HANAMI

HPC ALLIANCE FOR APPLICATIONS AND SUPERCOMPUTING INNOVATION: THE EUROPE - JAPAN COLLABORATION

Funded by the European Union

EuroHPC Joint Undertaking

This project received funding from the European High Performance Computing Joint Undertaking (EuroHPC JU) under the European Union's Horizon Europe framework program for research and innovation and Grant Agreement No.101136269. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or EuroHPC Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.

DELIVERABLE D5.2

Biomedical halftime report

Funded by the European Union
EuroHPC Joint Undertaking

Project Title	Hpc AlliaNce for Applications and supercoMputing Innovation: the Europe - Japan collaboration
Project Ref	EuroHPC International Cooperation (HORIZON-EUROHPC-JU-2022-INCO-04)
Project Acronym	HANAMI
Project Number	101139786
Type of Action	HORIZON JU Research and Innovation Actions
Topic	HORIZON JU Research and Innovation Actions
Starting Date of Project	2024-03-01
Ending Date of Project	2028-02-28
Duration of the Project	36 months
Website	http://hanami-project.com/

Work Package	WP5
Task	
Lead Authors	José Carbonell-Caballero (JCC), Oleksandr Krochak (OK), Mario Rüttgers (MR), Marco Ruscone (MR), Laxmana Yetukuri (LY),
Contributors	Miguel Vazquez (MV), Erik Lindahl (EL), Alfonso Valencia (AV), Andreas Lintermann (AL)
Peer Reviewers	
Version	1.0
Due Date	31/08/2025
Submission Date	31/08/2025

Dissemination Level

x	PU: Public
	SEN: Sensitive – limited under the conditions of the Grant Agreement
	EU-RES. Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON. Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC. Classified Information: SECRET UE (Commission Decision 2005/444/EC)

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EuroHPC Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.

Version History

Revision	Date	Editors	Comments
0.1	12/06/2025	JCC	Initial draft
0.2	18/07/2025	AK	Finalised JSC contribution (Project 3)
0.3	21/07/2025	LY	Contribution for project 2
0.4	28/07/2025	MR	Contribution for project 2
0.5	29/07/2025	JCC	Contribution for project 2
0.8	31/07/2025	JCC	Harmonizing styles
0.9	31/07/2025	AF	General evaluation
1.0	01/08/2025	JCC	Final draft before internal review
1.2	28/08/2025	JCC	Final documen

Glossary of Terms

Item	Description

Funded by the European Union
EuroHPC Joint Undertaking

Executive Summary

This midterm deliverable presents coordinated progress across the HANAMI life sciences work package (WP5), advancing personalised medicine through a combination of multiscale modelling and high-performance Computing (HPC). First, at the molecular scale, exascale biomolecular simulations have been enhanced by integrating the ExaFMM library into GROMACS. Second, at the cellular scale, genome analysis pipelines combining somatic mutation simulations, Boolean models, and AI tools have been developed and validated on leading HPC systems to support cancer research. Finally, at the flow scale, AI-augmented computational fluid dynamics simulations of respiratory flow and viral transport are performed to improve the success of rhinology treatment as well as to assess patient-specific infection risks. These efforts form a unified framework that leverages international collaboration and HPC to enable predictive, personalised health modelling.

Funded by
the European Union
EuroHPC
Joint Undertaking

Table of Contents

Funded by the European Union
EuroHPC Joint Undertaking

Funded by the European Union
EuroHPC Joint Undertaking

Table of Figures

Funded by the European Union
EuroHPC Joint Undertaking

1. Introduction

This deliverable presents the mid-term progress of three interconnected subprojects in WP5 within the HANAMI initiative, which aims to advance multiscale biomedical modelling through HPC and artificial intelligence (AI). The activities described scalable simulations across molecular, cellular, and organ levels, supporting the development of personalised medicine while fostering strategic international collaboration with Japanese institutes

Each presented project addresses distinct but complementary scientific and technical challenges. Project 1 focuses on replacing conventional long-range electrostatics methods in molecular dynamics with a scalable Fast Multipole Method (FMM). This effort involves integrating ExaFMM into GROMACS to enable exascale-capable simulations of cell-sized molecular systems. The work is carried out in collaboration with KTH and the Yokota Lab at ISCT. On the other hand, Project 2 integrates Boolean modelling, somatic mutation simulations, multiscale cellular simulations, and Large Language Models (LLMs) to advance the study of cancer biology. These developments are being carried out in close collaboration with the PRIME Institute and RIKEN. Notably, the containerised workflows under development are built around cellular simulators that are currently being refactored to fully exploit exascale computing architectures. Finally, Project 3 applies AI-assisted computational fluid dynamics (CFD) to simulate respiratory airflow and assess infection risk. This includes porting and optimizing the m-AIA code for heterogeneous HPC architectures (e.g., JURECA DC and Fugaku), and performing simulations that combine internal nasal and external geometries for comprehensive trans-nasal airflow analysis. This work is conducted in partnership with JSC, RIKEN R-CCS, and Kobe University.

Collectively, these projects contribute to HANAMI's mission of developing exascale-ready computational tools for biomedical research. This deliverable provides an overview of the current status, key achievements, and ongoing collaborative efforts, laying the groundwork for the second phase of the project.

Funded by the European Union
EuroHPC Joint Undertaking

2. Project 1: Exascale electrostatics & machine learning to enable molecular dynamics of cell-size systems

Prof. Berk Hess, Prof. Erik Lindahl, and Dr. Muhammad Umair Sadiq from KTH collaborated with Prof. Rio Yokota and Jiamian Huang from the Institute of Science, Tokyo (ISCT), Japan, on exascale electrostatics using ExaFMM; a high-performance Fast Multipole Method (FMM) library. The focus was on joint research, regular knowledge exchange, and scalable FMM development.

Below is an overview of the progress made on the project.

2.1 Introduction

The plan under the HANAMI project set out to address the parallel scaling limitations of long-range electrostatics in molecular dynamics by replacing the traditional PME (Particle Mesh Ewald) method with the Fast Multipole Method (FMM), which offers linear scaling and avoids the communication bottlenecks associated with 3D FFTs. A central objective was to develop high-performance CPU and GPU implementations of ExaFMM and integrate them with GROMACS through a general, modular, and future-proof interface. The project also aimed to exploit GROMACS's support for MPMD (Multiple Program, Multiple Data) execution to facilitate scalable MPI-based integration of FMM, improving performance on modern supercomputing architectures.

Significant progress has been made toward these goals. A multithreaded CPU version of ExaFMM has been successfully integrated with GROMACS and utilised for initial validation and performance testing, demonstrating energy conservation comparable to PME at a lower computational cost. A flexible and extensible FMM interface has been developed in GROMACS, supporting external FMM backends and multiple interaction configurations. Crucially, the development of the MPI-enabled version of ExaFMM is actively ongoing, with the design focusing on scalable domain decomposition and efficient communication patterns compatible with GROMACS's parallel infrastructure. This work is central to enabling exascale molecular dynamics and remains a key focus in the next development phase. GPU support is being developed in parallel, maintaining alignment with the original plan and preparing for deployment across diverse HPC platforms, including future portability to GENESIS.

The collaboration with the Japanese team, led by Prof. Rio Yokota and Jiamian Huang at the Yokota Lab, has played a key role in driving core developments in ExaFMM, especially in MPI scalability and GPU-focused advancements.

This report outlines the collaborative progress under the HANAMI project. **Section 2.2** covers the knowledge exchange between KTH and ISCT. **Section 2.3** highlights in-person visits and technical discussions. **Section 2.4** details development progress on ExaFMM and its integration with GROMACS. **Section 2.5** describes the new FMM interface in GROMACS.

Funded by the European Union
EuroHPC Joint Undertaking

Section 2.6 presents initial accuracy and performance results, and **Section 2.7** summarises the overall progress and future plans.

2.2 Knowledge Exchange for FMM Implementation between KTH and ISCT

The first phase of this project aimed to implement ExaFMM with the long-term goal of approaching the performance of PME (Particle Mesh Ewald), a widely used method for long-range electrostatics in molecular simulations. We collaborated closely with Prof. Rio Yokota and his team at the Yokota Lab, developers of the ExaFMM library, through **biweekly Zoom meetings** focused on knowledge exchange and aligning development efforts. These meetings also included discussions on PME, its performance characteristics, and future possibilities for FMM as a scalable alternative.

A shared code repository facilitated active contributions from both KTH and ISCT, with work divided and reviewed collaboratively before merging into the main branch. Technical decisions benefited from the insights and experience of both Prof. Berk Hess and Prof. Rio Yokota, helping guide the project toward more effective implementation strategies.

2.3 Scheduled Collaborative Visits

As part of the ongoing collaboration between KTH and the ISCT, the following in-person visits are planned to support technical exchange and integration efforts:

- **July 2025: Dr. Muhammad Umair Sadiq** visited Japan to work directly with the Yokota Lab on the work related to ExaFMM. Key activities during the visit included:
    - Visit to the Yokota Lab for direct collaboration on FMM-related development.
    - Meetings with researchers working on ExaFMM in the Yokota Lab.
    - Technical discussions with Prof. Yokota, including review of the current ExaFMM implementation and identification of improvement areas.
    - Discussion of integration plans for ExaFMM with GROMACS, along with the ongoing effort on the FMM interface in GROMACS.
    - Consideration of possible future collaborations between Europe and Japan in alignment with the HANAMI project.

- **Future visits: Prof. Berk Hess and Prof. Erik Lindahl** are scheduled to visit Japan to continue collaborative development and planning.

2.4 Progress on Scalable Electrostatics Infrastructure

Significant progress has been made on both GROMACS and ExaFMM integration. Berk Hess implemented support for FMM-based direct interactions within GROMACS, leveraging its

Funded by the European Union
EuroHPC Joint Undertaking

highly optimised short-range electrostatics kernels developed over decades. This allows GROMACS to handle all short-range calculations when FMM is enabled.

On the Japanese side, Prof. Yokota and his research team, including Jiamian Haung, have focused on a new implementation of the FMM known as Kernel-Independent FMM (KIFMM). The main advantage of KIFMM is its generality; it does not rely on problem-specific analytical kernels such as those for Laplace, Stokes, or Helmholtz equations, making it applicable to a wide range of problems. This marks the team's first attempt at implementing KIFMM, building on their prior experience with various other FMM formulations. Another key benefit of KIFMM is that most computations can be expressed as matrix–matrix multiplications, enabling efficient use of GPU tensor cores and improved utilisation of HPC infrastructure.

Dr. Muhammad Umair Sadiq also focused on contributing to ExaFMM and integrating it with GROMACS for computing long-range electrostatics. The goal is to replace PME with a scalable alternative by linking ExaFMM into GROMACS while retaining GROMACS's fast short-range kernel. Contributions from the KTH side in ExaFMM include reducing memory usage, optimising matrix operations, fixing bugs, implementing a multistep FMM variant, and adding a CMake option to separate CPU and GPU builds for smoother integration with GROMACS.

