Annual DDLS conference 2024

Poster abstracts

SciLifeLab

Knut and Alice Wallenberg Foundation

Contents

1. Tracing diatoms over space and time: a view from species distribution modeling
Baltic Sea; climate change; machine-learning algorithms; Ensemble; habitat suitability

Mohanad Abdelgadir & Anushree Sanyal. School of Natural Sciences, Technology and Environmental Studies, Södertörn University, Sweden

Diatoms are powerful bioindicators for anthropogenic impacts and environmental change in aquatic ecosystems. Yet how the geographical distribution of diatoms responds to ongoing and projected future environmental change across the Baltic Sea is not fully understood. We used a metadata-based modeling approach to predict the future spatial distribution and habitat suitability of selected diatoms across the Baltic Sea. Prediction was based on five different environmental variables: silicate, salinity, primary production of carbon, temperature, and four future scenarios for temperature and salinity in the years 2050 and 2100 using six machine-learning algorithms in species distribution modeling (SDM). Future predictions for the selected taxa of diatoms suggest a decreased distribution area in the Bothnian Sea and Bothnian Bay and an increased area in the Arkona basins. Predicted future spatial distribution and habitat suitability of selected taxa of diatoms in climate change scenarios are mainly driven by sea salinity and temperature in the year 2100.

---

2. A Machine Learning Approach for Novel Bacterial Exotoxin Discovery
machine learning; exotoxin; novelty; conditional random field

Albrecht, Florian, Umeå University, Sweden; Carroll, Laura, Umeå University, Sweden.

Exotoxins play important roles in bacterial pathogenesis. Using public data, I show that a machine learning model (i.e. a conditional random field; CRF) can be trained to recognize exotoxin-encoding genes in bacterial genomes. Leave-one-phylum-out cross-validation (CV) identified a best-performing CRF, which could detect exotoxin-encoding genes in unseen phyla (area-under-the-curve values of 0.443 and 0.882 for precision-recall [AUPR] and receiver operating characteristic [AUROC] curves, respectively). The optimal CRF was evaluated further via leave-one-contig-out CV, which assessed the CRF's ability to identify exotoxin-encoding genes in previously unseen contigs of previously seen species (AUPR = 0.9751, AUROC = 0.9940). Model introspection revealed PF08020, a protein domain of unknown function, as important for exotoxin discovery (CRF weight = 13.478), showcasing current knowledge gaps. These results highlight the utility of sequence segmentation models for exotoxin gene detection. Future research may utilize this model to discover novel exotoxins.

---

3. Interperetable models of black-box classifiers
Machine Learnig, interpretable classifiers, rule-based models

Maya Baghdy Sar, Girish Pulinkala, Mark Melzer, Jan Komorowski

The rapid rise of AI, Machine Learning (ML) in particular, offers a huge potential, but opaque models such as otherwise very successful Neural Networks or Support Vector Machines classifiers, often hinder trust and adoption due to their lack of interpretations. Compared to non-transparent neural networks, we create interpretable ML models, as opposed to explainable ones, using rule-based methods.

We designed a workflow for the creation of Rule-Based Mirrors (RBM) from Artificial Neural Network (ANN) classifiers. This classifier is called the mirror of the NN.

The support sets of the rules allow identification and interpretation of the TP, TN, FP, and FN decisions. We assessed the quality of predictions by performing several experiments on categorical, continuous and on mixed decision tables. Interestingly, the quality of the RBM classifiers was on par with the ANN classifiers for the discrete tables and only somewhat lower for the remaining cases (range over 85%-98%).

---

4. Quantifying cellular structures at the nanoscale with graph neural networks and super resolution microscopy

graph; machine learning; single molecule localisation microscopy; super resolution; actin

Bauer, Sebastian, Stockholm Unviersity, Sweden; Griffié, Juliette, Stockholm University, Sweden

The regulation of cellular processes depends on the dynamic spatial and temporal remodeling of key molecular components at the nanoscale, such as nanoclusters and the cytoskeleton. Actin filaments, in particular, are crucial for functions as cell migration and receptor dynamics, with even minor structural changes potentially contributing to various diseases. In this work, we present a graph neural network (GNN)-based method for quantifying the structural properties of the actin mesh using single-molecule localization microscopy (SMLM), which produces super-resolved spatial point patterns (SPPs) with a resolution of approximately 10 nm. By leveraging the shape-independent properties of graphs, our method facilitates seamless extension from 2D to 3D imaging and can be applied to other cellular structures. This approach offers deeper insights into actin mesh organization, with potential implications for understanding cytoskeletal disorders and cellular function.

---

5. SciLifeLab Precision Medicine Portal

Precision Medicine; Data Centre; Portal; Open science

Jan Lorenz1, Natashia Benzian Olsson2, Sebastian Lindbom Gunnari1, Maria Ahlsén1, Saman Rassam1, Cecilia Martinsson Björkdahl3, Janne Lehtiö3, Hanna Kultima2, & Johan Rung2 SciLifeLab Data Centre, Karolinska Institutet, Solna, Sweden SciLifeLab Data Centre, Uppsala University, Uppsala Sweden

The SciLifeLab Precision Medicine Portal, developed by the SciLifeLab Data Centre as part of the Data-Driven Life Science (DDLS) program, aims to support and accelerate data-driven life science research in Sweden. Accessible to all researchers, the portal promotes open science by helping research groups make their data FAIR and openly available. It will offer downloadable datasets through interactive data displays and provide comprehensive information on clinical quality registers, and disease-specific data sources. The portal also connects users to relevant events and training opportunities, fostering collaboration across the Swedish precision medicine community.

6. Rapid and scalable gene cluster family delineation with HTGCF
biosynthetic gene clusters; secondary metabolites

Larralde, Martin, EMBL, Germany; Blom, Josefin, Umeå University, Sweden; Gourlé, Hadrien, Umeå University, Sweden; Carroll, Laura M., Umeå University, Sweden; Zeller, Georg, EMBL, Germany

Biosynthetic gene clusters (BGCs), the genetic determinants of prokaryotic secondary metabolite synthesis, are a valuable resource for natural product-based drug discovery. Clustering BGCs into gene cluster families (GCFs) reduces dataset redundancy and allows for inferring the function of unknown BGCs. The available computational tools for this crucial step require users to sacrifice either scalability or accuracy, as none provide both. Here we introduce HTGCF (High-throughput creation of Gene Cluster Families), a novel tool which uses fragment mapping identification, nucleotide deduplication and protein representation steps to rapidly generate GCFs. Compared to the gold standard BGC clustering tools BiG-SCAPE and BiG-SLiCE, HTGCF clusters a manually curated dataset with comparable accuracy. Moreover, it outperforms in terms of memory requirements and computational time when clustering datasets of >11000 BGCs and could scale to a dataset of ~2.9 million BGCs to generate ~57000 GCFs. Thus, HTGCF represents a rapid and scalable method for GCF delineation.

---

7. Full-length multiplexed 16S sequencing in self-collected decade old dried cervical-vaginal fluid on paper cards: a pilot study
Self-sampling, Vaginal microbiota, 16S rRNA gene, Long read sequencing

Borgmästars Emmy, Department of Immunology, Genetics, and Pathology, Uppsala University; Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Hedlund Lindberg Julia, Department of Immunology, Genetics, and Pathology, Uppsala University; Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Stålberg Karin, Department of Women's and Children's Health, Uppsala University; SE-75185 Uppsala, Sweden; Sundfeldt Karin, Department of Obstetrics and Gynaecology, Institute of Clinical Sciences, Sahlgrenska Academy at Gothenburg University; SE-41685 Gothenburg, Sweden; Gyllensten Ulf, Department of Immunology, Genetics, and Pathology, Uppsala University; Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Enroth, Stefan, Department of Immunology, Genetics, and Pathology, Uppsala University; Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden

Vaginal microbiota is associated with various health conditions. We aimed to evaluate long read 16S rRNA gene sequencing for self-sampling paper cards stored for eleven years at room temperature. Dried cervico-vaginal fluids from 20 women were sequenced using the KinnexTM 16S multiplexing kit on a PacBio® Revio system. One negative and four positive controls with known bacterial compositions were also run. Short read 16S sequencing has previously been performed for 14 overlapping samples shortly after collection. The overall read quality was comparable for paper cards stored at room temperature for eleven years compared to eight years with all bacteria in the positive controls resolved at species level. Some bacteria resolved at Genus of Family level by short read sequencing data were resolved at species level by long read sequencing data. In conclusion, long read 16S sequencing was successful for paper cards stored at room temperature for over a decade.

