Helmholtz Foundation Model Initiative

Paul F. Jäger*, Stefan Bauer*, Stefan Kesselheim*, Fabian Isensee*, Rainer Kiko, Uwe Ohler, Oliver Stegle, Guido Grosse, Michael Bussmann, Frederik Tilmann, Peter Steinbach, Markus Götz, Guido Juckeland, Philipp Heuser, Ralf Mikut, Sören Lorenz, Mario Fritz, Jilles Vreeken, Klaus Maier-Hein, Lena Maier-Hein, Fabian Theis, Dagmar Kainmueller*

*coordination team

Mission

The Helmholtz Association with its 18 centers and large-scale facilities is a world leader in the generation of cutting edge research data. The challenge of processing this data at scale requires identifying and leveraging synergies throughout the entire data pipeline. This involves creating solutions that can be universally applied across various data sources and research tasks.

Recent advancements in AI research have given rise to a transformative paradigm designed precisely for such challenges: "Foundation models." These models are trained on vast and diverse datasets at scale, making them highly adaptable to solve a wide range of downstream tasks. This generalized approach has already demonstrated significant societal impact, evident in the form of user-facing AI applications like chatbots, revolutionizing interactions between humans and machines. Several recent examples showcase the potential of foundation models to also transform non-language-related research data into high-impact tools and publications, for instance RETFound, Segment Anything or IBM NASA Geospatial.

With its unparalleled data repositories across six research fields and the robust infrastructure of its five incubator platforms, the Helmholtz Association is uniquely positioned to forge cutting-edge foundation models. Our distinct advantage lies in the depth and diversity of our scientific data, which is beyond the reach of the datasets typically harnessed by large-scale AI in the tech industry.

Thus, we propose the **Helmholtz Foundation Model Initiative** (HFMI), a joint moonshot initiative of the Helmholtz research fields and the Helmholtz incubator. HFMI is a systematic approach to developing generalist AI models in a synergistic manner in various research domains. Specifically, HFMI's mission will be to provide an umbrella, infrastructure and systematic technical and conceptual support for identifying and developing foundation models in these various domains jointly with the respective researchers. These open-source models are anticipated to significantly expedite data analysis across all Helmholtz centers and are designed to be shared globally, ensuring that our association's pioneering role extends its benefits and leadership to the broader research community around the world.

In short, the following points describe the added value of HFMI:

- **Impactful applications.** From accelerating medical image analysis to revolutionizing environmental monitoring and supporting disaster management by extracting insights from remote sensing and *in situ* data, HFMI's potential applications promise profound impact.

- **Streamlined data analysis.** Foundational models streamline data analysis within and across research domains and centers. Their adaptability to a wide array of tasks eliminates redundancies and the costs associated with setting up separate analysis pipelines for each problem.

- **Significant reduction of annotation cost.** Large-scale data annotation for AI systems is a major bottleneck of current research endeavors. By leveraging foundational models, the Association can minimize this resource-intensive annotation process, leading to substantial financial savings.

- **Strong incentive for data curation.** HFMI acts as a tangible tool that incentivizes the practice of curating, sharing, and enhancing accessibility to valuable data within and across centers. This enhanced data culture unlocks the potential of Helmholtz data treasures and establishes a sustainable structure to nurture an expanding range of foundation models in the future.

- **Promotion of collaboration and interdisciplinarity.** The initiative serves as a tangible project that bridges the incubator's platforms and the various research centers. It fosters meaningful collaboration and interconnectivity, tying together distinct domains into a cohesive framework for the advancement of research, technologies, and applications.

- **Impact on the global community.** By emphasizing a commitment to open source principles, the Helmholtz Association secures a commanding global position as an innovative force in advancing research and technology. Through the development and dissemination of open source models that will be freely available, HFMI aims to extend its influence far beyond the Helmholtz community, fostering an expansive network of collaboration and driving progress on a worldwide scale.

These benefits collectively underline the transformative potential of the HFMI initiative, positioning the Helmholtz Association at the forefront of research and innovation, and driving collaboration, efficiency, and progress across its expansive network.

