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 D5.1

Biomedical project setup roadmap

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	5
Task	5.1
Lead Authors	Erik Lindahl (KTH), Mario Rüttgers (FZJ), Alfonso Valencia (BSC), José Carbonell Caballero (BSC), Arnau Montagud (BSC), Miguel Vazquez (BSC)
Contributors	Berk Hess (KTH), Andreas Lintermann (FZJ), Tommi Nyronen (CSC)
Peer Reviewers	France Boillod-Cerneux (CEA), Laure Caruso (CEA)
Version	2.0
Due Date	31/08/24
Submission Date	04/09/24

Funded by the European Union
EuroHPC Joint Undertaking

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
v0.1	02/07/2024	Erik Lindahl, Berk Hess (KTH)	Roadmap for Project 1
v0.2	05/07/2024	Mario Rüttgers, Andreas Lintermann (FZJ)	Roadmap for Project 3
v0.3	23/07/2024	Alfonso Valencia (BSC), José Carbonell Caballero (BSC),	Roadmap for "Project 2: Development of Genome Analysis Pipelines for Personalized Medicine"

Funded by the European Union
EuroHPC Joint Undertaking

		Arnau Montagud (BSC), Miguel Vazquez (BSC)	
v0.4	31/07/2024	Erik Lindahl (KTH)	Editing and formatting, executive summary
V0.5	15/08/2024	Erik Lindahl (KTH)	Editing
V0.6	29/08/2024	France Boillod-Cerneux (CEA)	Review and comments
V1.0	30/08/2024	All authors	Final revision
V2.0	04/09/2024	France Boillod-Cerneux (CEA)	Missing paragraph added.

Executive Summary

The biomedical research program of HANAMI consists of three project areas: (i) Advancing the state-of-the-art in Exascale-target molecular modelling through new algorithms for electrostatic interactions and AI/machine learning approaches to particle interactions, with the goal of simulating cellular-size systems; (ii) Better-scaling algorithms and pipelines for processing genomic information from real and synthetic data and their application in creating digital twins of tumour evolution through multiscale cellular simulations; (iii) method development and applications of fluid dynamics both inside the body (e.g. nasal cavities, used for surgical planning) and outside, e.g. in terms of air flows. All three pillars involve joint code development and research activities with RIKEN-CCS groups and associated Japanese teams, implementations target both Supercomputer Fugaku and EuroHPC leadership resources, and new algorithms will be made available as open-source software in major codes.

All collaborations have already started their active phase, there is dedicated staff recruited to the EU-based teams, joint planning meetings have been held with the RIKEN-based teams, and there are project meetings every few weeks between the

Funded by the European Union
EuroHPC Joint Undertaking

active-involvement researchers in each project, as well as plans for each code to target at least one major EuroHPC resource as well as Supercomputer Fugaku.

Table of Contents

Funded by the European Union
EuroHPC Joint Undertaking

Table of Figures

Funded by the European Union
EuroHPC Joint Undertaking

1 Introduction

The biomedical projects in HANAMI have been set up to target areas of strong excellence in software as well as scientific applications where we have identified

- Considerable overlap between the application scientific interests with the Japanese research ecosystem,
- Existing software codes/projects with strong scientific impact that can be evaluated in terms of e.g. usage citations,
- Interest in cross-cutting usage of software, i.e. value from running European codes in Japanese resources and vice versa,
- Key technological software development that can have substantial impact on scaling and performance with 24-36 months of effort, and
- Important scientific advances that can be realized once the improvements are implemented.

The first major initiative of the work package is centred on biomolecular modelling and simulations, where there is a strong history of collaborations in particular between the GROMACS teams and RIKEN – GROMACS was the first major MD code to be accelerated for the K Computer[1], and for Supercomputer Fugaku RIKEN early contributed the SVE acceleration that is now being expanded to EuroHPC supercomputers using e.g. modern Nvidia Arm-based CPUs.

For HANAMI, the work is specifically focused on extreme scaling of single simulations by breaking the impasse previously created by the reliance on 3D FFTs, and by implementing this as a library with both CPU and GPU support, we expect that both GROMACS and the RIKEN-developed GENESIS code (Kobayashi et al., 2017) will benefit significantly – in particular since future RIKEN supercomputers are likely to include GPU-based accelerators.

The biomolecular efforts will also use these advances to model whole-cell projects, and the teams have initiated collaborations for their respective codes to start sharing input/output data formats to promote interoperability and data sharing.

[1] https://www.riken.jp/en/collab/resources/kcomputer/

Funded by the European Union
EuroHPC Joint Undertaking

The second scientific problem targeted is personalized medicine. This field is currently growing extremely fast, although most of the tool development has previously not yet been focused on high-end parallel computers. For HANAMI, we have developed two use cases in particular needs of HPC resources due to their reliance on generative models to produce synthetic genomic data. In line with the EuroHPC focus on FAIR data policies, HANAMI will create reproducible pipelines operated by workflow managers and containerization technologies to ensure efficient HPC utilization. Second, we will leverage synthetic data generated from real cancer samples to personalise multiscale cellular simulations, creating digital counterparts of patient tissue organization at both molecular and cellular levels. As the data are synthetic, this excludes issues related to GDPR and sensitive data for the HANAMI project. Generative models will enable us to simulate thousands of potential tumorigenic trajectories and evaluate their effects on tumour growth from a cellular point of view, and we will particularly develop tools that enable more users in personalized medicine to move their utilization to the largest HPC resources in EU and Japan.

Finally, the third part of the work package is targeting method development in fluid dynamics applied to biological systems in general, and flow of air or blood in organs in particular. This area is currently less developed in life science (in terms of scientific impact), but it has the advantage of using HPC algorithms that often scale exceptionally well. Here, our efforts are focused both on jointly developing codes to make them available to researchers and attempt to develop cases that showcase how fluid dynamics HPC approaches can be used for impact in biomedical science.

Funded by the European Union
EuroHPC Joint Undertaking

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

Molecular dynamics (MD) simulations constitute an extremely powerful, but compute intensive, technique to investigate biomolecular systems in atomistic details. The method typically relies on using semi-empirical classical approximations for the potential function of interacting atoms and based on the force on each atom it is the possible to update the position of each atom a very short timestep later. For this to work, the integration time step must be in the order of femtoseconds, which means
billions of timesteps are needed to reach biologically relevant time scales. Thus, parallel scaling is critical to reduce the time to solution. In *theory* this is a straightforward problem since interactions between atoms can be calculated independently, and many simulation codes (including GROMACS (Abraham et al., 2015; Páll et al., 2020), developed within EuroHPC at KTH and collaborators) scale quite well. However, for *practical* applications, the need for billions of timesteps means users often target 1000-10,000 timesteps per second, which creates extreme demands on load balancing, low-latency communication, and algorithms with good scaling properties to thousands of nodes.

The goal of the work in HANAMI is to enable practical classical molecular dynamics simulations for biomolecular and material science systems to target tens to hundreds of thousands of nodes in a *single* simulation without resorting to ensemble parallelism. The two current critical bottlenecks we have identified to make this possible is to (1) replace the traditional methods for long-range electrostatics due to their non-ideal scaling behaviour, and (2) enable simulations with hundreds of millions to billions of particles to make it possible to model whole-cell systems, in particular combined with machine-learning-based potentials to describe interactions beyond atomic level.

Funded by the European Union
EuroHPC Joint Undertaking

2.1 Task 5.1: Exascale Fast Multipole Methods

Presently, the ultimate scaling in parallel simulations is usually limited by the need to calculate long-range electrostatics. This is traditionally performed by using the so-called "Particle Mesh Ewald" algorithm (Essmann et al., 1995), which in turn requires 3D Fast Fourier Transforms (3D FFT). Formally the computational aspect of this algorithm scales as O(N log N), where N is the number of atoms. However, the 3D FFT algorithms involve a transpose operation of the global grid, and for parallel implementations this unfortunately leads to $O(M^2)$ communication messages, where M is the number of nodes (Figure 1). The bandwidth can be handled since the amount of data per message is reduced as the level of parallelism increases, but since each message must be sent/received and handled, the inherent latency limits scaling in all simulations that depend on PME. The goal of the biomolecular work in the HANAMI project is to resolve this.

