PhD Retreat – October 14–16, 2025

Final program

Venue & Accommodation

Hótel Stracta
Rangárflatir 4, 850 Hella, Iceland
Website: https://www.stractahotels.is/
GPS: https://maps.app.goo.gl/ZfE6CzZSHaAS5oBX7

Organizers:
Morris Riedel and Gabriele Cavallaro

Day 1 – Monday, October 13, 2025

(Arrival day for international participants)

Transport from Keflavik Airport

Bus Options:

- **Reykjavik Excursion (Flybus)** – https://www.re.is/tour/flybus/
  Arrival: BSÍ Bus Terminal, Reykjavík
  Map: https://maps.app.goo.gl/5bhj2qbwRyhtznQU7
  Duration: ~45 minutes | Cost: 3,999 ISK

Optional Stop:

- **Blue Lagoon Transfer:** https://www.re.is/tour/blue-lagoon-transfer-from-keflavik-airport/

17:00 – 19:00

The 22nd IHPC Workshop
Venue: *Fenjamýri Meeting Room*, Gróska, Bjargargata 1, Reykjavík
Event link: https://www.ihpc.is/events/22nd-icelandic-hpc-community-workshop

Day 2 – Tuesday, October 14, 2025

Morning Transport

09:30 – 11:00 – Transport to Stracta Hotel (Hella) via bus and selected cars.
Departure: Reykjavik Center (Hallgrímskirkja).

https://maps.app.goo.gl/RZeXHF16NgjTGXW29

11:00 – 11:30 – Arrival at Stracta Hotel.
Luggage storage until check-in.

11:30 – 12:30 – Welcome Lunch at Stracta Hotel.

12:30 – 12:45

Welcome & Agenda
Morris Riedel & Gabriele Cavallaro

12:45 – 13:00

Scaling Data-Parallel Training – *Marcel Aach*
Scaling data-parallel training across hundreds or thousands of GPUs to reduce neural network training time goes hand in hand with an increase of the global batch size. However, it has been shown that training with batch sizes too large results in unstable training dynamics or leads to reduction of the generalization performance. This talk focuses on studying the phenomenon through the lens of the gradient noise scale metric.

13:00 – 13:45

Ecosystems for Data and Computing – *Dirk Baker*

Research practices all hinge on good scientific practices, not only for the benefit of the community but also the researcher themselves. Both research software and data engineering solutions have been developed at UFZ and HIFIS, enabling ad hoc and scalable research pipelines. This talk provides a brief overview of services and strategies in the interest of broadening the horizon a little bit. A bit of advice for projects and programming tasks is added in the mix.

13:45 – 14:00

HIPPO – Bringing EuroHPC to the Masses via DestinE – *Þorsteinn Elí*

HIPPO (HIgh Performance Processing Orchestration), a JSC-ESA collaboration within the DestinE framework, demonstrates the feasibility of modern workflow orchestration for Earth observation on EuroHPC resources. Leveraging Apache Airflow 3, HIPPO enables automated, reproducible, and scalable pipelines, from data ingestion to AI model retraining, using cron- and event-based scheduling, provenance tracking, and backfilling. I showcase an example of how Airflow 3 supports continual pre-training of geospatial models (e.g., TerraMind), enabling automatic updates to them with new satellite data every 6 months. Designed for both developers and non-technical users, HIPPO provides both guidelines to developers and ready-to-use workflows for tasks like embedding generation from Copernicus or HLS data. Through a demo of the Airflow UI, including failure recovery and logical date re-execution, we illustrate how HIPPO transforms complex HPC workflows into accessible, production-ready services, setting a new standard for reproducibility in Earth system science.

14:00 – 14:15

Large Language Models Can Self-Synthesize Training Data for Low-Resource Languages – *Mathias Stenlund*

A Large Language Model (LLM) learns about language and the world through pretraining on massive collections of text. Yet, this alone does not guarantee that the LLM will reliably follow user instructions. A method called instruction-tuning addresses this issue by further training the LLM on datasets containing pairs of instructions and desired responses. However, such datasets remain scarce, especially for low-resource languages. Recent studies explore generating synthetic instruction data as a cheaper and faster solution to this scarcity problem, but most methods are developed with only English in mind. We therefore ask ourselves: To what extent can these synthetic approaches also be applied to low-resource languages and what modifications need to be made to make them viable?

