Contribution ID: **101** Type: **Oral Presentation**

# The Helmholtz Model Zoo: Enabling AI Model Sharing and Inference in the Helmholtz Cloud

*Friday, 20 March 2026 09:22 (22 minutes)*

The Helmholtz Model Zoo (HMZ) is a cloud-based platform that provides remote access to deep learning models within the Helmholtz Association. It enables seamless inference execution via both a web interface and a REST API, lowering the barrier for scientists to integrate state-of-the-art AI models into their research.

Scientists from all 18 Helmholtz centers can contribute their models to HMZ through a streamlined, well-documented submission process on GitLab. This process minimizes effort for model providers while ensuring flexibility for diverse scientific use cases. Based on the information provided about the model, HMZ automatically generates the web interface and API, tests the model, and deploys it. The REST API further allows for easy integration of HMZ models into other computational pipelines.

With the launch of HMZ, researchers can now run AI models directly within the Helmholtz Cloud, ensuring that all data remain within the association and that our data sovereignty is preserved. The platform imposes no strict limits on the number of inferences or the volume of uploaded data, while Helmholtz Virtual Organizations (VOs) enable fine-grained access control for specialized models. External researchers can also access HMZ through Helmholtz VOs upon invitation by a Helmholtz representative, facilitating collaborative research beyond the association's boundaries. Data uploaded for inference is stored within HIFIS dCache InfiniteSpace and remains under the ownership of the uploading user.

HMZ is powered by GPU nodes, hosted as part of the DESY Hamburg HPC cluster. Model inference is managed through the NVIDIA Triton Inference Server, ensuring efficient GPU utilization. The development and maintenance of HMZ are led by the Helmholtz Imaging Support Team at DESY, with support from Helmholtz Federated IT Services (HIFIS) and the Helmholtz AI platform. Hardware and implementation have been supported by funds from the Haicore initiative.

Our presentation will provide an overview of HMZ architecture and its integration into a professional HPC environment. It will also address the scientific foundations of selected models and emphasise the benefits of operating them entirely within the Helmholtz infrastructure.

**Primary author:** Dr FUHRMANN, Patrick (DESY/dCache.org)

**Co-authors:** WERNERS, Hans (Deutsches Elektronen-Synchrotron (DESY)); Dr EREN, Engin (Deutsches Elektronen-Synchrotron (DESY)); Dr HEUSER, Philipp (DESY)

**Presenter:** Dr FUHRMANN, Patrick (DESY/dCache.org)

**Session Classification:** Virtual Research Environment (VRE) - I

**Track Classification:** Track 5: Virtual Research Environment (including tools, services, workflows, portals, ... etc.)