In the initial phase of the initiative, our focus is on individual domains—an approach that is crucial for building a suite of high-functioning foundation models. Recognizing generalization as a process that unfolds over time, we foresee a mid-term future where these models evolve into a cohesive system. This system will be characterized by the collective power of

foundation models transcending individual domain capabilities, laying the groundwork for a future where the interconnectedness of AI amplifies its utility and potential across disciplines.

To kick-off the HFMI initiative, we identified a diverse list of potential pilot foundations, where foundation models are expected to accelerate an entire research domain and generate high impact.

Key Strategies

HFMI is a unique moonshot initiative that coordinates and develops foundation models in the various research domains of Helmholtz. The following three key strategies will drive the success of the initiative:

Interdisciplinary team: After the successful establishment of the five platforms within the Helmholtz incubator for data science, the proposed initiative represents an ideal moonshot opportunity to show the collaborative value of these newly created structures combined with the established research fields of the association. In this new phase, HFMI will foster a constellation of project-specific teams, each **interweaving AI and domain expertise** to advance their unique research objectives. This ensures that AI model development and deployment are intimately connected with the data-rich contexts they serve. An integral part of domain expert involvement are the **data curators** who collect, refine, and manage the large-scale data sets, ensuring their AI-readiness and FAIR compliance.
Atop the projects sits the **HFMI Synergy Team**, a hub of coordination and strategic deployment, positioned to harness the strengths of the incubator platforms. These synergy team members are not confined to fixed locales but are instead dynamically aligned with **HI** and **HAI** for cross-project imaging and AI endeavors, **HIFIS** and **HAI** for computational infrastructure, and **HMC** for the principles of FAIR data stewardship. They serve as the connective tissue between individual projects and the broader infrastructure of the Helmholtz Association, including collaborations with external entities such as the NFDI.

Generalization as a process: The initial separation of domains is intentional, as we strongly believe that synergies across data sources and tasks need to be understood on a smaller scale first ("in-domain"). Similarly, and despite the success of multi-modal foundation models, we propose to initially focus on a single input modality. In the second step, foundation projects will aim to sustainably scale models to incorporate additional modalities and ultimately span the entire diversity of data in Helmholtz.

Acceleration through synergistic AI science: Foundation models, being a relatively new training paradigm, present numerous unanswered research questions. These include the extent to which an initial pre-training stage (before training of the domain-specific foundation) enhances performance, the ideal combination of data from related domains during training, and efficient language-based interaction with these domain-specific models. A dedicated team of Synergy AI scientists will tackle these questions. Insights gained will subsequently be applied to enhance the performance of HFMI foundation projects in years two and three of the initiative.

High-quality validation: Reliably assessing the added value of the developed foundation model is key for the success of this initiative. This requires an elaborate and specialized test-bed for generalization abilities, which can be **provided as part of the sister-initiative "HelmholtzNet"**. Further, while the development of foundations only relies on unlabelled data, validation will require strategies for high-quality annotations at scale. A dedicated strategy for high-quality validation also serves as a distinguishing factor among potential global competitive initiatives.

Implementation

Team Structure

The highly interdisciplinary initiative will comprise 3-5 Pilot Project Teams as well as an umbrella Synergy Team. The initiative will further be governed by the HFMI Steering Board.

HFMI Pilot Project Teams:

Following an analysis of and consulting from related successful moonshot initiatives like Deepmind's AlphaFold or KCL's RETFound, each HFMI Pilot Project is aimed to comprise around 7-8 FTEs located at the labs of the project PIs (e.g. "AI lead" and "Domain lead").

2-3 Data Engineers. Central to the creation of foundation models are vast, refined, and homogeneous data sets. Data Engineers are entrusted with the crucial responsibility of expeditiously and efficiently curating data from across Helmholtz centers as well as leveraging national and international data networks. Their tasks encompass the standardization of data formats across different devices and centers, the assurance of data quality, and rendering the AI-primed datasets accessible universally. The robustness of foundation models hinges significantly on this meticulous data curation process. In the broader spectrum, this role dovetails with Helmholtz's vision of a comprehensive, professional data curation strategy. Given the intricacies involved in understanding domain-specific data and its nuances, it is imperative that Data Engineers are stationed at the domain PI's lab to closely collaborate and tap into domain expertise.

