Grant agreement no. 675451

CompBioMed

Research and Innovation Action
H2020-EINFRA-2015-1
Topic: Centres of Excellence for Computing Applications

D5.2 Report on computing and data needs of the biomedical community

Work Package: 5

Due date of deliverable: Month 09

Actual submission date: 30 / June / 2017

Start date of project: October, 01 2016 Duration: 36 months

Lead beneficiary for this deliverable: *SARA*
Contributors: *UEDIN, UCL, USFD, UNIGE, BSC, UOXF*

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Dissemination Level		
PU	Public	YES
CO	Confidential, only for members of the consortium (including the Commission Services)	
CI	Classified, as referred to in Commission Decision 2001/844/EC	

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

Disclaimer

Table of Contents

List of Tables and Figures

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

Version Log

Version	Date	Released by	Nature of Change
V0.01	6/04/2017	Marco Verdicchio (SARA)	First Draft
V0.1	18/05/2017	Marco Verdicchio (SARA)	Formatted draft
V0.2	26/05/2017	Marco Verdicchio (SARA)	Added results from survey
V0.3	18/5/2017	Marco Verdicchio (SARA)	Edit
V0.4	29/5/2017	Toni Collis (EPCC) Narges Zarrabi (SARA) Andrew Narracott (USFD) Jonas Latt (UNIGE)	Added co-authors contributions
V0.5	31/5/2017	Marco Verdicchio (SARA)	Edit
V0.6	8/5/2017	Toni Collis (EPCC) Narges Zarrabi (SARA) Stefan Zasada (UCL)	Added co-authors contributions
V0.7	9/5/2017	Marco Verdicchio	Added editors contributions
V0.8	12/5/2017	Jazmin Aguado-Sierra (BSC) David Wright (UCL)	Review
V0.9	27/5/2017	Peter Coveney (UCL) Emily Lumley (UCL)	Review
V1.0	28/7/2017	Marco Verdicchio (SURFsara)	Final version

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

2 Contributors

Name	Institution	Role
Marco Verdicchio	SURFsara	Author
Narges Zarrabi	SURFsara	Co-author
Toni Collis	UEDIN	Co-author
Andrew Narracott	USFD	Co-author
Jonas Latt	UNIGE	Co-author
Stefan Zasada	UCL	Co-author
Francesc Levrero Florencio	UOXF	Editor
Mariano Vasquez	BSC	Editor
David Wright	UCL	Reviewer
Jazmin Aguado-Sierra	BSC	Reviewer
Peter Coveney	UCL	Reviewer
Emily Lumley	UCL	Reviewer

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

3 Definition and Acronyms

Acronyms	Definitions
CoE	Centre of Excellence
HPC	High Performance Computing
SME	Small and Medium Enterprises
WP	Work Package
CBM	CompBioMed
CPU	Central Processing Unit
GPU	Graphics Processing Unit
CFD	Computational Fluid Dynamics
ODE	Ordinary Differential Equations
GNU GPL	GNU General Public License
GNU LGPL	GNU Lesser General Public License
NSCCS	National Service for Computational Chemistry Software

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

4 Executive summary

One of main goals of the CompBioMed Centre of Excellence (CoE) is to strengthen and diversify the presence of biomedical applications in the High Performance Computing (HPC) landscape. In order to facilitate the development of specific services to support computational biomedicine, we need first to understand and characterize the immediate, medium and long term requirements of the main applications used by this community. The aim of this document, D5.2 "Report on computing and data needs of the biomedical community", is to investigate the characteristics of the main software used within the three CompBioMed research domains: cardiovascular/respiratory, molecularly-based and neuro-musculoskeletal medicine. For each application we identified the computational requirements, data access and storage models and infrastructure usage, for users both from academia and industry/SME. The results of this study are based on Fast Track developments taking place in Work Package 2, Work Package 6 as well as Work Package 5 for the identification and the deployment of HPC-ready CompBioMed applications (WP2), workflows (WP6) together with compute and data management solutions (WP1, WP5, WP6).

In this document we will not analyze the characteristics and functionalities of the software identified, nor the data management strategies and solutions adopted by the project partners, but rather we will focus our attention on the computational requirements of the codes and how the different tools are interconnected, to gain broad insight into what is required to boost applications in biomedicine to higher levels of research impact and computational performance.

The principal objectives of this deliverable are to:

- Identify the main Fast Track software components adopted within the computational biomedicine community.
- Describe the computational requirements (e.g. compilers, libraries, data formats) of the codes and of the different tools needed to run the simulation.
- Characterize the simulation infrastructure usage and the data storage and access models used within computational biomedicine.

This report contains the results of work conducted during months M1-M6 for the task T5.1:

Task 5.1: Characterizing Biomedicine Community Application Requirements (M1-M6)
Leader: SARA (6 PM);
Partners: UCL (1 PM), BSC (1), UOXF (1), USFD (1), UPF (1), UNIGE (1)
This task will identify and investigate the compute, data service and connectivity requirements for applications in the Biomedicine community. This includes requirements both from academia and from industry/SMEs, the latter of which may be more strict, e.g. with respect to confidentiality of methods and/or data. We will seek input from all partners to gain broad insight in what is required to boost applications in biomedicine to the higher levels of research impact and performance. In addition, we will conduct a review of widely used applications within the biomedical community at large. As part of this task, we will produce a set of requirements for academic usage but also for industrial/SME usage.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

This report will be used as a reference for the work conducted in task T5.2 “Access Mechanisms for Computing and Data resources” and task T5.3 “Efficient Infrastructure Usage by Community Codes”, and as input for the deliverables D5.3 “Report on access mechanisms to HPC systems” and D5.4 “Report on best practices for e-infrastructure application usage.

*“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement No **675451**“*

5 Introduction

This document outlines the application and data requirements currently determined by the CompBioMed community and its partners. It is focused on the Fast Track applications and their current usage patterns. We anticipate that this list will grow and as such we will retain the requirements in a document that is continuously updated throughout the lifetime of the Centre of Excellence, to reflect the evolving needs of those working inside the CompBioMed landscape. The data collected will also be made available through the CompBioMed Hub (www.compbiomed.eu/software-hub) which serves as a repository for all the information related to CompBioMed software. In particular, as part of the consortium's work is to develop Deep Track applications, we anticipate software to be added or developed during the Deep Track phase and therefore added to the application pipeline.

Analysing the software that is primarily employed within the three research areas of CompBioMed: cardiovascular, molecularly-based and neuro-musculoskeletal medicine, it is possible to identify a common structure in terms of the components and tools adopted to carry out the simulations. These three areas represent the most popular and relevant from a clinical and industrial translation perspective; however, other emerging areas such as computational oncology, computational immunology, etc. do rely on fundamental numerical techniques that are very similar to those used by the aforementioned application domains. Thus, by targeting those we can expect to produce a positive impact also on the broader generality of computational medicine.

