Funded by the European Union under the
HORIZON-EUROHPC-JU-2023-INCO-06 programme.
Grant Agreement No. 101196247

Deliverable D3.1
Life Sciences activities planning report

WP3 –Computational Life Sciences

Under Review

Funded by
the European Union

Document Information

Deliverable Number	D3.1
Deliverable Title	Life Sciences activities planning report
Due Date	2025-08-31 (PM7)
Deliverable Lead	UU
Authors	Alexandre Bonvin (UU) Leonardo Salicari (CINECA) Alessandra VIlla (KTH) Dmitry Morozov (JYI) Andrey Alekseenko (KTH) Szilárd Páll (KTH)
Keywords	GROMACS, HADDOCK, AI, ICE-CLOUD, immune complexes
WP	WP3
Nature	Report
Dissemination Level	Public
Final Version Date	2025-08-15
Reviewed by	Antsa Ratsima (BULL SAS), Jorge Macías Sánchez (UMA)
MGT Board Approval	2025-08-22

Document History

Partner	Date	Comments	Version
UU	2025-07-23	First draft	0.1
All	2025-08-05	Complete version for internal review	0.2
UU	2025-08-15	Updated version after internal review	0.3

Executive Summary

This document presents the planning and organization of activities within Work Package 3 (WP3) – Computational Life Sciences – of the GANANA project. WP3 is focused on advancing high-performance computing (HPC) applications in Life Sciences and contributes to the project's overarching goals: fostering scientific collaboration between the EU and India, and enhancing key scientific software tools. The definition of the work and the planning of activities have been carried out in close coordination between the EU and Indian partners to make sure both work and timelines are well aligned between the various teams.

The Computational Life Sciences workpackage is structured into three sub-projects:

- ***Sub-Project 1: Enhancing molecular dynamics simulations with GROMACS.*** Key tasks include improving performance through fixed-precision coordinates, developing modern communication algorithms, co-designing software/hardware solutions, and creating unified interfaces for hybrid simulations.
- ***Sub-Project 2: Advancing integrative modeling with HADDOCK.*** This includes modeling nucleic acids with modified bases, integrating AI for protein-protein interaction modeling, and optimizing HADDOCK for execution on Indian computing resources.
- ***Sub-Project 3: Creating a framework for AI-driven immunogenic peptide prediction.*** This involves integrating AI models for peptide-MHC-TCR interactions with GROMACS and HADDOCK simulations, along with building workflows for 3D modeling and ML-based binding prediction.

This document details WP3's organizational structure, milestones, deliverables, and a proposed timeline of the various tasks. It also highlights collaboration strategies and communication tools designed to support efficient EU-India cooperation. Finally, it provides an overview of the work completed or currently underway since the start of the project.

Table of Contents

1 Introduction

This document describes the planned activities within Work Package 3 (WP3) – *Computational Life Sciences*. WP3 focuses on high-performance computing (HPC) applications in the Life Sciences, contributing to the GANANA project's main objectives: strengthening of scientific collaborations between the EU and India and further improving leading software for scientific computing in the sciences.

The specific objectives of WP3 are to:

- Accelerate and enhance the GROMACS molecular dynamics simulation software.
- Extend the capabilities of HADDOCK as a framework for Integrative Modelling of Biomolecular Complexes, supporting a wider range of molecules and incorporating AI-based scoring functions.
- Develop a framework for AI-driven Immunogenic Peptide Prediction.
- Facilitate the deployment and use of software on Indian HPC/HTC resources.

In order to reach those objectives, WP3 is sub-divided into three sub-projects.

Sub-Project 1 focuses on *Accelerating and Enhancing Molecular Dynamics Simulations* with GROMACS as core software. This work will involve:

- Developing software for fixed-precision coordinates for improved performance of MD simulations (T3.1).
- Developing modernized communication algorithms for molecular dynamics simulations (T3.2).
- Co-designing software and hardware for molecular dynamics simulations (T3.3).
- Creating an Interface for arbitrary reaction coordinates and a unified interface for hybrid QM/MM and ML/MM enhanced sampling simulations (T3.4).

