical interactions

Jan Badura[1], Agnieszka Rybarczyk[1,2], Tomasz Żok[1]

[1]1 *Institute of Computing Science, Poznan University of Technology*

[2]2 *Institute of Bioorganic Chemistry, Polish Academy of Science*

Non-canonical base pair interactions play an important role in RNA structure and function, yet their systematic characterization remains challenging. Here, we present a pipeline for analyzing RNA secondary structures enriched with information on non-canonical base pairs. The analysis is performed on high-quality reference sets of RNA structures derived from the Protein Data Bank. Using the RNAPolis workflow, outputs from multiple annotation tools, including BPNet, DSSR, FR3D, MAXIT, RNAView, and DNATCO, can be integrated into a unified JSON-based representation of RNA 2D structures. This representation enables the identification of structural motifs such as stems, hairpins, loops, and single-stranded regions. We then detect non-canonical interactions and classify them according to the structural elements connected by the interacting nucleotides, considering intraloop, interloop, loop–hairpin, loop–single-strand, and loop–stem interactions. The proposed approach supports statistical analyses of non-canonical interactions across the dataset, including their distribution across different loop topologies and interaction classes, i.e., the frequency, mean, and median number of non-canonical base pairs by loop type.

Functional clustering of low-complexity regions across RNA-binding proteins

Abdeldjalil Bouziane[1], Marcin Grynberg[1], Aleksandra Gruca[2]

[1] *1- Institute of Biochemistry and Biophysics PAS, Warsaw, Poland*

[2] *2- Silesian University of Technology, Gliwice, Poland*

Low-complexity regions (LCRs) are common in RNA-binding proteins and are thought to contribute to molecular recognition, regulation, and subcellular organization, but a unified functional view of their sequence space is still lacking. In this study, we propose a map of LCR clusters in RNA-binding proteins, including transcription factors, splicing regulators, and broader RNA-associated proteins. We grouped them into distinct clusters and tested whether these clusters are enriched for specific functional annotations, domains, or cellular roles. We hypothesize that LCRs in RNA-binding proteins are organized into relevant clusters that correspond to different modes of RNA interaction and regulatory function. This work helps to define putative functions of many previously unknown members of the RNA-binding protein space, and provide a reusable framework for broader LCR classification across the proteome.

Identification of potentially amyloidogenic peptides in gut microbiota bacterial proteins using the updated AmyLoad database

Julia Cieszko[1], Michał Burdukiewicz[2,3], Jarosław Chilimoniuk[2,3], Valentín Iglesias[2], Ronja Tittel[2,4], Małgorzata Kotulska[1]

[1] *1 Wroclaw University of Science and Technology*

[2] *2 Medical University of Białystok, Clinical Research Centre*

[3] *3 Institute of Biotechnology, Vilnius University*

[4] *4 Brandenburg University of Technology Cottbus-Senftenberg*

The gut-brain axis is increasingly linked to neurodegenerative diseases, with bacterial amyloids potentially contributing to protein aggregation via cross-seeding . As amyloidogenesis knowledge evolves, bioinformatic tools require updates for accurate identification of amyloidogenic regions. The AmyLoad database was updated, removing β-structure-based annotations and introducing a revised classification of aggregating, non-aggregating, and fibril-forming peptides with supporting experimental evidence. The current version of the AmyLoad database is publicly available at: https://burdukiewicz.com/ramyload/. Peptides with confirmed fibril-forming ability were mapped to bacterial taxa enriched in neurodegenerative microbiota. Proteome analysis identified bacterial proteins matching AmyLoad peptide sequences, including known amyloidogenic hotspots potentially involved in fibril formation and cross-seeding. Amyloidogenic fragments alone are insufficient to infer aggregation in vivo, as structural accessibility and extracellular exposure are also required. We plan to assess IDRs and β-solenoid folds in proteins with AmyLoad hotspots using IUPred2A and predict localization with BUSCA to identify secreted/extracellular candidates. Preliminary analyses identified bacterial proteins with amyloidogenic hotspots and resolved cross-β structures. Foldseek comparison already revealed a protein structurally similar to the CsgA complex, supporting related β-sheet architectures. These findings suggest that integrating updated bioinformatic databases with structural and functional analyses improves understanding of bacterial amyloids in neurodegeneration. This work is a first step toward their potential role in amyloid-associated diseases and may support future therapeutic development.

G4Composer: A Web-Based Framework for Rapid Assembly of Complex G-Quadruplex Architectures

Dariusz Dziel[1], Mariusz Popenda[2], Joanna Sarzyńska[2], Tomasz Żok[1], Marta Szachniuk[1,2]

[1]*Poznan University of Technology, Poznan, Poland*

[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

A guanine-rich strand can fold back on itself into a unimolecular G-quadruplex (G4)—a four-stranded assembly of stacked G-tetrads that regulates telomere maintenance and gene expression and serves as a programmable building block for aptamers, biosensors, and DNA nanotechnology. Yet its marked structural polymorphism means that sequence alone rarely fixes a single three-dimensional fold, so reliable modelling often calls for direct control over the target topology rather than automated prediction.

G4Composer is a web-based framework for building unimolecular G4 models. Given a nucleotide sequence, it derives candidate topologies from two complementary sources: experimentally determined G4 structures selected by similarity to the query, and hypothetical folds proposed from the natural distribution of topologies across loop lengths. Each candidate is built to full-atom resolution through distance-restrained minimization and energy optimization, and the alternatives are ranked by computed energy. The same engine underpins an interactive design mode in which the user sets or edits every structural parameter by hand: glycosidic bond conformations (*syn*/*anti*), per-tetrad strand polarities, tetrad path, sugar puckers, and helical twist and rise. Any model can be reopened, adjusted, and rebuilt, and each is validated for tetrad integrity; both canonical and non-canonical DNA and RNA folds are supported.