8. The Human Disease Blood Atlas: profiling plasma proteomes across diverse disease cohorts
plasma proteomics, biomarker, precision medicine, pan-disease, machine learning

María Bueno Álvez (Science for Life Laboratory, Department of Protein Science, KTH-Royal Institute of Technology, Stockholm, Sweden), Fredrik Edfors (Science for Life Laboratory, Department of Protein Science, KTH-Royal Institute of Technology, Stockholm, Sweden), Mathias Uhlén (Science for Life Laboratory, Department of Protein Science, KTH-Royal Institute of Technology, Stockholm, Sweden)

We present the Human Disease Blood Atlas, a comprehensive open-access resource that explores the protein profiles in blood across 60 disease cohorts. By using a sensitive multiplex assay to analyze 1,463 proteins in plasma from 9,665 samples, we identified molecular signatures across a wide range of diseases, including cardiovascular, cancer, autoimmune, infectious, and pediatric disorders. Proteins associated with age, BMI, and sex differences were also highlighted. Notably, the study identified specific proteins elevated in cancers like acute myeloid leukemia and glioma, as well as cross-disease proteins, such as those involved in inflammation across cancer, autoimmune, and infectious diseases. These findings provide valuable insights for biomarker discovery, enabling deeper understanding of shared and unique molecular profiles across diverse pathological conditions, and supporting the advancement of diagnostic and personalized treatment strategies.

---

9. Agent Lens: LLM-Powered Autonomous Agent for Smart Microscopy
smart microscopy; AI agent; autonomous imaging; real-time feedback; biological research

Cheng, Songtao; Ouyang, Wei; Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Sweden.

Smart microscopy has the potential to revolutionize bioimaging by automating complex tasks, but current systems are limited by rigid, workflow-specific designs that prevent full autonomy. To overcome these limitations, we present Agent Lens, an AI agent system designed for smart microscopy. Leveraging large language models (LLMs) like GPT, Agent Lens offers a conversational interface that simplifies microscope control and acquisition workflows. The system uses GPT's function calling, code generation, and vision capabilities combined with a ReAct loop to enhance autonomous operations. Additionally, the system integrates real-time image analysis and dynamic feedback control, allowing for the adjustment of imaging parameters and error recovery. Future development includes incorporating advanced AI models like the Segment Anything Model (SAM) for precise cell targeting and few-shot learning, paving the way for high-throughput image acquisition and autonomous live-cell imaging.

---

10. Interpreting Microscopy Images with Machine Learning
microscopy, analysis, machine learning, explainable AI

Inês Cunha, Alix LeMarois, Erik Sahai, Juliette Griffié

Machine learning (ML) is becoming pivotal in life science research, offering powerful tools for interpreting complex biological data. In particular, explainable ML provides insights into

the reasoning behind model predictions, highlighting the data features that drove the model outcome. Our work focuses on building explainable ML models for microscopy images. These models not only classify cell fates but also reveal the underlying data patterns and features influencing these classifications. Specifically, we have developed models to classify individual lung cancer cell fates, such as proliferation and death, from live-cell microscopy data. By leveraging explainable ML techniques, we gained insights into the decision-making process of these models, revealing the key cellular features that determine whether a cell would proliferate or die. The combination of ML and specialised image acquisition will enable us to address specific biological questions and uncover novel insights about underlying cellular mechanisms. This work demonstrates the potential of explainable ML in enhancing our understanding of complex biological processes, and how we can gain novel knowledge from images.

---

11. Let's Treat Multi-Dataset Crystallography Data like a Brain

crystallography; fragments; drug discovery

DeBouver, Nicholas, Linköping University, Sweden; Strid Holmertz, Ylva, Linköping University, Sweden; Pearce, Nicholas, Linköping University, Sweden

Crystallographic Fragment Screening (CFS) identifies small molecules, or "fragments," that can be used in structure-based drug design. The PanDDA method employs a real-space approach to identify weakly binding molecules by analyzing electron densities statistically. However, the quality of PanDDA's analysis depends on data homogeneity, and heterogeneity is often observed even in crystals of the same protein grown under identical conditions. To overcome this, data sets are clustered to ensure homogeneity for analysis.

The new developments in PanDDA software improve real-space analysis by incorporating MRI image-alignment algorithms to reduce inter-dataset heterogeneity. These enhancements allow for better detection of local distortions in electron density, making it easier to distinguish significant local changes from large-scale effects. By improving the alignment of electron densities in CFS experiments, the sensitivity of PanDDA's analysis is significantly enhanced.

---

12. AIDA Data Hub Data Science Platform

sensitive data; long term primary storage; data science; precision health; artificial intelligence

Freyhult, Pontus, AIDA Data Hub, Linköping University, Sweden; Eren, Betul, AIDA Data Hub, Linköping University, Sweden; Balsever, Emre, AIDA Data Hub, Linköping University, Sweden; Ylipää, Erik, AIDA Data Hub, Linköping University, Sweden; Le, Minh-Ha, AIDA Data Hub, Linköping University, Sweden; Konda, Varshith, AIDA Data Hub, Linköping University, Sweden; Bivik Stadler, Caroline, AIDA Data Hub, Linköping University, Sweden; Lundström, Claes, AIDA Data Hub, Linköping University, Sweden; Hedlund, Joel, AIDA Data Hub, Linköping University, Sweden

The AIDA Data Hub Data Science Platform will provide a home for longitudinal research and clinical innovation in data-driven precision health, targeted to researchers, industry, and healthcare providers of national significance, co-located with national flagship compute systems. It will provide functionalities identified as critical in surveys to Swedish data-driven life science researchers. Key features include:

Data Science Platform: Computable long term primary storage, supporting advanced usage patterns including data collection, collaboration, visualization, enrichment, annotation, AI training, and federated analysis.

Close to flagship compute: Easy access from large scale national/European compute resources when extreme needs of computation arise.

Sensitive Data Services: Suitable for research on large amounts of data of extreme confidentiality including biomedical personal data.

Customizable security: The "closed by default and openable as needed" service model lets customers choose a security level appropriate for their processing without affecting other users.

---

13. OccuPy: Heterogeneity- and occupancy-aware cryo-EM analysis based on spatial filtering

cryo-EM; image analysis; structural biology; flexibility

Shah, Pranav, University of Oxford, UK; Burt, Alister, Genentech, USA; Forsberg, Bjoern, LiU, Sweden

Cryo-EM reconstruction averages images of individual particles to cancel noise and permit confident molecular modeling in 3D. Flexibility and heterogeneity in the averaged set of images however limit the fidelity of the reconstruction. We have developed a fast and simple spatial filter to estimate its local heterogeneity. In the absence of flexibility, this estimate approximates component occupancy. We demonstrate the utility of this method for cryo-EM map interpretation, particle-image signal subtraction, and to promote divergence of clusters during conventionally employed maximum-likelihood methods. This method affords both visualization the ability to focus on regions of variance in an unbiased way. This can be used to emulate homogeneous data, simplify hierarchical data clustering schemes, and perform accurate signal subtraction. Importantly, this quantifies heterogeneity from the start of processing, allows more reproducible processing procedures, and makes it easier for structural biologists to interpret their data reliably.