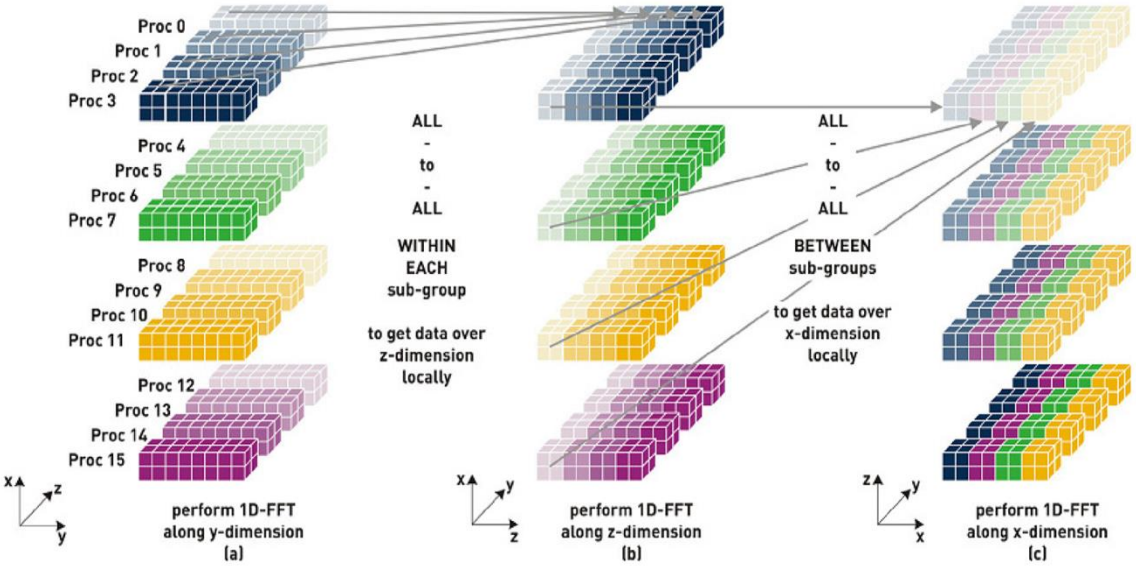


Figure 1. Illustration of a parallel implementation of the 3D Fast Fourier Transform, in particular the inherent grid transposition operations and the communication stages required. The number of messages sent increase as $O(M^2)$, where M is the number of nodes, which leads to unacceptable latencies in the limit of high-end parallelism. Reproduced from (Jagode, 2005).

Because of the parallelism limitation, the field has been investigating alternative solvers for electrostatics that provide better scaling properties. The most promising method is the Fast Multipole Method (FMM), which is widely used in N-body simulations. This method has O(N) arithmetic complexity and also O(N) computational complexity, which are the optimal scaling properties. But nearly all

Funded by the European Union
EuroHPC Joint Undertaking

available implementations have been optimized for the pure N-body problem, which means the memory limited regime, where molecular dynamics is in the latency limited regime. Thus, most codes require a complete rewrite, not to mention that existing implementations have had issues with energy conservation and absolute performance for specific-size problems when compared to traditional PME algorithms.

We have a long-standing collaboration with Prof. Rio Yokota from Tokyo Tech Institute (also affiliated with R-CCS), the developer of ExaFMM (Wang et al., 2021) , which is one of the fastest FMM codes in the world. Five years ago, we solved a fundamental issue for the application of FMM to molecular simulation, namely that standard FMM does not produce a conservative potential and forces (Shamshirgar et al., 2019).

Because there are discontinuities in the potential, very high accuracy was required in the FMM method to bring down the energy drift in MD simulations to an acceptable level. By instead applying a regularization, energy conservation can be guaranteed, and the accuracy requirements of the FMM method can be relaxed. This was a critical step to making FMM viable for use in MD simulations targeting e.g. biomolecular and material sciences systems (Figure 2, Figure 3).

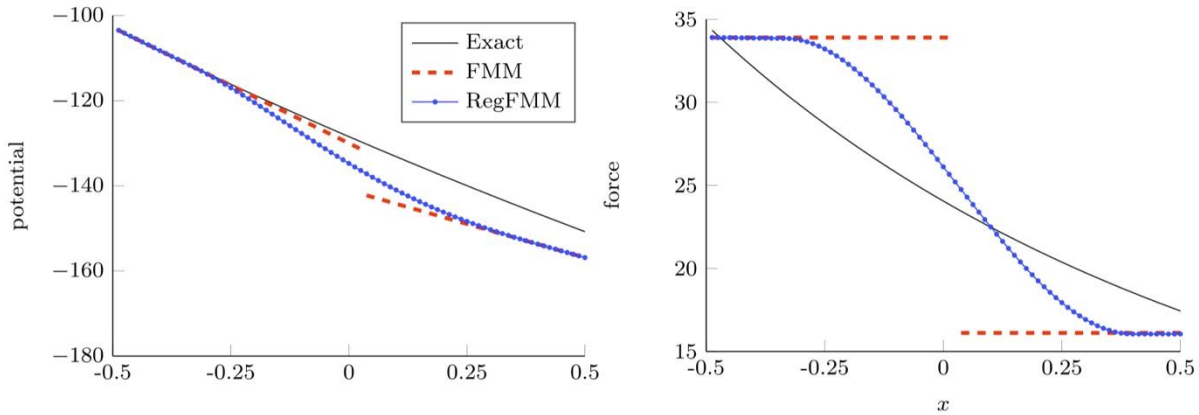


Figure 2. The challenge with discontinuities in potential/force between the short- and long-range regions for traditional FMM (red), and how the regularized FMM method developed by Yokota & Hess (Shamshirgar et al., 2019) solves this (blue).

Funded by the European Union
EuroHPC Joint Undertaking

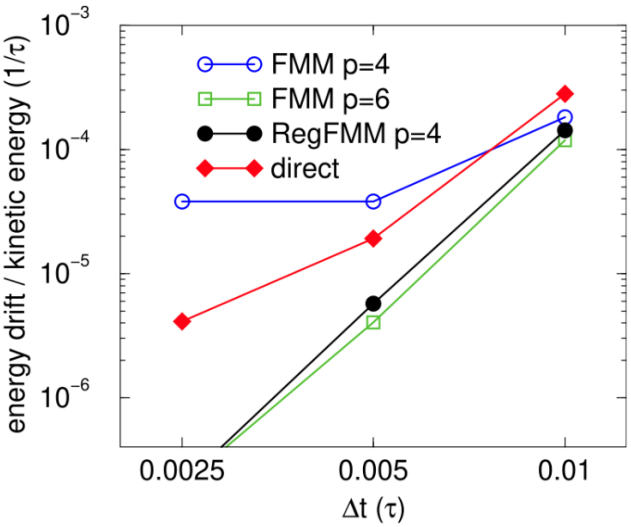


Figure 3. The regularized ExaFMM implementation provides much lower energy drift relative to the kinetic energy, which is important to retain a conservative potential and sample a single free energy landscape. Note how the regularized FMM with p=4 terms in the expansion achieves the same accuracy as traditional FMM with p=6 terms. Both of these drifts are below the statistical threshold, while both lower-accuracy traditional FMM and direct summation lead to much larger errors.

To make FMM usable in MD simulations in practice, significant work remains to be done. First, the ExaFMM code needs to be adapted for the atomistic molecular dynamics regime and regularization needs to be implemented in a production code. The group of Rio Yokota (Tokyo Tech Institute) is currently working on this, and as part of HANAMI the KTH team has hired Dr. Umair Sadiq as a new postdoctoral scholar, with additional assistance from Prof. Berk Hess (KTH).

The plan is to have two implementations, on targeting CPUs and one targeting GPUs, thus covering all important HPC platforms in both the EU and Japan.

Secondly, an interface (API) is required to transfer coordinates and charges from GROMACS to ExaFMM and to transfer the computed forces back.

Validation tests are required on the coupled code to check the potential and the forces and, in particular, energy conservation.

After that, performance test will be performed to study performance versus accuracy and to find the optimal parameter setup.

We are using a range of reference molecular systems for the design, ranging from large water boxes to complete-cell systems with billions of particles. As GROMACS already supports Multiple-Program Multiple-Data (MPMD) parallelization for electrostatics for the standard electrostatic interaction methods, we will reuse MPMD here. This means that a subset of the MPI-ranks will be dedicated to performing the FMM calculation only. This reduces the number of ranks

Funded by the European Union
EuroHPC Joint Undertaking

participating in the FFM part as well as the MD part, thereby reducing the parallel overhead. Or, alternatively, the scaling limit is pushed out further. The setup will be transparent to the user of the API (FMM). The FMM code will simply receive an MPI communicator and a division of space and particles. This in turn will make it relatively straightforward to also port this implementation to the RIKEN-developed GENESIS (Kobayashi et al., 2017) molecular dynamics code once it is working.

Finally, we will investigate the option of performing the direct pair-interactions of the FMM calculation the GROMACS non-bonded kernel. The ExaFMM code can compute both the direct particle-particle interactions and interactions involving multipoles. But as GROMACS already need to compute pair interactions for the Van der Waals interactions, computing the direct electrostatics interactions along with those is likely beneficial. With plain FMM this would be rather straightforward, the only complication is the MD and FMM code need to agree on which pair interactions should be computed. With the regularization, the pair interactions become much more complex and longer ranged. Their interaction pattern is rectangular, not spherical as for the Van der Waals interactions. In collaboration with the group of Prof. Rio Yokota, we will investigate strategies for computing the direct pair interaction in GROMACS.