For scalable FMM, achieving a functional MPI-enabled implementation is a key objective. To support this, we conducted extensive discussions on the specific data ExaFMM requires from GROMACS, particularly the structure and distribution of particle coordinates and charges needed for domain decomposition and inter-process communication. After validating the accuracy and performance of the serial version, which showed promising results, MPI development was initiated on the ISCT side. A preliminary MPI-enabled implementation is now in place, capable of handling domain partitioning and distributed computation, although further tuning and optimisation are needed to achieve full scalability.

Overall, on the ExaFMM side, we now have three working versions ready:

- A multithreaded CPU-based FMM implementation
- A GEMM-based GPU version
- A preliminary MPI-enabled version

2.5 Progress on FMM Interface in GROMACS

A flexible and extensible interface has been developed in GROMACS to support integration with various Fast Multipole Method (FMM) implementations, including ExaFMM. The design was shaped by ongoing input from both Japanese collaborators and German research groups working on FMM. Development efforts have been led by Dr. Muhammad Umair Sadiq, with continuous feedback from the German and Japanese sides guiding the direction and refinement of the implementation.

GROMACS now includes an option to enable support for external FMM libraries. Input formats compatible with well-known FMM implementations such as **FMSolvr** (by Jülich and Göttingen groups) and **ExaFMM** (by Yokota Lab) are accepted, even if the external library itself is not compiled. The interface allows FMM to be used for long-range electrostatics while offering flexible configuration options for short-range interactions. Users can choose between

Funded by the European Union
EuroHPC Joint Undertaking

using GROMACS's highly optimised short-range kernels, refined over decades, or letting the FMM implementation provide its own. Achieving comparable performance to GROMACS's short-range kernels within FMM remains a challenge, but the interface makes it possible to leverage GROMACS's strengths where appropriate. This setup enables testing and combining different short and long-range interaction kernels, allowing users to evaluate and choose the configuration, whether GROMACS or FMM-provided, that yields the best performance for a given system or architecture

The interface is designed to be general and modular, making it straightforward to extend support to additional FMM implementations beyond those currently integrated.

Moreover, the interface validates all necessary inputs before they are passed to a specific FMM implementation. The plan is to provide essential information, such as particle positions, charges, cell assignments, and tree rebuild triggers, to the FMM backend. GROMACS will also indicate whether the FMM tree should be rebuilt in a given step or retained from previous steps. This ensures consistency between the spatial grids used by FMM and GROMACS and helps maintain smooth and reliable force computations throughout the simulation.

All merge requests related to this interface are thoroughly reviewed by the GROMACS development team before integration, with a couple still under review and undergoing refinement.

The interface is now near completion and is planned to be part of the **GROMACS 2026.0** release.

Ongoing implementation of FMM Interface in GROMACS:
https://gitlab.com/gromacs/gromacs/-/tree/main/src/gromacs/fmm?ref_type=heads

Work on Integration of ExaFMM with GROMACS:
https://gitlab.com/mumairsadiq/gromacs-exafmm/-/tree/mus-fmm-exafmm-backend-multistep

2.6 Preliminary Accuracy and Performance Tests

Currently, the multithreaded CPU version of ExaFMM is integrated with GROMACS (link) and has been used for initial accuracy and performance testing. GPU and MPI versions are under active development and in the process of being integrated, but are not yet ready for evaluation. Initial tests were performed using a simple test system: an 8.5 nm water box containing 61,440 particles.

Findings are as follows for CPU-based multithreading implementation:

- The FMM implementation demonstrates high accuracy, achieving energy conservation comparable to PME (on the order of 10^{-3}) when using multipole order 4 and two direct neighbours, with short-range interactions handled by GROMACS (Figure 1). This is notable because FMSolvr (https://doi.org/10.1021/acs.jctc.0c00744), a well-known FMM implementation,

Funded by the European Union
EuroHPC Joint Undertaking

reports similar PME-level accuracy only at multipole order 8. Our findings indicate that a lower multipole order is sufficient when two direct neighbours are included, especially given the support of GROMACS's highly optimised short-range kernels, which further enhance performance.

- The same experiment was also performed with FMM handling both short-range and long-range electrostatics, while GROMACS computed only the Lennard-Jones interactions. Results from this setup indicate that the FMM direct kernel is approximately 20× slower than GROMACS's highly optimised short-range routines, supporting the choice to retain GROMACS for direct-space calculations. It's worth noting, however, that GROMACS kernels are tightly integrated into its internal execution machinery, while ExaFMM and other FMM libraries operate as external components. Some performance loss is expected from this separation, but even accounting for that, the ~20× gap remains a significant difference, and the choice of GROMACS direct kernels so far is superior.

- The MPI-enabled version of FMM is currently under active development, with Prof. Rio Yokota, Jiamian Huang, and their research team leading the effort. While a working version is ready, its communication pattern still requires optimisation, a topic we have discussed in detail. Further work is also needed on its integration with GROMACS. Nevertheless, since FMM relies primarily on local communication, unlike PME, which involves global all-to-all exchanges, it is expected to scale more efficiently on large parallel systems.

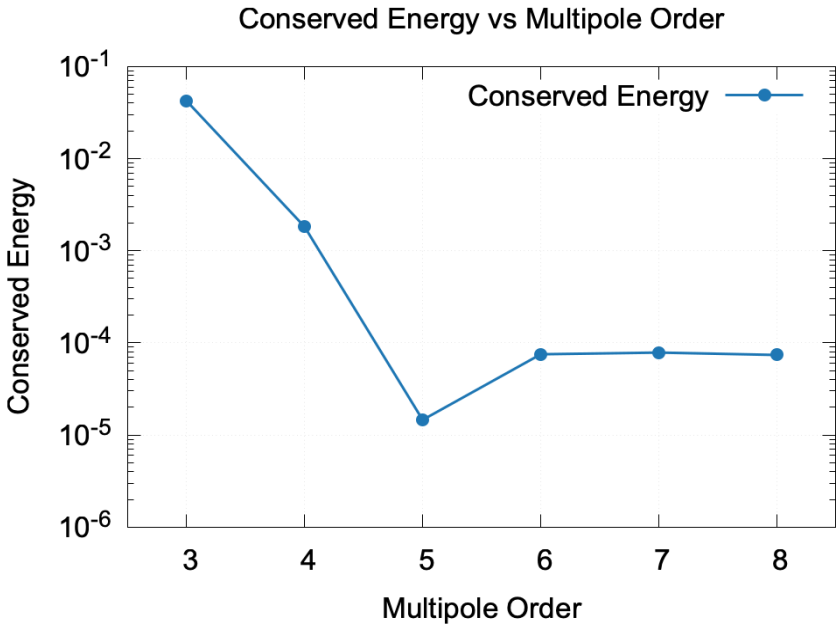


Figure 2.1: Energy conservation for 10,000 steps using ExaFMM (long-range) with GROMACS direct kernels for various multipole orders

2.7 Summary and Future Plans

Funded by the European Union
EuroHPC Joint Undertaking

- Substantial progress has been achieved in integrating ExaFMM with GROMACS for scalable electrostatics. A multithreaded CPU version is fully integrated and has demonstrated promising accuracy and performance, achieving PME-level energy conservation at lower multipole orders (of four). GPU and MPI versions are under active development, with correctness verified and integration ongoing.

- The newly developed FMM interface in GROMACS is general, modular, and nearly complete, enabling support for multiple FMM implementations and flexible configuration of long- and short-range interactions. Early benchmarks confirm the advantage of retaining GROMACS's optimised short-range kernels due to their superior performance.

- The MPI-based FMM implementation, led by the Yokota Lab, is still under active development and requires further optimisation and integration with GROMACS. Owing to its inherently local communication pattern, it holds the potential for better scalability compared to PME, which relies on global all-to-all communication.

- The full implementation is expected to be completed within six months, followed by tuning and scalability improvements. A joint publication with Japanese collaborators is planned. The next phase of the project, focused on applying machine learning for cell-scale molecular dynamics, is set to begin in early 2026.

Funded by the European Union
EuroHPC Joint Undertaking

3. Project 2: Development of genome analysis pipelines for Personalized Medicine

3.1 Introduction

A central objective of our project is to foster international collaboration with Japanese research teams by exploring the integration of molecular dynamics simulations, Boolean modelling, multiscale cellular simulations, and Large Language Models (LLMs). Our aim is to apply these complementary approaches to cancer research, with a particular focus on personalised medicine. To support this, we are prioritising the development of novel HPC solutions that lay the groundwork for advanced interdisciplinary research. These efforts have been significantly accelerated through dedicated collaborative events, including the BSC–RIKEN workshop and a follow-up Hackathon hosted by RIKEN, which have enabled rapid prototyping and knowledge exchange.

One of the major challenges in computational cancer biology lies in enhancing reproducibility when integrating a heterogeneous set of computational tools commonly used in molecular analysis for personalised treatment strategies. To address this, we are developing robust and portable computational workflows that leverage containerisation technologies, thereby improving both reproducibility and deployment efficiency. At the core of these workflows, we are refactoring multiscale cellular simulation frameworks to ensure they can efficiently scale across the compute resources of modern supercomputing infrastructures.

This report outlines the main advancements of this project under the HANAMI project, highlighting specific tasks we accomplish for fostering international collaboration between European and Japanese research institutions. Sections 3.1, 3.2, and 3.3 present an overview of the collaborative practices and knowledge exchange activities undertaken during the reporting period, including the organization of the BSC–RIKEN workshop and our participation in the RIKEN Hackathon. Sections 3.4 and 3.5 then focus on the ongoing efforts to simulate synthetic tumours and the refactoring of key components of the multiscale simulators, aimed at extending computational capabilities from the cellular to the tissue scale. Finally, Section 3.6 provides a synthesis of the overall impact of these activities and outlines future directions to further advance this line of work.

3.1 Collaborative Internships and Knowledge Exchange

As the initial phase of the project, an internship was conducted by Dr. Marco Ruscone from February 23 to March 12, 2025, at PRIME (Osaka University) and RIKEN (Kobe). The first part at PRIME, hosted by Prof. Ai Shinobu, focused on exploring the integration of molecular

Funded by the European Union
EuroHPC Joint Undertaking

dynamics (MD) simulations with Boolean modelling techniques. During this period, Dr. Ruscone acquired foundational knowledge in molecular dynamics, participating in seminars and using Prof. Shinobu's computational cluster to install and operate Rosetta software. He specifically performed ligand-protein interaction simulations using the Flex ddG method on the EGFR receptor, gaining valuable insights into MD databases and effective data extraction techniques.

In parallel, Dr. Ruscone actively disseminated multiscale modelling methodologies from the BSC. He conducted a hands-on training session on computational tools such as PhysiCell, MaBoSS, NeKo, and PhysiBoSS, generating significant engagement from PRIME researchers. A subsequent seminar highlighted practical applications of these tools, leading to additional discussions and collaborations.