---

14. The Swedish Reference Genome Portal: A new service facilitating access and discovery of genome data generated in Sweden

planning; conceptualisation; web design; implementation; scientist

Brink, Daniel P., SciLifeLab Data Centre, Uppsala University, Sweden; Crean, Rory, SciLifeLab Data Centre, Uppsala University, Sweden; Fuentes□Pardo, Angela P., SciLifeLab Data Centre, Uppsala University, Sweden; Ågren, Quentin, SciLifeLab Data Centre, Swedish Museum of Natural History, Sweden; Kochari, Arnold, SciLifeLab Data Centre, Uppsala University, Sweden; Lantz, Henrik, National Bioinformatics Infrastructure Sweden (NBIS), Uppsala University, Sweden; Kultima, Hanna, SciLifeLab Data Centre, Uppsala University, Sweden; Rung, Johan, SciLifeLab Data Centre, Uppsala University, Sweden; Persson, Bengt, National Bioinformatics Infrastructure Sweden (NBIS), Uppsala University, Sweden

The Data Science Node in Evolution and Biodiversity (DSN-EB) is part of the SciLifeLab and Wallenberg National Program for Data-Driven Life Science. The node aims to create an

environment where users can easily find, access, and utilize resources such as data, tools, and e-infrastructure that are essential for Swedish data-driven research in evolution and biodiversity. The first service that we will offer is a national genome portal that showcases genomic data from non-model eukaryotic species studied in Sweden. Each species will have a dedicated webpage with taxonomic, biological, and genome assembly information, as well as links to external resources such as the Swedish Biodiversity Data Infrastructure (SBDI), the European Nucleotide Archive (ENA), among others. The genome sequence and associated genomic data will be displayed in an embedded JBrowse genome browser, facilitating the visual representation of a variety of data types, session sharing, and export of publication-ready figures. Another important purpose of the genome portal is to promote and facilitate FAIR (Findable, Accessible, Interoperable, and Reusable) sharing of valuable data files produced during genomics data analysis that rarely get published. Examples of these include assembled transcripts, and annotation of features such as methylation patterns, non-coding RNA, and repetitive elements, to name a few. The poster will introduce key features of the genome portal and illustrate what researchers need to consider when submitting genomic data to the portal. With this service, we aim to make genomics data more accessible for all users regardless of their prior level of bioinformatics knowledge. We want more additions to the portal! If you are interested, please contact us at dsn-eb@scilifelab.se.

---

15. National Omero Resource
DDLS, Cell and Molecular Biology Node, Omero, Bioimaging

The Data Science Node for Cellular and Molecular Biology is dedicated to supporting life science researchers by providing accessible, cutting-edge data management and analysis solutions. Our primary goal is to streamline the research process by offering tools and services that help researchers efficiently handle, analyze, and interpret complex biological data. We specialize in facilitating easy access to scalable data storage, secure data sharing, and custom analysis pipelines tailored to the specific needs of cellular and molecular biology research. Our services ensure that data is well-organized, reproducible, and readily available for in-depth analysis, empowering researchers to draw meaningful insights and accelerate discoveries.

---

16. Data Delivery System (DDS): Secure and Simple Data Transfer for SciLifeLab Users
data; delivery; secure; encryption

Georgiev, Valentin, Data Centre, Sweden; Revuelta, Alvaro, Data Centre, Sweden; Österbo, Ina Odén, Data Centre, Sweden

The Data Delivery System (DDS) is a tool built for the simple and secure delivery of data from SciLifeLab platforms to their users. The system uses Safespring's object storage service as the delivery medium, thereby keeping the data within the Swedish borders. The DDS has built-in encryption and key management, enabling protection of the data without requiring the users to perform these steps themselves prior to delivery.

17. Discovering novel bacteriocins produced by Streptococcus mutans via genome mining approaches

biosynthetic gene clusters; secondary metabolites; Streptococcus mutans; bacteriocins; oral microbiome

Grimmer, Marlene, Umeå university, Sweden

The human oral microbiome, shaped by various host and environmental factors, can significantly impact oral and systemic health. The commensal bacterium streptococcus mutans, a key contributor to dental caries through formation of biofilms and lowering of pH, is known to produce bacteriocins (mutacins). However, due in part to its small genome size, its biosynthetic potential remains underexplored. This project utilized two tools, antiSMASH and GECCO, to mine all publicly available S. mutans genomes for biosynthetic gene clusters (BGCs) to identify putative novel bacteriocins. Comparative analysis with experimentally validated BGCs revealed several novel BGCs, including a previously undescribed putative Blp family class II bacteriocin. These findings reveal the hidden biosynthetic potential of S. mutans, suggesting that this bacterium can produce a wider array of secondary metabolites than previously recognized, and emphasize the need for further investigation into its secondary metabolites and their antimicrobial properties.

---

18. Metabarcoding vs Metagenomics; How lab work decisions can impact the identifiation of fungal biodiversity from eDNA samples

Metagenomics; metabarcoding; Fungi

Guilera Recoder, Monica; Andermann, Tobias; Rosling Anna; Uppsala University

During the past decade, high-throughput sequencing of environmental DNA (eDNA) has made possible the study of fungi without relying on fruitbody formation, increasing in an unprecedented way our knowledge about fungal species diversity. Two methods for species identification using eDNA are metabarcoding and metagenomics. These two methods detect species using different strategies, metabarcoding consists on PCR enriching a target locus while metagenomics consists on sequencing the full genome. In our study we test the extend of PCR-biases by amplifying a set of 7 soil eDNA samples using two different primer combinations, and comparing the sequencing results to the PCR-free approach. The tested primer pair combinations are the fungal-specific ITS1ngs+LR5 as well as the eukaryotic general primer pair ITS9MUNngs+TW14ngs. We present the differences in the retrieved species communities between these 3 approaches and discuss the taxonomic biases introduced by the two different PCR primer combinations.

---

19. Mapping the vascular walls of colorectal cancer

Spatial analysis, tumor microenviroment, angiogenesis

Linglong Huang1 Co-authors: Linglong Huang1, Mercedes Herrera1, Jonas Sjölund2, Vladimir Chocoloff 1, Simon Joost3, Rasul M Tabiev1, Lina Wik Leiss1, Carina Strell4, Luis Nunes4, Artur Mezheyeuski4, David Edler5, Anna Martling5, Fredrik Pontén4, Bengt Glimelius4, Tobias Sjöblom4, Maria Kasper3, Kristian Pietras2 and Arne Östman1 (1) Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden (2) Division of Translational Cancer Research, Department of Laboratory Medicine, Faculty of Medicine, Lund University, Lund, Sweden (3) Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden (4) Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden 5) Department of

Molecular Medicine and Surgery, Karolinska University Hospital Solna, Stockholm, Sweden.

Introduction: Properties of tumor angiogenesis impact on tumor growth and composition of the tumor microenvironment. Tumor vessels are composed of pericytes and other mural cells. Detailed characterization of the cellular composition of the microvasculature can suggest functional vessel and case properties relevant for outcome, response to treatment and immune surveillance.
Results: Based on single cell RNA-seq from three human stage II/III colon cancers and pilot in-situ multi-antibody staining of six tumors two main clusters of perivascular cells were identified; B1 (MCAM+, MYH11-) and B2 (MCAM+, MYH11+) cells. B1 cells displayed pericyte features including expression of RGS5, whereas B2 cells showed high expression of smooth muscle cell markers like ACTG2 and Desmin. Desmin-high B2 cells were predominantly located in muscular layers, whereas Desmin-low/MYH11-high B2 cells showed perivascular enrichment. An extended novel analyses workflow of digital image analysis was used to profile peri-endothelial composition in 19 intervals; 11 1-µm intervals from 0 to 10 µm and 8 5-µm intervals from 10 to 50 µm from endothelial areas. Each of these expansion areas was quantitated regarding density of PDGFRB+/- cells including B1/B2 cells, and a MYH11+/MCAM- cells. Density of macrophage subsets (CD68-positive subsets defined by CD163 and CD11c) were also quantitated. In general, B1 cells showed stronger spatial association with endothelial cells than B2 cells, and this spatial association was stronger for the PDGFRB+ B1 subset. Survival analysis indicated that prognostic relevance was strongest when PDGFRB+ (favorable prognosis) cells were closest to endothelial cells. Regarding perivascular macrophage density, M0 and transit M1/M2 macrophages were predominantly enriched in the peri-endothelial regions, whereas density of M1 and M2 increased gradually as distance from vessels increased. M2 density was overall associated with poor prognosis. Ongoing studies are using a clustering-based approach to identify vessel subsets, and prognosis associations are being consolidated in additional cohorts.
Summary: This study provides a novel approach for high resolution profiling of perivascular status in CRC, and possibly other tumor types, that could be of specific importance to identify tumor features associated with response to angiogenesis-directed therapies.

---

20. Computational Drug Discovery
Machine Learning; Deep Learning; Bioinformatics; Computational Drug Discovery; Data Curation

Stuti Jain, INRIA Saclay, France; Emilie Chouzenoux, INRIA Saclay, France; Angshul Majumdar, IIIT Delhi, India

Computational drug discovery has advanced significantly through innovative machine learning techniques for predicting drug-virus associations and drug repositioning. This work utilizes two key datasets: a Drug-Virus Association Dataset comprising 121 drugs and 38 viruses, and a Drug-Drug Interaction Dataset with 1200 drugs and over 36,000 known interactions. The research employs matrix completion techniques, leveraging drug and virus metadata based on chemical structures and genomic profiles.