14:15 – 14:45 – *Coffee Break*

14:45 – 15:00

ValEU: A Systematic Framework for Evaluating European Cultural Alignment in Large Language Models – *Annika Simonsen*

I present ValEU, a framework for evaluating LLM alignment with European cultural values. Based on the 2017–2022 Integrated Values Surveys with responses from 156,658 participants across 92 countries, my co-authors and I developed an evolutionary question selection methodology for identifying questions with high shared cultural importance to EU member states. Our optimization procedure identifies 53 core value questions and develops a weighted cultural alignment evaluation based on kernel density estimation fitted on EU survey responses. Factor analysis reveals 15 distinct value dimensions explaining 58.2% of variance, including social liberalism, civic engagement, electoral integrity, and gender equality attitudes. Evaluating 13 language models reveals that model size correlates with alignment, with top performers achieving over 96% EU alignment. However, we find substantial sensitivity to evaluation format: reformulating questions from traditional multiple-choice into situational prompts causes alignment scores to drop dramatically. This finding highlights critical methodological challenges in cultural alignment assessment and suggests that LLM survey evaluations should be used with caution when interpreting results for deployed systems.

15:00 – 15:15

AI for Acute Respiratory Distress Syndrome – A Systematic Review – *Simon Fonck*

Acute respiratory distress syndrome (ARDS) is a serious clinical condition that occurs in mechanically ventilated patients and has a high mortality rate despite many years of research. Epidemiological studies have shown that a major factor in the high mortality rate is the late or lack of diagnosis of the disease and thus inappropriate treatment of patients. One way of addressing this underdiagnosis is the use of artificial intelligence (AI) methods, which are becoming increasingly popular in medicine. In the context of ARDS, various AI methods are also being published and discussed in scientific literature. In order to provide a comprehensive overview of the current state of research, we conducted a systematic review in which various literature databases were included. A total of 114 papers dealing with AI for ARDS were ultimately identified. Various task categories, such as classification, prediction and mortality prediction, were proposed. It can be seen that AI methods have good, replicable performance and other the potential to support physicians in this context. However, it also shows that the methods are not very comparable due to different data bases and a lack of documentation. This is also reflected in low reproducibility, as in the majority of studies neither data, data preparation nor the developed code is made publicly available.

15:15 – 15:30

Quantum Machine Learning for Earth Observation: An Overview – *Amer Delilbasic*
Quantum Machine Learning (QML) is an emerging interdisciplinary field that leverages quantum computing principles to address complex computational challenges. Traditional machine learning models face major challenges in Earth Observation (EO) due to the increasing volume and complexity of data. These challenges include computational intensity, energy consumption, and data management constraints. QML emerges as a promising paradigm for overcoming these barriers. This presentation offers an overview of the state of the art and new trends in QML application to EO.

15:45 – 16:00

Applications of Quantum Computing to Climate Modeling: Overview of the Quantum Readiness for Earth Observation Project – *Edoardo Pasetto*
The Quantum Readiness for Earth Observation (QR4EO) project is aimed at studying the current possible applications of Quantum Computing to Earth Observation use-cases. The project will specifically focus on Statistical Downscaling for climate modelling, a task in which one aims at getting a finer resolution image related to a map of some quantity of interest starting from a lower resolution image. The network used for the task is a hybrid Diffusion model employing both quantum and classical layers in the implementation. The talk will then focus on an overview of the use-case and how the implementation of the model differs from a purely classical one.