By integrating these steps into one web application, G4Composer offers a fast, accessible alternative to molecular-dynamics folding for G4 structure

Benchmarking deep learning models for lncRNA subcellular localization prediction

Alicja Dzik[1], Aleksandra Świercz[1], Barbara Uszczyńska-Ratajczak[2]

[1]*Institute of Computing Science, Poznan University of Technology*

[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

Long non-coding RNAs (lncRNAs) are a functionally diverse group of transcripts, associated with the largest, but simultaneously the least understood part of the human genome. A key determinant of lncRNA function is its subcellular localization: molecules present in the nucleus are commonly involved in transcriptional and epigenetic regulation, while cytoplasmic lncRNAs often contribute to mRNA stability, translation and cellular signaling. Numerous deep learning based approaches have beed developed in an effort to accurately determine lncRNA localization, without the need for expensive and labor-intensive experimental techniques. The state-of-the-art models employ a variety of sequence representation strategies and hybrid architectures, most frequently integrating convolutional neural networks with transformers or other attention modules. Despite substantial progress, existing approaches are limited by incorporating strictly sequence data, difficulties capturing complex biological relationships and often only moderate performance. Comparisons between models are further complicated due to the use of small, imbalanced datasets, differences in target localization classes and evaluation protocols. To create a common basis for method evaluation, a unified benchmarking dataset was constructed by integrating publicly available lncRNA localization resources and applying a consistent preprocessing and redundancy-filtering pipeline. The resulting benchmark was then used for evaluating several representative prediction tools, to asses their performance and provide insights into their strengths and limitations.

Iscores: proper scoring rules for evaluating missing-value imputation methods

Krystyna Grzesiak[1], Jeffrey Näf[2]

[1] *University of Wrocław, Medical University of Białystok, 1*

[2] *University of Geneva 2*

Missing values are ubiquitous in modern bioinformatics applications, including genomics, transcriptomics, metabolomics, and clinical datasets. The Iscores package provides an implementation of proper scoring rules for evaluating and ranking imputation procedures based solely on partially observed data. The package implements the Energy-I-Score and the Density-Ratio I-Score (DR-I-Score), two measures designed to assess the agreement between observed and imputed distributions without requiring access to the true missing values.

THE AMYLOGRAPH AMYLOID-ANTIBODY DATABASE: DATA ACQUISITION, CURATION, EXTRACTION FOR AMYLOID-ANTIBODY INTERACTIONS

Valentín Iglesias 1,Mariia Rutkowska 1,Jarosław Chilimoniuk 2,Ronja Tittel 3,Martyna Podlasiek 4,Julia Szkóp 5,Damian Wilary 1,Dawid Kubkowski 6,Michał Burdukiewicz[1]

[1] *Clinical Research Center, Medical University of Białystok, Białystok, Poland., Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania., Brandenburg University of Technology Cottbus-Senftenberg, Senftenberg, Germany., i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal., Center of New Technologies, University of Warsaw, Warsaw, Poland., Institute of Mathematics, Faculty of Mathematics and Computer Science, University of Wrocław, Wrocław, Poland*

Notwithstanding the vast economic and time resources devoted over the last three decades in actively investigating treatments for neurodegenerative amyloidosis, their success remains scarce. Immunotherapies have shown substantial promise in reducing the pathologic and clinical signs of amyloidosis, a trend crystallised with the approval of the first disease-modifying treatments for Alzheimer's Disease (AD). Additionally, mAbs are in clinical trials for various amyloidoses with no current disease-modifying therapies, including Parkinson's Disease (PD), Huntington's disease or Amyloid light chain (AL) amyloidosis. Instead, multiple mAbs have shown insufficient results or are only useful for certain disease stages, genotypes or phenotypes. Understanding the molecular mechanisms of why certain antibody therapeutics succeed and others do not could allow for increasing the success rate and decreasing economic burden for these incurable conditions. We have applied FAIR guidelines to develop two-step manual curation protocols, trained a machine-learning model for assistance in manuscript selection, introduced standardised guidelines for annotation and description of the data, and applied state-of-the-art large language model (LLM)-assisted curation to avoid data leakage, to integrate previous knowledge into the first amyloid-antibody dedicated resource. We anticipate that the AmyloGraph Antibody Database will significantly advance the understanding of the mechanisms that underlie successful amyloid-targeting antibody therapeutics, thereby accelerating the development of more effective treatments for currently incurable amyloid diseases.

SCSM: Prior-Informed Genetic Demultiplexing for Multiplexed Single-Cell Sequencing of Clinical Cohorts

Katarzyna Jurkowska[1,2], Elżbieta Wierciak[1,2], Kees Van Bergen[2], Marek Kisiel-Dorohinicki[1], Gertjan Lugthart[2], Szymon M. Kiełbasa[2]

[1] *1. AGH University of Krakow*

[2] *2. Leiden University Medical Center*

Multiplexed single-cell sequencing, where cells from several individuals are pooled into one library, is now common in clinical cohorts, but assigning each cell back to its donor remains difficult when per-individual cell numbers are low or genotype coverage is sparse. Existing genotype-based tools, including Vireo, rely almost entirely on SNP profiles and lose accuracy as the number of pooled individuals grows or coverage per cell drops. We present SCSM, a probabilistic demultiplexing model extending the Vireo framework. SCSM jointly models three genetic signals: SNP-based genotype profiles, donor-specific HLA allele expression, and sex-linked read counts. It also incorporates experimental pool design as a structured prior, rather than treating each pool as an unconstrained mixture. We expect this prior to help most in low-cells-per-donor regimes, where genotype signal alone becomes uninformative - a question we are currently testing on public benchmark datasets. We show the model's clinical utility by demultiplexing bone marrow CITE-seq samples from allogeneic transplant recipients (thalassemia patients) and their stem cell donors. Across ten sequencing pools, SCSM gave accurate, scalable demultiplexing despite variable donor composition and cell yield per pool. SCSM is implemented as a modular extension of Vireo with configurable priors (HLA, sex, design), and can be applied beyond transplantation to any multiplexed single-cell cohort with auxiliary genetic or design information available.