The study highlights the efficacy of HyPALM and GRPDMF algorithms in predicting drug associations with SARS-CoV-2 variants . These methods achieved high AUC and AUPR scores, providing clinically relevant drug recommendations aligned with ongoing treatments. The HyPALM algorithm successfully predicted six drugs in its top ten that were under testing, with all top five predictions being clinically relevant. Similarly, GRPDMF's recommendations were consistent with current patient treatments. This research

demonstrates the potential of machine learning & deep learning in enhancing drug discovery processes and improving clinical outcomes through data-driven insights.

---

21. Blood proteome profiling using proximity extension assay in patients with acute myeloid leukemia

Biomarker discovery; plasma proteomics; acute myeloid leukemia; machine learning; proximity extension assay

Johansson, Emil, KTH, Sweden; Bueno Álvez, María, KTH, Sweden; Lehmann, Sören, Uppsala University, Sweden; Sjöblom, Tobias, Uppsala University, Sweden; Pontén, Fredrik, Uppsala University; Mardinouglu, Adil, KTH, Sweden; Uhlén, Mathias, KTH, Sweden; Edfors, Fredrik, KTH, Sweden

Acute Myeloid Leukemia (AML) is the most common form of acute leukemia in adults. Plasma proteomic profiling represents an attractive way to assess biomarkers and screen for early diagnosis in malignant diseases, but studies remain scarce in AML. This study was conducted by analyzing 1 463 plasma proteins in 52 AML patients at diagnosis using the Olink Explore 1536 platform. Both differential expression analysis and feature selection by machine learning were applied to find the most significant proteins to distinguish AML from 867 healthy individuals and 1 734 patients of varying cancer types, including different hematological malignancies. The analysis identified several proteins with significant altered expression in AML patients as compared to multiple controls, such as CDH15, EPO, FCGR3B, FLT3, FLT3LG, LCN2, PGLYRP1, and TNFRSF10C.

---

22. Profiling the Blood Proteome in Autoimmune Disease Using Proximity Extension Assay

plasma proteomics; autoimmune disease; biomarkers

Kenrick, Josefin 1 , Edfors, Fredrik 1 , Uhlen, Matthias. 1 , Nilsson, Peter. 1 , Bergström, Sofia. 1 , Pin, Elisa. 1 1 KTH, Protein Science, Solna, Sweden

Autoimmune diseases are heterogeneous diseases characterized by dysregulation of the immune system. They result in chronic inflammation and damage to overall health. There is a pressing need for the discovery of biomarkers to facilitate early diagnosis, stratification, and treatment evaluation for patients. In this study, five autoimmune diseases were selected for plasma profiling as part of the Human Disease Blood Atlas program, including myositis, rheumatoid arthritis, systemic sclerosis, Sjögren's syndrome, and systemic lupus erythematosus. In total, 592 plasma samples were analysed using the Olink Explore platform, resulting in expression data of 1159 unique proteins. Differential expression analysis identified potential prognostic biomarkers; some of these have previously been found to be associated with autoimmune disease, and others are novel. Pathway analysis provides further insights into the underlying biological processes involved. This study provides a comprehensive, exploratory analysis with the aim to identify distinct protein profiles within and across five autoimmune diseases.

23. Unveiling the Invisible: Machine Learning for Detecting Hidden Infections and Species in Metagenomic NGS Data
NGS, Metagenomics, Hidden Infection

Khakvar, Reza, Uppsala university, Sweden; Jan Komorowski, Uppsala University, Sweden

The emergence of metagenomic next-generation sequencing (NGS) has revolutionized our ability to probe complex microbial communities. However, detecting cryptic infections and rare species in these datasets remains elusive. It would be a major challenge. This study presents a new machine learning framework designed to increase the accuracy of detecting low-abundance rare pathogens and strains in metagenomic NGS data. Leveraging advanced algorithms and comprehensive training datasets Our method not only improves sensitivity and specificity; but also reduces computational costs. Results in previously unknown disclosures. Microbiology demonstrates the ability to learn. This work presents a profound impact on diagnostics, epidemiology, and environmental microbiology. It underscores the transformative power of artificial intelligence to advance our understanding of microbial diversity and infection dynamics.

---

24. Brownian motion data augmentation
Data augmentation; Nanopore; Time series; Neural networks;

Kipen, Javier, KTH, Sweden; Jalden, Joakim, KTH, Sweden

Nanopores are highly sensitive sensors that have achieved commercial success in DNA/RNA sequencing, with potential applications in protein sequencing and biomarker identification. Solid-state nanopores, in particular, face challenges such as instability and low signal-to-noise ratios (SNRs), which lead scientists to adopt data-driven methods for nanopore signal analysis, although data acquisition remains restrictive. In this paper, we augment training samples by simulating virtual Brownian motion based on dynamic models in the literature. We apply this method to a publicly available dataset of a classification task containing nanopore reads of DNA with encoded barcodes. A neural network named QuipuNet was previously published for this dataset, and we demonstrate that our augmentation method produces a noticeable increase in QuipuNet's accuracy. Furthermore, we introduce a novel neural network named YupanaNet, which achieves greater accuracy (95.8%) than QuipuNet (94.6%) on the same dataset. YupanaNet benefits from both the enhanced generalization provided by Brownian motion data augmentation and the incorporation of novel architectures, including skip connections and a self-attention mechanism.

---

25. Sharing data science applications and machine learning models with SciLifeLab Serve
machine learning, data science, application hosting

Kochari, Arnold, SciLifeLab, Sweden

SciLifeLab Serve (beta) is a platform offering data science application hosting (Shiny, Dash, Gradio, Streamlit, etc.) and machine learning model serving. The service is free to use for life science researchers affiliated with a Swedish research institute and their colleagues. Each application or model receives 2 CPU and 4 GB RAM by default, with the possibility to ask for more where there is need. SciLifeLab Serve is developed and operated by the SciLifeLab Data Centre. The poster will present the service. You can meet the team building the service and

ask any questions during the poster session.

---

26. Bacterial vaginosis: Understanding the effect of antibiotic-free treatment (pHyph) on the vaginal microbiome

bacterial vaginosis; vaginal microbiome; antibiotic-free treatment; Lactobacilli

Lahtinen, Emilia, Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Sweden & Gedea Biotech AB, Medicon Village, Sweden; Hugerth, Luisa, Centre for Translational Microbiome Research, Department of Microbiology, Tumor, and Cell Biology (MTC), Karolinska Institutet, Sweden & Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden; Edfeldt, Gabriella, Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Sweden; Strevens, Helena, Gedea Biotech AB, Medicon Village, Sweden; Kornfält, Sten, Gedea Biotech AB, Medicon Village, Sweden; Säfholm, Annette, Gedea Biotech AB, Medicon Village, Sweden; Engstrand, Lars, Centre for Translational Microbiome Research, Department of Microbiology, Tumor, and Cell Biology (MTC), Karolinska Institutet, Sweden; Schuppe Koistinen, Ina, Centre for Translational Microbiome Research, Department of Microbiology, Tumor, and Cell Biology (MTC), Karolinska Institutet, Sweden.

Bacterial vaginosis (BV) is the most common vaginal condition among reproductive age women, caused by an imbalance in the vaginal microbiome. Diagnosing and treating BV is challenging since the molecular mechanisms leading to BV development are unknown. Antibiotics are the first-line treatment, but they often lead to recurrence and increase antimicrobial resistance. pHyph is an antibiotic-free, intra-vaginal tablet by Gedea Biotech aimed at treating and preventing BV by disrupting pathogenic biofilm, restoring favorable vaginal pH, and promoting Lactobacilli growth. Two Phase 2 clinical trials have assessed pHyph's efficacy in treating BV and vulvovaginal candidiasis (VVC). Preliminary results show pHyph increases Lactobacillus abundance and decreases the abundance of BV-associated species, such as Atopobium vaginae, Gardnerella vaginalis, and BVAB2. Most women treated with pHyph shifted to a Lactobacillus-dominated microbiome. Ongoing analysis aims to further understand factors affecting microbiome composition, treatment success, and BV recurrence.

---

27. Structural Models and Refinement for Multi-Dataset Experiments

Structural Biology: Time-Resolved Crystallography: Computational Method Development

Lassinantti Lena, Pearce Nicholas M Physics, Chemistry and Biology (IFM), SciLifeLab, Linköping University, Sweden.