Dr. Ruscone also interacted with several research groups at PRIME, notably those led by Prof. Mariko Okada at the Institute for Protein Research and Prof. Imad Abugessaisa. Mariko Okada is a leading researcher in Japan in the field of systems biology and cellular biology and will be visiting our centre (BSC) during the month of September 2025, which will help to establish stronger ties with her laboratory and her network of collaborators in Japán. Additionally, Dr Ruscone met with Prof. Koji Nishida, Director of PRIME, facilitating broader institutional connections.

The second part at RIKEN, Kobe, hosted by Prof. Florence Tama, included another seminar on the PhysiBoSS modelling framework. This event facilitated interdisciplinary discussions and exchanges with researchers from various groups. The visit concluded with an insightful tour of the Fugaku supercomputer facility, providing valuable comparative insights with MareNostrum 4 at BSC.

Overall, this internship effectively established groundwork for future collaborative initiatives, significantly advancing knowledge exchange and integration of modelling methodologies between BSC, PRIME, and RIKEN.

3.2 Organisation of the Life Science session at the BSC-RIKEN Workshop

On May 29, Dr. Ruscone participated in the HANAMI workshop at RIKEN, Kobe, organised as part of the collaboration between RIKEN and BSC. The workshop opened with four flash talks by participants from both institutions, spanning multiple disciplines, including Life Science, Earth Science, and Social Science. Dr. Ruscone presented tools employed at BSC for modelling multiscale cellular simulations, specifically highlighting PhysiBoSS and PhysiMESS, and emphasising the high computational demands of such simulations.

Participants then engaged in breakout sessions focused on individual disciplines, with Dr. Ruscone organising and attending the Life Science session. This session included presentations from Dr. Eliott Jacopin (RIKEN), Dr. Cheng Tan (RIKEN), and Prof. Ai Shinobu (PRIME, Osaka University). Discussions addressed bridging molecular dynamics simulations

Funded by the European Union
EuroHPC Joint Undertaking

with mechanistic modelling, integrating AI agents using Model Context Protocol (MCP) with PhysiCell, MaBoSS, and Genesis software.

The session highlighted shared computational challenges, notably optimising spatially heterogeneous simulations on multiple nodes, a common issue for both MD and cellular simulations. Additionally, discussions revealed significant overlap between research lines, particularly in using MD simulations to inform signalling networks, an approach relevant for robust Boolean network modelling in PhysiBoSS.

3.3 Contributions to RIKEN's Hackathon and Software Enhancement

The Hackathon organised by RIKEN's Institute of Biomedical Research (IIB), led by Dr. Eliott Jacopin, provided an ideal setting to further explore integrating Large Language Models (LLMs) with diverse biological modelling techniques, as discussed previously at the BSC–RIKEN workshop. The primary platform used was Hatchling, developed by Dr. Jacopin. Hatchling is an interactive command-line interface (CLI)-based chat application that integrates local LLMs through Ollama with the Model Context Protocol (MCP), enabling advanced tool-calling capabilities. MCP enables structured, modular communication between LLMs and external computational tools, allowing LLMs to dynamically invoke specialised software, based on user queries or instructions (Figure 3.1).

```
[Tools enabled] You: which tools can you use?

Assistant:
2025-07-29 14:30:09,624 - ToolExecutionManager-openai-gpt-4o-mini - INFO - [Tool result: meta=None content
=[TextContent(type='text', text='Available tools:\n- create_network(list_of_initial_genes)\n- add_gene(gen
e)\n- remove_gene(gene)\n- export_network(format)\n- network_summary()\n- list_genes_and_interactions()\n-
 find_paths(source, target, maxlen)\n- reset_network()\n- get_help()\n', annotations=None, meta=None)] str
ucturedContent={'result': 'Available tools:\n- create_network(list_of_initial_genes)\n- add_gene(gene)\n-
remove_gene(gene)\n- export_network(format)\n- network_summary()\n- list_genes_and_interactions()\n- find_
paths(source, target, maxlen)\n- reset_network()\n- get_help()\n'} isError=False]

Final response based on tool results:
I have access to several tools that can assist with various tasks. Here is a list of the available tools:

- **create_network(list_of_initial_genes)**: Create a network using a list of initial genes.
- **add_gene(gene)**: Add a specific gene to the existing network.
- **remove_gene(gene)**: Remove a specific gene from the network.
- **export_network(format)**: Export the network in a specified format.
- **network_summary()**: Provide a summary of the current network.
- **list_genes_and_interactions()**: List all genes and their interactions in the network.
- **find_paths(source, target, maxlen)**: Find paths between two specified genes, with a maximum length li
mit.
- **reset_network()**: Reset the network to its initial state.
- **get_help()**: Get help or information regarding the tools.

These tools can be utilized for various functions related to gene networks and interactions.

### Citations
- NeKo
  Origin: Origin citation for NeKo
  Implementation: MCP citation for NeKo
```

Figure 3.1: Example of tools available for the LLM model once it gets access to the NeKo MCP server

Funded by the European Union
EuroHPC Joint Undertaking

In this context, the use of agentic systems and MCP provides a significant advantage over traditional workflows by enabling dynamic decision-making and seamless integration of multiple computational tools, enhancing automation, flexibility, and adaptability in constructing multiscale models.

During the Hackathon, Dr. Ruscone and Dr. Vazquez explored the Hatchling framework by developing MCP servers and approximately 30 specialised computational tools distributed across two distinct servers. These tools enable automated construction of gene regulatory networks (GRN) with the NeKo Python package and simulation of Boolean networks via MaBoSS (Figure 3.2).

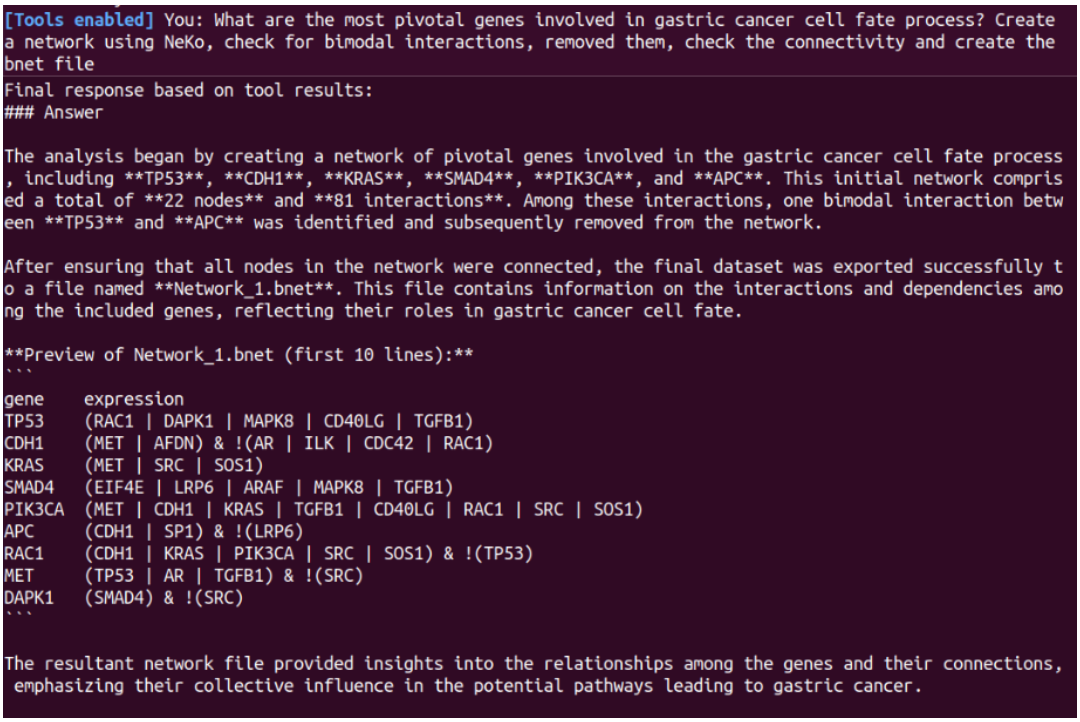


Figure 3.2: Example of prompt execution to create a network based on the results of the NeKo tools

The tools developed enable an LLM agent to:

- Using NeKo:
  - Fetch genes and interactions from the Omnipath and Signor database
  - Aggregate interactions to construct a fully connected network
  - Operate on the network by removing/adding genes and interactions
  - Verify basic network properties
  - Export the network in SIF and BNET format
- Using MaBoSS
  - Convert BNET file into MaBoSS input format (.bnd and .cfg)
  - Instantiate a MaBoSS model (Figure 3.3)
  - Change the input/output and different parameters of the model
  - Run the model
  - Get the output of the model and provide an analysis

Funded by the European Union
EuroHPC Joint Undertaking

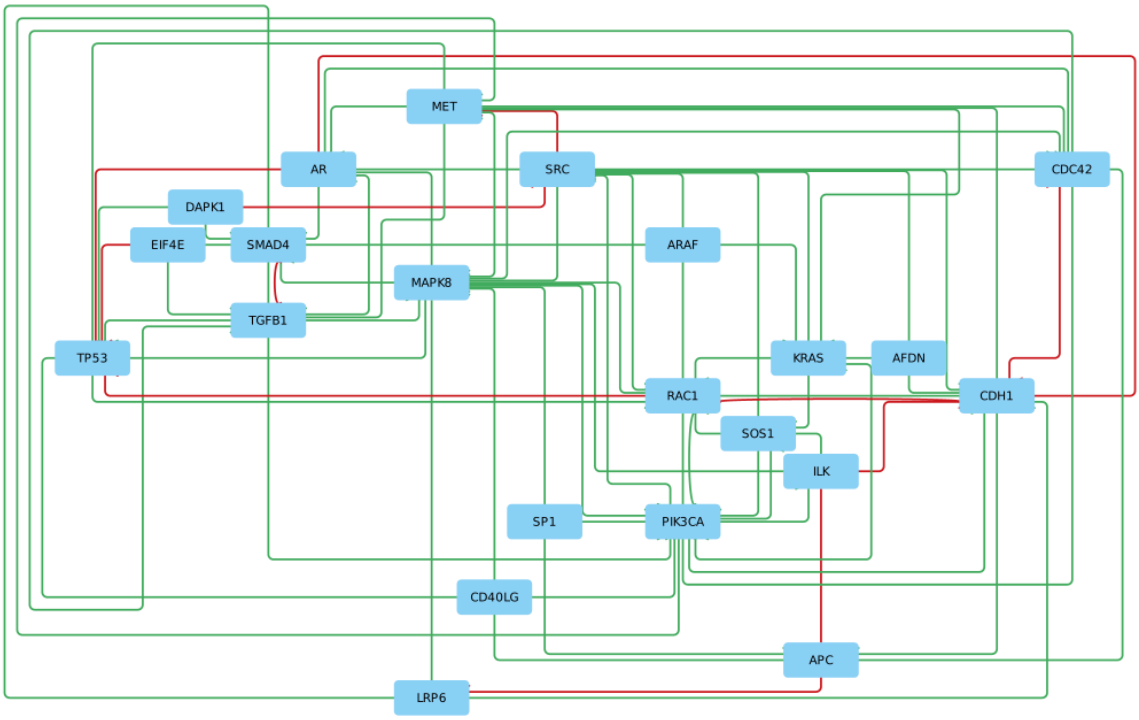


Figure 3.3: Final network created by the LLM agent through the Hathcling platform using the NeKo tools.