The current work division is that the Tokyo Tech Institute group is responsible for extending the present single-threaded implementation with strategies for kernel-independent FMM implementations that can handle both regularized and non-regularized versions in a common code, as well as supporting both periodic and non-periodic boundary conditions, while the KTH team is responsible for parallelization and creating the interface to GROMACS. The teams are jointly working on parallel scaling both in the library and GROMACS as well as benchmarking, with a target of having the initial internal versions working in GROMACS by project-month 12, a working version shared with external groups by project-month 18 that also supports multi-threading, and the supported release available in GROMACS release 2026. Based on the same library, the Sugita team at RIKEN R-CCS will support the library in the GENESIS code using the same API.

Funded by the European Union
EuroHPC Joint Undertaking

Benchmarks will be run on LUMI as well as HANAMI by project-month 30, where the EuroHPC/RIKEN teams currently have large allocations.

2.2 Task 5.2: Targeting cellular-scale systems with machine-learning methods

Historically, biomolecular MD simulations have been limited to relatively small systems. One important reason for this is that smaller systems (meaning: fewer atoms) made each time step faster on scalar or low-parallelism hardware, and since biological processes depend on reaching sufficiently long timescales there has been a strong driving force to limit the system size. The other limitation is that as the size of a system grows, the timescales of most motion and dynamics grow even faster, so despite potentially beautiful illustrations, the impact of performing very short simulations for gigantic system can be argued to be limited.

While these points still hold in general, the much-improved scaling and higher performance of modern MD codes means several groups want to use simulations to model highly complex systems involving multiple membranes, proteins, and dozens of other molecules to better understand interactions in biologically relevant surroundings.

One particularly important area for large-scale models is the recent emergence of cryo-electron tomogram data collection for entire cells, starting with the minimal cell JCVI-Syn3A (Hutchison et al., 2016), where the Elizabeth Villa (UCSD) and Zan Luthey-Schulten (UIUC) groups used tomograms to generate lattice-based models including placement of nucleic acids and chromosomes (Gilbert et al., 2021). As part of our joint centre for Quantitative Cell Biology (QCB), Thornburg et al. recently published the first-ever complete model of a cell that could predict simple time-dependent phenomena (Thornburg et al., 2022) (Figure 4, left).

To be able to parametrize the lattice models from physical interactions and simulate molecular interactions such as binding and diffusion, we have used these lattice models prepared the first-ever coarse-grained simulations of entire cells (Figure 4, right).

Presently there are numerous challenges with these simulations, including scaling limitations due to imperfect load balance and bottlenecks that have not been evident for smaller number of particles, including e.g. numerical issues when

Funded by the European Union
EuroHPC Joint Undertaking

summing interactions from millions of interaction pairs, or losses in accuracy when calculating differences in coordinates between atoms close to each other but far away from the origin – but we have been able to perform coarse-grained simulations up to 100 ns.

For the second task in the HANAMI project, we are systematically addressing all these scaling limitations to enable simulations of arbitrary-size systems – tentatively only limited by 64-bit integer enumerations. We also need to solve remaining problems with extensive processing times when preparing or starting simulations since these processes have previously been single-threaded and performed on a single node, and the trajectory writing has to be parallelized to avoid a bottleneck when running on large number of nodes. Since these efforts are co-funded by Swedish national resources, we expect to have addressed all of them together with Dr. Sadiq (KTH) by project-month 18, which is also when the ExaFMM module will be ready.

Finally, while it is possible to use both atomistic and coarse-grained classical force fields for these simulations, in materials science there is a strong emerging trend to rely more on knowledge-based neural network or AI force fields. In preliminary work we have added code to GROMACS to support forces provided e.g. by DeePMD-Kit (Zeng et al., 2023) and other external neural network force fields. In particular to target the whole-cell modelling, we need to extend this support to the regime of scaling to thousands of nodes, and ideally develop ways to combine knowledge-based potentials with physical interactions at long range, including e.g. the fast multipole methods of Task 5.1. Being able to perform whole-cell simulations using machine-learning force fields should also make it possible to train these force fields with either experimental data or lattice-model simulation results as the loss parameter, which would enable a new type of super-coarse-grained force fields that might make it possible to reach even longer timescales than today's coarse-grained force fields for large systems. We expect to support the first implementations of ML force fields in GROMACS internally by project-month 18, in public releases by project-month 24, and publish applications to whole-cell-scale systems by the end of HANAMI, project-month 36. In addition to the efforts focused on improved parallelization, the GENESIS (Japanese side) and

Funded by the European Union
EuroHPC Joint Undertaking

GROMACS (European side) teams have also initiated collaborations to implement shared formats for input, output and trajectory formats – with additional links to the Amber (Case et al., 2005) and NAMD (Phillips et al., 2005) developers – with the goal of establishing new de-facto world-wide standards for describing, sharing and archiving molecular simulation data and metadata between codes. The Sugita R-CCS team is working together with the KTH team on defining the trajectories and data onthologies, with the aim of having the initial specifications ready by project-month 12, implementations in GROMACS as well as GENESIS by project-month 24, and showing that the codes can exchange data and metadata by project-month 36.

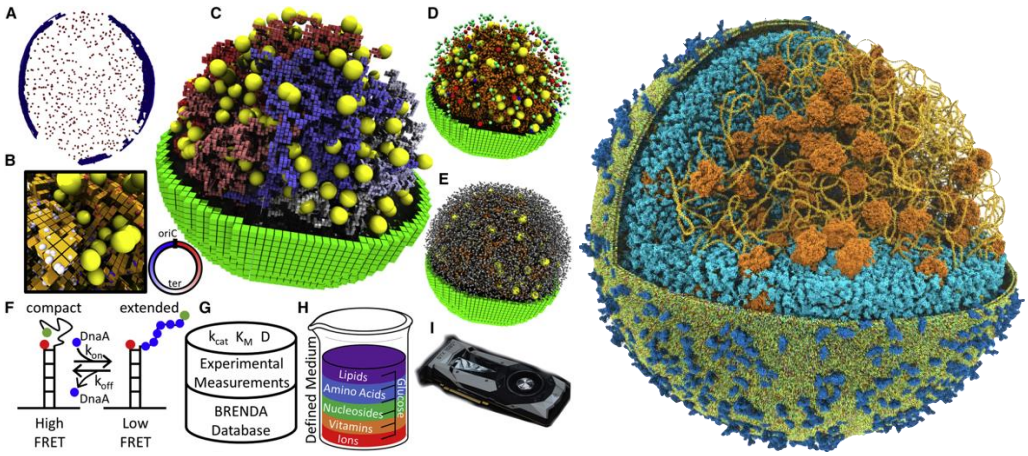


Figure 4. Left: Lattice-resolution simple models of the JCVI-Syn3A minimal cell, based on tomography data. Reproduced from (Thornburg et al., 2022). Right: The first-ever coarse-grained particle model of the minimal cell, using the MARTINI force field in GROMACS, using 880 million coarse-grained particles.

2.3 Conclusions

The RIKEN R-CCS and HANAMI biomolecular collaborations have been established and are running smoothly with regular online meetings, and there is a promising roadmap of significantly increased interoperability and exchange of code between two packages with extensive usage and strong application publications.

Prof. Erik Lindahl from KTH has already visited the Kobe laboratory of Prof. Yuji Sugita (RIKEN R-CCS biomolecular and AI4SCIENCE area co-director, PI of GENESIS) July 8-12, 2024, to initiate the collaborations. Dr. Sadiq (KTH) started his appointment in Stockholm summer 2024 and organizes roughly bi-weekly zoom meetings involving Prof. Rio Yokota (RIKEN & Tokyo Tech) as well as Prof. Berk Hess (KTH). Both Dr. Sadiq and Dr. Hess will visit Japan for the collaborations, currently targeting fall 2024 and spring 2025, respectively. Prof. Sugita visited

Funded by the European Union
EuroHPC Joint Undertaking

Europe quite recently and will be a recurring visitor with travel co-funded e.g. through CECAM conferences.

Both KTH & RIKEN are currently running strategic initiatives on AI4SCIENCE involving e.g. the development of foundation models for biomolecular research, and the HANAMI teams is a collaborator and supporting partner on a recent application by Profs. Mohamed Wahib (RIKEN R-CCS) and Yuji Sugita to the JST ASPIRE[2] call for "Foundational Computational Infrastructure for AI-driven Science", similar to the RIKEN teams' involvement in the HANAMI project.