Structural studies of macromolecules, in their apo forms and in complexes with ligands, have long been vital in biomedical research. With the rise of high-volume data from experiments like time-resolved (TR) crystallography, current methods of computational data processing—based on a one-model-per-dataset paradigm—are underdeveloped and impractical. To address this, we propose a core-model paradigm for TR experiments, where a single "core" model is refined across all time-points, ensuring consistency in model composition while adjusting for experimental variations. This approach requires generic and flexible merging tools to create arbitrarily complex multi-state models, and corresponding refinement protocols. Exhaustive sampling of state occupancies combined with electron density validation allows the identification of occupancies for each dataset. This makes modelling

and refinement tractable for large multi-dataset experiments, where no compositional heterogeneity between data exists. Our approach offers a novel way to think about structural approaches and opens new possibilities for future method development.

---

28. Investigating MS through the characterization of T and B cells molecular landscape with advanced light microscopy
MS; Machine Learning; Microscopy; Phenotyping

Emma Latron, Nicolas Ruffin, Inês Cunha, Chiara Starvaggi, Frederik Piehl, Juliette Griffié

Multiple Sclerosis (MS) is a chronic autoimmune disease with a complex etiology involving inflammation in the central nervous system causing neurodegeneration. MS affects almost 3 million people worldwide. The precise biological mechanisms behind the onset of the disease are still unknown. There is currently no cure, but disease-modifying treatments can delay the course of the disease. A variety of predisposing factors and pathological processes have been identified, shaping the pathogenesis of MS and giving rise to a wide spectrum of clinical manifestations.

We want to take a step towards personalized medicine by profiling individual patient's cells phenotype. Combining high-content microscopy with machine learning analysis pipelines, we explore the cellular phenotypes of pathological cells driving autoimmune responses in MS patients and identify the features that are relevant to each phenotype.
Ultimately, establishing pathological phenotypes would contribute to improving personalized treatment and earlier diagnosis of MS.

---

29. Spatial Dynamics of the Developing Human Heart
data analysis; interpretation; visualization

Eniko Lázár, Raphaël Mauron, Žaneta Andrusivová, Joakim Lundeberg; KTH Royal Institute of Technology, Department of Gene Technology, SciLifeLab

Cardiac structures develop through spatially defined interactions between various cell states during human cardiogenesis. To investigate early heart formation with a spatial perspective, we created a detailed cardiac cell atlas by integrating unbiased transcriptomics measurements from 80,000 cardiac single cells and 70,000 spatially barcoded heart regions from postconceptional weeks 5 to 14, complemented with imaging-based transcriptomics validation of 150 target genes. Mapping 72 fine-grained cell states to distinct structural and functional components of the developing human heart allowed us to resolve prominent developmental cardiac niches, such as the pacemaker-conduction system, cardiac valves, atrial septum, and autonomic cardiac innervation, and investigate their molecular dynamics and interactions. In summary, our study provides a comprehensive view of the cellular landscape of early human cardiogenesis, connecting the developing heart's molecular architecture to genetic causes of heart disease.

30. De-novo design of molecular glues with EvoBind-multimer
TPD, EvoBind-multimer, PROTAC, Molecular glue

Li, Qiuzhen, The Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Sweden; Vlachos, Efstathios Nikolaos, The Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Sweden; Bryant, Patrick, The Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Sweden

Targeted protein degradation has emerged as a promising therapeutic approach, leveraging technologies such as PROTACs and molecular glue degraders. These bifunctional molecules facilitate the linkage of target proteins to E3 ubiquitin ligases, inducing ubiquitination and subsequent proteasomal degradation. While this method demonstrates efficacy even with compounds exhibiting low affinity or bioavailability, several challenges persist. These include achieving high target protein specificity, limited targetable E3 ligases, and time-consuming optimization processes. To address these limitations, we introduce EvoBind-multimer (EBM), a novel AI-driven framework for molecular glue design. EBM focuses on cyclic peptides that mediate interactions between ligase domains and target proteins based exclusively on their sequences. This approach identifies optimal binding sites without reliance on predefined information and utilizes natural interfaces between the two target proteins. By circumventing the constraints of current methodologies, EBM presents a potential paradigm shift in targeted protein degradation strategies, offering new avenues for therapeutic development.

---

31. Direct optimisation of physical parameters in TLS models
ECHT, Structural Refinement, Parameter Optimisation, TLS, ADPs

Ying Luo, Nicholas M. Pearce, Bioinformatics Divison, IFM, SciLifeLab, Linköping University, SE

Translation-Libration-Screw (TLS)_models produce group anisotropic atomic displacement parameters (a-ADPs) using relatively few parameters compared to individual atomic a-ADPs. However, there has been a long-standing problem with the interpretation of refined TLS matrices in terms of physical motions, with refined TLS models often physically invalid. This problem stems from the refinement approach used in modern refinement programs, where they refine elements of the intermediate TLS matrices, rather than the underlying physical parameters.

To address this challenge, we restructured the TLS matrices in terms of physical motions in a spherical coordinate system. This allows for the direct optimisation of a TLS model by gradient descent with easily validated outputs. These changes are implemented in the ECHT model, a hierarchical TLS disorder analysis model, where the enhancements in parameter reliability and the streamlining of optimisation operations reduce the need for multiple optimisation regimes and a laborious validation process, while simultaneously providing more precise and reliable results.

---

32. On the hunt for novel virulence factors using comparative genomics
Comparative genomics, virulence factors, Pseudomonas aeruginosa

Marchi Agata (Chalmers University of Technology, Department of Life Sciences, Division of Systems and Synthetic Biology,Italy), Marcus Wenne (Chalmers University of Technology, Department of

Life Sciences, Division of Systems and Synthetic Biology, Centre for Antibiotic Resistance Research in Gothenburg (CARe), Sweden), Vi Varga (Chalmers University of Technology, Department of Life Sciences, Division of Systems and Synthetic Biology, Hungary), Johan Bengtsson-Palme (Chalmers University of Technology, Department of Life Sciences, Division of Systems and Synthetic Biology, Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Centre for Antibiotic Resistance Research in Gothenburg (CARe), Sweden)

Pseudomonas aeruginosa is an opportunistic pathogen that causes severe nosocomial infections. A large portion of the genome of P. aeruginosa consists of genes with unknown functions. This study aims to identify novel virulence factors among the numerous uncharacterized genes.
We compared P. aeruginosa genomes isolated from human and natural environment sources to identify clusters of homologous proteins that were significantly enriched with sequences coming from bacteria isolated from humans.
Results showed that approximately 3% of the clusters were significantly enriched with sequences from human isolates. Functional characterization of 45 selected clusters revealed roles primarily related to secretion systems, conjugation, and biofilm survival. Notably, 18% of the selected proteins had unknown functions, suggesting the existence of potentially novel virulence factors.
This study provides insights into previously uncharacterized genes in P. aeruginosa and identifies potential targets for further investigation in understanding and combating this pathogen's virulence mechanisms.

---

33. Linking Gut Microbiome Composition to Longevity and Mortality: Insights from the SIMPLER Cohort

gut microbiome, healthy aging

Cecilia Martinez Escobedo, Clemens Wittenbecher

The global population ages rapidly and understanding the key determinants associated with healthy aging is crucial. Emerging evidence suggests that the gut microbiome plays a critical role in determining human longevity and mortality, by modulating systemic inflammation, immune function, and metabolic regulation. With age, the individual gut microbiome becomes increasingly unique. Healthy aging is characterized by a decline in core taxa like Bacteroides and an increase in butyrate-producing microbes. In contrast, functional microbiome signatures related to the Enterobacteriaceae family are associated with higher mortality. Despite these findings, significant gaps remain in our understanding of the gut microbiome's role in longevity and mortality.

This study aims to examine the association of the gut microbiome composition and its functional capacity with longevity, healthy lifespan, and cause-specific mortality. In the prospective, population-based SIMPLER cohort, we analyzed shotgun metagenomics profiles, and health and mortality follow-up information of approximately 5,000 individuals. We performed survival analysis using Cox proportional hazards regression to link microbial features with mortality, adjusting for age, sex, and lifestyle. Our project aims to contribute to the understanding of the microbiome's impact on healthy aging and has the potential to inform precision healthcare strategies that promote beneficial gut microbiomes to enhance longevity.