Additionally, the team contributed directly to Hatchling by implementing an OpenAI API endpoint, thereby enabling direct access to OpenAI's sophisticated LLMs within the Hatchling framework.

Further, new integration modules for Hatchling were co-developed to seamlessly connect biological modelling tools such as NeKo and MaBoSS via LLM-based interactions. These modules represent a novel methodology, significantly streamlining biological modelling workflows through interactive, AI-assisted knowledge extraction and integration.

The software developments from this Hackathon have been publicly released and are accessible via:

- Hatchling: https://github.com/CrackingShells/Hatchling
- MCP server: https://github.com/marcorusc/Hatch_Pkg_Dev
- Future tool developments: https://github.com/CrackingShells/Hatch-MCP-Server

3.4 Simulated synthetic genomes to feed the cellular simulations

Funded by the European Union
EuroHPC Joint Undertaking

Developing synthetic genomes requires computational pipelines based on reproducibility, automation and portability of the various tools used in the analysis. The installation and maintenance of multiple software tools within complex analysis pipelines can be particularly challenging due to numerous and often conflicting software dependencies. Containerisation offers an infrastructure-agnostic solution for building bioinformatics pipelines that integrate multiple software applications, thereby enhancing reproducibility, portability and scalability. These containerised tools greatly facilitate the automation of pipelines through workflow managers such as Nextflow and Snakemake. These two workflow managers are widely adopted tools in the bioinformatics community.

One of the central tools in the synthetic genome analysis pipeline defined in Task 5.3 of HANAMI proposal is the sequencing read simulator. The current implementation of the read simulator uses NEATGenReads to produce the dataset of sequencing reads in the context of cancer genomes. All the preprocessing of ground truth data, which is specified as an evolutionary tree of clones, is done with in-house developed tools (https://github.com/Rbbt-Workflows/SyntheticCancerGenome) that are containerised as part of synthetic genome simulations. Multiple other software tools are used as part of the quality control and somatic variant detection in the pipeline (Table 3.1)

Table 3.1. Overview of software tools and their sources/documentation used in the synthetic genome analysis pipeline

Purpose	Tool	Source / Documentation
Simulate synthetic sequencing reads	NEATGenReads	https://github.com/zstephens/neat-genreads https://github.com/Rbbt-Workflows/SyntheticCancerGenome
Assess the quality of sequencing reads	FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Filter low-quality FASTQ reads	fastp	https://github.com/OpenGene/fastp
Align reads to a reference genome	BWA	https://github.com/lh3/bwa
Detect somatic variants from sequence data	Mutect2 (GATK)	https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2

Apptainer (formerly, Singularity) is a standalone container platform designed for compatibility with HPC environments. It enables the execution of applications without requiring a separate runtime or privileged (root) access, making it ideal for secure and scalable bioinformatics workflows. During the execution of the synthetic genome analysis workflow in an HPC environment, containers can be launched either from locally available Singularity/Apptainer images (e.g., NEATGenReads for cancer genomes) or dynamically generated on the fly by converting existing Docker images (e.g., for FastQC) from container registries such as

Funded by the European Union
EuroHPC Joint Undertaking

https://quay.io/ and https://hub.docker.com/ The synthetic genome analysis pipeline is based on Nextflow DSL2 best practices (https://nextflow.io/), leveraging standardised nf-core modules and sub-workflows**.** These modules are designed to be fully portable and do not require system-level dependency installations. Instead, they are executed within version-locked Singularity/Apptainer containers**,** ensuring reproducibility and compatibility across computing environments.

To ensure efficient use of resources, we employed an HPC batch scheduler to execute the containerised Synthetic Genome pipeline (Rbbt-Workflows/SyntheticCancerGenome). Computations were parallelised at the chromosome level per clone and optimised to fit within a three-day wall-clock time limit on the CPU computing environment of the LUMI supercomputer.

Synthetic sequencing data for one genome was generated by simulating reads from each clone separately and subsequently mixing them according to specified clonal fractions. Pilot simulation experiments involving tumour-in-normal and normal-in-tumour scenarios were successfully conducted at full genome scale on LUMI. Two batch jobs were deployed (to fit within a three-day wall-clock period on a single 128-core CPU node) on the LUMI CPU partition to produce synthetic read data for the full genome (Table 3.2).

Table 3.2: Computational resources for generating a synthetic genome based on pilot experiments on the LUMI supercomputer

Job/id	Partition	CPUTimeRAW (core-h)	wall-clock time (d-hh:mm:ss)
Batch job for processing 1-9 Chromosemes (job id: 11949222)	small	9046	2-22:40:44
Batch job for processing 10-22, X and Y chromosomes (job id: 11949223)	small	6101	1-23:40:17

These pilot runs provided valuable estimates of the computational resources required to perform all 1000 genome simulation runs on the LUMI supercomputer.

3.5 Scaling-up multiscale simulations

As part of our contribution to advancing multiscale biomedical simulations within this project, we developed and evaluated BioFVM-B, a novel, scalable high-performance diffusion solver designed for large-scale agent-based cellular models. BioFVM-B addresses a major computational bottleneck in simulating the diffusion and decay of substrates (e.g., nutrients, drugs, signaling molecules), which is a critical component of modeling tissue

Funded by the European Union
EuroHPC Joint Undertaking

microenvironments and a limiting factor in the construction of Human Digital Twins. The solver builds upon existing frameworks (BioFVM and BioFVM-X) by introducing a distributed-memory implementation based on MPI and OpenMP, alongside optimized data structures and a block-based domain decomposition strategy that supports non-blocking communication between compute nodes.

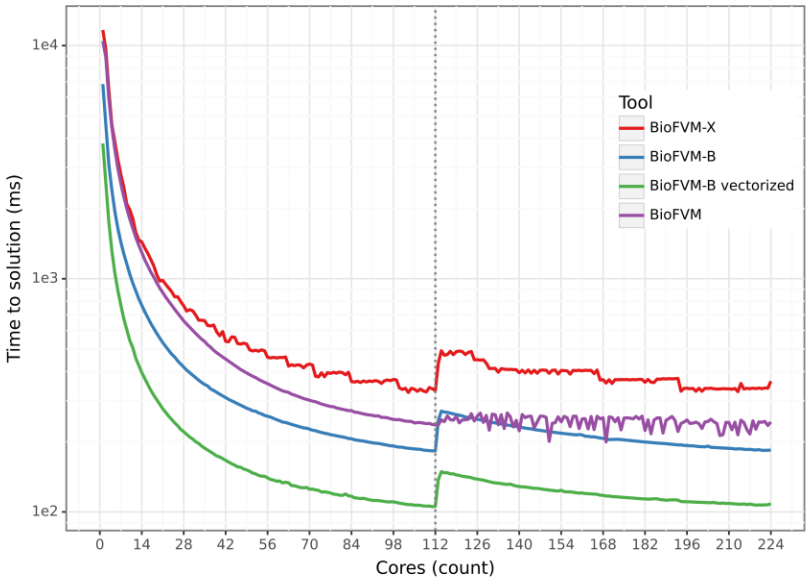


Figure 3.4: *Time to solution per simulation time step for the diffusion-decay solver across the BioFVM, BioFVM-X, and BioFVM-B implementations.*

Through systematic benchmarking on the MareNostrum 5 supercomputer, BioFVM-B demonstrated up to 196.8× speedup and a 36% reduction in memory usage compared to existing state-of-the-art solvers. Crucially, it enables simulations at centimeter scale with micrometer resolution, supporting the modeling of full organ microenvironments composed of billions of cells. This represents a key technical enabler for next-generation personalized medicine pipelines, allowing researchers to simulate realistic, organ-level dynamics with high spatial and temporal resolution. BioFVM-B is fully open source and will be integrated into distributed versions of widely used simulation tools such as PhysiCell, forming the foundation for future work on large-scale digital twin frameworks within the project. Original code of this development can be found at https://gitlab.bsc.es/ jestragu/biofvm-b

3.6 Summary and Future Plans

We will continue advancing collaborations with PRIME by developing workflows to integrate molecular dynamics information into Boolean models, and with RIKEN by further enhancing the Hatchling platform and creating additional MCP servers for PhysiCell and PhysiBoSS.
While the integration of OpenAI's API into Hatchling enabled initial experimentation with tool-calling through proprietary LLMs, the long-term goal is to host open-source or fine-tuned proprietary LLMs locally, specifically trained for biological analysis. Achieving this objective will require access to substantial HPC resources.

Funded by the European Union
EuroHPC Joint Undertaking

Moreover, future development will aim not only to use LLMs to generate and configure multiscale models, but also to enable deployment and execution of large-scale simulations. These simulations, based on highly scalable frameworks such as PhysiCell and PhysiBoSS, are being designed to run efficiently on HPC platforms. This integration of AI-guided modeling and HPC execution is expected to further automate and accelerate biological discovery processes.

Significant progress has been made in the development of our simulator for somatic mutations in cancer, enhancing its realism and flexibility. The simulator enabled parallelisation of computations at the chromosome level for each clone. And, also major effort was put into containerisation and automation of key building blocks involved in the computational workflows of Task 5.3, improving reproducibility and deployment efficiency. Based on the resource estimates from the piloting experiments, we managed to get resource estimates for performing large-scale simulation experiments.

In future, we will focus on the following topics:

- Apply for the computing resources on LUMI for generating 1000 synthetic genomes. Once the resources are available, production-scale simulations will be performed.
- Optimise the overall pipeline to improve the portability, scalability of the synthetic genome pipeline.
- Build containerised pipeline for tumour evolution simulation pipelines (Task 5.4)

Finally, we plan to extend the capabilities of BioFVM-B along several key directions to support future project milestones. First, we aim to enable compatibility with hardware accelerators such as GPUs and vectorized architectures like RISC-V, allowing for further improvements in execution speed and energy efficiency on heterogeneous HPC platforms. We also intend to develop adaptive heuristics for optimizing block partitioning during distributed execution, making the solver more responsive to variations in problem size and computing resources. Another important step will be the full integration of BioFVM-B into distributed versions of PhysiCell, enabling large-scale, parallel simulations of agent-based models at organ-level resolution. This integration will make it possible to simulate complex biological processes in centimeter-scale tissues with micrometer precision, critical for advancing digital twin technologies and in silico biomedical studies.