The application pipeline, which goes from the first step of the creation of input to analysis of the output results, can be described as an interconnected layered structure:

- Layer 1: Data processing: signal processing, image processing, etc.
- Layer 2: Data digestion: big data analytics, Kalman filtering, etc.
- Layer 3: Model pre-processing: model generation, meshing, etc.
- Layer 4: Simulation ("solvers"): model solution
- Layer 5: Model post-Processing: scientific visualisation, results data analytics, reporting, Kalman filtering, etc.
- Layer 6: Workflows: orchestration of multiple models to capture a multi-step, multiscale, or multi-physics structure of the application

Following this approach, which results from the combined activities of Work Packages 2, 5 and 6, we can better identify the specific technical requirements for each software and tool adopted in the different CompBioMed applications (cardiovascular, respiratory, neuro-musculoskeletal and molecular). This allows for interchanging and reuse of components across applications. The aforementioned six layers can be considered as a general classification which can be identified in all domains; however, it is possible that some of the layers might not be manifested in all of the CompBioMed research exemplars.

In Section 6 we describe the components of the layers identified above and present an analysis of the characteristics and requirements of the main software applications adopted in the Fast Track of the project.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

Section 7 presents information on the computational schemes used by the different simulations that have been identified, and the requirements of each application, in terms of HPC infrastructure usage and data access.

The information and the data presented in this document have been collected during the first 9 months of the project using the form presented in Annex 1 and by direct interaction with three main groups of users:

- CompBioMed core partners
- CompBioMed associate partners
- Users in biomedical fields outside CompBioMed.

These groups include partners from both academic and industrial/medical fields. In addition, the data collected for the preparation of the D1.3 Data Management Plan have been taken into consideration for the analysis provided in this document.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

6 Application requirements

The analysis of the components of the CompBioMed applications pipeline enables the identification and categorisation of the computational requirements for the main applications and tools used by the community. In this section we describe the following main components:

- Data formats;
- Model simulation codes;
- Model Pre- and Post-processing;
- Workflow management.

6.1 Data formats

Data processing and data digestion are the first two layers in the CompBioMed software pipeline as described in Section 5, above. These two steps involve the treatment and management of the data used and produced by the simulation codes. This section describes the main data formats adopted within the biomedical community and how they correspond to the specific software used within CompBioMed. This is then considered in more detail in Subsection 6.2, Model simulation codes, and Subsection 6.3, Pre- and Post-processing tools.

It is important to note that here we are only analysing the data types adopted, while data management requirements such as data access, preservation and long-term storage have been addressed in the Data Management Plan (Deliverable 1.3) with relevant discussion of specific details contained in Section 7 of the present document.

CompBioMed works with a wide range of data types for a variety of research applications. Each data type may require different software or tools installed on the target computational platform, to enable accessing and editing.

Table 1 provides a description of the data types and formats of the data being generated from, or processed by, CompBioMed applications. This includes the main adopted data types within the project partners, but in addition we have identified many other formats that could be used in the future. Data types are categorized based on their field of use: visual (image or video), scientific, and other (i.e. documents, not formatted data, etc.). The table also provides a short description of each data type and the main CompBioMed software that it is associated with.

Table 1. CompBioMed adopted data types.

Visualization Data		
Type	**Description**	**CBM Associated software**
Video file formats	Proprietary or open file formats for the storage of digital videos (i.e. mov, mp4).	VMD, Paraview, VTK, Video editing software packages.
Image file formats	Proprietary or open file formats for the storage of digital images. Different file format may have different properties: lossy compression (JPG), lossless compression (PNG), vector graphics (EPS, SVG, PDF).	VMD, Paraview, VTK, Image editor software packages.
Scientific Data		
Type	**Description**	**CBM Associated software**
PDB	The Protein Data Bank file format is a textual file format	Commonly used within all

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

	for the storage of experimentally determined three-dimensional structures of biological macromolecules including atomic coordinates, crystallographic structure factors and NMR data. **Website:** www.wwpdb.org	molecular dynamics codes.
PSF	The Protein Structure File, is a file format for the description of molecular force fields used in molecular dynamics calculations.	NAMD, CHARMM
PRMTOP	Amber specific file format Parameter/Topology file specification, used extensively by the AMBER software suite for biomolecular simulation and analysis. **Website:** ambermd.org/formats.html	AMBER
VTK	ASCII and binary files designed to work with the Visualization Toolkit (VTK) offer a consistent data representation scheme for a variety of dataset types, and to provide a simple method to communicate data between software. **Website:** www.vtk.org	Visualization Toolkit (VTK)
XTC	XTC is a compressed trajectory format from GROMACS. These files require significantly less disk space, but results in a loss of precision. **Website:** manual.gromacs.org/online/xtc.html	GROMACS, VMD,
DICOM	Digital Imaging and Communications in Medicine, is a standard for handling, storing, printing, and transmitting information in medical imaging. DICOM is both a communications protocol and a file format, which means it can store medical information, such as ultrasound and MRI images, along with a patient's information, all in one file.	Alya, BONEMAT, ShIRT
C3D	The coordinate 3D is a file format originally developed for the storage of the result of photogrammetry software. In CompBioMed is used in relation to gait analysis.	BuilderM2O
Other Data		
Type	**Description**	**CBM Associated software**
Binary	Binary file is a computer-readable, but not human-readable non formatted data type.	Multipurpose (e.g. programs executables, numerical data)
Documents data files	Proprietary or open file formats to store processed results (i.e. xlst, csv) or formatted textual information (e.g. pdf, tex).	Microsoft Office Suite, Latex interpreters, PDF readers.
HDF5	Hierarchical Data Format (HDF) is a set of file formats (HDF4, HDF5) designed to store and organize large amounts of data. **Website:** www.hdfgroup.org	HDF5 libraries and tools.

6.2 Model simulation codes

The simulation of computational models (i.e. model simulation) requires the execution of numerical solvers to predict behaviour that is complementary to quantification and qualitative

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

observation. These models are often complex, and are the result of many years or even decades of development from theoretical, algorithmic and software development points of view. Therefore, this step in the CompBioMed pipeline does not consider the development or improvement of such models and application code, but instead the effective use and the required resources to execute the models. Table 2 lists the main applications that have been identified as key software stacks used by the CompBioMed community. The table summarises the characteristics of the code, the type of licence adopted, the application domain and a link to the developer's website and/or a contact list in the case of an application developed by a CompBioMed partner which is not otherwise publicly available.
The software listed below may be available in different versions, sometimes with specific features restricted to a particular distribution. In this document we do not take into account this aspect, since many of the requirements remain the same between different distributions, but it is important to note that the way the software is used and how it is interfaced with external tools may be limited, in some cases, to a specific version of the code.