34. Building a Human Cell Simulator: Foundations in Data Streaming and Collaborative Annotation

human cell simulator; bioimage analysis; decentralised storage; data streaming; collaborative annotation

Mechtel, Nils, KTH, Sweden; Gómez de Mariscal, Estibaliz, Instituto Gulbenkian de Ciência (IGC), Portugal; Pape, Constantin, Georg August University of Göttingen, Germany; Reder, Gabriel, University of Cambridge, United Kingdom; Ouyang, Wei, KTH, Sweden.

The creation of a human cell simulator represents a transformative step toward understanding complex cellular behaviors. Achieving this requires a robust infrastructure that supports both scalable data management and efficient data generation. Our research establishes a foundation by integrating decentralized storage, data streaming, and crowd-sourced annotation. A flexible streaming dataloader connects to S3, enabling rapid access to large datasets across decentralized locations. Complementing this, our collaborative annotation platform uses the Segment Anything Model (SAM) in a human-in-the-loop approach to refine existing bioimaging data. AI agents play a crucial role in enhancing data generation and automating scientific workflows, with future integration into automated microscopy. Together, these tools form the building blocks for constructing a comprehensive human cell simulator, providing deeper insights into cellular processes and accelerating in-silico drug discovery.

---

35. Deep plasma proteome characterization in two independent clinical cohorts identifies clusters of biomarkers separate benign and malignant tumours in women with suspicion of ovarian cancer

Ovarian cancer; Proteomics; Machine learning; Olink

Moskov, Mikaela, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Hedlund, Lindberg, Julia, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Popova, Svetlana, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden, Department of Clinical Pathology, Akademiska University Hospital, Uppsala, Sweden; Forsberg, KG, Simon, Olink Proteomics AB, Uppsala, Sweden; Diamanti, Klev, Olink Proteomics AB, Uppsala, Sweden; Lycke, Maria, Department of Obstetrics and Gynaecology, Institute of Clinical Sciences, Sahlgrenska Academy at Gothenburg University, SE-41685 Gothenburg, Sweden; Ivansson, Emma, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Tolf, Anna, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden, Department of Clinical Pathology, Akademiska University Hospital, Uppsala, Sweden; Sundfeldth, Karin, Department of Obstetrics and Gynaecology, Institute of Clinical Sciences, Sahlgrenska Academy at Gothenburg University, SE-41685 Gothenburg, Sweden; Stålberg, Karin, Department of Women's and Children's Health, Uppsala University, SE-75185 Uppsala, Sweden; Gyllensten, Ulf, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden; Enroth, Stefan, Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden.

Ovarian cancer is the deadliest of all gynecological cancers with surgery often being necessary for final diagnosis. We analyzed 5,416 plasma proteins using Olink® proximity extension assays from two independent cohorts (N=171+233) of women with benign or

malignant tumors confirmed through surgery.
The analysis identified 191 significant proteins, with 99.7% of the findings replicated. Only 21/191 proteins showed a significant correlation to tumor gene expression. Strong correlations between proteins could be found for 62/191 proteins, suggesting many of the observed associations could be secondary effects. A predictive model trained on the discovery cohort using eight proteins, showed an AUC of 0.96, achieving 97% sensitivity and 68% specificity when tested in the replication cohort. For early-stage tumors, the sensitivity reached 91%, surpassing the performance of the clinical biomarker CA-125.
These results suggest that molecular tests could correctly identify up to one third of benign cases, reducing the need for surgical diagnosis.

---

36. Efficient Protein-Protein Interaction Prediction Using AlphaFold2
Protein-protein interactions, AlphaFold, Machine Learning

Sarah Narrowe Danielsson, Arne Elofsson* *Department of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, Sweden.*

Proteins are the workforce and fundamental building blocks in living organisms, consisting of amino acids arranged into complex chains, often working collaboratively. Understanding which proteins interact is pivotal for fully understanding their function. Experimental methods for investigating protein-protein interactions (PPIs) are available but suffer from high false positive rates, conflicting results as well as being time-consuming and expensive. To address these challenges, complementary computational approaches like AlphaFold by Google Deepmind, have emerged, which excels in predicting protein structures and PPIs. However, AlphaFold's major limitation is its time-consuming process, taking hours for a single interaction, making it impractical for investigating millions of interactions. We are currently working on ways to improve the efficiency of AlphaFold, with OpenFold, without sacrificing too much in performance. The current idea is to simplify the architecture of OpenFold in order to allow for screening of pairwise combinations of entire proteomes. The work is still in progress.

---

37. A generalised protein aggregation simulator for predicting cellular signal transduction
Computational Biology; Bioinformatics; Super-resolution Microscopy; Signal Transduction; Transmembrane Proteins.

L. Panconi, O. P. L. Dalby, C. Zaza, D.M. Owen, S. Simoncelli, J. Griffié

The plasma membrane acts as the primary site for signalling pathway activation and inter-cellular communication. Protein nanoscale spatial organisation, interaction and aggregation are largely responsible for regulating these processes. Studies have shown that even relatively small changes in receptor distribution can initiate downstream signalling. Dynamic characterisation of protein clustering is essential for hypothesis testing, quantification of spatial statistics and investigating perturbation conditions. In order to interrogate the molecular distribution dynamically, an in silico modelling platform is required. Drawing from biophysical diffusion properties, agent-based modelling and spatial point pattern statistics, we present an inexpensive and modular framework for simulating protein aggregation dynamics, quantifying bulk parameters and evaluating emergent behaviour. With this modality, we quantify the impact of protein aggregate geometry on the probability of inducing signal transduction. We apply our methodology to the specific case of TCR-CD3

dynamics to derive novel insights into the prerequisites of T cell activation.

---

38. Blurry Bacillus Boundaries: Using pangenomics to improve species delineation within Bacillus cereus sensu lato

Microbiology, genomics, database, foodborne pathogen

Raghuram, Vishnu, Department of Clinical Microbiology Umeå University, Sweden; Ramnath, Vignesh, Department of Clinical Microbiology Umeå University, Sweden; Larralde, Martin, Structural and Computational Biology Unit EMBL, Germany; Carroll, Laura, Department of Clinical Microbiology Umeå University, Sweden.

The Bacillus cereus group is a complex of Gram-positive, spore forming bacteria. Members of the B. cereus group include, among others: B. anthracis - the causative agent of anthrax; B. cereus sensu stricto - a foodborne pathogen; and B. thuringiensis, a biocontrol agent. Species identification in the B. cereus group is currently a challenge, as there is no universally accepted taxonomy. Current taxonomic classification methods for the B. cereus group vary in the number and composition of the species they define (12 - 58 "species").

Biological and environmental barriers can lead to each species unit having a distinct gene content "signature". We began with 5,976 B. cereus group genomes from the BTyperDB database and estimated the pangenome of a subset. We then investigated distinct patterns of phylogenetic dispersion of genes associated with specific functions. Overall, this work attempts to provide a deeper understanding of the genetic diversity within the B. cereus group.

---

39. BTyperDB: A community-curated, global atlas of Bacillus cereus sensu lato genomes and metadata for epidemiological surveillance

Pathogen surveillance; computational microbiology; bacterial genomics; web application

Ramnath, Vignesh, Umeå University, Sweden; Henriksson, Johan, Umeå University, Sweden; Carroll, Laura, Umeå University, Sweden

Existing public pathogen databases can lead to dangerous pathogen misidentifications when applied to Bacillus cereus sensu lato (s.l.). Current taxonomic assignments are inadequate for evaluating B. cereus s.l. pathogenic potential at the strain level, as many important virulence factors can be gained, lost, and variably present within and across species boundaries. Moreover, incomplete metadata in aforementioned databases also makes epidemiological surveillance challenging.

To combat these issues, we developed BTyperDB, an atlas of B. cereus s.l. genomes with standardized, community-curated metadata. 6,901 public B. cereus s.l. reads/genomes underwent genome assembly, taxonomic assignment and in silico typing using a variety of methods, resulting in 5,976 high-quality genomes. To help users rapidly access and download genomic (meta)data associated with B. cereus s.l. strains, genomes within BTyperDB can be queried via an interactive web application (www.btyper.app). In summary, BTyperDB can help improve B. cereus s.l. surveillance, source tracking, outbreak detection and response efforts.

40. When cows fly: tracking the geographic spread of broad- and narrow-host range Salmonella enterica serotypes using whole-genome sequencing
Whole-genome sequencing

Santos Bandos Rodrigues, Laura, Umeå University, Umeå, Sweden, Raghuram, Vishnu, Umeå University, Umeå Sweden, Carroll, Laura, Umeå University, Umeå, Sweden.