Sub-Project 2 focuses on a F*ramework for Integrative Modelling of Biomolecular Complexes* with HADDOCK as core software. This work will consists in:

- Extending the capabilities of HADDOCK in modeling nucleic acids with modified bases (T3.5).
- Developing improved AI models for protein-protein interactions (T3.6)., which will result in new modules added to HADDOCK and building an efficient execution machinery for HADDOCK on Indian resources (especially ICE-CLOUD) (T3.7).

Sub-Project 3 focuses on developing a *Framework for AI-driven Immunogenic Peptide Prediction*, which will combine AI models for predicting peptide- Major Histocompatibility complex (MHC)- T-cell receptor (TCR) interactions and HADDOCK and Gromacs for modelling those. The work will train *AI models for Genomic based identification of peptides and TCR loops* (T3.8), develop *Workflows for 3D modelling of peptide MHC/TCR*

complexes (T3.9), construct *automation workflows* to simulate the dynamics of the generated models (T3.10) and develop Structural and sequence based machine learning models to predict binding specificities of TCRs.

In the following we first describe the overall organization of the work package with its sub-projects, milestones and deliverables, together with the planned timeline of work. As the EU and India proposals differ in task numbering, both are reported. We then describe the communication and collaborative strategy and tools that have been put in place to facilitate the work and stimulate the collaboration between the European and Indian teams. Finally, we provide an overview of the current work in each of the three sub-projects.

Note that the planning of the work reported in this deliverable brought to light some inconsistencies between the two projects in terms of timeline of the planned activities. Also the starting dates of the EU and Indian project differ by 2 ½ months (EU project starting date is February 1st, official approval of the Indian proposal is April 17th). Accordingly, the delivery date of some deliverables and milestones has been adjusted (pending approval by the project officer).

2 WP3 project organization and timeline

2.1 Project1: Accelerating and Enhancing Molecular Dynamics Simulations

2.1.1 Description of work

GROMACS (www.gromacs.org) is a very efficient and versatile molecular dynamics (MD) engine. It is used worldwide for both atomistic and coarse-grained simulations of systems going from biological macromolecules, polymers to many other types of molecules. It has become one of the most used HPC applications worldwide, due to its good performance on all relevant CPU and GPU architectures and its ease of use. In order to expand GROMACS capabilities to enhancing and accelerating MD simulations, we will work on **four main tasks**, namely:

- **Improving performance by implementing fixed precision coordinates** The interest in simulating large biological systems is continuously increasing, posing particular challenges for simulation software. Large systems cover larger coordinates in space, which means less absolute precision is available. The main effect that deteriorates the quality of the results is larger noise, due to rounding errors in the integration. One remedy is compiling GROMACS in double precision (for small systems single precision has proved sufficient), but that affects performance significantly and excludes the use of general-purpose GPUs. To avoid the overhead, in both memory and computation, of double precision, an elegant solution is to use fixed precision instead. Fixed precision coordinates have not yet been implemented in GROMACS, and many other molecular dynamics packages, because all code dealing with coordinates needs to be adapted.
  The EU team will design a class that holds fixed precision coordinates and has methods to produce distances in floating point format given pairs of atom indices. EU and India partners will implement this class in the most important functions, so that standard MD simulations can make use of this. India partners will evaluate the performance of the resulting code both on currently available hardware as well as on the indigenous hardware, when that becomes available.

- **Improving scalability of molecular dynamics simulation by modernising communication algorithm** Modern HPC architectures expose an increasing amount of parallelism. Making use of this requires exposing an increased amount of concurrency in both computation and data movement. This necessitates not only expressing finer-grained parallel computation but also

fine-grained communication with the possibility of overlapping communication and computation. Core parallel MD algorithms such as pair interaction calculation have traditionally been optimised to avoid computation and reduce communication volume as much as possible. However, on modern architectures these approaches are often not optimal. For scalability it is critical to optimize data movement and to employ latency-hiding techniques (reduce or overlap) often at the cost of additional computation or increased volume. In addition, thanks to the widespread use of accelerators, algorithms need to become GPU-centric and make better use of the throughput-oriented hardware latency hiding capabilities.
GROMACS has mature support for GPU-resident parallelization and direct-GPU communication; this still uses a CPU-initiated approach which suffers from latencies.
The EU-team will modernise communication algorithms to target current CPU-only and GPU-accelerated architectures and implement fine-grained GPU-resident communication; latency hiding techniques will be considered by using accelerator-centric optimisation approaches. Indian partners will evaluate the implementation and will provide feedback for potential optimisations. This evaluation will represent one of the primary targets for the **software/hardware co-design** efforts (see description below). Results of which will be considered in improving code or documentation, when applicable.