[2] https://www.jst.go.jp/aspire/en/program_e/announce_e/announce_aspire2024_e.html

Funded by the European Union
EuroHPC Joint Undertaking

3 Project II: Development of genome analysis pipelines for Personalized Medicine

Personalized Medicine is becoming a reality rooted in the vast data availability and high computing needs. Among the many challenges in this area, we propose in this use case to address two of the more intense ones in terms of HPC demands and complexity.

The first one is related to the use of generative models to produce synthetic genomic data. These models are gaining significant attention as they allow developers to efficiently benchmark their analysis tools without the ethical and legal restrictions associated with real genomic data, and at the same time can be used to explore biological scenarios that are not accessible to the real data. These models require intensive HPC resources to massively explore the space of potential genomic features and produce the corresponding sequencing data including classical artefacts found in real samples. The developments will be encapsulated in reproducible pipelines operated by workflow managers and containerization technologies, which ensure efficient utilization in HPC environments.

In the second part of this use case, we will leverage the synthetic data generated from real cancer samples to personalise multiscale cellular simulations, creating digital counterparts of patient tissue organization at both molecular and cellular levels. Generative models will enable us to simulate thousands of potential tumorigenic trajectories and evaluate their effects on tumour growth from a cellular point of view. Multiscale simulations represent one of the most promising approaches to implementing digital twins, where the use of parallel and distributed computing becomes essential. By simulating billions of cells, we will accurately model tumour masses and their surrounding microenvironment, thereby capturing the differential behaviour of different parts of the tumours in patient-specific characteristics with high fidelity.

Funded by the European Union
EuroHPC Joint Undertaking

3.1 Task 5.3: Genome analysis pipeline

The computational challenges in personalised medicine have been amplified by the growing volume of sequencing data, also represented by the adoption of synthetic generative data models that allow developers to test their software while avoiding the special legal and ethical requirements of patient data. The importance of synthetic omics data generation lies in its ability to replicate the characteristics and patterns of real-world data without exposing confidential information. This is particularly relevant in the field of biology where the collection of data, such as clinical data containing sensitive patient information, is not straightforward (Selvarajoo and Maurer-Stroh, 2024). Synthetic data are generated using statistical methodologies or machine learning and are used for a wide range of applications, including as test data for new products and tools, and for model training and validation without compromising consumer privacy. It has also been shown to be less expensive, as it reduces the number and time taken for experiments and combines well with the real data to increase the overall number of observations. Likewise, its HPC needs are rooted in the algorithms used for the generation of synthetic genomes as well as the many genomes needed to realistically replicate a population of patients.

In the context of the European Health Data Space (EHDS) and General Data Protection Regulation (GDPR), synthetic omics data generation becomes even more significant (Ahmed et al., 2024). The EHDS aims to empower individuals to take control of their health data and facilitate the exchange of data for the delivery of healthcare across the EU (Marcus et al., 2022). Synthetic omics data, by preserving data security while still allowing researchers, analysts, and decision-makers to gain valuable insights, aligns perfectly with the goals of the EHDS.

One of the areas of biomedicine that has benefited most from synthetic data generation is cancer research. Based on sequencing data, there are several alterations that can be identified, such as single nucleotide variants, structural variants, copy number changes, etc. The methods that identify these alterations have different approaches that may make different calls. Developing and evaluating

Funded by the European Union
EuroHPC Joint Undertaking

these approaches requires datasets of sequencing data, which face two main issues: sensitivity of the data and availability of ground truth. Synthetic tumour sequencing data can address both issues.

The BSC has implemented an approach that uses a sequencing read simulator, NEATGenReads (Stephens et al., 2016) to turn a specification of the ground truth into a dataset of sequencing reads. These simulators produce realistic reads by using a model trained on real data that can reflect sequencing errors, changes in coverage-based GC content, etc. The ground truth is specified as an evolutionary tree of clones, each having acquired a set of somatic mutations while inheriting the mutations of the parent. Additionally, a germline genotype is also specified. The pipeline turns this ground truth into one or more synthetic sequencing datasets by simulating reads from each clone separately and then mixing them according to specified clonal fractions. The simulation pipeline supports short variants (SNVs and Indels) as well as structural variants (SVs). Also, it keeps track of variant phasing in different chromosomes even through duplications and deletion events arising from SVs and produces realistic breakpoints and fusions in the DNA reads. It can also simulate tumour-in-normal and normal-in-tumour contamination at any specified level of purity. These simulations can be very costly computationally, the process has been parallelized so each chromosome copy is simulated in parallel for each clone, but they still can take hundreds of CPU hours per simulation. The underlying read simulator could be swapped with another or have its efficiency improved.

Simulating the ground truth faithfully as DNA sequencing reads is relatively well solved. However, determining what is a realistic ground truth to simulate remains an open question that tracks our own understanding of the underlying biology. The complexity of the ground truth can range from small scenarios around a particular alteration e.g. a variant with low cellularity or a complex structural rearrangement, designed to test a very specific detection tool, to more comprehensive datasets that represent complete clonal evolution scenarios with several sequencing rounds following a simulated patient journey, and where clones present different mutational processes, respond differently to treatment or are able to establish metastases. There are different elements that build into that ground

Funded by the European Union
EuroHPC Joint Undertaking

truth, such as clonal structure, driver and passenger mutations, germline genotype, small variants, copy number changes and other structural variations, mutational signatures, mutations leading to genomic instability or resistance to treatment of metastatic potential, etc. Currently, we use data from knowledge bases and from real cohorts to source some of these elements, either by sampling randomly or by using generative approaches to model them from real data using AI.

Within the entire process of generating synthetic tumour data, there are several places that are potentially interesting scientifically, especially in the generation of the ground truth: germline and somatic genotypes, clonal structure and drivers, mutational signatures, sourcing driver and passenger SVs, etc. Connecting DNA sequencing to RNA sequencing is a complicated task. Reflecting mutations in DNA into RNA may be straightforward, but determining how mutations in DNA affect RNA quantities, splice variants, or allele frequencies, is still largely an open problem.

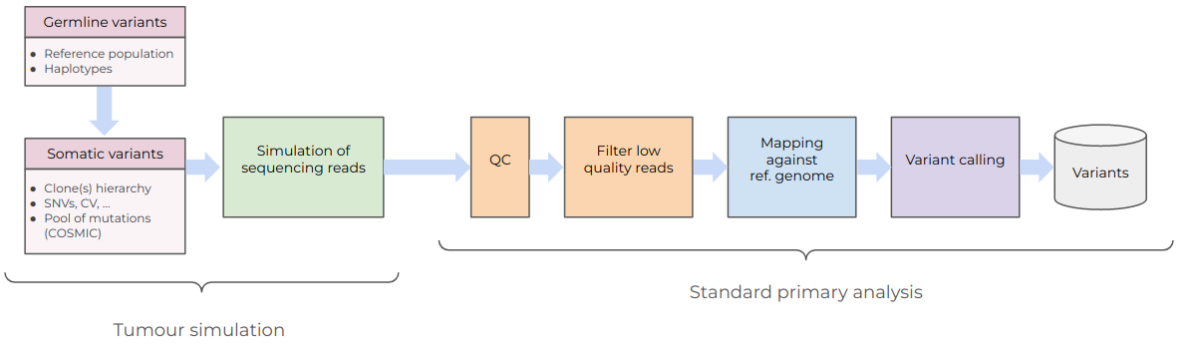


Figure 5. Computational pipeline representing tumour genome simulations and the subsequent primary analysis also compatible with real samples.

Figure 5 illustrates the computational workflow designed to simulate synthetic tumour samples. The workflow starts by simulating both germline and somatic mutations from CSC and BSC in-house developed tools. Then, NEATGenReads is used to generate the corresponding *fastq* files to represent the sequencing protocol. It also includes a standard set of primary processing steps typically used to identify germline and somatic mutations in a tumour sample. In particular, it includes QC of reads with FastQC[3], filtering low-quality reads with Samtools (Danecek et al., 2021), mapping of filtered reads against the human reference

[3] http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Funded by the European Union
EuroHPC Joint Undertaking

genome with BWA (Li and Durbin, 2010) and variant calling with Mutect2 (Cibulskis et al., 2013).

Although the mutations in the synthetic genomes are already known, this primary analysis is incorporated to replicate the same procedures usually applied to real tumour samples. This approach allows for the evaluation of how sequencing artefacts introduced by the read simulator and the primary analysis impact the pipeline's ability to detect existing mutations. From a computational perspective, the workflow can be parallelized to process thousands of synthetic genomes simultaneously. During a single processing run, the sequencing reads simulation, and the mapping steps are the most computationally intensive. These steps natively support parallel execution on all available CPU cores on a given compute node, ensuring maximum use of available computing resources for each execution.