Foodborne pathogen Salmonella enterica can be transmitted between animals and humans. However, Salmonella serotypes differ in terms of the range of hosts they infect. Here, we used whole-genome sequencing (WGS) data from open-access databases to identify relationships between Salmonella host range and geographic spread. Briefly, raw reads from (i) two bovine-adapted Salmonella serotypes (S. Dublin and S. Cerro) and (ii) one broad host range serotype (S. Newport) were pre-processed and assembled using Bactopia (n = 162, 144, and 122 genomes, respectively). For each serotype, Parsnp was used to call SNPs and construct phylogenies. Notably, our approach identified geographically distinct clades within the phylogenies of serotypes Dublin and Cerro and a lack thereof for serotype Newport. This supports our hypothesis that narrow host range serotypes may not spread as easily geographically compared to broad host range serotypes. In the future, these results may be used to improve Salmonella surveillance and intervention strategies.

---

41. Tracing diatoms over space and time: a view from species distribution modeling
Baltic Sea; climate change; machine-learning algorithms; Ensemble; habitat suitability

Abdelgadir, Mohanad, Sanyal, Anushree, Södertörn University, Sweden

Diatoms are powerful bioindicators for anthropogenic impacts and environmental change in aquatic ecosystems. Yet how the geographical distribution of diatoms responds to ongoing and projected future environmental change across the Baltic Sea is not fully understood. We used a metadata-based modeling approach to predict the future spatial distribution and habitat suitability of selected diatoms across the Baltic Sea. Prediction was based on five different environmental variables; silicate, salinity, primary production of carbon, temperature, and four future scenarios for temperature and salinity in the years 2050 and 2100 using six machine-learning algorithms in species distribution modeling (SDM). Future predictions for the selected taxa of diatoms suggest a decreased distribution area in the Bothnian Sea and Bothnian Bay and increased area in the Arkona basins. Predicted future spatial distribution and habitat suitability of selected taxa of diatoms in climate change scenarios are mainly driven by sea salinity and temperature in the year 2100.

---

42. Integrating Graph Neural Networks to Analyze Drug Effects in Cancer-Fibroblast Cocultures
tumor microenvironment , cell-cell interaction, coculture, graph convolutional networks, chemical structures

Osheen Sharma[1], Greta Gudoityte[1], Olli P. Kallioniemi[1 3], Flavio Ballante[2], Lassi Paavolainen[3], Brinton Seashore-Ludlow[1] [1]Department of Oncology and Pathology, Karolinska Institute, Science for Life Laboratory, Solna, Sweden [2]Department of Medical Biochemistry and Biophysics, Karolinska Institute, Science for Life Laboratory, Solna, Sweden [3]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland

Understanding the tumor microenvironment (TME) is vital for developing effective cancer therapies, as the complex cell-cell interactions within the TME significantly influence disease progression and treatment response. Traditional drug testing relies on monoculture assays, which fail to replicate the dynamic interactions between cancer cells and surrounding fibroblasts. To address this, we developed a 2D coculture system that combines ovarian cancer cells (Ovcar3 and Kuramochi) with fibroblasts (BJhTERT) to better simulate these interactions. This approach enables us to explore how small molecules interact with cells in both monoculture and coculture, providing deeper insights into drug mechanisms of action (MOA). By leveraging graph convolutional networks (GCN) and chemical structures of the compounds, we aim to predict MOAs more accurately and assess how cell-cell interactions modulate drug response. This study bridges the gap between simplified models and the complexity of the TME, advancing our understanding of cancer drug discovery.

---

43. Clonal hematopoiesis of indeterminate potential is associated with pro-inflammatory proteomics markers in coronary artery disease patients
Clonal hematopoiesis; Cardiovascular disease; Genomic; Proteomics;

Siamala Sinnadurai, Department of Cardiology, Erasmus MC, Rotterdam, The Netherlands

Clonal hematopoiesis (CH), a somatic mutations in blood cells without hematologic malignancies, is associated with the progression of atherosclerosis. Patients under 80 hospitalized for acute coronary syndrome (ACS) or myocardial revascularization was studied. Targeted next-generation sequencing (NGS) was used to detect CH mutations in 92 genes. Multiplex proteomic profiling via extension proximity assays (EPA) measured circulating inflammatory biomarkers. Of the 130 patients analyzed, 17.6% had CH mutations. Patients with DNMT3A or TET2 mutations exhibited higher body mass index (BMI) and elevated inflammatory markers, such as high-sensitivity C-reactive protein (hs-CRP), interleukin-6 (IL-6), along with increased myocardial stress markers like troponin T and NT-proBNP. After adjusting for factors such as BMI, age, and gender, certain inflammatory markers (e.g., HGF, IL10RB, IL5RA, and TNFRSF9) remained significantly associated with CH. The study suggests that mutations in TET2 and DNMT3A may drive inflammatory responses by engaging cytokine receptor activity and leukocyte activation pathways.

---

44. Improving AlphaFold for Efficient Protein Structure Prediction
Protein Structure Prediction, Alphafold, Machine Learning Optimization

Tadiello Matteo, Stockholm University

AlphaFold has transformed protein structure prediction, but its attention-based architecture poses challenges for large protein predictions due to high computational and memory demands. This project presents a mamba-based pair representation module to replace the attention mechanism in the Evoformer/Pairformer block. This should reduce computational complexity, optimizing AlphaFold's performance for large proteins. This improvement aims to accelerate prediction time and reduce memory usage, enhancing the scalability of AlphaFold for complex biological systems.

45. Imputing complex cell features to spatially and temporally dissect cancer heterogeneity

single cell; cancer heterogeneity; micro-environment; cell lineage relationships; computational inference

Marcel Tarbier, Uppsala University / SciLifeLab, Sweden

One of the most powerful and convenient ways to study a cells state is transcriptomics. But dissecting cancer heterogeneity and phenotype switches requires - above all - resolution: single-cell, spatial, temporal, molecular. Current technologies can resolve some of these layers, but not all, especially for human biopsies. Fortunately, many things we can't measure directly, we can infer computationally! Here, we present strategies to infer cell lineage relationships and micro-environments from single-cell RNA-sequencing, and how these proxies can help in understanding cancer phenotype switches.

---

46. Data centre Scilifelab services

serve, sclílifelab data repository

Data stewards, Scilifelab

This poster describes in brief serbices being offered by Data centre at scilifelab for reseachers, PhDs, postdocs, early stage researchers, scilfelab and DDLS fellows for carrying out data driven life science research.

---

47. Predictive Power of the Prenatal Microbiome: Species-Level Insights into Postpartum Depression Risk

Machine Learning; Postnatal Depression; Gut Microbiome; Gut Brain Axis

Bangzhuo Tong, Uppsala University; Andrey Shternshis, Uppsala University; Emma Fransson, Uppsala University; Alkistis Skalkidou, Uppsala University; Lars Engstrand, Karolinska Institutet; Prashant Singh, Uppsala University; Luisa W. Hugerth, Uppsala University.

Postnatal depression (PND) is defined as depression diagnosed up to 8 weeks postpartum and can affect c. 20% of pregnancies. PND is associated with poor maternal health after delivery and adverse offspring outcomes. The Edinburgh Postnatal Depression Scale (EPDS) is commonly used to define PPD symptomatically, with a score of 12 or higher indicating depression. Emerging researches suggest a bidirectional communication between the gut microbiota and the gastrointestinal function, central nervous system circuitry, autonomic nervous system and immune system via the gut-brain-axis. Few studies have explored the predictive potential of prenatal microbiome data on postpartum depression. Using Random Forest, we predicted PPD at 6 weeks postpartum, as defined by EPDS scores, with species-level microbiome data from the second and third trimesters in a Swedish cohort. We eventually achieved a model accuracy of 0.57 and an AUC of 0.63 using only second-trimester microbiome data.

48. Machine Learning Classification of Multivariate Time Series Data for Clinical Neuroscience Applications

Machine Learning; Precision Medicine; Clinical Neuroscience; Time Series; EEG/MEG

Uribarri, Gonzalo, KTH & SciLifeLab, Sweden; Solana, Adrià, KTH, Sweden; Fransén, Erik, KTH & SciLifeLab, Sweden

Effective classification of multivariate time series data is critical to the advancement of precision medicine, enabling analysis and diagnosis in a wide variety of data modalities such as brain activity, microbial time series, and time-lapse microscopy. Traditional time series classification models, such as convolutional neural networks or ROCKET algorithms, struggle when the number of instances is not very large and the dimensionality of the data is high. In addition, they are generally not easy to interpret. We present Detach-Rocket Ensemble, a novel machine learning algorithm that improves ROCKET for the case of multivariate data, enhancing its performance and providing an accurate estimation of channel relevance. We demonstrate its effectiveness on real-world EEG and MEG datasets, including one where the task is to discriminate between healthy controls and patients with Alzheimer's disease, highlighting the potential of our methodology for clinical neuroscience applications.