Table 2. CompBioMed model simulation codes.

Name	Website CompBioMed Contacts	Licence	Research domain	Description
Alya	**Website:** www.bsc.es/es/computer-applications/alya-system **Contacts:** Mariano Vazques mariano.vazquez@bsc.es Gauillaume Hozeaux guillaume.houzeaux@bsc.es	Limited	Cardiovascular Respiratory	The Alya System is the BSC simulation code for multi-physics problems, specifically designed to run efficiently using supercomputers. The code is particularly well-suited for the simulation of complex problems in different domains of Science and Technology. The purpose of the code in the CompBioMed project is to solve electromechanical simulations of the heart.
CHASTE	**Website:** www.cs.ox.ac.uk/chaste/ **Contacts:** chaste-users@maillist.ox.ac.uk	GNU LGPL (versions 3.1 and later are being made available under the 3-clause BSD licence)	Cardiovascular	CHASTE (Cancer, Heart and Soft Tissue Environment) is a general purpose simulation package aimed at multi-scale, computationally demanding problems arising in biology and physiology.
HemeLB	**Contacts:** Peter Coveney p.v.coveney@ucl.ac.uk Robin Richardson robin.richardson@ucl.ac.uk	GNU LGPL	Cardiovascular	Software for rapid, patient specific simulations of blood flow in intracranial cerebral aneurysms to assess the short and long range effects of their treatment via the introduction of a stent or flow diverter.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

				The simulations are executed over a full range of physiological states and provide clinicians with information on the long term consequences of a choice of treatment.
Hemocell	**Contacts:** Alfons Hoekstra a.g.hoekstra@uva.nl Gabor Zavodszky g.zavodszky@uva.nl	Limited	Cardiovascular	An extensible HPC framework to simulate the flows of dense cellular suspensions. Extension modules include different cell material models, advection-diffusion of additional biochemical agents, platelet binding formation mechanisms.
Palabos	**Website:** www.palabos.org **Contacts:** Jonas Latt jonas.latt@unige.ch	Open-source licence AGPLbv3.0	Cardiovascular	Palabos is a general library for fluid flow simulation using the Lattice Boltzmann method. In CompBioMed, Palabos is used to simulate blood flow in arteries.
openBF	**Website:** alemelis.github.io/openbf.jl **Contacts:** Alberto Marzo a.marzo@sheffield.ac.uk Alessandro Melis a.melis@sheffield.ac.uk	Limited	Cardiovascular	A library written in Julia, and developed at The University of Sheffield. OpenBF simulates the flow within networks of elastic vessels. The partial differential equation (PDE) solver is based on the finite-volume method. It is first-order and second-order accurate in time and space, respectively.
ShIRT	**Contacts:** Alberto Marzo a.marzo@sheffield.ac.uk Alessandro Melis a.melis@sheffield.ac.uk	Limited	Neuro-musculoskeletal	The Sheffield image registration toolkit (ShIRT) is an image processing software developed in C language and uses MATLAB for images pre-processing. ShIRT employs an elastic registration algorithm to map an image (the moved image) onto a reference image (reference image). This mapping represents the displacement of each

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

				pixel (or voxel, for 3D images) from the reference image to the moved image. ShIRT is part of the BoneDVC workflow, where the computed displacement field is further differentiated to calculate the strain field within bone tissue specimens.
FLIBRA	**Contacts:** Julien Favier Julien.favier@univ-amu.fr	Limited	Respiratory	Simulation of mucociliary transport in human lungs.
OpenFOAM	**Website:** openfoam.org	GNU GPL	Cardiovascular	OpenFOAM (Open source Field Operation And Manipulation) is a C++ toolbox for the solution of continuum mechanics problems, including computational fluid dynamics (CFD).
NAMD	**Website:** www.ks.uiuc.edu	Free for non-commercial use.	Molecular	Software for molecular dynamics simulations of biological molecules.
GROMACS	**Website:** www.gromacs.org	GNU Lesser General Public License	Molecular	Software for molecular dynamics simulations of biological molecules.
AmberTools	**Website:** ambermd.org	GNU GPL	Molecular	Suite of molecular dynamics simulation tools for molecular model building and analysis. Primarily for biological molecules.
CHARMM	**Website:** www.charmm.org	Free for academic, government and non-profit use only.	Molecular	A molecular simulation program with a broad range of application to many-particle systems.
GAMESS-US	**Website:** www.msg.ameslab.gov	Limited	Molecular	The General Atomic and Molecular Electronic Structure System (GAMESS) is a general *ab initio* quantum chemistry package for the calculation of the electronic structure of atoms and molecules.
Gaussian	**Website:** gaussian.com	Commercial	Molecular	Widely used commercial code for electronic structure calculations. In the CompBioMed project

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

				It is used for the parameterization of small molecules, such as drugs and inhibitors, for MD simulations, and for the analysis of drug-protein binding interactions.
AMBER	**Website:** ambermd.org	Commercial	Molecular	Software for molecular dynamics simulations of biological molecules.
MATLAB	**Website:** mathworks.com	Commercial	Neuro-musculoskeletal	Numerical computing environment optimized for engineering and scientific problems.
Ansys software's suite	**Website:** www.ansys.com	Commercial	Cardiovascular, Neuro-musculoskeletal	General purpose computational tools for engineering simulations.
ACEMD	**Website:** www.acellera.com	Commercial	Molecular	ACEMD is a production molecular dynamics software specially optimized to run on NVIDIA graphics processing units (GPUs).

6.2.1 Software prerequisites

The computational execution of simulations is commonly the most computationally expensive part of the application pipeline in any project. The simulation step also often requires the usage of specialized computational libraries and compilers in order to achieve acceptable/good performance. Table 3 lists prerequisites for the software packages identified in Table 2 (above). This includes the requisite compilers, mathematical/scientific libraries and computational tools needed for their compilation and execution.

Table 3. Model simulation software prerequisites.