4. Project 3. Personalized Medicine on Macroscopic Level: Respiratory Flows and Infection Risk Analyses

The HANAMI partner JUELICH, more specifically the Simulation and Data Laboratory Fluids and Solids Engineering (SDL FSE) from the Jülich Supercomputing Centre (JSC) works in Work Package 5 (WP5) on two biomedical use cases (Tasks 5.5 and 5.6) that relate to the domain of

Funded by the European Union
EuroHPC Joint Undertaking

Computational Fluid Dynamics (CFD). In contrast to Project 1, see Section 2, and Project 2, see Section 3, the investigations concentrate on analyses and simulations on a macroscopic level. They introduce physical models and corresponding simulation methods, such as the Lattice-Boltzmann method, on a microscopic level. The work is jointly performed with the Japanese partner from the RIKEN Center for Computational Science (R-CCS), i.e., Prof. Makoto Tsubokura, the head of Complex Phenomena Unified Simulation Research Team and his colleagues at R-CCS. The simulations make use of the multi-physics open-source simulation code m-AIA[1], mainly developed by the Institute of Aerodynamics and Chair of Fluid Mechanics, RWTH Aachen University, and co-developed by JUELICH. Furthermore, the group at R-CCS develops and uses the CUBE code, which is an R-CCS in-house closed-source code.

This section reports on Project 3 “Personalized Medicine on Macroscopic Level: Respiratory Flows and Infection Risk Analyses” in WP5 and its activities in the corresponding tasks, i.e., Task 5.5 “AI-assisted automated CFD pipelines and acceleration of CFD computations” and Task 5.6 “AI-assisted surgery planning and risk assessment of exhaled infectious aerosols”.

4.1 Introduction

This project aims to advance the current state of the art in personalised medicine by leveraging a combination of existing CFD methodologies, i.e., the Lattice-Boltzmann method implemented in the m-AIA code, and various rapidly evolving AI components. Specifically, this is done through two tasks, which define two related but distinct use cases in the field of respiratory flows.

Task 5.5 concerns accelerating and enhancing the various steps of CFD workflows for personalised medicine with the help of AI components. JSC brings its expertise in super-resolution techniques utilising convolutional neural networks for increasing the resolution of computer tomography data, as well as physics-informed neural networks for flow initialisation. These methodologies are applied to the problem of nasal cavity flows. The application of the super-resolution technique allows for increased fidelity of the simulation, as the numerical accuracy of the underlying nasal cavity geometry is less limited by the spatial resolution of CT imaging. Additionally, the total simulation time can be further shortened by using physics-informed neural networks for specifying an initial condition, which is closer to the converged state. Thus, fewer iterations of the solution process are required, leading to a reduction in the total computational time and effort.

The work in Task 5.6 further promotes the use of AI components and allows for an AI-assisted surgical geometry modification by utilising a reinforcement learning (RL) algorithm that suggests a modified nasal cavity geometry. This modified geometry is a solution to an optimisation problem, where the pressure loss and temperature difference are used simultaneously as objective functions. This physics-driven modification potentially improves the overall comfort of the patient. The optimisation problem can be further extended to also account for other objective functions, such as balanced volume flow rate through both nasal passages and surface wall-shear stress.

[1] m-AIA repository https://git.rwth-aachen.de/aia/m-AIA/m-AIA

Funded by the European Union
EuroHPC Joint Undertaking

These two tasks primarily build on the existing experience of the SDL FSE with various AI methodologies. However, by leveraging the experience of RIKEN R-CCS with COVID-19 simulations and quantitative risk assessment, there is an opportunity for a fruitful collaboration on risk assessment of infectious aerosol particles, utilising the nasal cavity geometry of the previous two tasks. This would provide a more personalised view on the problem of infection risk reduction, building on the previous award-winning research carried out by the Complex Phenomena Unified Simulation Research Team during the COVID-19 pandemic[2]. Corresponding activities are foreseen in the second period for Task 5.6.

While there are two capable simulation codes - m-AIA and CUBE, employed by the SDL FSE and the Complex Phenomena Unified Simulation Research Team, respectively, the results reported in this periodic report are achieved by using m-AIA. The primary reason is that, in contrast to CUBE, m-AIA is an open-source code, and it is easier to disseminate the research results and associated methodologies to a wider community. Nevertheless, the CUBE code has been optimised over the past decades for optimal performance on a series of supercomputers at R-CCS. The second phase of the project will therefore also look at CUBE and benchmark its performance on European systems as far as licences permit.

More details about m-AIA's computing performance can be found in Section 4.2. Subsequently, Section 4.3 presents the highlighted scientific results achieved in the first reporting period. It should be noted that this section only includes the results directly related to the HANAMI project, while further associated publications are presented in Section 4.8. As the exchange and collaboration with the Japanese partner is key to the success of this project, Section 4.4 expands on the details of collaborative activities. Next, a brief summary of work already achieved and the plan for the next phase of the project are described in Section 4.5. Computational resources that have been acquired within the reporting period to work on the tasks are presented in Section 4.6. Finally, Section 4.7 provides a summary and some conclusions, and Section 4.8 lists the full list of publications.

4.2 m-AIA performance on HPC systems

The multi-physics simulation framework m-AIA has been extensively tested on several JSC supercomputers[3]. Tables 4.1 - 4.4 show the scaling test results across different systems, while Figures 4.1 and 4.2 display the scaling for CPU and GPU-ported variants of the code, respectively.

Table 4.1 m-AIA performance on Hazel Hen, JURECA-DC and CLAIX supercomputers.

CFD Method	Number of cells (millions)	HPC system	number of cores / runtime [s]					
			384	758	1,536	3,072	6,144	12,288

[2] Gordon Bell Special Prize for HPC-Based COVID-19 Research Awarded to Japanese Team for Novel Aerosolized Droplet Simulation https://www.acm.org/media-center/2021/november/gordon-bell-special-prize-covid-research-2021

[3] JSC Supercomputers https://www.fz-juelich.de/en/ias/jsc/systems/supercomputers

Funded by the European Union
EuroHPC Joint Undertaking

Lattice-Boltzmann	1070	HAZEL HEN	110.49	58.70	30.65	16.41	8.96	4.97
		JURECA-DC	156.28	87.43	46.21	21.52	-	-
Thermal Lattice-Boltzmann	1070	JURECA-DC	334.14	183.09	95.95	47.84	-	-
		CLAIX	233.20	120.69	63.22	32.56	17.56	-

Table 4.2 m-AIA performance on the JUQUEEN supercomputer.

CFD Method	no. cells (M)	HPC system	number of cores / runtime [s]								
			1,024	2,048	4,096	8,192	16,384	32,768	65,536	131,072	262,144
Lattice-Boltzmann	1230	JUQUEEN	633.85	332.76	169.87	88.79	47.29	25.25	12.60	7.26	6.19

Table 4.3 m-AIA performance on the JUWELS Booster supercomputer.

CFD method	no. cells (million)	HPC system	number of threads / runtime [s]					
			256	512	1,024	2,048	4,096	8,192
Lattice-Boltzmann	158	JUWELS Booster	117.03	44.66	20.83	12.20	7.75	5.62
	1230		-	-	-	120.94	48.06	22.42

Table 4.4 m-AIA performance on the JUWELS Cluster supercomputer.

CFD method	no. cells (M)	HPC system	number of nodes (4 GPUs per node) / runtime [s]								
			1	2	4	8	16	32	64	128	256
Lattice-Boltzmann	134	JUWELS Cluster	135.61	71.54	45.93	27.82	16.38	11.06	-	-	-
	2097		-	-	-	-	307.76	174.24	93.27	58.79	38.01

The newest addition to JSC's supercomputers, JUPITER, underwent some initial scaling tests. More specifically, a small scaling test was carried out on JEDI, the first module of JUPITER. More thorough and large-scale testing will be carried out when JUPITER is officially opened and the performance is stabilized.

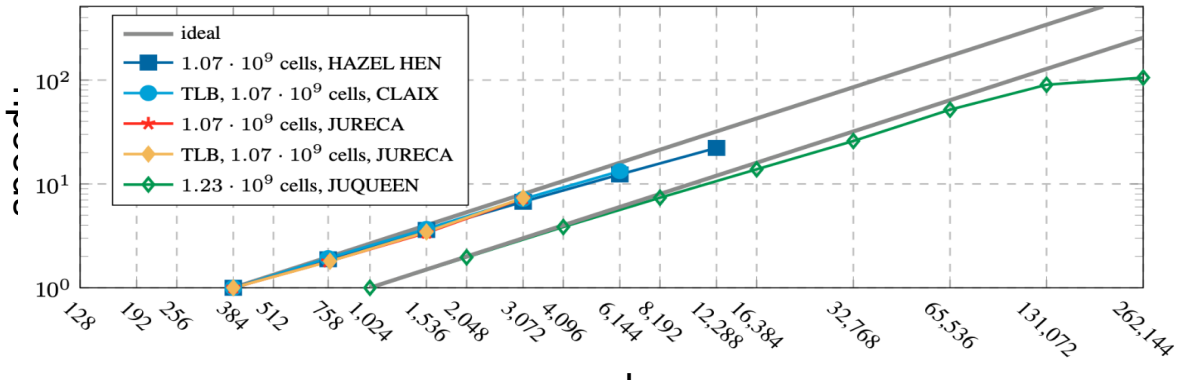


Figure 4.1 m-AIA scaling on CPU JSC systems for the (Thermal) Lattice-Boltzmann method.

Funded by the European Union
EuroHPC Joint Undertaking

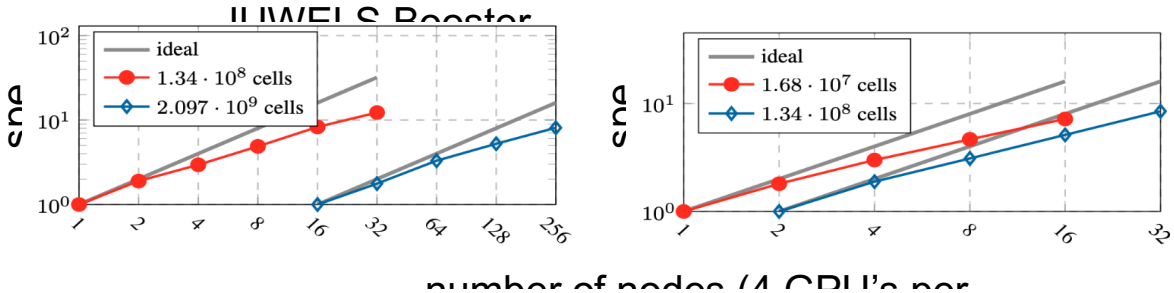


Figure 4.2 m-AIA scaling on GPU JSC systems for the Lattice-Boltzmann method.

Just recently, m-AIA has also been successfully ported to Fugaku, a flagship Japanese HPC system. No extensive scaling test has been done yet due to limited computational budget on the system, but this will be done in the foreseeable future.