- **Improving efficiency of molecular simulation by software/hardware co-design** Co-design of HPC applications with HPC hardware and software stacks has become key to further improving time- and energy to solution. Modernising and optimising algorithms benefits from a collaborative approach between application software and HPC hardware / software teams. By exploring ways to specialise and tune HPC software stacks for key applications like GROMACS, new opportunities open up to exploring application needs and optimising HPC software stacks early. This can help ensure that HPC software stacks are ready before machine deployment. In addition, codesign helps inform the design of HPC systems to be better-suited for certain applications, e.g. by the choice of the ratio of CPU to GPU resources, compute to network, or network topologies.
  The EU team will design test cases and identify key optimisation targets, taking into account current and planned EU and Indian HPC platforms. EU partners will support the Indian partner in their evaluation of CDAC communication software stacks as well as hardware (depending on availability), including the evaluation of **modernising communication algorithm** algorithms described above . Indian partners will evaluate and optimize the C-DAC communication library and SYCL software stack for MD applications using GROMACS.

- **Enhancing simulation sampling via ML based reaction coordinates and hybrid QM/MM and ML/MM approach.** Molecular dynamics is limited to timescales of microseconds, because the integration time step is on the order of femtoseconds, but Biologically relevant processes usually take many orders of magnitude longer. However, these processes can be studied, when the system is biased along a properly chosen reaction coordinate. GROMACS has basic (distance and angle based) reaction coordinates, which can be used for biassing with the accelerated weight histogram method. However, often more complex reaction coordinates are required to sample processes of interest, in particular for conformational transitions, folding, or chemical reactions. Whereas with molecular mechanics force fields, molecular dynamics is limited to microsecond timescales, simulations based on hybrid quantum mechanics / molecular mechanics (QM/MM) are limited to nano-seconds, which is too short to observe enzymatic reactions, which are rare events due to the high energy barrier that separates the reactant from the product. Therefore, to understand the mechanism and predict the effect of mutations, enhanced sampling along the reaction coordinate is essential.
  The Indian partners will investigate new, more complex ML-based reaction coordinates. EU partners will facilitate the reading of the reaction coordinates in GROMACS, such that India-developed reaction coordinates can be used to bias MD simulations. Indian partners will validate and apply the new type of reaction coordinate for computing reaction free energies and barriers in enzymes at the QM/MM level of theory. Because the current QM/MM implementation connects GROMACS and CP2K, a plane-wave based density-functional theory (DFT) code, the description of the QM subsystem is restricted to DFT methods. Yet, many problems in biology require more accurate descriptions that are based on correlated ab initio quantum chemistry, in particular for excited states. To provide access to such methods, EU-team will restructure the interface into a general API to arbitrary quantum chemistry software packages. In addition, because modern ML approaches provide potential energy functions that can compete with quantum chemistry for accuracy at only a fraction of the computational costs (after training), EU-team will also optimize the API for message passing between GROMACS and various ML softwares. The validity of these developments will be verified together with the Indian partners on specific use-cases of relevance to immunology.

2.1.2 Timeline

The following gantt chart presents the timeline of activities in sub-project1, both on the EU and India side. Task numbers correspond to the respective numbering in the EU and India proposals. Since the Indian project only started two and half months after the start of the EU one (EU project starting date is February 1st, official approval of the Indian proposal is April 17th), the timeline runs up to month 38 as the EU project has already been extended by two months to match the activities on the Indian side. The Indian due dates are reported in the 1 to 38 months scheme (which means 2 should be subtracted to match the due date in the Indian proposal).