Computational languages		
Name	**Description**	**CBM associated software**
C, C++, Fortran	**GNU compilers** Set of compilers created under the GNU project and distributed under GNU GPL license. **Intel compilers** Commercial set of compilers highly optimized for intel architectures. **Other** IBM XL, PGI, Platform specific compilers (e.g. Cray)	Alya, CHASTE, HemeLB, Hemocell, GROMACS, NAMD, Amber, Ambertools, GAMESSUS, OpenFOAM, CHARMM, ShIRT
Python	Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.	CHASTE, Palabos (for compilation), ACEMD,

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

	Website: www.python.org	openBF , HemeLB
Java	Java is a general-purpose, concurrent, class based, object-oriented programming language. **Website:** www.java.com	
Julia	Julia is a high-level, high-performance dynamic programming language for numerical computing. **Website:** julialang.org	openBF
Tcl	Tcl (Tool Command Language) is a high-level, interpreted, dynamic programming language, suitable for a very wide range of uses, including web and desktop applications, networking, administration, testing and many more. **Website:** www.tcl.tk	NAMD
CUDA	CUDA is a parallel computing platform and programming model, developed by NVIDIA, enabling the exploitation of the power of graphics process units (GPU). **Website:** www.nvidia.com	GROMACS, NAMD, AmberTools, CHARMM, ACEMD
Computational libraries and tools		
Name	**Description**	**CBM associated software**
MPI	The Message Passing Interface Standard (MPI) is the de-facto standard for the implementation of message passing parallel computing. MPI is available in a variety of libraries, including MPICH, OpenMPI, MVAPICH as well as bespoke libraries for specific hardware such as the Cray-MPI library. The standardisation of MPI provides a portable, efficient, and flexible standard for the message passing parallel programming model. **Libraries:** OpenMPI, MPICH, MVAPICH. **Website:** mpi-forum.org	Alya, LUNGS, Palabos, HemeLB, Hemocell, GROMACS, NAMD, Amber, AmberTools, GAMESSUS, OpenFOAM, CHARMM, CHASTE
OpenMP	OpenMP is an API designed to allow the use of shared-memory processing using threading. Using compiler directives and library routines OpenMP provides a portable, scalable model for parallel shared0memory programming and is frequently used in hybrid programming to provide shared memory support. **Supports:** C, C++ and Fortran codes. **Website:** http://www.openmp.org/	Alya, CHASTE, Hemocell
Boost	C++ library providing support for a variety of application functions including linear algebra, image processing and random number generation.	CHASTE, OpenFOAM
Make, CMake, Autotools	Set of open-source tools designed to control and manage the building and testing process of executables and non-source files of a program from the program's source file. **Website:** www.gnu.org/software/autoconf www.gnu.org/software/make	Alya, Hemocell, GROMACS, OpenFOAM

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

	cmake.org	
SCons	Open source software construction tool based on python. **Website:** scons.org	CHASTE
HDF5	Suite of tools, libraries and data format optimized for the management of extremely large and complex data collections. **Website:** www.hdfgroup.org	HemeLB, CHASTE, Hemocell
Git, svn	Distributed control version systems for the development of computational projects. **Website:** git-scm.com subversion.apache.org	CHASTE, HemeLB, Alya.
Mathematical and scientific libraries		
Name	**Description**	**Applications**
Blas, Lapack, MKL, Atlas	Specialized mathematical libraries optimized for performing linear algebra operations.	GROMACS, NAMD, LUNGS
Metis, Parmetis	Serial (METIS) and MPI-based parallel (parMETIS) libraries that implement a variety of algorithms for partitioning unstructured graphs, meshes, and for computing fill-reducing orderings of sparse matrices.	Alya, CHASTE, Hemocell
PETSc	The Portable, Extensible Toolkit for Scientific Computation (PETSc), is a suite of data structures and routines for the scalable (parallel) solution of scientific applications modelled by partial differential equations. **Website:** www.mcs.anl.gov/petsc	CHASTE
FFTW	Free C library for the computation of the discrete Fourier transform (DFT) for both serial and parallel architectures.	GROMACS, NAMD, CHARMM
CVODE	C library for the solution of stiff and nonstiff ordinary differential equation (ODE) systems	CHASTE

6.2.2 Software licensing

The availability of software is often governed by licencing and is therefore of concern for the CompBioMed community. Licences provide a legal framework for the distribution, adaptation and copyrighting of software. These also typically contain information on liability and the responsibility of user and developer. This ensures that developers and owners receive appropriate recognition, but also crucially controls how software can be used by third parties.

Within CompBioMed we have identified two key types of licence:
- Free and Open Source: This provides licence rights for use and development of the software by the customer (i.e. the end user), facilitating both contributions to the software community from beyond the initial development team, and also free for use.
- Limited/Proprietary: The software publisher and/or owner grant the user the right to use but not to develop or re-distribute the software.

Crucially with proprietary licences, these may also have fees attached (in commercial licences), and additional restrictions on how this is made available on shared platforms, such as the cloud or supercomputers, or for specific type of usage (academic vs commercial use). For

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

example, within an HPC centre, software is often centrally provided but, due to the licencing restrictions, access is limited to those who can show they have a valid personal licence or who use the software for non-commercial purposes. Other platforms may have bought a licence that provides usage access to anyone with access to the platform, but this may often come at great expense to the service providers, so is often limited to very specialist providers, such as the UK's National Service for Computational Chemistry Software (NSCCS) which provides a dedicated computational chemistry platform and therefore most of their users have very similar requirements, limiting the number and therefore cost of individual products that need to be purchased under licence. Commercial licence charging models, can be a barrier for both academic and SME users. When the developers have not anticipated a new user group and do not provide clear procedures, or the cost of the license depends on the number cores used by the code, that must be taken into consideration during planning and it can become prohibitively expensive in the case of high-end HPC simulations.

6.3 Model Pre- and Post-processing tools

The pre- and post-processing of simulation data are crucial steps in computational modelling, often forgotten with respect to the simulation itself. In the pre-processing stage, the input for the computationally expensive simulation is generated from experimental data or a previous simulation technique. During post-processing, results are analysed and handled in a variety of ways, to extract and visualize valuable information from the simulation outputs.

The characteristics and functionalities of pre- and post-processing packages and tools can be very diverse (e.g. mesh generators, visualization tools, molecular structures generators, etc.) and they are often distributed and designed to work with a specific software package or data format. A comprehensive list of the main software adopted within the community is presented in Table 4. The table contains a general description of the software capabilities and requirements, the context in which the software is used and, where possible, the associated data types with which they work (see Table 1 for more information about data types employed in CompBioMed).

Table 4. Pre- and Post- processing tools adopted.

