The Helmholtz Model Zoo: A Cloud-Based Platform for AI Model Sharing and Inference in the Helmholtz Association

Hans Werners, Engin Eren, Patrick Fuhrmann and **Philipp Heuser**
Helmholtz Imaging, Deutsches Elektronensynchrotron - DESY, Hamburg
{hans.werners, engin.eren, patrick.fuhrmann, philipp.heuser}@desy.de

Abstract

The Helmholtz Model Zoo (HMZ) is a cloud-based platform enabling seamless sharing and inference of deep learning models across the Helmholtz Association's 18 research centers. By automating model deployment and providing both web and programmatic interfaces, the HMZ lowers technical barriers to AI adoption in scientific research. Integrated with Helmholtz infrastructure (Helmholtz ID authentication, dCache storage, DESY's HPC cluster with NVIDIA L40S GPUs), the platform ensures secure, scalable inference while maintaining data sovereignty. NVIDIA Triton Inference Server and Slurm manage GPU resources efficiently, supporting data-sets from gigabytes to terabytes. Virtual organizations enable fine-grained access control for specialized models. Launched in July 2025 in beta, the HMZ focuses on domain-specific applications, with future plans for model quality metrics and agentic capabilities.

Keywords: Model Zoo, AI, Inference, Deployment, Platform, Helmholtz

1 Introduction

While deep learning holds transformative potential for scientific research, its adoption is often constrained by deployment complexities, data sovereignty concerns, and limited access to specialized computing resources. The Helmholtz Association's 18 research centers face these challenges particularly acutely, where the demand for advanced AI tools frequently exceeds individual groups' technical capabilities.

The Helmholtz Model Zoo addresses these issues by providing a centralized platform that simplifies the entire AI model deployment and usage. By abstracting technical complexities, it enables researchers to focus on scientific objectives rather than infrastructure management. The platform serves as a bridge between computational experts developing sophisticated models and domain scientists seeking to apply them, thereby democratizing access to AI-driven analysis across diverse research disciplines.

2 Platform Design and Objectives

The HMZ was designed with three core objectives: simplifying model sharing, providing scalable inference infrastructure, and seamless integration with existing Helmholtz systems. Its architecture reflects these goals through several key design decisions.

At its heart lies an automated deployment pipeline where submitted models are automatically containerized and exposed through both web interfaces and REST APIs. This automation significantly reduces the technical overhead for model contributors while providing intuitive access points for end-users. The platform generates these interfaces dynamically based on model metadata, thus abstracting technical complexities.

Data security and sovereignty were paramount considerations. All processing occurs within the Helmholtz Cloud [https://helmholtz.cloud/] environment, with data stored in HIFIS's [Konrad et al. [2022]] dCache [Mkrtchyan et al. [2021]] InfiniteSpace system. This ensures compliance with institutional data policies while providing the scalability required for scientific datasets. The platform imposes no arbitrary limits on data volume or inference operations, supporting analyses from pilot studies to production-scale workflows.

3 User Access and Authentication Framework

Access to the HMZ is managed through the Helmholtz ID, the association's centralized authentication system. This system not only eliminates the need for separate credentials but also enables sophisticated access control through virtual organizations. These VOs allow research groups to securely include external collaborators while maintaining precise control over resource sharing. During the ramp up phase access is restricted to the 47,500 employees of Helmholtz and their collaborators.

The platform serves two user groups with distinct needs. Domain scientists interact with models through automatically generated intuitive web interfaces that require no programming expertise. For programmatic access, REST APIs enable integration with existing analysis pipelines. Model contributors, by contrast, submit their trained models through a standardized GitLab workflow, adapting provided Python templates to their specific requirements. While this process has been streamlined, it assumes basic Python proficiency to ensure model compatibility.

4 Technical Implementation

The HMZ's technical architecture leverages existing Helmholtz infrastructure to ensure both compatibility and scalability. The platform integrates the Helmholtz ID for authentication, institutional PostgreSQL databases for metadata management, and DESY's Maxwell HPC cluster for computational resources. This foundation minimizes deployment complexity while providing robust performance.

The user interface employs a modified JupyterHub deployment that spawns dedicated Flask servers for each session rather than traditional notebooks. This approach executes all jobs under institutional credentials, enabling seamless integration with existing account management systems for resource allocation and data access control. The JupyterHub also provides built-in authentication tokens for secure API access.