In this task, we will simulate **a set of approximately 1,000 synthetic genomes** using the methodologies described above, representing different mutational trajectories in a population of cancer patients. The generative model will be informed by the somatic mutations of the most prevalent solid cancers, such as breast cancer. In addition, the computational workflow will be adapted and extensively benchmarked both in European infrastructures (such as Marenostrum5 and LUMI) and Supercomputer Fugaku. The evaluation will include standard metrics such as CPU time, memory consumption or I/O load, both at the general and building block level. In addition, specific metrics will be obtained to measure parallel and communication efficiency and computational scalability, as well as energy metrics collected to evaluate power consumption at each step. Finally, the generated synthetic genomes will then be used in the next task (Task 5.4) to generate multiscale simulations, thus creating digital twins of real patients.

3.2 Task 5.4: Tumour evolution simulation pipelines

Computational simulations are revolutionising biomedical research by enabling scientists to explore the dynamics of complex biological systems, thereby enhancing tasks such as large-scale screening of potential new drug compounds. In this context, simulations of cellular populations extend traditional mechanistic

Funded by the European Union
EuroHPC Joint Undertaking

models in systems biology by addressing crucial aspects of tissue organisation, such as cell-to-cell communication. These simulations are typically represented as multiscale systems, concurrently simulating various spatial and temporal scales to accurately predict real system responses in a highly optimised manner. Despite challenges related to parameterization and computational complexity, these techniques are proving invaluable for building patient-specific digital twins, enhancing our understanding of prevalent complex diseases such as cancer while paving the way for a new generation of computational medicine techniques.

PhysiBoSS, a multiscale modelling framework, is used to create digital twins of tumours by simulating intracellular signalling and multicellular behaviour, which are key aspects of tumour growth and development (Ponce-de-Leon et al., 2023). By creating a digital twin, or a virtual replica, of a tumour, researchers can study its growth patterns, cellular interactions, and response to various treatments in a controlled, virtual environment. This approach allows for a more detailed understanding of the tumour's characteristics and which mechanics commit the cells to abnormal growth and invasion.

We have lately been working on MPI (Message Passing Interface) and GPU (Graphics Processing Unit) methods that have the potential to significantly enhance the capabilities of PhysiBoSS. MPI is a standardised and portable message-passing system designed to function on a wide variety of parallel computing architectures. It can help in distributing the computational load across multiple machines, thereby speeding up the simulation process and enabling the simulation of real-sized tumours of billions of cells using hundreds of CPUs. GPUs, known for their high computational power, can also be leveraged to handle the complex calculations involved in tumour simulations. The use of GPUs can lead to a substantial reduction in computation time, making it feasible to run larger and more complex simulations. We are currently working on these expansions to integrate them into PhysiBoSS to enhance the scalability and efficiency of the simulation of digital twins of tumours.

On the other hand, PROFILE is a tool designed to personalise patient-specific Boolean models using omics data (Béal et al., 2019). Boolean models are a type of

Funded by the European Union
EuroHPC Joint Undertaking

mathematical model used to represent biological systems, where each element (e.g., a gene or protein) can exist in one of two states (e.g., on or off). In the context of personalising intracellular models based on simulated genomes, PROFILE can be used to tailor these models to reflect the unique characteristics of a patient's cellular biology by integrating patient-specific omics data into the Boolean model variables. For instance, if the genomic data indicates a certain gene is mutated in a patient's cells, this information can be used to adjust the state of the corresponding element in the Boolean model. Similarly, transcriptomics data, which provides information about gene expression levels, can be used to further refine the model. By personalising the Boolean models in this way, PROFILE enables the creation of patient-specific intracellular models that reflect the biological characteristics of a patient's cells (Montagud et al., 2022). This can provide valuable insights into the cellular processes driving disease in individual patients and may ultimately guide the development of personalised treatment strategies.

Tumour evolution simulation pipelines that include HPC-intensive synthetic genome generation, personalisation of Boolean models using PROFILE, and HPC-intensive multiscale simulations using PhysiBoSS can face several computational bottlenecks. The first bottleneck can occur during the synthetic genome generation process. This process involves simulating the entire genome of a cell, which can be computationally intensive due to the large size and complexity of the human genome. Additionally, the generation of multiple synthetic genomes to represent the genetic heterogeneity within a tumour can further increase the computational load. The second potential bottleneck is that the multiscale simulations using PhysiBoSS can also pose a significant computational challenge. These simulations aim to model the behaviour of millions of cells over time, which requires substantial computational resources. Furthermore, the integration of intracellular models (from PROFILE) into the multicellular framework of PhysiBoSS adds another layer of complexity, potentially slowing down the simulation process. Therefore, efficient computational strategies and HPC resources are crucial for the successful implementation of such a comprehensive tumour evolution simulation pipeline.

Funded by the European Union
EuroHPC Joint Undertaking

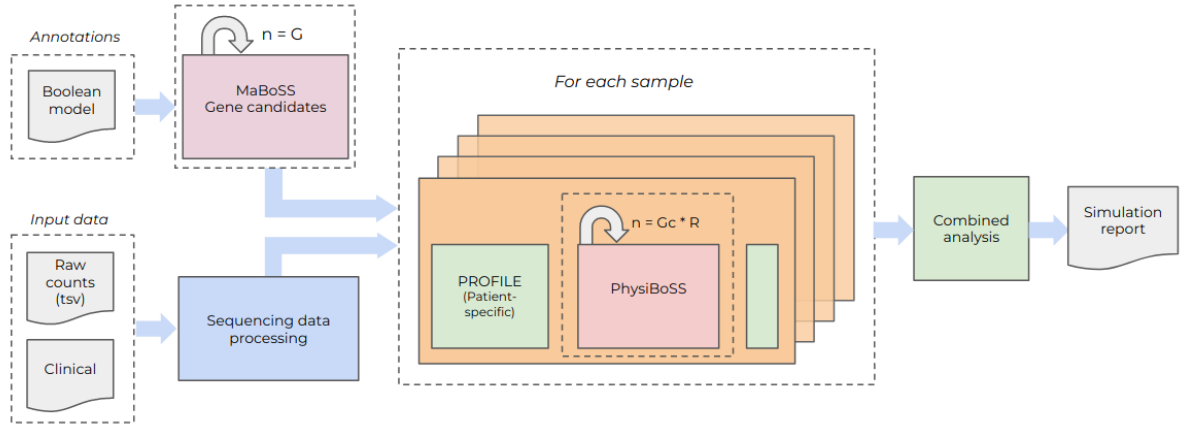


Figure 6. Computational workflow describing sample-specific personalisation and the downstream multiscale cell-level simulations.

Figure 6 illustrates the computational workflow designed to create digital twin representations both of synthetic and real patients. The workflow begins with processing molecular information from samples to generate a list of potential somatic alterations affecting molecular pathways. Simultaneously, selected pathways, modelled as Boolean networks, are analysed to identify potential protein candidates for drug interventions. Patient-specific customizations of these pathways are then prepared using PROFILE, followed by multiscale cellular simulations with PhysiBoSS. Finally, all patient-specific results are integrated to provide a comprehensive overview of the sample population. As with the previous section, the computational workflow will undergo extensive testing both at European facilities and on Supercomputer Fugaku, employing similar metrics. This evaluation is particularly significant for assessing the distributed computing capabilities of PhysiBoSS using MPI. It will provide insights into how effectively the MPI-enabled version of PhysiBoSS scales with increased simulation resources, such as a higher number of cells or substrates in the medium.

CSC is focusing its efforts on Task 5.4 to enhance container support on HPC clusters, extending the reach of its solutions beyond Europe to organizations such as RIKEN (Fugaku) in Japan. The effort aims to improve container support by designing and prototyping a solution covering container building, provisioning, and software catalogues. The blueprint and prototype services will enable users to easily build custom containers targeting multiple architectures (LUMI, RIKEN, MN5) and provide the ability to publish software catalogues with optimized containers. These

Funded by the European Union
EuroHPC Joint Undertaking

containers are a key cornerstone for portable workflows, providing a common set of software.

The developed container technologies will be documented and made accessible online (e.g., at docs.csc.fi) and will be transversal, meaning they will be applicable to all scientific workflows within the HANAMI use cases and beyond. The HANAMI project by CSC primarily focuses on HPC containers for scientific software developed by the Permedcoe.eu and Bioexcel.eu HPC Centres of Excellence. While computations in HANAMI use cases are not intended to process sensitive data, the design principles of any new solutions will incorporate data privacy and security compliance measures for the target HPC systems

To achieve this objective, the HANAMI resources will be used to collaborate with consortia in the field of Personalized Medicine, including the European 1+ Million Genomes initiative, EOSC4Cancer, EOSC-Life, GDI, and EUCAIM, while leveraging standards such as GA4GH.org (e.g., TES Task Execution Service).