Name	Type	Description	Data types
fpocket	Post-processing	Fpocket is an open source code for structure based virtual screening code for protein pocket detection, pocket descriptor extraction, or drug-ability prediction. **Prerequisites**: visualization software (VMD or PyMol) Website: fpocket.sourceforge.net	PDB, ASCII
LIGSIFT	Post-processing	LIGSIDT is an open-source tool for ligand structural alignment and virtual screening. **Website:** http://cssb.biology.gatech.edu/LIGSIFT	PDB, ASCII
BONEMAT	Pre-processing	Bonemat is a freeware software package that maps a Finite Element mesh onto a bone to measure elastic properties derived from Computed Tomography images. **Prerequisites:** VTK, Microsoft Windows **Website:** www.bonemat.org	DICOM

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

BuilderM2O	Pre-processing	Builder Mark II – Organ (BuilderM2O) allows the importing of imaging, gait analysis, EMG, segmentation, and finite element data, and fuse them toward the creation of organ scale subject-specific models of bones, joints, muscles, ligaments, and cartilages. **Prerequisites:** Microsoft Windows 32 bits **Website:** www.builderm2o.org	C3D, DICOM
ITK-Snap	Pre-processing	Open-source software for the segmentation of 3D medical images. **Website:** www.itksnap.org	DICOM
VMTK	Post-processing	Collection of libraries and tools for 3D reconstruction, geometric analysis, mesh generation and surface data analysis for image-based modeling of blood vessels. **Prerequisites:** Python, CMake, Git, C++ compiler **Website:** www.vmtk.org	VTK
GiD	Pre-processing Post-processing	GiD is a universal, adaptive and user-friendly pre- and post-processor for numerical simulations in science and engineering. It has been designed to cover all the common needs in the numerical simulations field from mesh generation to output analysis and visualization. **Website:** www.gidhome.com	
IRIS Mesh	Pre-processing	Mesh generator developed and provided to CompBioMed by BSC.	
ParaView	Post-processing	Open-source, multi-platform data analysis and scientific visualization application. **Prerequisites:** CMake,Qt libraries, Python. **Website:** www.paraview.org	
Visit	Post-processing	Open Source, interactive, scalable, visualization, animation and analysis tool. **Website:** wci.llnl.gov/simulation/computer-codes/visit/	

Scripts and tools that have been developed 'in-house' by CompBioMed partners are routinely used for the preparation of inputs and the analysis of output files especially in the case of workflows interconnecting different software. These solutions, along with others from across the international community are widely adopted (especially among computationally expert users), but their requirements will not be analysed here since these tools are usually specific to a problem and not easily portable.

6.4 Workflows management tools

Computational modelling of complex systems is pursued in many aspects of biomedical research, medicine and technology. These modelling approaches often require, and benefit, from the automated processing of heterogeneous data and the execution of complex networks of tasks, or a workflow, including the offloading of the data processing to remote resources

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

and enabling the execution of even larger and more complex workflows. Many workflow management systems are available in the scientific domain. Some of these will be adopted by the CompBioMed Centre of Excellence to execute applications of complex human physiology.

Requirements for workflow management can be divided into two broad classes:

i. Automation of a number of tasks in serial or the operation of several independent tasks in parallel;
ii. Highly interactive orchestration of several tightly-coupled computational tasks.

These two classes of requirement will be exemplified within CompBioMed using two distinct general purpose workflow management tools: Taverna and MUSCLE2. Taverna provides functionality to enable management of complex workflows within the first class and MUSCLE2 provides a framework to integrate the orchestration of complex tightly-coupled workflows.

It is also possible to identify application specific workflow management tools. For example, the Binding Affinity Calculator (BAC), developed by UCL, and HTMD, developed by Acellera, are software interfaces used in molecular medicine and drug screening simulations to help in model construction and optimization of multistep procedures.

A summary of how these workflow tools map onto the application pipeline delivered within CompBioMed is provided in Table 5, below. Experience of deployment of these workflow management tools on the HPC platforms leveraged by CompBioMed will be documented to provide case studies for workflow management in this environment. Beyond the applications in Table 5, this expertise will also be used to inform training delivered to researchers to enable them to extend this approach to additional biomedical simulation tasks and to address the needs of the broader community in this area.

Table 5. CompBioMed workflow management tools.

Name	Description	Prerequisites	CBM associated service
Taverna	Taverna is an open source workflow management system, written in Java and available both as a desktop client interface and command-line tool. **Website:** taverna.incubator.apache.org	Java	boneDVC (USFD) CT2STRENGTH (USFD)
Muscle2	MUSCLE 2 - The Multiscale Coupling Library and Environment is a portable framework to do multiscale modelling and simulation on distributed computing resources. Distributed under LGPL version 3 license. **Website:** apps.man.poznan.pl/trac/muscle	Java, Ruby, CMake, C compilers	3D In-stent restenosis (UvA)
BAC	Specialized workflow management tools for the automated calculation of drug-protein binding affinities, in some instantiations based on the RADICAL CyberTools workload execution tool. **CompBioMed Contacts:** David Wright dave.wright@ucl.ac.uk Shunzhou Wan shunzhou.wan@ucl.ac.uk	RADICAL CyberTools, Python	Protein−ligand binding affinities (UCL)

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

HTMD	HTMD is a molecular-specific programmable environment to prepare, handle, simulate, visualize and analyze molecular systems. HTMD is based on Python, and in a single script, it is possible to plan an entire computational experiment, from manipulating PDBs, building, executing and analyzing simulations, computing Markov state models, kinetic rates, affinities and pathways. **Website:** www.htmd.org	Python	Molecular dynamics calculations (UPF)

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

7 Infrastructures and data access requirements

One of the main goals of this CoE is to raise awareness of the innovation possibilities created by high performance computational biomedicine and provide researchers and SMEs with the expertise necessary to run their simulations on the infrastructure that is most appropriate.

The access to computational resources for biomedical applications is often hindered by lack of knowledge of the potential of HPC and advanced simulation. This section provides information on the requirements of simulation tools, in terms of infrastructure usage and data access, for the three different research exemplars in CompBioMed. We have identified the primary methods of biomedical simulation as performed by the community, the computational resources required, and data management tools and storage services for saving simulated and non-simulated data.

7.1 Infrastructure usage requirements

Here we present an analysis of the main computational strategies adopted by the CompBioMed core and associate partners to run their simulations. Table 6 summarizes the characteristics in terms of the total number of processors, types of parallelism adopted and memory usage of the main code identified in CompBioMed's three principal research fields of cardiovascular/respiratory, molecularly-based and neuro-musculoskeletal medicine. In order to facilitate the usage and the diffusion of the software listed here, for each case described we propose a pool of optimal target systems, based on the expertise of the computing centre in the deployment and optimization of scientific software and on the current usage made by the CompBioMed partners. These resources can be grouped into three main categories: supercomputers, accelerators and cloud computing.

Supercomputers

The term 'supercomputer' is often applied to a broad range of architectures from the most powerful computers in the world (e.g. the Pan-European HPC supercomputers - Tier 0, National HPC supercomputers- Tier 1) to computer clusters (Regional HPC facilities - Tier 2, Local infrastructures - Tier 3) resources. Supercomputers provide dedicated CPU time. Supercomputer nodes are generally similar or identical architectures, enabling the tuning and optimisation of software for the architecture provided. In addition, supercomputers are a dedicated resource with minimal software overheads and minimised user access to ensure optimal use of the computational resource for software execution. Supercomputers are usually massively parallel with distributed architectures and memory. They require specialist programming methods, such as use of message passing (e.g. via the Message Passing Interface - MPI)) or Partitioned Global Address Space (PGAS) to make effective use of the distributed memory. In addition, many of the top 20 fastest machines in the world (www.top500.org), and increasing numbers of smaller machines, include 'accelerators'.