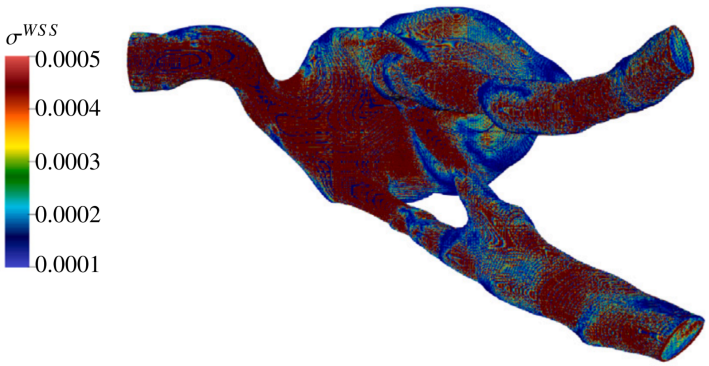
4.3 Key scientific results achieved in the first reporting period

In the first reporting period, several ongoing research efforts led to the following publications, which fully align with the aforementioned Tasks 5.5 and 5.6 of HANAMI:

[P3_1] Patient-specific lattice-Boltzmann simulations with inflow conditions from magnetic resonance velocimetry measurements for analyzing cerebral aneurysms -
M. Rüttgers et al. (doi:10.1016/j.compbiomed.2025.109794)

This publication demonstrates the use of the m-AIA code to perform patient-specific CFD simulations of cerebral aneurysm hemodynamics by incorporating realistic inflow conditions obtained from magnetic resonance velocimetry. The methodology shows how MRI-derived velocity data can be directly coupled with high-fidelity Lattice Boltzmann simulations to capture complex, patient-specific flow fields in medical contexts.

While this study focuses on cerebral aneurysms, it represents an important extension of personalised CFD analysis from respiratory flows to complex vascular flows. By demonstrating patient-specific modelling of blood flow dynamics with realistic inflow data, it lays the foundation for developing AI-assisted surgery planning tools for aneurysm treatment, aligning directly with HANAMI's objective of advancing personalised medicine through high-fidelity, AI-enhanced simulations.



Funded by the European Union
EuroHPC Joint Undertaking

Figure 4.3. Wall shear stress inside the nasal cavity for a Lattice Boltzmann model based on the Carreau–Yasuda equation.

[P3_2] Towards a widespread usage of computational fluid dynamics simulations for automated virtual nasal surgery planning -
M. Rüttgers et al. (doi:10.1016/j.future.2025.107935)

This paper introduces a computationally efficient framework for CFD-based virtual planning of nasal cavity surgeries, such as septoplasty and turbinectomy. A hybrid method combining the lattice-Boltzmann approach with level-set-based interpolation allows simulation of surgical variations through a discrete set of geometry steps at single surgery interventions. This reduces the number of required simulations to as few as 21 per intervention. Previous methods rely on more costly coupling strategies, such as reinforcement learning or thermal modelling, which may still be appropriate for complex planning scenarios involving multiple intervention sites or thermal flow analysis. Despite this simplification, the method retains key predictive capabilities, optimising pressure loss and flow balance while reducing unnecessary tissue removal. Case studies demonstrate tissue savings of 12–25% and the use of a regression model to further reduce computational cost by 50%, with less than 4% prediction error.

This work supports HANAMI WP5 by delivering a clinically viable, automated CFD framework tailored for personalised medicine at the macroscopic level. By enabling efficient prediction of nasal airflow outcomes from surgical interventions, the study demonstrates how CFD can be integrated into routine clinical workflows. The hybrid method lays foundational work toward AI-assisted, patient-specific surgery planning tools envisioned in HANAMI.

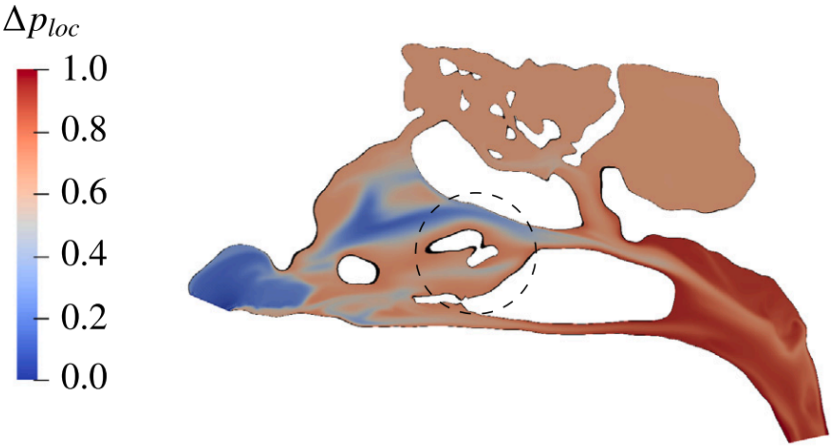


Figure 4.4. Normalised pressure loss after geometry modification between the inlets (nostrils) and the outlet (pharynx).

[P3_3] HydroGym-GPU: From 2D to 3D Benchmark Environments for Reinforcement Learning in Fluid Flows -
C. Lagemann, M. Rüttgers, et al. (doi:10.34734/FZJ-2025-02455)

This work introduces HydroGym-GPU, a GPU-accelerated extension of the HydroGym platform that integrates the highly parallelised lattice-Boltzmann solver of the m-AIA framework with RL workflows for fluid flow control. By enabling large-scale, three-dimensional CFD environments on modern GPU architectures, the platform removes a major barrier to studying realistic flow control problems. It provides standardised, open benchmark

Funded by the European Union
EuroHPC Joint Undertaking

environments that support advanced RL algorithms in scenarios of significantly increased physical and computational complexity, addressing a critical gap between 2D academic examples and real-world 3D control challenges.

This development supports HANAMI WP5's vision of advancing AI-assisted, patient-specific simulations by laying the groundwork for integrating RL-based surgical planning tools into a robust, open-source framework. The close collaboration with Steven Brunton's group from the University of Washington ensures the application of state-of-the-art RL methodologies to large-scale biomedical CFD problems. Notably, the RL-based nasal surgery planning tools developed in HANAMI are intended to become part of this open platform, ensuring their long-term impact and accessibility for future clinical and research use.

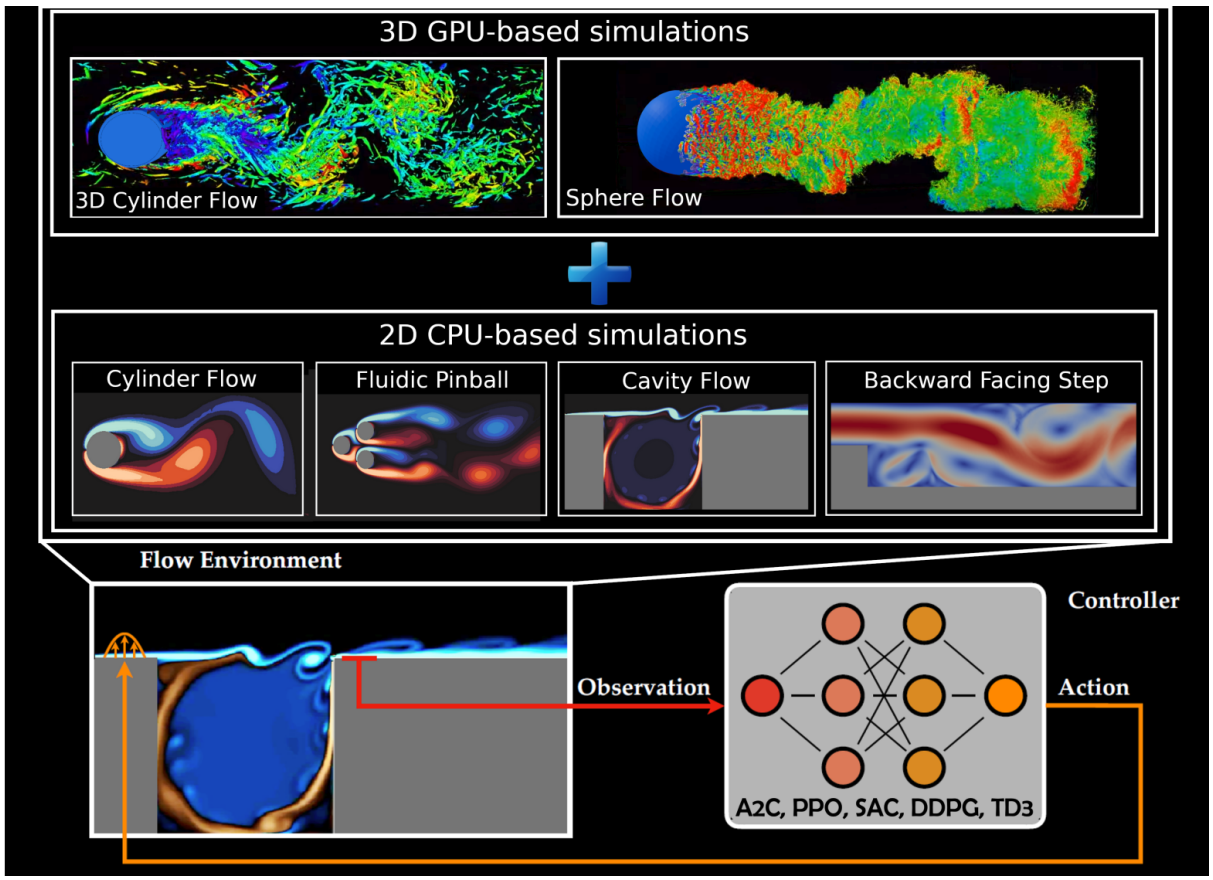


Figure 4.5. Flow configurations and benchmark algorithms included in the initial and extended HydroGym-GPU platform

[P3_4] Wet-Surface Modeling in Lattice-Boltzmann Simulations for Evaluating Surgery Impacts on the Humidity Transfer in Nasal Flows -
S. Ito, M. Rüttgers et al. (DOI: 10.34734/FZJ-2025-02479)

This study presents the first lattice-Boltzmann simulations using the m-AIA framework to explicitly model humidity fields in patient-specific nasal airflow, providing a more physiologically meaningful measure of the nose's conditioning ability than temperature alone. A novel wet-surface boundary treatment was implemented to account for latent heat effects from mucosal evaporation, dynamically influencing wall temperature and water vapour concentration. The study compares pre- and post-surgical (turbinectomy) nasal

Funded by the European Union
EuroHPC Joint Undertaking

cavities, revealing that surgery reduces inspiratory pressure loss but leads to drier air reaching the back of the throat, matching the patient's reported symptoms.

This development represents an important step for HANAMI WP5. By moving beyond temperature-based conditioning metrics to predict full humidity fields, the approach offers a more accurate and relevant assessment of nasal surgery impacts on airflow and patient comfort. The validated wet-surface modelling approach will be integrated into the AI-assisted nasal surgery planning tool being developed in HANAMI, improving its ability to recommend surgical interventions that balance airflow efficiency with humidification performance.