2 AI Engineers. AI Engineers play a pivotal role in shaping the initiative's technical direction. They oversee the implementation and training of models, and they architect the core software framework, creating a robust platform that effectively integrates the innovative techniques introduced by AI Scientists. Collaborating closely with Data Engineers they provide direct feedback to refine the quality and format of datasets, ensuring that data is streamlined for AI tasks. Interactions with AI Scientists also ensure that the latest AI methodologies are precisely tailored to each research domain's specifics. Being situated at the AI PI's lab allows them immediate access to deep AI expertise, fostering the consistent infusion of cutting-edge

AI solutions. This role also presents a promising career avenue for HI/HAI consultants, emphasizing Helmholtz's dedication to talent development.

2-3 AI Scientists. AI Scientists are at the forefront of innovation, dedicated to exploring and experimenting with multiple research directions within foundation model research. By keeping a keen eye on the evolving landscape of AI techniques, they ensure the projects integrate the latest, most effective methods. Their tasks span from implementing to evaluating and refining the most recent training and adaptation strategies tailored to each project. Collaborating closely with the AI Domain Engineers ensures a smooth flow of pioneering methodologies into the core projects. Being based at the AI PI's lab ensures a fertile ground for their work, as being surrounded by a wealth of AI expertise is crucial for the diverse and experimental nature of their role.

HFMI Synergy Team:

To ensure the success of the initiative as a whole and leverage synergies across projects, additional core personnel in the form of 8 FTE is required.

HFMI Coordinators (no additional cost). The four driving Principal Investigators of HFMI work diligently to translate the strategic direction provided by the steering board into actionable plans, ensuring that the initiative's objectives seamlessly align with broader organizational aspirations. As coordinators, they not only monitor the initiative's progress and address high-level challenges but also guarantee the optimal allocation of resources. Externally, they represent the initiative, advocating its significance, sharing its accomplishments, and persistently exploring opportunities for expansion and collaboration.

1 Project manager. An overarching project manager is pivotal to the initiative, ensuring that synergies across various projects are effectively harnessed. Beyond orchestrating systematic virtual and on-site engagements such as meetings, workshops, hackathons, and retreats, the project manager will also be responsible for monitoring project milestones, facilitating resource allocation, and fostering a cohesive environment that promotes knowledge exchange and collaboration. This centralized role serves as a keystone, ensuring that all projects align with the initiative's broader objectives while addressing any challenges or roadblocks that may arise.

2 Computational engineers. Two dedicated computational engineers are essential to this initiative, given the complexity of training foundation models, which necessitates vast amounts of top-tier computational infrastructure. While state-of-the-art resources are available within the Helmholtz Association (e.g., at FZJ and KIT) and nationally (e.g., at hessian.ai, west.ai, etc.), it's imperative to not only identify the best machines and deployment strategies for each training session but also ensure dynamic, efficient access to these resources. This task demands specific expertise, as leveraging such high-performance computing infrastructures is intricate and requires specialized knowledge. Therefore, on-site personnel with this proficiency are indispensable. A second aspect of this task is to ensure a seamless interface

between the prototyping environment of developers and the large-scale HPC environment allowing for efficient and iterative monitoring and testing of models.

4 Synergy AI scientists. A dedicated team of Synergy AI scientists, strategically located at the labs of the HFMI coordinators, will address the pressing, fundamental questions about foundation models, many of which emerge directly from the challenges encountered in ongoing HFMI projects. This centralized positioning ensures a close synergy with overarching goals, facilitating rapid knowledge exchange and alignment with the initiative's objectives. These insights will not only deepen our understanding but will also be harnessed to enhance the performance of HFMI foundation projects in the second and third years of the initiative.

1 Data curation coordinator. The Data Curation Coordinator will oversee all data curation efforts across the initiative, ensuring comprehensive management and accessibility of datasets across centers. They will champion the application of FAIR data principles, ensuring that all data assets are Findable, Accessible, Interoperable, and Reusable. This role is pivotal in maintaining data quality and facilitating data-driven advancements within the initiative

HFMI Steering Board:

The HFMI Steering Board will be composed of one PI from each participating incubator platform. Additionally, each platform may invite one additional expert to the board. The board will provide crucial oversight and direction for HFMI, ensuring alignment of goals and resources across the involved centers. With representatives from each platform, the board can facilitate effective communication, coordinate collaborative efforts, and address challenges in a unified manner. Their collective expertise and leadership will be instrumental in setting strategic priorities, making informed decisions, and fostering a cohesive approach to advancing foundation models.