Development of confomational ensemble embedding method for flexible target diffusion based generative models

Mikołaj Kikolski[1], Sebastian Kraszewski[1]

[1]*Department of Biomedical Engineering, Wrocław University of Science and Technology*

SBDD has been transformed over the past five years by deep generative models that propose ligands directly in the 3D pocket of a target protein. Equivariant diffusion frameworks such as DiffBP, SurfDock and GCDM now routinely produce chemically valid, geometrically plausible ligands conditioned on a receptor structure. A conformational ensemble is the set of 3D structures a protein populates at physiological conditions, together with the relative weights of each state. Such ensembles are most commonly obtained from MD simulations, but can also be drawn from enhanced-sampling or imaging-derived methods. For targets including GPCRs, kinases, and viral proteases, the bound conformation often differs substantially from any single deposited structure. Designing ligands against snapshots risks producing molecules that fit a state the protein rarely visits, or missing pockets that open only transiently. Conditioning a generative model on the ensemble, rather than on one frame, should yield ligands whose predicted affinity and pose are robust across the conformations the target actually samples. The outcome is a pocket-centred, equivariant ensemble encoder that maps an MD trajectory of a target to a fixed-dimensional latent vector usable as the input for diffusion based SBDD models. The encoder can be trained on MD trajectories drawn from public repositories (GPCRmd, mdCATH) for three benchmark targets spanning the relevant modes of flexibility - the β_2-adrenergic receptor, CDK2 and the SARS-CoV-2 main protease. The current state of work requires further evaluation of strategy using existing models and benchmarks to provide quantitative results validating the selected approach. The analyses, even though already planned, have not been completed yet on the implemented encoder.

Somatic piRNA expression and their putative regulatory roles in two insect models

Krystian Komenda[1,2], Guillem Ylla Bou[2]

[1]1 *Doctoral School of Exact and Natural Sciences, Jagiellonian University*

[2]2 *Laboratory of Bioinformatics and Genome Biology, Jagiellonian University*

PIWI-interacting RNAs (piRNAs) are small non-coding RNAs best known for protecting the germline genome from transposable elements but their presence and potential functions in somatic tissues of arthropods remain barely understood. Here, we investigated somatic piRNA candidates and their putative mRNA regulatory interactions in two insect models, the cockroach Blattella germanica and the cricket Gryllus bimaculatus.

For each insect model, we performed paired smallRNA-seq and RNA-seq of brain, sperm, and carcass tissues, under control conditions and knockdowns of the two piRNA biogenesis pathways. With the small RNA-seq data we annotated pathway-specific piRNA candidates and studied they expression across tissues and conditions. With the RNA-seq data, we analyzed the expression profiles of the putative piRNA targets (including both, transposable elements and mRNAs), identifying potential piRNA–target interactions across tissues and species.

Both species exhibited robust piRNA expression in the two non-gonadal tissues studied (particularly brain and carcass), supporting abundant somatic piRNA expression. The study of piRNA-target reveled novel putative interactions, including potential piRNA-mRNA interactions with potential roles controlling biological processes.

In summary, these results provide a comparative catalog of somatic piRNAs in these two insect species and new i insights about the potential role of piRNAs to regulate biological functions in animals.

Investigating the Impact of Protonation-State on the Conformational Dynamics of the Glucose Facilitated Transporter (GLF)

Nitheeskumar Kumar[1], Abraham Muñiz-Chicharro[1], Jan Brezovsky[1]

[1]*1 Laboratory of Biomolecular Interactions and Transport, Department of Gene Expression, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznań*

The glucose facilitated diffusion (GLF) protein belongs to the major facilitator superfamily (MFS) of transporter proteins. GLF is involved in transporting sugar molecules across biological membranes. The efficient transport of sugars, such as glucose, fructose, and xylose, is particularly relevant to microbial metabolism and biotechnological applications involving biomass utilization and biofuel production. Protonation states of ionizable residues can strongly influence protein structure, dynamics, and function. While proton-coupled transporters within the MFS family, such as the GLF homologue XylE, utilize protonation-dependent conformational changes during transport. However, the role of protonation-state changes in facilitated transporters, which do not rely on proton gradients as an energy source. To investigate the role of protonation-state effects in GLF, molecular dynamics (MD) simulations have been performed using multiple fixed protonation-state models, including protonation of residues D27 and E196 individually and in combination. In parallel, constant-pH molecular dynamics (CpHMD) simulations have been performed to capture dynamic protonation behavior under physiologically relevant conditions. Comparative analyses of structural stability, conformational dynamics and helical behaviour of key gating regions are being carried out to evaluate the impact of protonation-state on transporter function. By integrating fixed-protonation and constant-pH simulation approaches, this work aims to assess how protonation-state affects the conformational landscape of GLF and to evaluate the importance of dynamic protonation effects in computational studies of membrane transport proteins.

Interface Usability Barriers for Older Adults and Their Implications for Health-Related Applications

Julia A. Manikowska[1], Julia Samp[1], Piotr Lukasiak[1]

[1] *1 Poznan University of Technology*

Healthcare is becoming increasingly digital. Patients are expected to use applications to communicate with doctors, schedule appointments, and manage medication. However, they may create barriers for older adults. They appear when interfaces are difficult to understand at first glance. Hidden actions, icon-only buttons, low contrast, and small clickable elements may make digital health services harder to use independently. Our study examined usability barriers in application prototypes that reflected common functions used in health-related services. Participants completed task-based scenarios. The evaluation included task difficulty, need for assistance, task completion, and user feedback. Additional colour variants were included to better understand contrast-related difficulties. After the experiment, participants indicated which interface elements were the most problematic. The results showed that only 48.3The study indicates that health-related applications for older adults should better support independent use. Interfaces should offer clearer navigation, visible and explicitly labelled actions, stronger colour contrast, and larger clickable elements. As healthcare systems are increasingly moving toward digital patient engagement, UX/UI design should take into account the real interaction needs of older users.