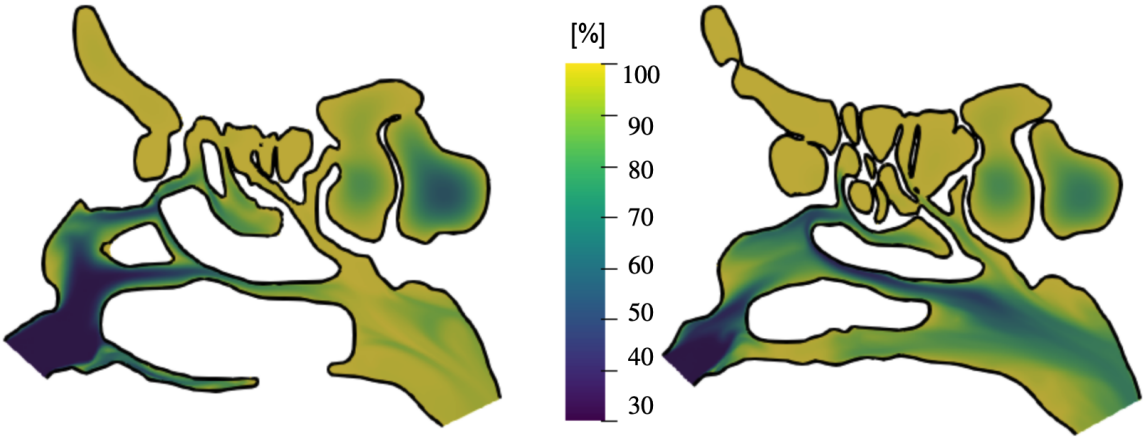


Figure 4.6. Humidity distribution for a cross-sectional area through the left nasal passage (from the patient's view) of the pre-surgical (left) and the post-surgical (right) cases.

[P3_5] Parallel Reinforcement Learning and Gaussian Process Regression for Improved Physics-Based Nasal Surgery Planning -
M. Rüttgers et al. (doi:10.1007/978-3-031-85703-4_6)

This paper presents a novel approach for physics-based virtual nasal surgery planning that combines parallel RL with Gaussian process regression (GPR) to optimise surgical outcomes efficiently. By leveraging parallel RL agents to explore surgical parameter spaces and training GPR surrogate models for rapid prediction of fluid-dynamic outcomes, the method significantly reduces the number of expensive CFD simulations required for surgery planning. Demonstrated on patient-specific nasal geometries, this approach delivers high-quality surgical recommendations with reduced computational cost, making AI-driven CFD planning more practical for clinical use. The work was recognised with the **Best Paper Award** at the PPAM conference in Ostrava, September 2024.

The integration of GPR into the workflow benefited substantially from the expertise of the Japanese project partners, enabling efficient surrogate modelling of CFD results. This close collaboration showcases the value of combining domain knowledge, HPC resources, and AI methodologies across institutions to deliver practical, clinically relevant tools for personalised medicine.

Funded by the European Union
EuroHPC Joint Undertaking

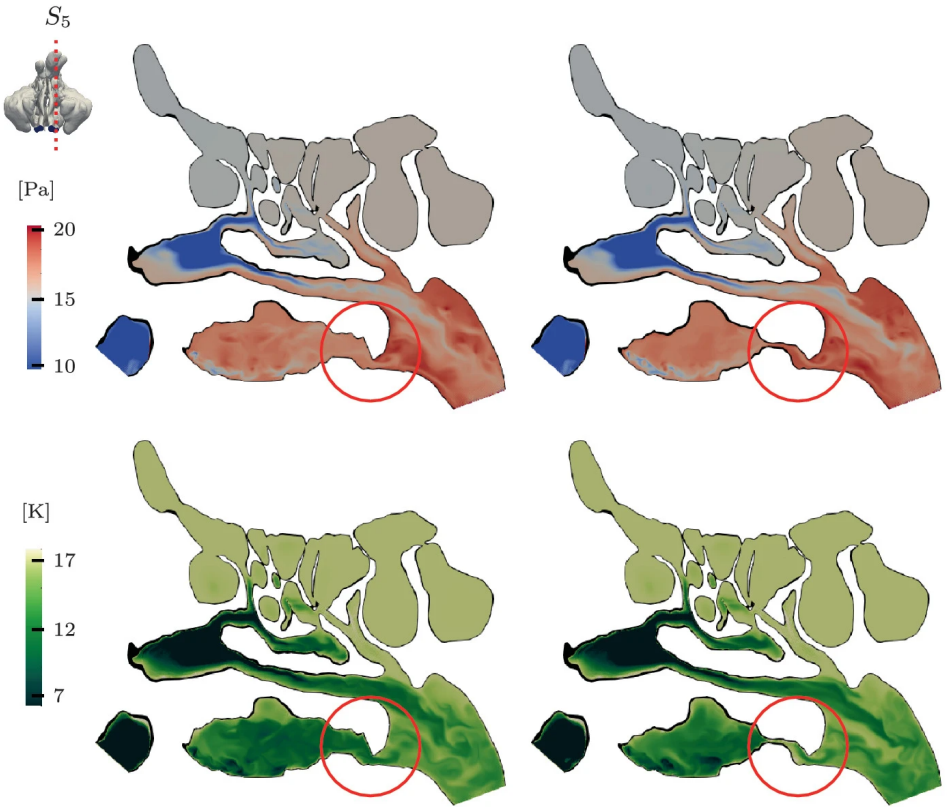


Figure 4.7. Total pressure loss [top] and temperature increase [bottom] between the inlets and locations inside of the nasal cavity for (αC,αD)=(0.85, 0.25) [left] and (αC,αD)=(0.6, 0.15) [right] at the cross-sectional area $S5$, which is illustrated by the dashed line in this figure. The red circles highlight the region of large deviations between the two cases.

4.4 Knowledge and software exchange between JUELICH and R-CCS

Monthly meetings take place to update on the progress of the Joint Laboratory for Extreme Scale Computing (JLESC) Project “Deep Neural Networks for CFD Simulations”. The groups exchange primarily on the topics of PINNs and Graph Neural Networks (GNNs). This collaboration has been fruitful, with both groups contributing to one joint publication in a JLESC special issue [P3_6] and another manuscript currently under review. For PINNs, the research contributed to benchmarking the performance of PINNs to traditional Deep Neural Networks (DNNs) for classical flow problems when data is sparse and localised. Also, the generalizability of Physics-Informed Deep Operator Network (PI-DeepONet) is investigated in the joint work. For GNNs, the groups are looking into the accuracy and performance of physics-aware and data-free GNNs for problems in urban flows and combustion. This work is expected to contribute to a better understanding of the effect of integration of physical laws in training Neural Networks (NNs) for CFD applications.

In addition, the SDL FSE team member Alex Krochak visited R-CCS in June 2025 to compile, set up, and test the multi-physics simulation code m-AIA from RWTH Aachen on Fugaku. As Fugaku features a different chip architecture for its login and compute nodes, cross-

Funded by the European Union
EuroHPC Joint Undertaking

compilation proved to be difficult. Nonetheless, the code was eventually successfully compiled with the use of the Fujitsu C++ compiler together with the Clang compilation flag option to allow for the C++17 standard that is used by m-AIA (the standard Fujitsu compiler unfortunately does not offer full C++17 support). A merge request to include the full configuration steps necessary for Fugaku compilation is currently in progress as m-AIA is transitioning to C++20 standard support. It is planned to retry the compilation with C++20 standard libraries and update the merge request accordingly in the next few months.

In addition to the m-AIA compilation, part of the research stay at R-CCS was specifically dedicated to the knowledge exchange with Complex Phenomena Unified Simulation Research Team members. It is planned to include Dr. Rahul Bale, one of RIKEN's experts on CFD simulation for quantitative infection risk assessment, as a collaborator for future HANAMI-related personalised medicine work at JSC, as part of Task 5.6. Additionally, Krochak gave a presentation to the members of the R-CCS team and M.Sc. students under Prof. Tsubokura's supervision at Kobe University. The topic of the presentation was the use of AI methods for large-eddy simulation closure modelling in CFD.

4.5 Future work in the second HANAMI period

Initial efforts by JUELICH in the first reporting period were directed towards setting up the required tools necessary for the successful integration of AI methodologies into m-AIA CFD workflows. JUELICH has successfully achieved the main goals of this task. The publications presented in Section 4.2 [P3_1, P3_2, P3_3, P3_4] underline the improvements in physical modelling and in enabling AI components, such as RL, to improve simulation performance via accurate AI-augmentation. This foundation allowed for ultimate additions of various AI components to the CFD pipeline for nasal cavity simulations, as it is described in the first part of Task 5.6, see for more detail [P3_5].

Thus, the remainder of the project will focus on the second part of Task 5.6, namely, providing a quantitative CFD-based infection risk assessment, as well as improving the structure and ease-of-use of previously developed tools. To this end, a software repository has been hosted on the AI4HPC open-source library of models[4]. This repository is currently being developed to allow these tools to be accessed and used by anyone interested.

For the infection-risk assessment, a combined geometry of both the internal nasal cavity and the external surroundings is used to compute the trajectory of viral droplets and their likelihood of entering the nasal cavity, see Figure 4.8 for reference. The initial conditions of these droplets, i.e., their initial position, velocity, diameter, and density, are varied to study the effect of these parameters on the likelihood of particle deposition and thus on an infection. Using the initial particle condition, such as its position and velocity, together with its ultimate probability of deposition, can be used to enhance already existing simulations which used simpler deposition models, as well as to provide physical insights on how to reduce infection risks by accounting for such effects. For example, some simulations assume that the particle will be deposited inside the airway as soon as it's within a certain distance of

[4]AI4HPC Nasal Cavity repository https://gitlab.jsc.fz-juelich.de/CoE-RAISE/FZJ/ai4hpc/ai4hpc-dev/-/tree/nasal-cavity?ref_type=heads

Funded by the European Union
EuroHPC Joint Undertaking

a nasal opening, and do not account for particle velocity. The reader is referred to the next section, particularly to the 2025 compute time sub-project, for more details of the work planned.

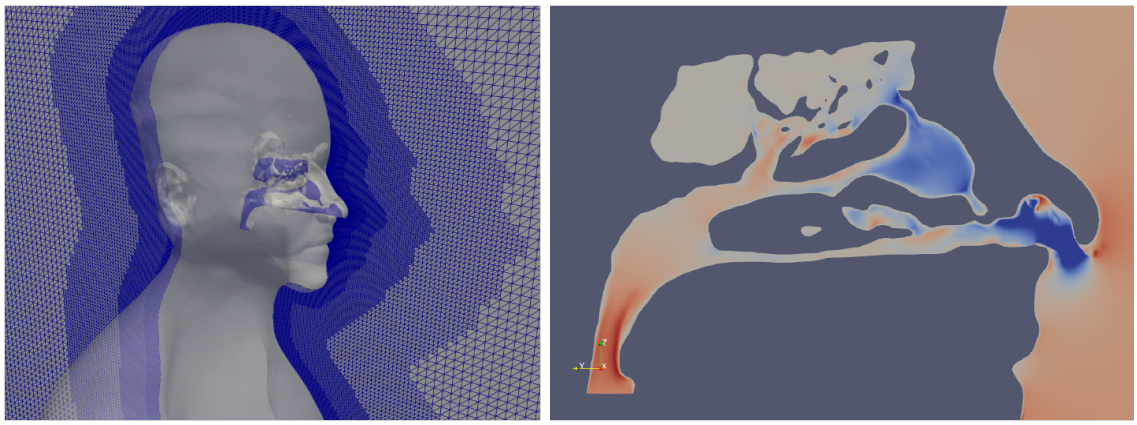


Figure 4.8 - Combination of external and internal geometries for future HANAMI simulations (a). Velocity magnitude through a slice of nasal cavity internal geometry, preliminary simulation (b).