Foundation Project Structure

Each foundation project is set up to run for three years and comprises three project stages.

Foundation Stage (year 1)

In the first year, the focus centers on achieving the paramount objective: crafting a fully functional prototype foundation model. This entails a parallel effort in data curation and model development, culminating in large-scale model training. Moreover, this stage includes devising strategies to validate the model's benefits, accentuating its transformative potential. This validation process will involve labelled data curation for model fine-tuning and validation, ensuring that the model is rigorously tested and refined using high-quality, accurately labelled datasets. Integral to the foundation stage is the strategy to "**harvest the low hanging fruit**", ensuring that easily attainable gains are captured to facilitate early successes and build momentum for the project.

Readiness Stage (year 2)

The second year centers on enhancing the technological readiness of the prototype model. This stage involves transforming the model into a practical tool and rolling it out in the community, adhering to the open-source ethos of "**from Helmholtz for the world**." This approach ensures the model's accessibility and adaptability, fostering collaboration and innovation across various domains. The focus will be on deploying the model to address real-world challenges within and beyond the Helmholtz association, solidifying its role as a valuable asset in the scientific and technological landscape.

Visionary Stage (year 3)

The third stage aligns with a "high risk-high gain" philosophy and explores advanced concepts such as interactive features via language understanding. It also delves into the model's generalization across diverse data domains and emerging modalities like time-series and scalar data. Central to this phase is the "**Explain and Expand**" initiative, focusing on understanding what the model has learned and conducting an in-depth assessment of its behavior and trustworthiness. The Visionary Stage aims to push the boundaries of current foundation models and unveil novel capabilities, while also establishing a robust framework for understanding and enhancing the model's effectiveness, transparency, and ethical considerations.

Computational Infrastructure

The computational infrastructure at the centers KIT and FZJ is ideal for training the foundation models. The systems HoreKa and JUWELS Booster represent the state-of-the art in computational infrastructure for Artificial Intelligence application. Access to both systems are established as Helmholtz AI Compute Resources (HAICORE) and already in use by many scientists, including most of the groups mentioned in the list of pilots below. Scientists from the Helmholtz association are eligible for applying for large computational resources of hundreds of thousands of GPU-Hours in the context of Nationales Hochleistungsrechnen (NHR) or the Gauss Centre for Supercomputing (GCS). Typically, such applications will be written during the application process of potential projects or the initial phase of approved projects. Starting late 2024, JUPITER, the first European Exascale Supercomputer will be available via a GCS application procedure. With more than 10.000 GPUs of the newest generation, this machine can sustain even the largest model developments.

Prospective project formation and selection process

With the rapid progress in AI, a swift initiation of projects is crucial for the success of this initiative. To quickly set HFMI in motion, while effectively utilizing Helmholtz resources and pinpointing the most impactful projects, a streamlined multi-step process will be executed:

1. A "HMFI Call for Pilot Foundations" will be released within the Helmholtz Association. The call will comprise two parts, aiming at Domain scientists and AI experts, respectively.
2. HFMI facilitates the matching of Domain- and AI experts where needed. To this end, domain scientists may, optionally, pre-submit only the domain-focused part of the proposal. This will serve as basis for a matchmaking event with AI experts, organized by HFMI.
3. An independent panel of reviewers will then select the standout projects for the inaugural funding round.

List of potential pilot foundations

We identified a diverse list of potential pilot foundations, where foundation models are expected to accelerate an entire research domain and generate high impact::

Helmholtz Plankton Foundation

"The plankton imaging community at Helmholtz spans five centers and millions of acquired images per year from a diverse set of devices. Equally, the relevant downstream tasks range from classification at varying granularities, to trait extraction or instance segmentation. Thus, our domain suffers from a severe data processing bottleneck and in return poses an enormous synergistic potential to be leveraged by a foundation model."