Accelerators

Accelerators come in a variety of types, from GPGPUs (general purpose graphics processing units) to the Intel Xeon Phi. Accelerators can be used as a standalone item, with increasing numbers of personal computing facilities including programmable GPUs, or attached in small or large numbers to the massively parallel architectures of supercomputers. For specific computational problems, accelerators enable better performances than standard CPU at a lower energy usage (e.g. floating point operations/Watt). To achieve a high level of performance requires a specialised parallelism to expose the required level of vectorisation in

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

the software. With some codes it may be impossible to expose the required level of vectorisation to benefit from using accelerators because the algorithm does not have a high enough level of operational intensity (i.e. enough floating point operations for each memory access). In addition to exposing parallelism, some accelerators (e.g. GPUs), require the use of additional programming techniques in the software (e.g. a new language such as CUDA, or a new API such as OpenCL).

Cloud computing

Cloud computing enables on-demand access to resources, without the overhead of their ownership. Cloud computing presents the required computational resources to the user by the use of one or more virtual machines. However, this can have a negative impact on performance, depending on the particular problem under consideration. Although virtualisation is now a minimal overhead, due to the use of multiple virtual machines, communication latency in cloud computing is higher than in a dedicated HPC resource, resulting in reduced performance. The use of virtual machines means that the software is not running natively on the hardware and therefore communication is not tightly coupled. However, the performance gap is closing, as cloud service providers adopt hardware capable of the performance of HPC, such as Azure where the interconnect now uses Infiniband, a networking standard commonly used in HPC system. Therefore, if an application has tightly coupled processes, requiring significant communication, cloud computing is likely to perform less well than dedicated supercomputing facilities.

Table 6. CompBioMed applications infrastructure usage.

Research exemplar	Typical mode of operation in CompBioMed	Memory requirements	Simulation codes used	Target System
Cardiovascular/ Respiratory	**Highly parallel single run** #Cores: 1K – 1M cores Type of parallelism: - MPI - Hybrid MPI/OpenMP	10 - 200 GB total memory Memory depends on sparsity and size of the problem.	Alya CHASTE FLIBRA HemeLB Hemocell Palabos Ansys	- Tier 0
	Intermediate parallel concurrent runs #Cores: 10 – 100 cores per run Type of parallelism: - OpenMP - MPI - Hybrid MPI/OpenMP	1 - 100 GB per run Total memory depends on the number of concurrent jobs and size of the problem.	Chaste CHASTE Palabos Ansys	- Tier 1 - Tier 2 - Cloud
Molecurlaly-based medicine	**Highly parallel single run** #Cores: 100 – 1K cores 10-100 GPUs Type of parallelism: - MPI - Hybrid MPI/OpenMP	100 MB – 50 GB total memory. Memory depends on number of atoms in the system and level of theory adopted.	GROMACS NAMD AMBER Desmond	- Tier 0 - Tier 1 - GPUs

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

	- CUDA			
	Intermediate parallel concurrent runs #Cores: 10 – 100 cores per run, 1-10 GPUs Type of parallelism: - OpenMP - MPI - Hybrid MPI/OpenMP - CUDA	Typical usage: 100 MB – 1 GB per run Total memory depends on number of concurrent jobs and size of the simulated system.	GROMACS NAMD AmberTools ACEMD GAMESS-US Gaussian	- Tier 1 - Tier 2 - GPUs - Cloud
Neuro-muscoloskeletal medicine	**Serial concurrent runs** #Cores: 1 core per run Type of parallelism: - Workflow managed	Typical usage: 1GB – 100 GB total memory Memory depends on the size of input images.	MATLAB Ansys	- Tier 2 - Cloud
	Weakly parallel concurrent runs Cores per run: 2 – 10 cores per run Type of parallelism: - Pure OpenMP - Pure MPI - Hybrid MPI / OpenMP	1 MB – 1 GB total memory Memory depends on the size of input images.	MATLAB Ansys	- Tier 1, Tier 2 - Cloud

The CompBioMed Centre of Excellence provides a range of computational resources to the consortium members. These include a range of European Tier 0 (supra-national European HPC resource, e.g. Piz Daint), Tier 1 (national HPC resource, e.g. ARCHER and Cartesius) and Tier 2 (regional HPC resource, e.g. Cirrus, Lisa) supercomputers and Cloud-based computing access for both academic (SURFsara's HPC Cloud) and industrial users (Amazon Web Services AWS, Microsoft's HPC Azure and DNA Nexus). Access to the cloud infrastructures is available for all the partners in the consortium and through CompBioMed it will be possible to get special rates or easier access paths for these services. The facilities available provide a range of hardware and software capabilities, covering the full range of requirements as determined by the current software usage as described in Section 6. The full range of HPC resources available to CompBioMed is presented below in Table 7.

Table 7. CompBioMed available HPC resources.

Name	Owner	Annual allocation	Specifications	Notes
Cirrus	UEDIN	400,000 core hours	SGI ICE XA cluster equipped with 10080 Intel Xeon E5-2695 processors. Infiniband interconnections and 256 GB of memory per node.	
ARCHER	UEDIN	1,900,000	Cray XC30, 118080 cores, Aries interconnect,	

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

		core hours	2.67Gb per core	
Bull CEPP, Extreme Factory and other resources	Bull/ATOS	847,720 core hours	Initially bullX B520 supercomputers with 14 core Ivybridge Xeon processors, FDR Infiniband and varying memory/node. These systems are regularly refreshed and new technology will become available during the course of the project. Resources will be managed by Bull Centre for Excellence in Parallel Computing (CEPP). This has a monetary value of €68k/yr.	
Cartesius	SURFsara	1,000,000 core hours	40,960 cores + 132 GPUs: 1.559 pflop/s (peak performance). 117 TB memory.	
LISA	SURFsara	700,000 core hours	8960 cores, 30TB RAM, Total peak performance 158 tflop/sec.	
Lemanicus Blue Gene/Q	EPFL	15,000,000 core hours	209 tflops peak performance. 16 TB of RAM, 2.1 PB of disk space.	
Scylla	UNIGE	300,000 core hours	300 Intel Sandy Bridge cores. Infiniband Interconnect. Total RAM 800 GB	
Baobab	UNIGE	300,000 core hours	Cluster with 912 Intel Sandy Bridge cores. Infiniband interconnect. Total RAM 3300 GB. Storage space 40 TB	
Iceberg	USFD	300,000 core hours	Cluster with 3440 cores, 16 GPU units, 31.8 TB RAM, 300 TB filestore.	Replaced by ShARC in Q1/2017
Legion	UCL	600,000 core hours	Cluster with 7864 cores, 342TB of RAID 6 storage, Infiniband interconnect	
Computational Chemistry Cluster	Evotec	1,858,351 core hours	424 core Dell cluster.	
GPUGRID	UPF	5,000,000 GPU core hours	Volunteer distributed computing platform for molecular dynamics simulation. Equivalent in capacity to 1000 K40 GPUs.	
Marenostrum	BSC	150,000 core hours	Peak performance 1.1 petaflops. 48,896 Intel Sandy Bridge processors in 3,056 nodes, and 84 Xeon Phi 5110P in 42 nodes. Over 104.6 TB of main memory and 2 PB of GPFS disk storage.	
GPU Cluster	Acellera	350,000 GPU core hours	40 GTX780 GPU total, equivalent to 40 Tesla K40	
ShARC	USFD	300,000 core hours	Dell Precision Rack 7910 with 128 node each including 2 CPUs Intel Xeon E5-2630 v3 (2.40GHz) with 128 GiB of RAM, for a total of 2016 core and 8,800 GiB	Replaces Iceberg from Q1/2017