---

49. SPOT-BGC: A Snakemake Pipeline to Output meTagenomics-derived Biosynthetic Gene Clusters

biosynthetic gene clusters, pipeline, Snakemake, metagenomics, microbiome

Varga, Virág (Division of Systems and Synthetic Biology, Department of Life Sciences, SciLifeLab, Chalmers University of Technology, Gothenburg, Sweden); Hugerth, Luisa Warchavchik (Department of Medical Biochemistry and Microbiology, SciLifeLab, Uppsala University, Uppsala, Sweden); Bengtsson-Palme, Johan (Division of Systems and Synthetic Biology, Department of Life Sciences, SciLifeLab, Chalmers University of Technology, Gothenburg, Sweden)

Biosynthetic gene clusters (BGCs) are operonic sets of microbial genes that synthesize specialized metabolites including antimicrobials, siderophores and other secondary metabolites. The ability to accurately predict BGCs from metagenomic samples is particularly important for studying the interactions of unculturable microbes. However, such analysis is complicated by the sheer magnitude of data that requires processing. As new datasets become available, it is also important to be able to integrate them into existing models. Here, we present a pipeline for the prediction of BGCs from metagenomic data using Snakemake.

The pipeline handles quality trimming, filtration of human-derived reads, per-sample and per-cohort assembly, taxonomic assignment and BGC predictions (using GECCO and AntiSMASH), but is also flexible for integration of data at multiple points of the process. We envision that this pipeline will be useful for streamlining the identification of BGCs across microbiomes.

---

50. Pathogens Portal

Open Data; FAIR; Data Visualisations; Infectious Diseases; Pandemic Preparedness; Antibiotic Resistance

Hughes, Liane, SciLifeLab Data Centre. BMC, Uppsala Universitet Husargatan 3, 751 22 Uppsala University, Uppsala, Sweden; Öjefors Stark, Katarina,SciLifeLab Data Centre. BMC, Uppsala

Universitet Husargatan 3, 751 22 Uppsala University, Uppsala, Sweden; Panneerselvam, SenthilKumar,SciLifeLab Data Centre. BMC, Uppsala Universitet Husargatan 3, 751 22 Uppsala University, Uppsala, Sweden; Aziz, Abdullah,SciLifeLab Data Centre. HPC2N, Umeå University, 901 87, Umeå, Sweden; Venkatesh, Nalina Hamsaiyni, SciLifeLab Data Centre. HPC2N, Umeå University, 901 87, Umeå, Sweden; Dulaud, Paul, SciLifeLab Data Centre. HPC2N, Umeå University, 901 87, Umeå, Sweden; Kultima, Hanna, SciLifeLab Data Centre. BMC, Uppsala Universitet Husargatan 3, 751 22 Uppsala University, Uppsala, Sweden; Rung, Johan, SciLifeLab Data Centre. BMC, Uppsala Universitet Husargatan 3, 751 22 Uppsala University, Uppsala, Sweden.

The Swedish Pathogens Portal aims to support research groups and organisations in making data and other research resources as open and FAIR as possible. In particular, the Portal publishes data-centric articles (data highlights) and editorials describing recent research, data dashboards (including custom, dynamic data visualisations), as well as resources related to community building (e.g. pages on events and funding opportunities, and ongoing projects). The scope of the Portal includes infectious disease and other topics related to pandemic preparedness, such as antibiotic resistance. The Portal is seeking feedback from the pandemic preparedness research community on their current needs in order to create resources that can accelerate and support their research activities.

---

51. Longitudinal analysis of genetic and environmental interplay in human metabolic profiles and the implication for metabolic health

metabolomics; genetics; lifestyle; protein-metabolite network; metabolic health

Jing Wang1#, Alberto Zenere1#, Xingyue Wang1#, Göran Bergström2, Fredrik Edfors3, Mathias Uhlén3, Wen Zhong1 1 Science for Life Laboratory, Department of Biomedical and Clinical Sciences (BKV), Linköping University, Linköping, Sweden. 2 Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, and Clinical Physiology, Sahlgrenska University Hospital, Gothenburg, Sweden 3 Department of Protein Science, KTH - Royal Institute of Technology, Stockholm, Sweden*

To understand how genetics and environmental factors shape human metabolic profiles, we performed a comprehensive two-year longitudinal analysis of 101 clinically healthy individuals aged 50 to 65, integrating genomics, metabolomics, proteomics, clinical measurements, and lifestyle questionnaire data. Our findings demonstrated the significant role of genetics in determining metabolic variability, identifying 22 plasma metabolites as genetically predetermined and stable over time. Environmental factors such as seasonal variations, weight management, smoking, and stress were also found to substantially influence metabolite levels. We developed an integrative metabolite-protein network comprising 5,649 significant protein-metabolite pairs and identified 87 causal metabolite-protein associations under genetic regulation. This network revealed that each individual has stable and unique protein-metabolite profiles, emphasizing metabolic individuality. Despite the general stability in metabolic profiles, the interplay between genetics and environmental factors drove individualized metabolic dynamics. Key proteins and metabolites have been identified for advancing metabolic risk assessment, offering new insights for monitoring metabolic health.

52. HUBMet: An integrative database and analytical platform for human blood metabolites and metabolite-protein associations
Metabolome; Human Blood; Database; Analytical platform

Blood metabolites can reflect both instantaneous and stable individual profiles, making them potential biomarkers of health status. There have been well-constructed metabolome databases and analytical platforms, including Human Metabolome Database and MetaboAnalyst. However, none is specifically designed for human blood metabolites. Besides, dynamic nature and complex tissue origins increase complexity of blood metabolome analysis. Here, we presented an integrative human blood metabolite database and analytical platform (HUBMet), including 3,950 unique metabolites annotated with four term types, including Class, Pathway, Disease, and Drug. A metabolite-protein network comprising 129,814 links between 4,455 proteins and 1,744 blood metabolites was constructed, and a genetic-based algorithm was developed for tissue specificity analysis of metabolite-associated proteins. Based on the well-constructed database, a web server was developed with single metabolite query and three analytical modules: Enrichment Analysis, Metabolite-Protein Association & Tissue Specificity Analysis, and Metabolite-Protein Network Analysis, with its analytical capacity validated through a COVID-19 cohort study.

---

53. Unveiling Biases in Insect Diversity: A Comparative Study of eDNA Substrates and Malaise Traps
Insect Diversity, Environmental DNA (eDNA), Malaise Traps, Substrate Bias, Biodiversity Assessment

Beilun Zhao, Tobias Andermann

This study examines the biases introduced by different substrates in assessing insect diversity, focusing on environmental DNA (eDNA) substrates and traditional Malaise traps. While Malaise traps are widely used to capture flying insects, they often miss species from cryptic or less accessible habitats. eDNA sampling has emerged as a complementary approach, but its effectiveness varies depending on the substrates used. We conducted a comparative analysis of various eDNA substrates, including anthill, sediment, soil, spiderweb, water, and rot wood, alongside Malaise traps to assess their ability to capture a more comprehensive picture of insect diversity. Our results reveal that each substrate captures a distinct taxonomic composition, with substantial variation in both diversity levels and species representation. This study emphasizes the need to integrate multiple methods to reduce biases and enhance the accuracy of insect biodiversity assessments, providing valuable insights for ecological monitoring and conservation strategies.

---

54. Decoding the genetics and lifestyle influence on human metabolic health
genetics, proteomics, metabolomics, data-driven, precision medicine

Zhong, Wen, LiU, Sweden

Metabolic disorders, such as type 2 diabetes, obesity, and cardiovascular diseases, are rising globally. Research has shown that both genetic predisposition and lifestyle factors, including diet, physical activity, and stress, contribute significantly to the risk of developing these conditions. However, the extent to which each factor influences metabolic health varies among individuals, and still requires deeper exploration. Proteins and metabolites are increasingly recognized as promising biomarkers for metabolic health, serving both to