The anticipated outcome is an overall blueprint for providing portable multi-architecture software catalogues based on containers, and facilitate the support for automatic building and provisioning of containers, along with improved and comprehensive expert user documentation that will be openly available to facilitate global collaborations.

3.3 Conclusions

The described tasks on tumour simulations illustrate a use case of interest in personalised medicine, where the somatic alterations found in a particular patient constitute the starting point for designing a digital twin. In this context, the use of HPC becomes crucial since the computational requirements for processing tumour samples and their posterior cellular simulations require high-end HPC clusters. One of the objectives of these projects is benchmarking described use cases in different architectures in Europe and Japan to adapt the tools to the machines with the aim of providing inputs for the co-design of future computer architectures. Furthermore, the forthcoming availability of massive amounts of data in biomedicine will enable the modelling and simulation of tissue and organ dynamics with unprecedented precision. This surge in data will necessitate a substantial increase in computational resources, as well as the development of a new

Funded by the European Union
EuroHPC Joint Undertaking

generation of computational tools adapted to a more sophisticated computational environment, positioning HPC as a critical pillar for the future of biomedicine.

4 Project III: Fluids

This roadmap describes the collaborative work on software between the Simulation and Data Laboratory 'Highly Scalable Fluids and Solids Engineering' (SDL FSE) at the Jülich Supercomputing Centre (JSC), Foschungszentrum Jülich GmbH (FZJ), and the 'Complex Phenomena Unified Simulation Reseach Team' at RIKEN R-CCS to successfully realize Task 5.5 and Task 5.6 of the work package. It provides details about specific software features, use cases, and testing requirements that are identified to employ the software.

4.1 Task 5.5: AI-assisted automated CFD pipelines and acceleration of CFD computations

In this task, flow fields of numerical simulations will be initialized with a meaningful approximation coming from Physics-Aware Graph Convolutional Neural Networks (PA-GCNNs) that are jointly developed between JSC and RIKEN R-CCS. GCNNs in general are capable of predicting flow fields around irregularly shaped bodies using computational meshes that can easily be converted into graphs (Chen et al., 2021). With the physics-aware component, the governing equations of fluid mechanics are embedded into the loss function, which allows the prediction of flow fields solely based on geometric information of the computational domain. That is, the loss is computed based on boundary information and the residual of the governing equations, and no ground truth data from numerical simulations is required.

In a first step until project month 18, a general GCNN is developed that reads information about boundary conditions and neighbouring nodes from computational meshes of the open-source multi-physics simulation framework m-AIA (Lintermann et al., 2020) (formerly known as Zonal Flow Solver - ZFS). This GCNN is mainly developed by the Jülich side within the frame of the Joint Laboratory for Extreme-Scale Computing (JLESC) project 'Deep Neural Networks for

Funded by the European Union
EuroHPC Joint Undertaking

CFD Simulations'[4], which is an ongoing long-term collaboration between JSC and R-CCS. Bi-weekly meetings support an extensive exchange and joint developments. In a second step until project month 24, the physical loss is integrated into the GCNN. The integration of the physical loss is mainly done by the Japanese partners, who have already shown their expertise about physical loss functions in a recent joint publication about the choice of physical constraints when using physics-informed neural networks for flow predictions (Puri et al., 2024).

In the final step until project month 36, a flow initialization technique is tested for generic flow fields and nasal cavity flows. For the generic cases, three types of domains are investigated that are referred to as cases I-III. For each of these case, nine test flow domains are analysed not belonging to the training data of the PA-GCNN. In case I, flow through randomly constricted variations of a 2D channel featuring a channel height of D_I and a channel length of $L_I = 4\ D_I$ are analysed. An example configuration is shown in Fig. 1a. In case II, rectangular channels with a channel height of D_{II} and a channel length of $L_{II} = 2\ D_{II}$ that incorporate a star-shaped obstacle are investigated. This obstacle is generated by a random selection of a central point within the channel, around which a pre-defined number of vertices is randomly distributed. A corresponding example configuration is shown in Figure 7, panel b. In case III, flow through constricted 90° L-bend pipes are analysed. The pipes are constructed to have the diameter of D_{III} and long straight sections of length $L_{III} = 7\ D_{III}$ joined by a 90° bend. The constrictions are placed at fixed points along the straight sections, up to two before and up to two after the bend. The constrictions are created by reducing the radius of the pipe by up to 50%. An example configuration is illustrated in Figure 7, panel c. For the nasal cavity flows, the flow domains of the use case for Task 5.6 are used, which are described below.

4.2 Task 5.6: AI-assisted surgery planning and risk assessment of exhaled aerosols

This work will target surgery planning and waiting room scenarios analyzed by means of respiratory flow simulations. For the former a previously developed

[4] https://jlesc.github.io/projects/dnn_cfd/

Funded by the European Union
EuroHPC Joint Undertaking

Reinforcement Learning (RL) algorithm suggests modifications to the human airway to balance the objective functions of simultaneously minimizing the pressure loss and increasing the temperature between inflow and outflow regions (Rüttgers et al., 2024). After a modification, the algorithm receives feedback from m-AIA. To improve the predictive capabilities, the following two approaches are investigated until project month 18: (i) training of parallel environments are executed on multiple MPI ranks and the RL agents of each environment share their experience in a pre-defined interval and (ii) for some of the geometry modifications the expensive numerical solver is replaced by predictions from a Gaussian Process Regression (GPR) model for an improved computational efficiency. The GPR model is developed by the Japanese partners and integrated into the workflow of the surgery planning tool.

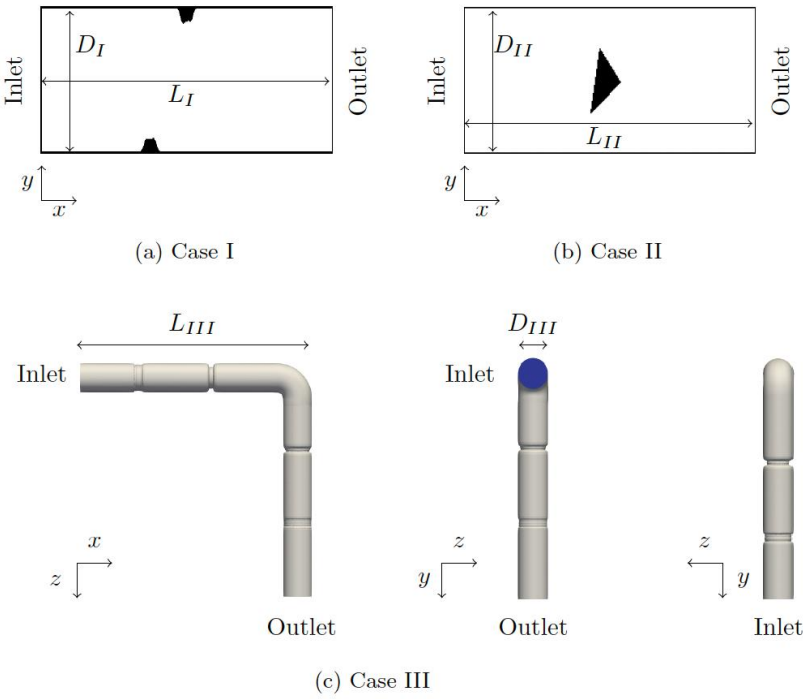


Figure 7. Computational domains for (a) 2D flows through a constricted channel (case I), (b) 2D flow around polygonial obstacles (case II), and (c) 3D flow through an L-bend with constrictions (case III). Each subfigure shows one example geometry for the corresponding case.

Funded by the European Union
EuroHPC Joint Undertaking

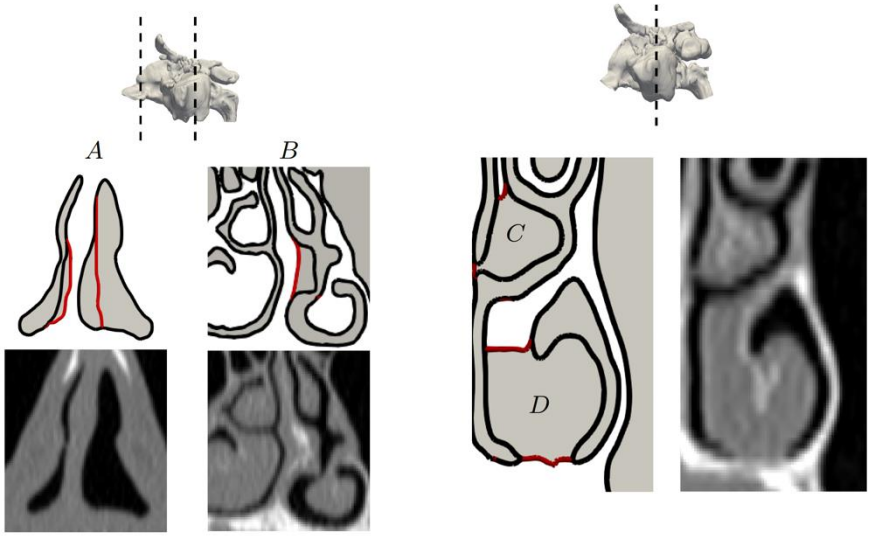


Figure 8. The nasal cavity of two patients, the first suffering from a deviated septum (A) accompanied by a bony spur (B), and the second from enlarged middle (C) and inferior (D) turbinates. The close-ups on cross-sectional areas illustrate the pre-surgical state (black), and planned interventions (red). They are juxtaposed to the corresponding pre-surgical CT images. The RL algorithm finds the optimal geometry modification between the black and the red states. Note that these are only representative cross-sectional areas. The RL agent modifies 3D regions.