16:00 – 16:30 – *Coffee Break*

16:30 – 17:15 - Day 2 closing and round table discussions - PostDoc experiences- (Rocco Sedona, Chadi Barakat, Dirk Baker, Marcel Aach, Su Nan, Ehsan Zandi)

Day 2 Closing & Round Table Discussions – Postdoc Experiences
Participants: *Rocco Sedona, Chadi Barakat, Dirk Baker, Marcel Aach, Su Nan, Ehsan Zandi*

17:15 – 20:00 – Free Time
20:00 – 21:30 – Dinner at Hotel Stracta

Day 3 – Wednesday, October 15, 2025

09:00 – 14:45 – PhD Student Presentations

09:00 – 09:15

A Hybrid Training Strategy Involving Forward Gradient Descent for the Training of Deep Neural Networks – *Joseph Xavier Arnold*
Backpropagation remains the most widely used method to compute the exact gradients to train a neural network. However, it is computationally expensive besides being biologically implausible. Also, the computation of the forward pass followed by a backward pass hinders parallelization. The forward gradient algorithm, which approximates the gradient by computing the directional derivative in just the forward mode makes it a computationally efficient alternative to backpropagation. However, forward gradient being a noisy estimate of the actual gradient, the similarity between the forward gradient and the actual gradient decreases with the increase in the number of learnable parameters due to high variance. This makes forward gradient-based approaches impractical in the training of large models. We propose a hybrid training strategy that involves backpropagation and forward gradients to address the issue of sequential backpropagation of the gradients.

09:15 – 09:30

Scale Matters: How Spatial Resolution Impacts Remote Sensing–Based Landscape Complexity Assessment – *Liang Tian*
This study examines how spatial resolution influences the assessment of landscape complexity from satellite data. Using Sentinel-2, Landsat 8, and MODIS imagery, we compared texture- and land-cover-based metrics across multiple scales. Fine-resolution data from Sentinel-2 captured ecological patterns most effectively, showing the strongest correlations with biodiversity observations from NEON. The results highlight that aligning spatial resolution and analysis scale with ecological processes is essential for accurate landscape complexity assessment.

09:30 – 09:45

Beyond a Pixel: Identifying Landscape Features at a Sub-Pixel Level – *Surbhi Sharma*
In Remote Sensing, accurate identification of landscape features such as buildings, forests, rivers, roads, etc., is key to understanding ecological processes, monitoring environmental changes, and supporting sustainable land management. However, conventional landscape features mapping methods are often constrained by pixel-level resolution, which can obscure small or mixed features within heterogeneous landscapes. Even with openly available high-resolution datasets such as Sentinel-2, the accurate identification of landscape features is often affected by the discrepancy between the spatial resolution of the satellite images and the extent of the feature present at a pixel level. By applying a sub-pixel mapping approach, it is possible to look "beyond a pixel" to capture the hidden complexity of landscapes such as buildings surrounded by gardens, narrow water bodies, and small grasslands, offering new opportunities for environmental monitoring, resource management, and spatial decision-making.

09:45 – 10:00

Exploration of Spatial-Temporal Embeddings for Crop Yield Prediction– *Adriko Kennedy*

This study explores the use of spatial-temporal embeddings from recent geospatial foundation models for crop yield prediction. Building on baseline experiments with traditional regression and explainable models, we integrate Sentinel-2 and Sentinel-1 imagery with climate data (ERA5) to evaluate how pretrained models such as Clay, Prithvi-EO, SatMAE, and AlphaEarth capture yield-relevant dynamics across space and time. Using the Embed2Train framework, embeddings are extracted and compared to assess their ability to represent crop growth patterns and predict yield under varying environmental conditions. The results demonstrate the potential of foundation model embeddings to enhance large-scale, data-efficient, and explainable yield prediction workflows.

10:00 – 10:15

Enhanced Understanding of Crop Field Photos Through Visual Question Answering and Captioning – *Stefano Maurogiovanni*
The integration of ground-level data from field photographs and temporal data from Satellite Image Time Series (SITS) can provide a comprehensive understanding of crop development stages and the impact of climatic events on crop status. As part of the ESA-funded "Agrovision" project, this research aims at developing a vision-language model capable of generating informative descriptions of crop images and effective visual question answering (VQA). By providing rich and context-grounded descriptions of crop field pictures, the Agrovision VLM is envisioned as a powerful AI assistant for domain scientists working with agricultural data.