Data management relies on dCache, a high-performance storage system originally developed at DESY. This integration supports datasets from gigabytes to terabytes without arbitrary limits, with multi-protocol access via WebDAV and rclone. The dCache infrastructure ensures all user data remains under institutional control while delivering the performance required for scientific analyses.

The computational backbone consists of six dedicated nodes in the Maxwell cluster, each equipped with four NVIDIA L40S GPUs. NVIDIA Triton manages inference operations through a high-performance C++ architecture, efficiently distributing GPU resources across concurrent users. It interfaces directly with frameworks' native C++ engines, bypassing Python overhead.

Slurm handles job scheduling, allocating CPU resources for I/O and pre- or postprocessing. This architecture optimizes resource utilization through shared model instances, where a single loaded model serves numerous parallel users simultaneously, maximizing GPU memory efficiency without performance degradation.

Model submission begins with automatic GitLab project creation through the HMZ website. Contributors adapt a template repository to their model's requirements, after which CI/CD pipelines verify compatibility. Following technical review, approved models are automatically deployed. The built container images are deployed to CernVM File System (CVMFS) [Blomer et al.], a dedicated software distribution service mounted on Maxwell.

5 Current Status and Future Development

Since its July 2025 beta launch, the HMZ has onboarded an initial collection of domain-specific models tailored exclusively to scientific applications, deliberately excluding general-purpose LLMs which are provided through separate Helmholtz services [Strube]. The platform's full potential will be realized as adoption grows and the community contributes more diverse models.

The development roadmap includes implementing model quality indicators to help users evaluate performance, and integrating agentic capabilities to enable autonomous workflow operations. The team will also monitor usage patterns to optimize resource allocation, particularly for large-scale datasets.

6 Conclusion

The Helmholtz Model Zoo represents a significant step toward democratizing AI in scientific research. By automating complex deployment processes and integrating with existing infrastructure, it enables researchers across disciplines to leverage advanced AI tools without specialized technical knowledge. This bridges the gap between computational experts and domain scientists, accelerating data-driven discovery.

The platform's integration with Helmholtz systems ensures both immediate usability and long-term sustainability. As it evolves through community contributions and targeted enhancements, the HMZ is poised to become an essential resource for AI-driven research across the association. Future developments in quality metrics and agentic capabilities will further enhance its utility, solidifying its role as a cornerstone of Helmholtz's digital research infrastructure.

The Helmholtz Model Zoo is available at https://hmz-hub.desy.de. Support and feedback can be directed to support@helmholtz-imaging.de [Imaging].

7 Acknowledgments

We acknowledge the contributions of DESY IT, Helmholtz Federated IT Services (HIFIS), and Helmholtz AI teams whose expertise was instrumental in developing the platform. Special thanks to the HAICORE initiative for financing the GPU nodes powering the HMZ. This work utilized Maxwell computational resources at DESY, Hamburg, Germany.

References

J. Blomer, B. Bockelman, P. Buncic, B. Couturier, D.-F. Dosaru, D. Dykstra, G. Ganis, M. Giffels, H. Nikola, N. Hazekamp, S. Heikkila, S. Isgandarli, R. Meusel, S. Mosciatti, R. Popescu, J. Randall, T. Shaffer, S. Teuber, D. Thain, B. Tovar, S. Traylen, and D. Weitzel. The CernVM file system: v2.7.5. URL `https://zenodo.org/records/4114078`. Language: eng.

https://helmholtz.cloud/. Helmholtz Cloud. URL `https://helmholtz.cloud/`.

H. Imaging. Helmholtz Imaging. URL `https://helmholtz-imaging.de/`.

U. Konrad, T. Huste, U. Jandt, and T. Schlauch. Helmholtz Federated IT Services: Innovative Cloud and Software Services for Science. Feb. 2022. doi: 10.5281/zenodo.15235978. URL `https://zenodo.org/records/15235978`.

T. Mkrtchyan, K. Chitrapu, V. Garonne, D. Litvintsev, S. Meyer, P. Millar, L. Morschel, A. Rossi, and M. Sahakyan. dCache: Inter-disciplinary storage system. volume 251, page 02010, 2021. doi: 10.1051/epjconf/202125102010. Publisher: EDP Sciences.

A. Strube. Blablador. URL https://helmholtz-blablador.fz-juelich.de/url.