Anonymized Computer Tomography (CT) data of two patients are used, see Figure 8. 2. The first patient suffers from a deviated septum (location A) and a bony spur (location B), and the second patient from enlarged turbinates (locations C and D). The patients gave informed consent for inclusion of their data. The CT data of the first patient are composed of 119 axial slices with 512 x 512 pixels each. The pixel spacing is 0.5 mm, and the space between the axial slices is 0.7 mm. The CT recordings of the second patient have 103 axial slices, again with 512 x 512 pixels each. The pixel spacing is 0.326 mm, and the space between the axial slices is 1.0 mm. The 3D model of the pre-surgical upper airway is extracted from the Digital Imaging and Communications in Medicine (DICOM) files of the CT data with the pipeline described in (Rüttgers et al., 2022). For the first patient, surgery planning of a septoplasty is investigated, and for the second patient, the surgical potential of a turbinectomy is analysed. More details about the medical data, types of surgeries, and motivation from a medical background are given in (Rüttgers et al., 2024; Waldmann et al., 2022).

With an improved computational efficiency, the action space can be increased, and more data can be transferred between m-AIA and the RL algorithm. However, currently, the coupling is realized by overwriting and reading a Boolean-array, which informs m-AIA or the RL about the action to perform next, i.e., a simulation

Funded by the European Union
EuroHPC Joint Undertaking

of learning step. To allow for a more efficient data transfer between the two instances, an improved coupling with an MPI-based library, e.g, the Physics Deep Learning Coupler (PhyDLL)[5], at the interface is implemented until project month 24 by the Jülich side.

Finally, it is investigated whether m-AIA can be replaced in three steps by a variation of the GCNN mentioned in T.5.5, which is jointly developed with the Japanese partners from RIKEN R-CCS until project month 36:

- In the first step, a pre-defined number of flow fields for varying geometries is computed by m-AIA
- In the second step, the flow fields of these configurations are used as training data for the GCNN which is trained to predict the flow fields of varying geometries for the same patient.
- In the third step, the trained GCNN is coupled to the parallel RL algorithm to determine the optimized shape.

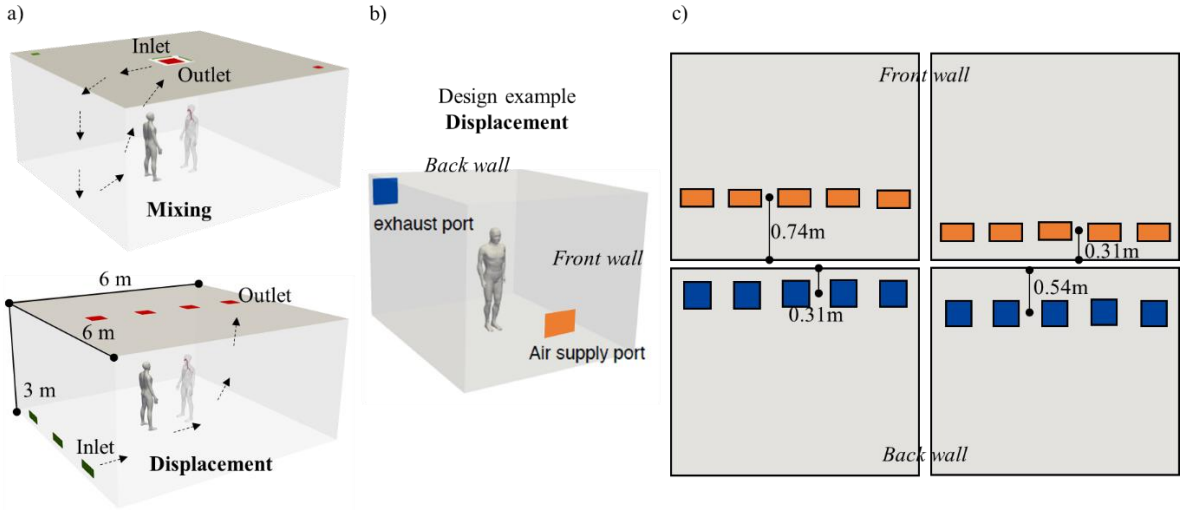
This approach allows to consider a much larger number of geometry modifications, and, therefore, leading to an increased action space. The goal is to train agents in modifying the CT data directly to explore geometry variations that go beyond the action space which is pre-defined by a surgeon. The flow simulations conducted for surgery planning currently require days, even if the number of simulations are reduced by the first step. Therefore, an important factor of the practicability of the proposed method in daily clinical environments is the access to Exascale HPC resources to reduce the computation time to hours rather than days.

The waiting room scenarios evaluate exposure to airborne pollutants, which is a major environmental health challenge, especially in a society that spends most of the day indoors. In the built environment, ventilation systems are essential to maintain air quality, but their performance is highly influenced by multiple factors, such as temperature, wind speed, or endogenous human emissions through

[5] https://phydll.readthedocs.io/en/latest/

Funded by the European Union
EuroHPC Joint Undertaking

respiratory events. Numerical flow simulations conducted by the Japanese partners are used to predict indoor flow and the distribution of pollutants. Specifically, Large-Eddy Simulation (LES), which precisely predicts particle tracking from the larger to the smallest scales of turbulence (Murga et al., 2023), are used. To model the complex interactions between the human envelope through its physiological functions and local indoor flow produced by the ventilation system, HPC is essential



in terms of data analysis, processing, and speed.

Figure 9. Use case for the analysis of waiting room scenarios. (a) Target room design, (b) Example of simplified room for optimization, (c) Position of inlets "air support port" (orange) and outlets "exhaust port" (blue).

This subtask evaluates the performance of various ventilation systems in the built environment by minimizing transmission of airborne particles measured through the deposition on human tissue, aided by two different types of complex virtual manikins: source-occupants and a receptor-occupants (until project month 24). Thereafter, the layout and size of the supply-exhaust vents and ventilation flow rate of the most advantageous system are optimized through a genetic algorithm that also considers energy consumption until project month 36.

A specific use case is shown in Figure 9 for two mannequins. The target room has dimensions of 6 × 6 × 3 meters and the ventilation rate correspond to the outdoor air requirement of 10 L/s per person. Outside, the room target as well as supply air temperature are set to 33 °C, 25 °C, and 20 °C, respectively. Two types of widely used ventilation systems are considered: 1) mixing ventilation, where the air is supplied-exhausted at ceiling level to form a well-mixed environment; and 2)

Funded by the European Union
EuroHPC Joint Undertaking

displacement ventilation, where the air is supplied at a low height and exhausted at ceiling level, creating a stratified distribution to carry pollutants above occupant-level. Building design and exemplified flow direction are shown in Figure 9, panel a. The two complex virtual mannequins are standing face-to-face in the middle of the room, separated by a distance of 1 meter. The source-occupant is constantly talking based on a 1-to-10 enumeration model, scaled 50% to simulate "loud" speaking (Gupta et al., 2010). Particles are released from its mouth at peak velocities during speech. The receptor-occupant has an integrated respiratory tract, attached to the body by the nares' surface until the 7th generation of bronchial tubes. A continuous breathing cycle based on a sinusoidal function is applied at the nasal openings and at the end of each bronchus. The main objective is to enhance the built environment by integrating human-built flow envelopes.

The testing requirements for the JSC and RIKEN R-CCS machines (Supercomputer Fugaku) are taken from successfully approved compute time proposals. On the Jülich side, e.g., for the project "Improved Diagnostics of Respiratory Flows Using a Lattice-Boltzmann Method and Machine Learning Techniques" and the granting period from 1 May 2024 to 30 April 2025 the following resources have been granted:

- JURECA DC Module CPU: 1.76 million core-h
- JURECA DC Module GPU: 2.60 million core-h
- FUGAKU CPU: 1.5 million core-h

After the end of the granting period, this project will be extended for the activities in HANAMI.