10:15 – 10:45 – *Coffee Break*

10:45 – 11:00

On the Generalizability of Artificial Neural Networks in Spin Models – *Su Nan*
The applicability of artificial neural networks (ANNs) is typically limited to the models they are trained with and little is known about their generalizability, which is a pressing issue in the practical application of trained ANNs to unseen problems. Here, by using the task of identifying phase transitions in spin models, we establish a systematic generalizability such that simple ANNs trained with the two-dimensional ferromagnetic Ising model can be applied to the ferromagnetic q-state Potts model in different dimensions for $q \geq 2$. We demonstrate that similar results can be obtained by reducing the exponentially large state space spanned by the training data to one that comprises only three representative configurations artificially constructed through symmetry considerations. We expect our findings to help improving the interpretability and simplifying the development of related tasks in (quantum) machine learning across disciplines.

11:00 – 11:15

Alternative Byte-Pair Encoding for DNA-Specific Classification Tasks with Neural Networks – *Josephine Busch*
Traditional tokenization algorithms used for Deoxyribonucleic Acid (DNA) classification tasks are optimized for Natural Language Processing (NLP) tasks and ignore key-characteristics of DNA sequence data. A specialized tokenizer for DNA data could utilize those characteristics to perform better on DNA classification tasks. This master's thesis describes the implementation and evaluation of HelixSwap, a novel tokenization algorithm based on Byte Pair Encoding (BPE). In two evaluation scenarios, Machine Learning (ML) models trained using HelixSwap are compared to models using BPE for tokenization. HelixSwap leads to an improved pretraining performance in all investigated use cases while using less compute resources on average. The models using HelixSwap achieve comparable results during fine-tuning to the BPE tokenizer models.

11:15 – 11:30

Hackathon presentation – *Ingólfur Hjörleifsson*

11:30 – 11:45

Remote Sensing Assisted Digital Twin Networks for 5G/6G Mobile Networks – *Ehsan Zandi*
In 5G and emerging 6G mobile networks, particularly above 6 GHz, accurate and adaptive Ray Tracer (RT) is an essential tool for creating a Digital Twin Network (DTN), i.e., a virtual replica of the physical network. DTNs allow rapid deployment in new scenarios without extensive measurement campaigns and can accurately simulate complex real-world environments, including highly intricate propagation channels. This capability is crucial for 6G technologies (e.g., millimetre wave, sub-THz communication, integrated sensing and communication). RT in DTNs is essential for the planning, optimization, and parameter tuning of 5G/6G radio access networks (RANs). Current RT engines neglect high level of detail (LoD) information about the scene such as street-level information or building facades and their material. This has a huge impact on the quality of ray tracer, particularly in mmW bands, where it can result in significant degradation of QoS. To cope with this shortcoming, we aim at integrating remote sensing data — including high-resolution satellite imagery, point clouds, and semantic scene information — into ray tracers such as Sionna.

12:00 – 13:15 – *Lunch Break*

13:15 – 13:30

Identifying Brain Imaging Biomarkers for Early Diagnosis of Parkinson-plus Syndromes – *Mengyu Li*

Parkinson-plus syndromes, also known as atypical parkinsonism, are a group of neurodegenerative diseases, including progressive supranuclear palsy (PSP), multiple system atrophy (MSA), and dementia with Lewy bodies (DLB). Many overlapping clinical features lead to misdiagnosis with AD, psychiatric problems, vascular dementia, or PD. Therefore, a major diagnostic challenge remains to identify and optimize the specific characteristics that can robustly separate individual dementias at early stages. Currently, there is no treatment for PSP, MSA, or DLB and they continue to be under-recognized. The generation of detailed and reliable evidence is the first step toward therapeutic development. This research project will develop novel MRI analysis tools that will help address these important challenges. In this project, a deep learning-based detailed automated brain segmentation and labeling algorithm specific for Parkinson-plus imaging biomarkers will be developed and evaluated. Based on brain morphology, this method computes the structural measures of the brain (volume and MRI) as imaging biomarkers, that robustly characterize the brain atrophy seen in DLB, PSP, and MSA. In the end, a pilot study on Parkinson-plus patients with a known diagnosis will be carried on and a comparison between our biomarkers and the current gold standard will be shown.