7.2 Data access requirements

The CompBioMed consortium will hold and generate a large amount of heterogeneous data, from the research data held by the consortium partners, to data generated and required as input by modelling and simulation activities. Despite the difference in the types (medical

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

images, proteins structures, etc.) and origins (medical instruments, crystallographic measures) of the data used within the three research exemplars in CompBioMed, the requirements in terms of data access and storage are very similar across the different domains and largely depend on the problem size. The data access requirements for secure storage and data management explained in this Section, hold for data in both academia and industry/SMEs. The details of the data type adopted within the project and the tools used for the pre-processing of the model and the post-processing of the output, are discussed in Section 6. In the following subsections we will focus on the needs of the CompBioMed community in terms of data storage (archive space, long term storage, etc.) and data management requirements (data access and retrieving, handling of metadata and identifiers, etc.) emerged from the survey conducted within core and associate partners of the project. In addition, we will provide a short description of the data storage facilities and the data management tools available within the consortium.

7.2.1 Data storage

Data storage requirements refer to disk space usage and long term storage needs, for both non-simulated (generally used to build the model) and simulated data (generated from computational models).

The size of non-simulated data that is typical of that used in computational biomedicine may vary from a few Megabytes (e.g. molecular structures files and/or molecular dynamics input files), to a few hundred Gigabytes (e.g. medical images processed in cardiovascular and neuro-musculoskeletal modelling) and the dimensions of a single file are directly proportional to the size of the investigated problem. The total disk requirements also depend on the total number of input files used simultaneously in a single experiment. Many users within the CompBioMed community, require the storage of a large number of input files for each simulation, resulting in significant disk space requirements even in the case of small input files. The storage of the simulation output data also depends on the nature of the modelled problem. The format of the data output, the number of the output files and the way the results are post-processed are all factors that can determine the size of the stored data. For the simulations considered in this report, the typical total output volume ranges between a few Gigabytes and tens of Terabytes.
In some circumstances, identifiable medical data needs to be operated on within a Data Safe Haven, a technical environment for receiving, handling and storing sensitive data securely. Deploying such a solution (which is very often institution specific) is beyond the scope of CompBioMed, but we expect to work with existing institutional data safe havens to integrate the tools and the data management policies adopted within the consortium.

One of the main requirements for users within the CompBioMed community is the ability to correctly and safely store the modelling data for a significant period of time. The results of the survey have shown that both research and industrial partners would benefit from a storage system which allows the archiving and preservation of large data sets for several years and the possibility of shared access to these files with collaborators and/or customers.

Data storage services provided by the HPC centres present in CompBioMed are one potential solution for the data storage requirements identified above. These infrastructures are, in fact, designed for the long term storage of large quantities of files (usually orders of magnitude larger than the dimensions identified above) and provide specialized tools for the backup and preservation of the archived data. In addition, HPC centres provide access to specialized tools for remote data visualization and high performance data analysis which are usually not available at the user's site. Moreover, the archive and storage facilities at HPC centres are

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

usually connected via a fast network to the HPC. This facilitates the transfer and staging of large amount of data to the compute infrastructure for computing, processing and post-processing of the data.

The characteristics and the capabilities of the main data storage services available within the CompBioMed community are described in Table 8. We will leverage facilities offered by EUDAT for the long-term presentation of data (www.eudat.eu/). UCL has previously developed a relationship with the EUDAT data nodes RZG and EPCC to provide long term B2SHARE and B2SAFE provision, which we will aim to make use of in this project. In addition, SURFSara is a partner in the EUDAT consortium, leading the work package that develops and maintains the B2SAFE EUDAT service.

Additional information about the policies that will be adopted and the plans for the medium and long term storage of CompBioMed data can be found in the data management plan D1.3 and will be reported in the deliverable "D6.2 Deployment of project informatics platform" that will be released in month 12.

Table 8. CompBioMed available data storage services.

Name	Owner	Storage medium	Specifications	Notes
Data Archive	SURFsara	Tape	Long-term storage, bit-wise preservation, backup.	NFS mounts to Lisa and Cartesius supercomputer
Scratch file system Lisa/Cartesius	SURFsara	Disk	Short-term storage space for simulation run. No backup.	parallel I/O
BeeHub	SURFsara	Disk	Mid-term storage, data sharing, backup.	100 GB for free and more upon request
SURFdrive	SURFsara	Cloud storage	Short-term storage, data sharing with synchronization, backup.	100 GB for free
Data ingest service	SURFsara	-	Ingesting data from external storage media to SURFsara domain	
UK-RDF and DAC	EPCC	Disk, backup tape	Archiving service. Medium and long-term storage. Independent service, but directly mounted on ARCHER login nodes. The DAC provides specialised data pre- and post-processing hardware for data on the RDF.	7.89PB disk, 19.5PB backup tape.
Scratch file system Cirrus	EPCC	Disk	Lustre parallel file system with high read/write bandwidth. Not backup.	406 TiB disk, parallel I/O
HSM Tivoli	BSC	Tape	Backup system.	
Active Archive	BSC	Disk	Mid-long term GPFS storage system.	
Scratch file system Marenostrum	BSC	Disk	Short-term, fast access filesystem for HPC executions, mounted in all BSC supercomputers.	

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

7.2.2 Data management

Data management requirements, refer here to the ensemble of tools and policies needed to access, move and categorize the input and output data files created and used during the simulation pipeline. The requirements, in terms of data access and transfer, mainly concern the simulation data. Whereas the transfer of the input files to remote hosts for the simulation can be easily achieved with regular data transfer tools (i.e. scp, rsync, ftp), the ability to efficiently access, retrieve and share archived simulation's outputs is of great importance for both industrial and academic partners.