LCR-Dataset: A Benchmark for Language Models to Analyze Functions of Low-Complexity Regions in Scientific Literature

Sylwia Morawiec[1], Aleksandra Gruca[1]

[1]1 *Department of Computer Networks and Systems, Silesian University of Technology, Gliwice, 44-100 Poland*

Low complexity regions (LCRs) in proteins are fragments with limited amino acid composition that are important for protein functions. Finding literature on LCRs can be challenging due to the small number of papers, but machine learning methods, especially language models, may help researchers retrieve relevant articles. In this research, we present a manually curated collection of titles and abstracts from PubMed representing publications that describe seven functions associated with LCRs. This LCR-dataset serves two purposes: (1) it is a comprehensive repository of papers describing functions of LCRs and (2) it may serve as a benchmark for training and fine-tuning machine learning models for scientific literature retrieval. We divided LCR-dataset into training, testing, fine-tuning, and validation subsets. We used transformer-based language models such as BERT and BioBERT for token-level embeddings, and SPECTER for document-level analysis. To address the issue of high-dimensional token-level analysis we applied PCA to get a single representative vector for each paper. We provide baseline results for each task and evaluate fine-tuned transformer models using Random Forest classifier. Binary models effectively identify publications related to LCRs and their functions. For multiclass classification, we define classes based on the presence of LCR and their specific function, and compare pretrained models against baselines. Our dataset offers a range of classification tasks, making it a valuable resource for developing machine learning methods that can broaden and systematize our understanding of LCRs and their role in proteins. It can also serve more generally as a benchmark for evaluating the ability of language models to classify scientific papers by a specific topic.

Computational analysis of sugar–Glf interactions to improve sugar transport efficiency

Dr. Abraham Muñiz-Chicharro[1], Nitheeshkumar Kumar[1], Prof. Dr. Jan Brezovsky[1]

[1] *1. Laboratory of Biomolecular Interactions and Transport, Department of Gene Expression, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznań*

Glf is a membrane sugar transporter that facilitates sugar uptake through a conformational cycle. During this cycle, Glf transitions between outward-facing open, outward-facing occluded, fully occluded, inward-facing occluded, and inward-facing open states, allowing sugars to be recruited from the extracellular environment, translocated through the transporter, and released into the cytoplasm. Because sugar uptake is directly linked to carbon availability and metabolic activity, improving Glf transport efficiency could enhance bacterial growth and increase biomass production. In this work, we used molecular dynamics (MD) simulations to analyse the distribution of glucose, xylose, and fructose between water and the membrane, showing that sugars interact substantially with the membrane surface rather than remaining only in the aqueous phase. Brownian dynamics (BD) simulations were then used to evaluate sugar–Glf encounter events in different conformational states. Preliminary results indicate that sugars tend to occupy membrane regions with mixed electrostatic potential. Although this observation remains to be further validated, it suggests that the mixed electrostatic environment displayed at the entrance of the outward-facing open conformation might contribute to ligand recruitment and sugar–protein interactions. In contrast, in the inward-facing open and inward-facing occluded conformations, the entrance region is mainly negative, which may push sugars away from the transporter and favor their interaction with the membrane. These findings suggest that the conformationally dependent electrostatic landscape of Glf may play a key role in sugar recruitment and transport efficiency.

A Comparative Genomic Framework for Identifying Candidate Functional Sites in the European Bison Genome

Kamil Patan[1], Mateusz Konczal[1]

[1]*Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland*

The tribe Bovini (cattle, bison, yak and relatives) is attractive for comparative genomics: its species share a recent common ancestor yet span very different demographic histories. The European bison (*Bison bonasus*) is the focal case, having survived near-total extinction in the early 20th century and been rebuilt from only twelve founders, which makes it a compelling target for a genome-scale characterization of candidate changes at evolutionarily constrained loci. We build a progressive Cactus/HAL whole-genome alignment of eight bovid genomes, using the cattle bosTau9 assembly as the coordinate backbone, and project mammalian phyloP conservation scores from the Zoonomia 241-way alignment onto the European bison genome, restricted to high-confidence 1:1 orthologous regions. From these tracks we screen for candidate constrained-site disruptions: deeply conserved positions or elements that differ, are disrupted, or project poorly in the European bison. Candidates are classified by phylogenetic distribution using two contrasts: European versus American bison as the primary contrast, separating European bison-specific events from bison-wide ones, and domestic versus wild yak as a benchmark. We will test whether European bison-specific candidates are enriched in fast-evolving categories such as immune defence and chemosensation. Together, the alignment, projected scores and functional annotations form a reusable feature set for constraint-aware variant annotation. In future work we plan to train a variant-effect model following CADD-style scoring and benchmark it against existing human-trained and genomic foundation predictors, testing whether a lineage-specific model improves prioritization where human-trained predictors transfer poorly.

FROM HEALTH TO ADENOMA TO COLORECTAL CANCER: EXPLORING GUT MICROBIOME SIGNATURES

Sofiya Aniskevich[1], Maja Piątek[1], Michalina Kuzio[1], Hanna Marazevich[1], Ksenia Prybylskaya[1], Sylwia Bożek[2]

[1] *1 Student Scientific Society of Bioinformatics 'In Silico' Faculty of Biochemistry Biophysics and Biotechnology Jagiellonian University in Kraków ul. Gronostajowa 7 30-387 Kraków Poland*

[2] *2 Sano Centre for Computational Medicine Czarnowiejska 36 30-054 Kraków Poland*

Gut microbiome plays a key role in maintaining homeostasis and regulating proper intestinal function. The shifts in its composition may indicate a severe disease with a high mortality rate, such as colorectal cancer (CRC). The early diagnosis of the precancerous stage - adenoma, requires a colonoscopy, which is an invasive procedure. Understanding the changes that happen inside the gut microbiome will aid prevention and early detection by using more accessible non-invasive methods. Despite significant progress in colorectal cancer research, robust microbiome markers based on quantitative and qualitative analyses have not yet been firmly established. We performed a cross-study analysis on 2165 microbiome samples collected within 24 independent studies, in order to identify microbiome-derived markers that could play a fundamental role in colorectal cancer progression. Preliminary results did not reveal significant differences in the microbiome composition between the study groups. However, we observed that the microbiome functions are more stable than taxonomic composition, suggesting that functional profiles should be analysed further. Moreover, the effect of geographic location where one lives may have a greater influence on colorectal cancer development. Identifying gradual microbial changes may improve screening strategies and enable early intervention, helping reduce the number of colorectal cancer cases and deaths.