4.6 Computational resources

The activities described in Section 4.2 and Section 4.3, as well as those foreseen in the future, see Section 4.5, can only be performed with HPC compute time. Much of the development involved training of AI models and, hence, for better efficiency, required the usage of Graphics Processing Units (GPUs). Therefore, compute time proposals were filed containing requests for performing simulations on Central Processing Unit (CPU)-based systems, and for performing training on GPU partitions.

Specifically, two VSR compute time proposals at JSC were successful in 2024 and 2025, each with a runtime of one year[5]. They received compute time grants amounting to the budgets listed in Table 4.3 below. The corresponding proposals were jointly prepared between JSC and R-CCS. The main Principal Investigator (PI) is the leader of the SDL FSE, Dr. Andreas Lintermann. As co-PI, Dr. Junya Onishi from R-CCS contributed to the proposal and to the implementation of the work together with the team members from the SDL FSE.

While the 2024 compute time project mainly concentrated on the activities in Task 5.5, the 2025 compute time project (05.2025 - 04.2026) also contains a sub-project targeted at the infection-risk analysis for trans-nasal-cavity flow as planned in Task 5.6. Specifically, the following tasks are part of the compute time project:

- A combination of the internal and external flow domains to numerically simulate the process of inhalation of viral particles, i.e., droplets containing viral matter, in the form of Lagrangian particles transported by the flow.

[5] VSR Compute Time call https://www.fz-juelich.de/en/ias/jsc/systems/supercomputers/apply-for-computing-time/vsr

Funded by the European Union
EuroHPC Joint Undertaking

- The particles are distributed with varying density, diameter, initial velocity, and initial position to provide a realistic setup for analysis.
- The simulations are performed with the lattice-Boltzmann fluid mechanics solver and the Lagrangian particle tracking modules of the multi-physics open-source simulation framework m-AIA.
- A successful implementation of these tasks will provide data-driven assessment on the probability of inhalation of viral particles by healthy patients in clinical environments, such as waiting rooms, or in public transportation, depending on the initial position of the particles.

For the VSR projects, the Jülich Research on Exascale Cluster Architectures (JURECA) DC system is used, which consists of multiple partitions, however, only the JURECA DC CPU and GPU partitions are of relevance for the implementation of Project 3. The JURECA DC module containing both the CPU and GPU partitions has (as of 07/2025) 480 standard compute nodes, each equipped with 2 AMD EPYC 7742 with 2 x 64 cores, clocked at 2.25 GHz. The 192 accelerator nodes are additionally equipped with four NVIDIA A100 GPUs, each with 40 GB HBM2e. The interconnect is a Mellanox InfiniBand HDR (HDR100/HDR) DragonFly+ network[6].

For the activities that have taken place during the research visit described in Section 4.3 and beyond, the compute time allocation on Fugaku of the Complex Phenomena Unified Simulation Research Team at R-CCS is used. The Fugaku system has 158,976 nodes, each node containing 48 Armv8.2-A SVE 512-bit cores and additional assistant cores, as well as 32 HBM2 memory. The network topology is a Tofu Interconnect D[7].

Table 4.3: Compute time grants received during the first reporting period.

System	Budget	Timeline	Access
Improved Diagnostics of Respiratory Flows Using a Lattice-Boltzmann Method and Machine Learning Techniques			
JURECA DC CPU	1.76 M. core-h	from 05.2024 to 04.2025	SDL FSE and Dr. Junya Onishi
JURECA DC GPU	2.60 M. core-h	from 05.2024 to 04.2025	SDL FSE and Dr. Junya Onishi
Improved Diagnostics of Respiratory Flows Using a Lattice-Boltzmann Method and Machine Learning Techniques - extension			

[6] More information on JURECA https://apps.fz-juelich.de/jsc/hps/jureca/configuration.html

[7] More information on Fugaku https://www.fujitsu.com/global/about/innovation/fugaku/specifications/

Funded by the European Union
EuroHPC Joint Undertaking

JURECA DC CPU	1.22 M. core-h	from 05.2025 to 04.2026	SDL FSE and Dr. Junya Onishi
JURECA DC GPU	1.57 M. core-h	from 05.2025 to 04.2026	SDL FSE and Dr. Junya Onishi
Establishment of a unified method for solving complex and composite phenomena and innovative manufacturing through data science integration			
Fugaku	5 M. node-h	from 04.2025 to 03.2026	RIKEN Complex team and Alex Krochak

4.7 Summary and conclusions

In the first period of HANAMI, Project 3 has made strong progress in using and improving CFD workflows for personalised medicine with AI components, as well as improving the overall physical modelling and computation performance. New AI methods, such as super-resolution imaging, PINNs, RL, and wet-surface modelling, were integrated into the simulation pipeline, improving the physical accuracy and reducing the overall simulation effort. This improved the technology readiness level of CFD workflows for personalised medicine.

The open-source m-AIA code performed well across several JSC supercomputers in Germany and was successfully ported to Fugaku in Japan. An evaluation of the code performance on Fugaku's Fujitsu A64FX ARM processors is planned soon.

Close collaboration with R-CCS, especially with the Complex Phenomena Unified Simulation Research Team, played a key role. Outcomes included software integration, regular knowledge exchange, and collaboration towards an infection-risk model based on Lagrangian particle tracking simulation(s) in combined internal/external geometries.

The next steps include refining the simulation tools via the AI4HPC platform and working towards the infection-risk model.

Overall, main goals of the project have already been achieved and documented through various peer-reviewed publications, and this project successfully demonstrates how the combination of CFD and data-driven methods can enhance patient-specific treatment.

4.8 Full list of publications

The following lists the key peer-reviewed articles that have been published during the first reporting period in renowned journals (papers [P3_1, P3_2]). The paper [P3_5] won the

Funded by the European Union
EuroHPC Joint Undertaking

Diploma Best Paper Award (winner for workshops) at the Parallel Processing and Applied Mathematics (PPAM 2024) Conference, which took place in Ostrava, Czech Republic, September 8 - 11, 2024.

[P3_1] Rüttgers, M., Waldmann, M., Ito, S., Wüstenhagen, C., Grundmann, S., Brede, M., & Lintermann, A. (2025). Patient-specific lattice-Boltzmann simulations with inflow conditions from magnetic resonance velocimetry measurements for analyzing cerebral aneurysms. *Computers in Biology and Medicine*, Vol. *187*, 109794. https://doi.org/10.1016/j.compbiomed.2025.109794

[P3_2] Rüttgers, M., Waldmann, M., Hübenthal, F., Vogt, K., Tsubokura, M., Lee, S., & Lintermann, A. (2026). Towards a widespread usage of computational fluid dynamics simulations for automated virtual nasal surgery planning. *Future Generation Computer Systems*, Vol. *174*, 107935. https://doi.org/10.1016/j.future.2025.107935

[P3_3] Lagemann, C., Rüttgers, M., Gondrum, M., Meinke, M., Schröder, W., Lintermann, A., & Brunton, S. L. (2025). HydroGym-GPU: From 2D to 3D Benchmark Environments for Reinforcement Learning in Fluid Flows. *Forschungszentrum Jülich Technical Report*. https://doi.org/10.34734/FZJ-2025-02455

[P3_4] Ito, S., Rüttgers, M., Waldmann, M., & Lintermann, A.(2025). Wet-Surface Modeling in Lattice-Boltzmann Simulations for Evaluating Surgery Impacts on the Humidity Transfer in Nasal Flows. *Forschungszentrum Jülich Technical Report*. https://doi.org/10.34734/FZJ-2025-02479

[P3_5] Rüttgers, M., Hübenthal, F., Tsubokura, M., & Lintermann, A. (2025). Parallel Reinforcement Learning and Gaussian Process Regression for Improved Physics-Based Nasal Surgery Planning. In R. Wyrzykowski, J. Dongarra, E. Deelman, & K. Karczewski (Eds.), *Parallel Processing and Applied Mathematics (PPAM 2024). Lecture Notes in Computer Science (LNCS)* Vol. 15581, pp. 79–96. Springer. https://doi.org/10.1007/978-3-031-85703-4_6

[P3_6] Puri, R., Onishi, J., Rüttgers, M., Sarma, R., Tsubokura, M., & Lintermann, A. (2024). On the choice of physical constraints in artificial neural networks for predicting flow fields. *Future Generation Computer Systems, Vol. 161*, pp. 361–375. North-Holland. https://doi.org/10.1016/j.future.2024.07.009

[P3_7] Liu, X., Rüttgers, M., Quercia, A., Egele, R., Pfaehler, E., Shende, R., Aach, M., Schröder, W., Balaprakash, P., & Lintermann, A. (2024). Refining computer tomography data with super-resolution networks to increase the accuracy of respiratory flow simulations. *Future Generation Computer Systems, Vol. 159*, pp. 474–488. North-Holland. https://doi.org/10.1016/j.future.2024.05.020

[P3_8] Rüttgers, M., Waldmann, M., Vogt, K., Ilgner, J., Schröder, W., & Lintermann, A. (2024). Automated surgery planning for an obstructed nose by combining computational fluid

Funded by the European Union
EuroHPC Joint Undertaking

dynamics with reinforcement learning. *Computers in Biology and Medicine, Vol. 173*, 108383. Pergamon. https://doi.org/10.1016/j.compbiomed.2024.108383

5. Conclusions

This mid-term deliverable highlights the coordinated progress achieved across the three WP5 projects, reflecting the strong collaborative efforts between European and Japanese partners. The work completed demonstrates HANAMI's capacity to co-develop advanced computational tools and methodologies that operate across molecular, cellular, and organ-level biological modelling in life sciences domain.

These collaborations have laid the groundwork in Project 1 for exascale-capable molecular dynamics simulations, through the integration of a scalable Fast Multipole Method (FMM) into GROMACS, enabling efficient long-range electrostatics computations at the scale of entire cells. In Project 2, joint activities with Japanese partners have enabled the full integration of Large Language Models (LLMs) for network reconstruction and pathway modelling that will be essential for scaling mulsticale cellular simulations. Project 3 has advanced AI-assisted CFD workflows for personalised medicine by integrating methods like super-resolution imaging, PINNs, reinforcement learning, and wet-surface modelling. These enhancements improved physical accuracy while reducing the computational cost of simulation. The open-source m-AIA code showed strong performance on JSC systems and was successfully ported to Fugaku. A simulation for risk quantification of viral infection through inhalation is currently in progress.

Together, these achievements provide a strong technical and collaborative foundation for the second half of the project, which will focus on further scaling, deeper integration across projects, and the continued strengthening of European-Japanese collaboration to advance personalised medicine through exascale-ready biomedical modelling tools.

Funded by the European Union
EuroHPC Joint Undertaking