Rainer Kiko, GEOMAR

Helmholtz Airborne Optical Imaging Foundation

"High-resolution Airborne Optical Imaging (AOI) data is widely acquired within several Helmholtz centers using airplane and drone platforms. Flagship sensors such as the Modular Aerial Camera System (MACS) from DLR are used to capture dozens of TB with millions of multispectral images onboard AWI research aircraft in rapidly changing polar regions every year. Drone-based MACS is being employed for rapid response and mission support in coping with natural disasters , requiring near-real time data processing and automated analysis. Additionally, the rapidly evolving fleet of drone platforms in Helmholtz and their various optical imaging systems are acquiring millions of images annually of natural ecosystems, cities and infrastructure, and agricultural landscapes. Automated processing and analysis of vast amounts of data across diverse platforms and optical sensor systems requires a Helmholtz AOI Foundation Model to rapidly transform big AOI data into actionable knowledge, for example identifying points of interest and using this knowledge to guide the drone and adopt the operation of MACS.."

Guido Grosse, AWI; Heinz-Wilhelm Hübers, DLR

Helmholtz Plasma Foundation

"In FB Matter, novel plasma accelerators play an extremely important role. With the help of high-power lasers, matter is transformed into a plasma of electrons and ions and such strong fields are generated that particles are accelerated to high energies on a few centimeters instead of kilometers. Within the ATHENA platform, these novel accelerators for ions and electrons are being intensively studied. The highly nonlinear dependence of laser and plasma properties makes optimization of the accelerators for applications an important field of research. In addition to a large amount of experimental data, there is also a large variety and volume of simulational data from exascale simulation codes such as PIConGPU. Linking all these data from experiments and simulations and the different plasma accelerator platforms at Helmholtz in a foundation model would be unique in the world. In addition to data from accelerator research, this foundation model can also contain data from matter under extreme conditions, such as those produced and studied at the Helmholtz-International Beamline for Extreme Fields at the European XFEL, at HI Jena and at GSI. Thus, the Foundation Model can help answer questions from a variety of application areas ranging from astrophysics and fusion to materials research and radiation therapy."

Michael Bussmann / Peter Steinbach, HZDR

Helmholtz Human-Anatomy Foundation

"Current AI models for radiological analysis are mostly trained on individual body parts and imaging modalities. I am convinced that collecting and curating the vast quantity of available radiological data such as PET, MRI, or CT images of the entire human anatomy will create a foundational understanding that can catalyze a plethora of clinically relevant applications.."

Klaus Maier-Hein, DKFZ

Helmholtz Electron Tomography Foundation

"Electron tomography enables nanometer-scale visualization of molecules and cellular features inside cells. High noise levels and complexity of biological systems can make it challenging to extract this information. An extendable foundation model for analyzing electron tomographic data, will provide a deeper understanding of the molecular mechanisms underlying both health and disease"

Mikhail Kudryashev, MDC

Helmholtz Genome Foundation

"For the past decades, probabilistic language models have been the leading approach to parse and annotate genomes, including the human genome. The rise of large language models opens a genuine new direction to identify and annotate functional DNA sequence elements, and to extract dependencies between interacting elements that are far apart in the genome. The resulting models hold outstanding potential to advance the ability to predict phenotypes from patients' genomes, and solve common and rare genetic disease cases in data-driven manner. The genome foundation will also provide the basis to generalize and translate between modalities, e.g. from genome to health record. Finally, by extending the scope beyond human genome data, the genome foundation will also catalyze research on bacterial and virus genome sequences, with implications in a wide range of use cases, including the monitoring and prediction of bacterial outbreaks."

Uwe Ohler, MDC / Oliver Stegle, DKFZ

Helmholtz Cellular & Spatial Genomics Foundation

"Molecular profiling of biological samples has entered a new phase, in which the activity of genes or their switches can now be measured for individual cells across millions of observations, and often also in their spatial context. First transformer models for single cell variation are being built, including in Helmholtz. Adding the spatial angle for tissue state quantification, the impact expands further - the resulting data can be compared to images, with thousands of channels (genes) albeit at lower resolution and massive missing data. Spatial -omics data is expected to be the premier diagnostic approach to extract information from biopsies and provide interceptive, personalized healthcare. In our context, it provides a missing link between molecular and histological image data and the genome sequence. It will thus provide an ideal testing ground on how to combine or integrate multiple foundation models sketched above. Ultimately we expect a molecular biology foundation model to be multi-scale, spanning the genome foundation up to the cellular, tissue and organ level."