Automated and autonomous experimental setup (AAES) for long-term and multi-site behavioural experiments

Kacper Roszczak[1]

[1] *UAM Population Ecology Lab 1*

Understanding biotic interactions is one of the core challenges in modern ecology. The direct interference between animals has traditionally been studied through playback experiments, which involve broadcasting stimuli to quantify animal responses. However, conventional methodologies require human presence, which introduces observer bias and limits spatial and temporal scalability. To address these constraints, this project developed and field-validated an Automated and Autonomous Experimental Setup (AAES) for long-term, observer-free behavioural studies. The AAES uses a single-board computer with an integrated Neural Processing Unit (NPU). This robust hardware platform integrates a high-definition camera and a speaker within a durable, battery-powered housing supplemented by a solar panel for continuous field endurance. A custom software framework coordinates autonomous operations, combining scheduled playbacks with an energy-efficient wake logic: a Passive Infrared (PIR) motion sensor wakes the device from sleep mode upon detecting movement, after which the NPU instantly analyses the camera feed to verify avian presence and initiate recordings. The prototype underwent field validation over several weeks to evaluate its reliability in natural environments. Successful trials confirmed that the system can maintain strict schedules, manage power effectively through solar charging, and execute AI-driven triggers without human intervention. By providing a highly scalable, edge-AI capable tool, the AAES significantly minimises human disturbance, opening new horizons for automated ecological playback networks.

Identification of double-stranded RNA molecules formed by 5'-overlapping protein-coding gene transcripts

Antoni Sawa[1], Adam Szczyrba[1], Izabela Makałowska[1], Joanna Kozłowska-Masłoń[1]

[1]*1 Institute of Human Biology and Evolution, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland*

Opposite-strand overlapping protein-coding genes remain a poorly understood element of genome structure. In the case of 5'-overlapping genes, transcripts contain complementary sequences, which can form RNA-RNA duplexes. Despite long-standing awareness of the existence of such genes, the prevalence and biological importance of their interactions remain unclear. The aim of this study was to identify RNA-RNA duplexes formed by protein-coding gene transcripts and to determine whether the secondary structures of these transcripts can limit sense-antisense interactions. Around 900 pairs of identified 5'-overlapping genes were analyzed in the transcriptome of HeLa cells using the PARIS2 and LIGR-seq methods, allowing for in vivo RNA interaction detection. cDNA libraries prepared from these methods were sequenced using ONT technology, and sequencing data was analyzed using the DuplexDiscovereR package for R. This analysis should allow for determination of the frequency of duplex formation and identification of cases in which secondary RNA structures interfere with duplex formation, and its results will allow for a better understanding of the role of overlapping genes in gene expression regulation and human transcriptome organization.

Democratising structure-based protein function prediction with metagenomic-deepFRI

Valentyn Bezshapkin[1], Filip Schymik[2,3], Piotr Kucharski[4], Paweł Szczerbiak[2,3], Jakub W. Wojciechowski[2], Lukasz M. Szydlowski[2,5], Tomasz Kosciolek[2,3]

[1] *1. Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland*

[2] *2. Sano Centre for Computational Medicine, Krakow, Poland*

[3] *3. Faculty of Computer Science, AGH University of Krakow, Krakow, Poland*

[4] *4. Aiformatics, Krakow, Poland*

[5] *5. IRA 3P Medicine Lab; Center for Space Biomedicine and Radiological Protection, Gdansk Medical University, Gdansk, Poland*

The microbiome is the source of many times more genes than our own body cells and this "second genome" influences metabolism, immunity, and even cognition. Characterizing it beyond simple taxonomy faces three challenges: large dataset sizes, uncharacterised "dark" proteins, and difficulty in reasoning over gene-level annotations. A potential solution of the first two is development of bioinformatics tools that allow high-throughput inference without relying on direct orthology information.

One example is deepFRI, a GO term prediction tool that leverages both sequence and structure of a protein. The inclusion of structure improves both the confidence and specificity of its prediction. For large scale metagenomic experiments however, obtaining high quality protein structures for all query sequences may be computationally challenging.

Metagenomic-deepFRI (mdF) is an extension of the original pipeline that allows for automatic retrieval of homologous protein structures from the reference databases. It uses Foldcomp for lightweight storage, MMseqs2 for scalable template searches, and PyOpal for global alignment and contact-map construction. Retrieved templates reach a median TM-score >0.8 across all sequence identities, and reconstructed contact maps hold a median F1 0.9 even in the 20-30

Metagenomic-deepFRI is under ongoing development, aiming to transform it into a comprehensive pipeline for both bioinformaticians and biology oriented scientists looking for more ease of use.

LaRNAl: a web platform for RNA structure analysis

Maksim Serdakov[1], Davyd Bohdan[2], Grigory Nikolaev[2], Janusz Bujnicki[2], Eugene Baulin[1,2]

[1]*Laboratory of RNA Algorithms, IMol Polish Academy of Sciences, ul. M. Flisa 6, PL-02-247 Warsaw, Poland*

[2]*Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw, Poland*

RNA secondary structure is a key determinant of the overall RNA fold, and its understanding is vital for discerning RNA function. We recently introduced SQUARNA, a de novo prediction approach that revisits the concept of base pair maximization and develops it into a stem maximization idea coupled with the free energy minimization framework. SQUARNA can predict alternative structures and handle pseudoknots of arbitrary complexity. Benchmarking shows that SQUARNA outperforms existing methods, including deep learning models, in both single-sequence and alignment-based RNA secondary structure prediction.

Here we present LaRNAl (https://larnal.imol.institute), a web platform for RNA secondary structure analysis of individual RNA sequences powered by SQUARNA. In addition to the RNA sequence, users may optionally provide a multiple sequence alignment, structural restraints, residue reactivities from chemical probing, and a reference structure. Automated searches for structural restraints are enabled by default: Rfam family templates identified with Infernal (cmscan), G-quadruplex patterns filtered by the G4Hunter score, and the binding motifs of six RNA-binding proteins. The results page reports ranked alternative structures with their scores and interactive 2D diagrams powered by RFviewJS and RNAcanvas. The platform will soon host R3FOLD, our tool for RNA 3D structure modeling, along with other complementary tools, together forming a unified environment for RNA structure analysis.