The results of the survey conducted, revealed that users from different CompBioMed domains have common needs in terms of accessing and labelling simulation related data. These data, as demonstrated in the previous section, need to be stored for a significant period of time, and is usually required for publishing purposes or needed to be reused for collaborations or further analysis. The ability to attach metadata to the input and output files is also important for biomedical simulation. Many users have expressed the needs of having a system of data annotations and DOIs which helps users to parse and analyse large datasets.

Particular attention should also be paid to the issue of data protection for the cases where sensitive information is handled. Although our CoE is designed to operate exclusively with anonymised data (pseudo-anonymised data would be released only through a trusted third party), protection of confidentiality is a general issue in medical data that needs to be addressed. Support for data security is essential for many CoE partners both for handling of patient data and also for protecting intellectual property (e.g. in drug discovery). To address these issues, CompBioMed works with data and HPC centres that are certified to the international standard for information security, ISO27001, such as SURFSara. In addition, one of the CoE Associate Partners, DNA Nexus, provides encryption and security explicitly designed around ISO27001 standards to provide cloud-based secure and encrypted data services The data security aspects of the project are addressed in more detail in the Data Management Plan D1.3, where a specific section is dedicated to data recovery as well as secure storage and transfer of sensitive data. This topic will also be addressed more comprehensively with the work carried out in Task 5.5 where we plan to formulate and undertake a set of tasks to improve the working of these secure data services with large-scale resources.

As already stated in the Data Management Plan, in CompBioMed we will make use of the services provided by EUDAT (the European Research Data Services, Expertise & Technology Solutions). This collaborative European Project, in which several of the CompBioMed partners have an active role, provides a set of tools and data storage services for an efficient access, reuse and interoperability of scientific data, including tools for the treatment of sensitive data. The main characteristics of the tools provided and their functionalities are presented in Table 9, below.

Table 9. EUDAT services for data management.

Name	Description	Features
B2DROP	A secure and trusted data exchange service for researchers and scientists to keep their research data synchronized and up-to-date and to exchange with other researchers.	
B2SHARE	A user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store, share and preserve small-scale research data from diverse contexts.	- Persistent identifiers assigned to data - Openly accessible and

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

		harvestable metadata
B2SAFE	A robust, safe and highly available service which allows community and departmental repositories to implement data management policies on their research data across multiple administrative domains in a trustworthy manner.	Based on iRODS technology and persistent identifiers (PIDs).
B2STAGE	a reliable, efficient, light-weight and easy-to-use service to transfer research data sets between EUDAT storage resources and high-performance computing (HPC) workspaces.	Transfer large data collections from EUDAT storage facilities to external HPC facilities for processing.
B2ACCESS	An easy-to-use and secure Authentication and Authorization platform.	B2Access can be integrated with any B2service
B2FIND	a simple, user-friendly metadata catalogue of research data collections stored in EUDAT data centres and other repositories.	

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

8 Conclusions

In this report we analysed the main applications and tools used within the CompBioMed community and characterized the requirements they have to run efficiently on High Performance Computing infrastructures. This analysis includes the data formats used, pre- and post processing tools, workflow management tools and model simulation codes routinely adopted by computational biomedicine users coming from academy, industry and medical institutions.

In addition to the computational characteristics of each application we also presented an analysis of the infrastructure usage and the data access approaches followed by users. The results highlight that common strategies are adopted to run simulations and many of the codes and tools used are already deployed or will be deployed on dedicated HPC and cloud infrastructures. This collection of CompBioMed applications is also useful to identify and, where appropriate, provide alternative options where usage is limited due to software performance or restricted commercial licencing. This document contains an exhaustive list of the main codes used in biomedical modelling, with a description of their capabilities and the requirements to run them. As such, it should be valuable for users who are new to the field or would like to exploit high performance computational biomedicine simulation tools.

The outcome of the present analysis will be used as a baseline for the developments foreseen in the Deep Track activities of the project. The information contained here will also inform the characterization of the Fast Track applications and the research carried out in CompBioMed ("D2.1: Report on Application Software Readiness and Fast Track Exploitation (M12)" and "D6.1: Report on existing solutions in support of biomedical applications (M12)") conducted in WP2 and WP6.

The analysis and the information reported here will also serve as a reference for many of activities of WP5 (see Tasks 5.2, 5.3, 5.4) which is focused on the provision of support to biomedical researchers to run their applications on HPC resources and to increase the use of existing compute and data infrastructures.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

9 Annexes

Annex 1. Example of the form distributed for the collection of the application descriptions and usage.

Application info

Name	*Code name and version number if apply (i.e. GROMACS > 5.1.1)*
Domain	*Biomedical research domain (i.e. Cardiovascular… if used within multiple domain, please specify)*
Description	*A free text description explaining the purpose of the codes in the CompBioMed context (some of the codes may be multidisciplinary codes, please here refer to their use inside CompBioMed project).*
License	*Open, limited, commercial*
Contacts	*Name, email of contact person(s) and/or website if code developed externally.*

Computational info

Programming Language	*Programming language in which the code is written. If multiple languages are used, please list all. (i.e. Fortran – main code / C++ / Python).*	
Dependencies	*Dependencies on external libraries / tools (i.e. metis, scons, cmake…)*	
Parallel	*Yes / No*	
	Type of parallelism	*Type of parallelism. (i.e. MPI, Hybrid MPI+OpenMP, GPUs)*
	Scalability	*Typical run: #cores / #nodes / #gpus* *Large run: #cores / #nodes / #gpus*
Memory requirements	*Memory requirement for single execution (approximately…)*	
Disk requirements	*Disk I/O requirement (parallel I/O needed, large disk space and/or remote databases access needed during execution, etc.)*	
Other	*Any other comment on the computational characteristics of the code that doesn't fit above.*	

Run info

System where it runs	*On which HPC system the code is used.*	
Mode of operation	*Brief description on how a production run of the code is structured (i.e. single extreme parallel run for each investigated problem / sensitivity analysis: multiple independent runs /multiple intermediate parallel runs interconnected etc.).*	
Input description	**Format**	*Specify if specific data format is used (i.e. pdb, hdf5, images).*
	Coming from	*If applicable please specify if input file are generated by other codes.*
	Disk use	*Dimension and number of input file(s).*
Output description	**Format**	*Specify if specific data format is used (i.e. pdb, hdf5, images).*
	Used by	*If applicable please specify if output files are used as input by other codes or post processed for visualization or analysis purposes.*
	Disk use	*Typical dimension and number of output file(s).*
Complimentary Tools	*List here any other tool/code used with the application for production runs (i.e. Taverna for workflow management, spark/Hadoop for HPDA, etc.).*	

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

Potential Development

OPTIONAL *(any comment on potential development of the application/workflow are welcome but not needed at this stage)*

Additional info

OPTIONAL

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*