Joachim Schulze, DZNE / Fabian Theis, HM

Helmholtz Smart Grid Foundation: "AlphaGrid"

"The energy grid is rapidly transforming towards an increased share of renewable and clean energy sources. Yet, their integration is challenging due to their volatile nature, e.g. wind or solar, and dynamic participation. As a result grid operation is becoming increasingly complex and costly. The aim of this foundation is to create a digital twin of the electrical grid and an AI reinforcement agent to guarantee safe and robust operation. The foundational model will rely on publicly available and multi-modal data of the grid's graph topology, congestion, production and consumption."

Markus Götz, KIT

Helmholtz Atmospheric Dynamics Foundation

„Large-scale AI models are rapidly transforming weather and climate modeling with AI-based forecasting tools that show better performance than operational weather models in many aspects. With AtmoRep, a first abstract foundation model of atmospheric dynamics has been built. AtmoRep has demonstrated very good skill in a range of zero shot applications and in more difficult tasks, which required finetuning. Using a foundation model for the atmosphere opens up many new opportunities for research and for novel relevant weather and climate applications, including enhanced analyses of atmospheric observations, compression of huge climate model datasets, spatial downscaling of weather forecasts to achieve more local predictions, interpolation between climate scenarios, and more."

Martin Schultz, FZJ

Helmholtz Synchrotron Radiation Tomography Foundation

Tomography with synchrotron radiation is a growing field where organic and inorganic matter is imaged in 3D and 4D reaching unmatched spatial and temporal resolution. This yields a very heterogeneous collection of volumetric data, where diverse samples of different size and composition with impact in material science, medicine, life science, geology, and even archaeology are studied. Within the Helmholtz Association such data is collected by a broad community of user groups at Hereon (DESY), KIT, and HZB. The diversity of the tomograms with respect to the measured samples, the used x-ray energy, the contrast mechanisms as well as the sample environments (in situ/in vivo/operando), necessitates usually the specific retraining of any used neural network per sample. Thus pooling the available data from more than 10 years of measurements at PETRA III, from KIT and HZB for the training of a foundation model will be extremely beneficial for the downstream data-analysis pipelines. Today manual annotation of data leads to months and years spent on the analysis, which will be significantly more efficient by the use of a foundation model. The availability of such a foundation model will not only benefit the huge number of external academic and industrial users at the Helmholtz facilities, but will find its application also at all major synchrotron radiation facilities across the world.

Julian Moosmann, Hereon / Philipp Heuser, DESY

Helmholtz Brain Foundation

Decoding the structural and functional organization of the human brain is the key to a deeper understanding of its underlying working mechanisms, providing the foundation for the development of novel therapies for neurodegenerative diseases, as well as the development of brain-inspired technologies. Today's research on the brain's organizational principles relies on the analysis of large multimodal and multiscale imaging datasets that capture the

variability of brain organization. A foundation model for the human brain is the next step towards establishing a unified framework for multimodal data integration and analysis across scales.

Christian Schiffer, FZJ

Helmholtz Natural Hazard from Ground Motion Foundation

Sudden geohazards such as earthquakes, volcanic eruptions, landslides, flash floods etc. cause enormous loss of life and property. They cause seismic disturbances of the ground as they are ongoing, and in some cases still poorly understood preparatory processes can be observed in seismic data. Machine learning has revolutionised the treatment and analysis of seismic data, however, many models exist for different tasks and their transfer to other settings poses challenges. Hundreds of TB of seismological and GNSS time series from tens of thousands of sites are available within the databases of Helmholtz institutions and their partners, having recorded hundreds of thousands of these events. Emergent new technologies such as fibre-optic sensing enable sampling at high spatial resolution, allowing new insights but also presenting the challenge of a large data set with sparse labelling. The time is ripe to provide the community with a large-scale foundational model that will deliver enhanced early warning capability, revolutionise our ability to analyse patterns, and carry the potential to identify new preparatory phenomena. This foundational model will benefit from inclusion of multi-modal datasets such as remote sensing data in a later stage, and can straightforwardly transferred/merged to a closely related use case, the understanding of geothermal systems for geoenergy production, a crucial component of the energy transition.

Frederik Tilmann, GFZ