TickPacket

Mariia Rutkowska[1], Joanna Oklińska[1], Damian Wilary[1], Anna Moniuszko-Malinowska[1], Michał Burdukiewicz[1,2]

[1]*1 Medical University of Białystok, Białystok, Poland*

[2]*2 Vilnius University, Vilnius, Lithuania*

Tick-borne diseases (TBDs), including tick-borne encephalitis (TBE), Lyme borreliosis, anaplasmosis, and rickettsiosis, are a growing public health and data harmonization challenge in Europe. Their rising incidence and geographic spread are linked to climate change and exurbanization, which may alter tick distribution and increase human exposure to tick habitats. Diagnostic assessment is challenging due to heterogeneous manifestations, co-infections, exposure factors, and long-term sequelae, while local formats and non-standardized terminology limit cross-site comparison.

To address this gap, we are developing TickPacket, a standardized, ontology-driven data model for TBDs based on the GA4GH Phenopacket Schema version 2.0. This work is conducted within the OneTick consortium, which provides multi-center cohorts covering TBE, Lyme borreliosis, anaplasmosis, rickettsiosis, and follow-up data.

As a first implementation case, we focused on TBE. We identified relevant Phenopacket elements, including Individual, Biosample, Phenotypic Features, Measurements, Disease, and Metadata. We curated TBE-specific phenotypic features using HPO, selected laboratory measurements using LOINC and SNOMED CT, and defined disease representation using MONDO and SNOMED CT. We then created test patient records representing TBE clinical scenarios and transformed them into Phenopacket-compatible JSON files. Validation with the Phenopacket validator confirmed schema compliance and highlighted TBD-specific elements requiring further modeling, including environmental exposure, co-infections, third-party evidence, TBE clinical forms, and long-term sequelae.

TickPacket aims to improve interoperability across clinical centers and support comparative analyses of TBD manifestations and outcomes.

Explainable Artificial Intelligence methods for DeepVariant

Piotr Suszyński[1]

[1] *Centre for Credible Artificial Intelligence, Warsaw University of Technology*

DeepVariant is a widely used variant caller based on a CNN. It uses read pileups as its input, utilizing the information about the neighborhood of a suspected variant location. As a black-box machine learning method it fails to deliver justification for the decision it makes. While for the majority of cases the model is correct some examples remain challenging, but most importantly some variants may be of special interest, for example variants identified in GWAS. To give insight into the reasons behind DeepVariant's decisions we applied Explainable AI (XAI) techniques, including LIME and Integrated Gradients, which allowed us to highlight input regions most important to the final decision. We also developed a multi-objective method based on evolutionary algorithm and using read-level mutation operators to generate biologically-plausible, non-adversarial counterfactual examples which minimize changes relative to the original while flipping the model decision. We conducted experiments using Genome in a Bottle data, producing counterfactuals for examples incorrectly classified by DeepVariant. We also checked a constrained setup, where our method was allowed to make modifications only to the variant neighborhood, without touching the variant location itself. Even with such constraint our method was still able to produce pileups for which DeepVariant decision flipped to another genotype, potentially exposing undesirable model behavior. We hope that the tools we produced will prove useful in understanding the reasons behind DeepVariant outputs, increasing trust in its calls but also helping to identify likely mistakes. In our continued research we aim to also identify cases where the model might consistently make mistakes or base its decision on spurious correlations.

Secretome: finding and characterizing secreted proteins in the human gut microbiome

Alicja Wojciechowska[1], Emilia Stacherczak[1], Filip Schymik[1], Maskymilian Gmur[1], Tomasz Kosciolek[1]

[1] *1 Sano Centre for Computational Medicine*

Secreted proteins in bacteria perform a variety of functions such as enzymatic activity, facilitating quorum sensing or extracellular matrix formation. They are the key players which allow bacteria to communicate with each other and modulate the environment. This makes secreted proteins a promising resource for increasing our knowledge on host-microbiome interactions. In this study, we benchmark different cellular localization prediction methods to detect secreted proteins in the human gut microbiome proteome. We compare four different approaches: 1) homology-based annotations obtained with EggNOG, 2) cellular localization predictions from a neural network DeepLocPro, 3) prediction of signalling peptides with SignalP (5.0 and 6.0) followed by predictions of transmembrane regions with DeepTMHMM and 4) cellular localization gene ontology terms from mDeepFRI. We additionally characterize each of the predicted sets of secreted proteins with respect to their sequence, structure and function. We find a low overlap between the identified datasets of secreted proteins with different methods, which implies different biological interpretation of the secretome features. Identified discrepancies point towards development of consistent methodology for secreted protein detection and data analysis pipelines that account for the prediction tool biases.

Beyond Binary ORA: Preliminary Results of Unsupervised Clustering on LLM Embeddings for Omics Data

Klaudia Skutnik[1], Dawid Zamojski[1,2], Joanna Zyla[1]

[1]*Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*

[2]*Genetic Laboratory, Gyncentrum Sp. z o.o., Sosnowiec, Poland*

Understanding molecular mechanisms in omics studies has traditionally relied pathway enrichment analysis like ORA algorithm. The development of large language models (LLMs) opens up new possibilities for analysis based on deep semantic similarity. The aim of this study was to investigate and compare the effectiveness of the traditional ORA method with a approach based on multidimensional embeddings. Vectorization was performed using the PubMedBERT medical language model for NCBI genes summaries and pathway descriptions from the KEGG database. An LLM Enrichment Score (ES) was defined as the sum from the cosine similarity matrix between DEGs and pathways, taking background genes into account as normalization factor. The resulting one-dimensional ES values were subjected to unsupervised clustering using k-means (KM), hierarchical clustering (HC), and Gaussian Mixture Models (GMM) approach. Compared to ORA, GMM and HC clustering of LLM embeddings achieved sensitivity (0.667 and 0.633, respectively) but low precision, as they incorporated many additional rejected The KM showed the poorest outcome due to large number of clusters (k=7), which increase granularity of results. Overall, LLM clustering redistributes ORA's rigid binary classifications into broader semantic spaces. LLM embeddings, particularly with GMM or HC, preserve core ORA signals while significantly expanding the biological context. Notably, in presented analysis of lung cancer data, these methods identified many more cancer-related pathways compared to traditional ORA. Although the embedding-based enrichment methodology requires further refinement, it is highly promising for further research and deeper omics data interpretation.