13:30 – 13:45

Coreset Selection for Scalable Remote Sensing: Energy Efficient Sampling Strategies on the Sen4Map Dataset – *Samy Hashim*

The exponential growth of high-resolution satellite imagery poses significant ecological challenges for training large-scale machine learning models in remote sensing. Coreset selection provides a promising solution by identifying representative subsets of data that preserve essential information while reducing redundancy and training time. In this work, we investigate clustering-based coreset selection methods (such as K-means clustering) on the Sen4Map dataset, a large-scale Sentinel-2 benchmark for land-cover and land-use mapping. We compare approaches from literature with our own methods, evaluating them in terms of model accuracy, energy usage, training efficiency, and carbon emissions on high-performance computing (HPC) infrastructure. It is expected that the results will demonstrate that dataset size can be decreased in the magnitude of 40–50% without a major reduction in performance, and that emissions will be reduced proportionally to the size reduction of the dataset.

13:45 – 14:00

Improving the Equivariance of Embeddings Using Image World Models – *Erik Scheurer*
Embedding datasets are increasingly used in Earth observation but often lack equivariance to geometric transformations such as rotation or scaling. We train a Joint-Embedding Predictive Architecture (JEPA) with a world-model-style predictor that learns to forecast embeddings of transformed image patches. On multispectral (HLS) data, conditioning on transformation parameters and fine-tuning improve equivariance. A class token reduces class-specific noise when trained on ImageNet1k which enables improved interpolation but shows no benefit on a non-classification dataset.

14:00 – 14:15

TN-PRESTO: Towards Onboard Geospatial Inference via Tensorized PRESTO – *Rui Wang*
We present a plan to develop TN-PRESTO, a compact geospatial foundation model aiming for onboard satellite inference. Building on PRESTO, a lightweight transformer pre-trained on multi-sensor pixel time-series, we exploit its small footprint and robustness to missing modalities to enable dependable, low-latency onboard processing. To further shrink memory and compute, we reshape PRESTO's attention and MLP weights into high-order tensors and apply low-rank tensor network decompositions (e.g., Tucker, Tensor-Train), producing factorized layers that retain core information while cutting parameters and inference times. We plan to validate TN-PRESTO on onboard-critical tasks—cloud masking, rapid flood/change mapping, and crop condition alerts—showing that it preserves accuracy while reducing computational resources. Next, we would analyze deployment viability on budget-constrained satellites by tying model needs to concrete payload specs—supporting our argument with hardware-backed evidence.

14:15 – 14:30

Quantum-Enhanced Consensus Clustering through Quantum Annealing and QAOA – *Daniele Franch*
Consensus clustering integrates multiple partitions into a unified solution, enhancing stability across clustering algorithms. However, ensuring global consistency remains computationally demanding. Quadratic Unconstrained Binary Optimization (QUBO) formulations provide a natural framework for this task, making consensus clustering amenable to quantum optimization methods such as quantum annealing and the Quantum Approximate Optimization Algorithm (QAOA). We extend the established QUBO formulation for correlation clustering with a refinement protocol that allows for improvement of suboptimal solutions. To address current quantum hardware limitations, we implement a subproblem decomposition strategy that iteratively resolves constraint violations on manageable subsets. Validation on benchmark and real-world datasets demonstrates competitive accuracy compared to established methods.

14:30 – 14:45

Quantum-Enhanced Variational Inference: Towards Practical Applications – *Hao Geng*
This presentation outlines the early stages of my doctoral research. After a brief introduction to my academic background and master's thesis, I will discuss my current focus, which builds on the recent theoretical work of Miyahara & Roychowdhury (2023), which demonstrates a quantum advantage in variational Bayes inference. My work seeks to explore how such quantum-enhanced approaches can be adapted for real-world applications. While the methodology is still under active study, the research aims to identify pathways by which theoretical advances in quantum variational inference may translate into practical computational benefits.

14:45 – 17:00 – Open discussions & networking session.

20:00 – 21:30 – Dinner at Hotel Stracta

Day 4 – Thursday, October 16, 2025

07:30 – 09:00 – Breakfast at Hotel Stracta Restaurant

09:00 – 10:00 – Check-Out

10:00 – 12:00 – Closing Session (*Morris Riedel & Gabriele Cavallaro*)
Summary of PhD Retreat objectives and challenge areas; discussion of executive summary and next retreat.

12:00 – 13:30 – Lunch at Hotel Stracta

14:00 – 15:30 – Transfer back to Reykjavik center

15:30 – End of Program