4.3 Conclusions

The described tasks build the foundation for a better understanding of respiratory diseases. This includes the understanding of nasal obstructions inside of the nasal cavity, but also infections risks in indoor scenarios that stem from expiration. Numerical simulations allow detailed analysis of flow fields, supported by machine learning techniques that help to accelerate processes or find optimal solutions

Funded by the European Union
EuroHPC Joint Undertaking

more efficient. In this context, the use of HPC becomes crucial since the computational requirements for highly resolved simulations and deep neural networks require high-end HPC clusters.

5 Long-term impact assessment

For each project area, we will track both the technical, infrastructure, and scientific impact of the work, with the three categories respectively characterized by (1) algorithmic and computer science publication advances, (2) benchmarks showing the progress compared to pre-HANAMI versions as well as the number of groups outside HANAMI using the improvements, and (3) joint publications between HANAMI/RIKEN teams on scientific applications as well as impact on the respective national computational/scientific roadmaps.

We will also specifically promote initiation of new collaborations targeting e.g. post-HANAMI development of new joint foundation models and similar generative AI methods that can be applied to the biomolecular scientific area.

6 Bibliography

Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B., Lindahl, E., 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Ahmed, Z., Wan, S., Zhang, F., Zhong, W., 2024. Artificial intelligence for omics data analysis. BMC Methods 1, 4, s44330-024-00004–00005. https://doi.org/10.1186/s44330-024-00004-5

Béal, J., Montagud, A., Traynard, P., Barillot, E., Calzone, L., 2019. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. Front. Physiol. 9, 1965. https://doi.org/10.3389/fphys.2018.01965

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J., 2005. The Amber biomolecular simulation programs. J. Comput. Chem. 26, 1668–1688. https://doi.org/10.1002/jcc.20290

Funded by the European Union

EuroHPC Joint Undertaking

Chen, J., Hachem, E., Viquerat, J., 2021. Graph neural networks for laminar flow prediction around random two-dimensional shapes. Phys. Fluids 33, 123607. https://doi.org/10.1063/5.0064108

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219. https://doi.org/10.1038/nbt.2514

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., Pedersen, L.G., 1995. A smooth particle mesh Ewald method. J. Chem. Phys. 103, 8577–8593. https://doi.org/10.1063/1.470117

Gilbert, B.R., Thornburg, Z.R., Lam, V., Rashid, F.-Z.M., Glass, J.I., Villa, E., Dame, R.T., Luthey-Schulten, Z., 2021. Generating Chromosome Geometries in a Minimal Cell From Cryo-Electron Tomograms and Chromosome Conformation Capture Maps. Front. Mol. Biosci. 8, 644133. https://doi.org/10.3389/fmolb.2021.644133

Gupta, J.K., Lin, C.-H., Chen, Q., 2010. Characterizing exhaled airflow from breathing and talking. Indoor Air 20, 31–39. https://doi.org/10.1111/j.1600-0668.2009.00623.x

Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science 351, aad6253. https://doi.org/10.1126/science.aad6253

Jagode, H., 2005. Fourier transforms for the bluegene/L communication network. (Master's thesis). The University of Edinburgh, Edinburgh.

Kobayashi, C., Jung, J., Matsunaga, Y., Mori, T., Ando, T., Tamura, K., Kamiya, M., Sugita, Y., 2017. GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational

Funded by the European Union
EuroHPC Joint Undertaking

platforms. J. Comput. Chem. 38, 2193–2206. https://doi.org/10.1002/jcc.24874

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595. https://doi.org/10.1093/bioinformatics/btp698

Lintermann, A., Meinke, M., Schröder, W., 2020. Zonal Flow Solver (ZFS): a highly efficient multi-physics simulation framework. Int. J. Comput. Fluid Dyn. 34, 458–485. https://doi.org/10.1080/10618562.2020.1742328

Marcus, J.S., Martens, B., Carugati, C., Bucher, A., Godlovitch, I., 2022. The European Health Data Space. SSRN Electron. J. https://doi.org/10.2139/ssrn.4300393

Montagud, A., Béal, J., Tobalina, L., Traynard, P., Subramanian, V., Szalai, B., Alföldi, R., Puskás, L., Valencia, A., Barillot, E., Saez-Rodriguez, J., Calzone, L., 2022. Patient-specific Boolean models of signalling networks guide personalised treatments. eLife 11, e72626. https://doi.org/10.7554/eLife.72626

Murga, A., Bale, R., Li, C.-G., Ito, K., Tsubokura, M., 2023. Large eddy simulation of droplet transport and deposition in the human respiratory tract to evaluate inhalation risk. PLOS Comput. Biol. 19, e1010972. https://doi.org/10.1371/journal.pcbi.1010972

Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., Hess, B., Lindahl, E., 2020. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J. Chem. Phys. 153, 134110. https://doi.org/10.1063/5.0018516

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., Schulten, K., 2005. Scalable molecular dynamics with NAMD. J. Comput. Chem. 26, 1781–1802. https://doi.org/10.1002/jcc.20289

Ponce-de-Leon, M., Montagud, A., Noël, V., Meert, A., Pradas, G., Barillot, E., Calzone, L., Valencia, A., 2023. PhysiBoSS 2.0: a sustainable integration of stochastic Boolean and agent-based modelling frameworks. Npj Syst. Biol. Appl. 9, 54. https://doi.org/10.1038/s41540-023-00314-4

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

Funded by the European Union
EuroHPC Joint Undertaking

flow
fields, Future Generation Computer Systems, Volume 161. https://doi.org/10.1016/j.future.2024.07.009

Rüttgers, M., Waldmann, M., Schröder, W., Lintermann, A., 2022. A machine-learning-based method for automatizing lattice-Boltzmann simulations of respiratory flows. Appl. Intell. 52, 9080–9100. https://doi.org/10.1007/s10489-021-02808-2

Rüttgers, M., Waldmann, M., Vogt, K., Ilgner, J., Schröder, W., Lintermann, A., 2024. Automated surgery planning for an obstructed nose by combining computational fluid dynamics with reinforcement learning. Comput. Biol. Med. 173, 108383. https://doi.org/10.1016/j.compbiomed.2024.108383

Selvarajoo, K., Maurer-Stroh, S., 2024. Towards multi-omics synthetic data integration. Brief. Bioinform. 25, bbae213. https://doi.org/10.1093/bib/bbae213

Shamshirgar, D.S., Yokota, R., Tornberg, A.-K., Hess, B., 2019. Regularizing the fast multipole method for use in molecular simulation. J. Chem. Phys. 151, 234113. https://doi.org/10.1063/1.5122859

Stephens, Z.D., Hudson, M.E., Mainzer, L.S., Taschuk, M., Weber, M.R., Iyer, R.K., 2016. Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. PLOS ONE 11, e0167047. https://doi.org/10.1371/journal.pone.0167047

Thornburg, Z.R., Bianchi, D.M., Brier, T.A., Gilbert, B.R., Earnest, T.M., Melo, M.C.R., Safronova, N., Sáenz, J.P., Cook, A.T., Wise, K.S., Hutchison, C.A., Smith, H.O., Glass, J.I., Luthey-Schulten, Z., 2022. Fundamental behaviors emerge from simulations of a living minimal cell. Cell 185, 345-360.e28. https://doi.org/10.1016/j.cell.2021.12.025

Waldmann, M., Rüttgers, M., Lintermann, A., Schröder, W., 2022. Virtual Surgeries of Nasal Cavities Using a Coupled Lattice-Boltzmann–Level-Set Approach. J. Eng. Sci. Med. Diagn. Ther. 5, 031104. https://doi.org/10.1115/1.4054042

Wang, T., Yokota, R., Barba, L., 2021. ExaFMM: a high-performance fast multipole method library with C++ and Python interfaces. J. Open Source Softw. 6, 3145. https://doi.org/10.21105/joss.03145

Funded by the European Union
EuroHPC Joint Undertaking

Zeng, J., Zhang, D., Lu, D., Mo, P., Li, Zeyu, Chen, Y., Rynik, M., Huang, L., Li, Ziyao, Shi, S., Wang, Yingze, Ye, H., Tuo, P., Yang, J., Ding, Y., Li, Y., Tisi, D., Zeng, Q., Bao, H., Xia, Y., Huang, J., Muraoka, K., Wang, Yibo, Chang, J., Yuan, F., Bore, S.L., Cai, C., Lin, Y., Wang, B., Xu, J., Zhu, J.-X., Luo, C., Zhang, Y., Goodall, R.E.A., Liang, W., Singh, A.K., Yao, S., Zhang, J., Wentzcovitch, R., Han, J., Liu, J., Jia, W., York, D.M., E, W., Car, R., Zhang, L., Wang, H., 2023. DeePMD-kit v2: A software package for deep potential models. J. Chem. Phys. 159, 054801. https://doi.org/10.1063/5.0155600

HANAMI
EU-Japan Alliance in HPC

Funded by the European Union
EuroHPC Joint Undertaking