Modeling the JAK/STAT Signaling Pathway in Uveal Melanoma Metastasis Using Petri Nets

Zuzanna Kałużna[1]

[1] *Faculty of Computing and Telecommunications, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland*

Petri net modeling serves as a powerful mathematical framework for unraveling complex signaling cascades in cancer biology. By representing cellular mechanisms as formal graph topologies, these models allow researchers to analyze regulatory structures and signal transmission pathways that are challenging to evaluate in clinical settings. To elucidate the molecular drivers of eye-to-liver metastasis, this study developed a discrete Petri net model of the JAK/STAT signaling pathway involved in uveal melanoma progression. Grounded in curated biological databases and literature, the network captures detailed protein states, phosphorylation events, and transcriptional activations driving tumor cell survival and migration. Structural analysis in the Holmes software environment confirmed that the reconstructed Petri net is fully covered by t-invariants. This validation demonstrates the mathematical correctness and biological consistency of the network, proving that all modeled subprocesses can operate in closed, sustainable modes. Ultimately, this structural framework establishes a reliable foundation for future dynamic simulations, perturbation analyses, and multi-pathway integration to uncover potential therapeutic targets against metastatic uveal melanoma.

Adapting Deep Learning to Site-Specific IHC Variations with Limited Annotations

Marcin Kuźniar[1,2,3], Krzysztof Pysz[1], Mateusz Maniewski[4,5], Artur Bartczak[6,7], Izabela Wasiak[8], Jarosław Kwiecień[3], Piotr Giedziun[3,9], Piotr Krajewski[3], Witold Dyrka[1]

[1]*Department of Biomedical Engineering, Wrocław University of Science and Technology, Poland*

[2]*Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Poland*

[3]*Technology and R&D Team, Cancer Center Sp. z o. o., Poland*

[4]*Department of Tumor Pathology and Pathomorphology, Oncology Centre Prof. Franciszek Łukaszczyk Memorial Hospital, Poland*

[5]*Doctoral School of Medical and Health Sciences, Nicolaus Copernicus University in Toruń, Poland*

[6]*Department of Pathology, Klara Jelska Odrodzenie Specialist Hospital for Lung Diseases, Poland*

[7]*Department of Tumor Pathology, Maria Sklodowska-Curie National Research Institute of Oncology, Kraków Branch, Poland*

[8]*Department of Pathology, 10th Military Research Hospital in Bydgoszcz, Poland*

[9]*Department of Artificial Intelligence, Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Poland*

The availability of high-quality multi-scale annotations remains a limiting factor for deep learning. This challenge cannot be fully addressed with growing publicly available datasets, as models often require adaptation to site-specific variations. Here, we present two approaches to leverage limited annotations for improved semantic segmentation and prediction of diagnostic indices. We analyzed two small cohorts from Polish oncology centers. First, we targeted Ki-67 index prediction in breast cancer (53 WSI), prioritizing interpretability. For tissue segmentation, we combined sparse manual annotations with pseudo-masks from a teacher model to train UNets. For nuclei segmentation, we coupled generic masks with sparse expert key-points to train a UNet using an SE-ResNet50 encoder and custom loss. Second, we addressed TPS prediction in PD-L1 stained NSCLC (162 WSI). Using only slide-level TPS, we employed a Multiple Instance Learning approach fitting a Zero-Inflated Beta (ZIBeta) distribution. The proposed strategies substantially improved upon respective baselines. For breast cancer tissue segmentation, our model improved overall weighted IoU by 6.5% and reduced ECE by 12% compared to the teacher model. The final Ki-67 index predictor achieved a patch-level MAE of 0.04 while maintaining transparency via nuclei segmentation. In the lung cancer study, the combined MIL and ZIBeta approach outperformed conventional regression baselines, reaching an MAE of 0.14 (a 27% improvement). Furthermore, it offered an additional confidence metric via estimated distribution precision. Overall, our results demonstrate efficient strategies to tackle the limited availability of high-quality training data, facilitating model adaptation to specific laboratory staining and imaging conditions.

Allergies gone viral – a machine learning analysis of the phageome in allergies

Radoslaw Plawecki[1], Kamil Szyc[1], Krystyna Dąbrowska[1,2], Izabela Rybicka[2], Zuzanna Kaźmierczak[2], Aleksandra Wilczak[2], Sophia Baldysz[2]

[1]*1 Wroclaw University of Science and Technology, Wroclaw, Poland*

[2]*2 Hirszfeld Institute of Immunology and Experimental Therapy Polish Academy of Sciences, Wroclaw, Poland*

The prevalence of allergies has increased in recent decades, potentially influenced by factors such as diet and gut microbiota. Artificial intelligence (AI) offers powerful tools for analysing complex biological datasets. This study aimed to investigate gut phageome data from healthy individuals and allergy patients using AI methods. Putative viral genomes were identified from assembled contigs using geNomad (geN), VIBRANT (VIB), and VirSorter2 (VS2). Three complementary data modalities were constructed: taxonomy, host prediction, and functional annotation. Features were analysed using unsupervised, supervised, and ensemble learning. Random Forest and CatBoost models were evaluated using Leave-One-Out and Repeated Stratified 5-Fold Cross-Validation, while early- and late-fusion approaches were compared for multimodal integration. GeNomad performed comparably to VIBRANT, whereas VirSorter2 performed weakest. Random Forest and CatBoost achieved similar results. Multi-omic integration (geN: 0.72, VIB: 0.75, VS2: 0.69) consistently outperformed single modalities (geN: 0.66, VIB: 0.63, VS2: 0.58). Functional annotation provided the strongest predictive signal, followed by taxonomy, while host prediction performed least effectively. Feature importance suggested a potential association between altered short-chain fatty acid metabolism and allergic disease. Healthy individuals showed a higher proportion of proteins with unknown or other functions, potentially indicating greater functional diversity. Overall, the findings support a potential link between the gut phageome and allergic disease, warranting validation in larger and more balanced datasets.

From gene scoring to trio analysis: a new application framework for the GORDIAN score

Sara Pini[1], Aleksandra Suwalska[2], Andi Nuredini[1], Joanna Polańska[2], Rossella Tupler[1]

[1] *University of Modena and Reggio Emilia, Department of Biomedical, Metabolic and Neural Sciences, Modena, Italy*

[2] *Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*

Facioscapulohumeral muscular dystrophy (FSHD) is typically caused by transcriptional derepression of 4q35 locus. However, clinical variability and molecularly unsolved cases suggest complex polygenic drivers. GORDIAN (Gene scORing for inDividualized rIsk AssessmeNt) is a gene-level scoring framework for candidate gene prioritization in hereditary disease developed by the CoMPaSS-NMD project. Here, we present the application of the GORDIAN score to identify impaired molecular pathways in an unsolved FSHD trio (affected proband, healthy parents). GORDIAN scores were calculated across 12,796 genes per individual from standard GATK-processed WES data. Two computational strategies compared the proband to both parents: 1. Selection of genes with higher GORDIAN scores in the proband than both parents (n=2,764) yielded 17 primary STRING clusters ($\geq$ 9 genes; 276 total genes). DAVID annotation highlighted hubs in ECM organization, sarcomere organization, and microtubule structures. 2. CERNO analysis of ranked GORDIAN profiles isolated proband-specific driver pathways (higher AUC than both parents), revealing impaired cell-cell and cell-ECM adhesion, muscle constituents, Ca2+ binding, and microtubule dynamics. Significant overrepresentation (224 genes) was found (Fisher's exact test, $p < 0.001$) cross-referencing unique genes from proband-impaired pathways (n=2,842) against known neuromuscular disorder (NMD) genes. By capturing cumulative variant burdens, this pathway-centric framework provides a scalable paradigm to decode missing heritability and explain phenotypic variability across unsolved genetic conditions.

Probing the Embedding Space: A Multi-Model Semantic Analysis of KEGG Pathways

Karolina Widzisz[1], Klaudia Skutnik[2], Andrzej Polanski[1], Joanna Zyla[3]

[1]*Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland*

[2]*Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland*

[3]*Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*

Biological pathway databases such as KEGG provide essential descriptions of pathway functions and biological processes. However, gene sets and pathway definitions are frequently criticized for their inherent redundancy. Numerical embeddings offer a method to quantify semantic relationships and evaluate signal redundancy within them. In this study, embeddings for 271 KEGG pathways descriptions were generated using 8 models: SapBERT, PubMedBERT, BioLORD, BioGPT, BioBERT, MedCPT, MiniLM, and MPNet. Next, mean-centered was applied to reduce anisotropy (anisotropy ratio: 0.48–0.99), strongest in PubMedBERT and BioBERT and weakest in MPNet and MiniLM. Redundancy was calculated as pairwise cosine similarities for all 36,585 pathway pairs per model. Its distributions were compared using Friedman test, Kendall's W, Conover post-hoc with Benjamini-Hochberg correction, and matched-pairs rank-biserial correlations. Similarity within and between KEGG categories was also examined. Similarity distributions differed significantly between models ($p<0.001$), with BioBERT showing the highest variability (SD=0.16) and SapBERT the lowest (SD=0.11). The effect size was negligible (Kendall's W: 0.0006; rank-biserial correlations: -0.02 to 0.015), indicating no consistent tendency for any model to assign higher similarity to the same pathway pairs relative to others. For all models, pathways within the same KEGG category had higher mean similarity than pathways from different categories (within- vs between-category similarity of 0.098 vs -0.036 for BioBERT and 0.039 vs -0.018 for SapBERT). Overall, model choice affected how well similarities distinguished related from unrelated pathways but did not bias which pairs were rated as more or less similar.

Deep Generative Learning for De Novo Antifungal Peptide Design

Sosina Sewunet[1], Moshe Steyn[2], Witold Dryka[1]

[1] *1. Department of Biomedical Engineering, Wroclaw University of Science and Technology, Poland*

[2] *2. School of Molecular and Cellular Biology, University of Arizona, US*

Invasive mycoses remain clinically constrained by only three antifungal drug classes. Antimicrobial peptides (AMPs) offer a promising scaffold through selective disruption of ergosterol-rich fungal membranes. To develop a deep generative pipeline for computational design of novel antifungal peptides. 7,962 antifungal peptides (DBAASP, DRAMP, APD6, curated) were encoded via ESM2 (35M parameters) into 480-dimensional embeddings. A VAE (480→32→480; val. loss ≈90) generated 5,000 synthetic embeddings. WGAN-GP (500 epochs; $n_c = 5$; $\lambda = 10$; 12,962 embeddings) refined the latent space, reducing Wasserstein Distance by 85% (14.29→1.93) and Gradient Penalty by 94% (0.328→0.020). Candidates were retrieved via cosine similarity ($\bar{x} = 0.9753$), filtered by length (8–55 aa), charge (+1 to +10), and hydrophobicity (0.20–0.70), then scored by a 7-criterion antifungal framework (max 100 pts). Filtering yielded 2,926 candidates; 1,572 scored ≥80/100 and 196 achieved 100/100. All top-100 contained Tryptophan, zero cysteines, charge +2–+6, hydrophobicity 0.33–0.54 (mean 28.3 aa). Of 704 novel candidates (cosine < 0.95), 6 scored 100/100; RRLTLWLTLRR (11 aa; cosine = 0.862) ranked first. GAN_02911 (cosine = 0.000) scored 90/100, confirming exploration beyond the training corpus. Cecropin P4 (APD: AP01527; *Ascaris suum*) was independently retrieved, validating the pipeline. The pipeline identified high-confidence and genuinely novel antifungal candidates. RRLTLWLTLRR and GAN_02911 are priority leads for MIC validation against azole-resistant *A. fumigatus*.