		Project months (EU started at M1, India at M3)																																				
Task	**Activity**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Fixed precision coordinates for improved performance of MD simulation																																						
T3.1.1 (India)	Literature search on methods for fixed precision approach. Study of Gromacs Source Code.																																					
T3.1.1 (India)	implementation of the defined algorithm to GROMACS codebase																																					
T3.1.1 (India)	Evaluate accurancy and perfomance																																					
T3.1 (EU)	Design a class and a methods for fixed precision approachs and implementation																																					
T3.1 (EU)	Make production-ready implementation																																					
Modernised communication algorithm for molecular dynamics simulations																																						
T3.1.4 (India)	Design & Implement the application based topology awareness																																					
T3.2 (EU)	Design and implementation of modern communication algorithms to target current CPU-only and GPU-accelerated architectures																																					
Co-design software/hardware for molecular dynamics simulations																																						
T3.1.3 (India)	Benchmark & optimize the C-DAC software Stack for GROMACS																																					
T3.1.3 (India)	Porting C-DAC SYCL Ecosystem for MD applications																																					
T3.3 (EU)	Define and set up GROMACS test cases																																					
T3.3 (EU)	Identify key optimisation targets, taking into account current and planned EU and Indian HPC platforms																																					
Interface for arbitrary reaction coordinates and unified interface for hydrid QM/MM and ML/MM enhanced sampling simulations																																						
T3.1.2 (India)	Design ML-based methods for reaction coordinate																																					
T3.1.2 (India)	Validation of QM/ML/MM reaction coordinates																																					
T3.4 (EU)	Reading Indian ML-based reaction coordinates in GROMACS																																					
T3.4 (EU)	Generalized QM/ML/MM interface for QM/ML engines																																					

Caption: The color code represents which team is responsible for the activity: light orange for the Indian Team and green for the EU team.

2.1.3 Associated Milestones

Milestones, which are checkpoints throughout the project demonstrating that the work is on track (only defined in the EU proposal), are listed in the following table together with the associated task number (both EU and India). In case a Milestone's due date has been changed because of the alignment of activities between EU and India, the original due date is indicated between brackets.

Milestones			
Associated Task	**Name (Milestone EU project number)**	**Means of verification**	**Due date (Month)**
Task 3. 1.1 (India) **Task 3.1 (EU)**	EU GROMACS publicly available implementation of fixed precision (**M11**)	Code available on GitLab and registered with DOI in Zenodo	**36**
Task 3.1.4 (India) **Task 3.2 (EU)**	GROMACS publicly available implementation of modernised communication algorithms to target CPU-only and GPU-accelerated systems (**M12**)	Code available on GitLab and registered with DOI in Zenodo	**24**
Task 3. 1.3 (India) **Task 3.3 (EU)**	Define and set up GROMACS test cases to be used by C-DAC (**M13**)	Test cases available on-line	**14**
Task 3.1.2 (India) **Task 3.4 (EU)**	GROMACS AI-based reaction coordinate and ML/QM/MM interfaces prototype (**M14**)	Code available on GitLab and registered with DOI in Zenodo	**32** (was M30)

2.1.4 Associated deliverables

The following table lists deliverables (EU) or technical reports (India) and their respective due dates. The due dates slightly differ with the India reports being due toward the end of the project. Their content will however be aligned with the final Indian report which will be also presenting the activities in the last phase of the project, including dissemination and uptake in the Indian research community. The Indian due dates are reported in the 1 to 38 months scheme (which means 2 should be subtracted to match the due date in the Indian proposal). In case a Deliverable's due date has been changed because of the alignment of activities between EU and India, the original due date is indicated between brackets.

Deliverables			
Associated Task	**Deliverable Name (number)**	**Due date (India)**	**Due date (EU)**
Task 3.1.3 & 3.1.4 (India) **Task 3.2 & 3.3 (EU)**	India: Design & Implement the application based topology awareness . India: Porting SYCL Ecosystem and Benchmark & optimize CDAC Software Stack for MD EU: Modernized communication and co-design for Molecular Dynamics Simulations (**D3.2**)	**26** **38**	**24** (was 16)
Task 3.1.1 & 3.1.2 (India) **Task 3.1 and 3.4 (EU)**	India: 1. Prototype for fixed point precision in Gromacs code for coordinate calculations. 2. API's to process MD trajectory data and generate reaction coordinates using ML techniques EU: Hybrid QM/MM interface and fixed precision Molecular Dynamics Simulations (**D3.6**)	**38**	**36**

2.2 Project2: Framework for Integrative Modelling of Biomolecular Complexes

2.2.1 Description of work

This specific sub-project aims at developing a framework for integrative modelling of biomolecular complexes centered around the HADDOCK software. HADDOCK (High Ambiguity Driven protein-protein DOCKing) is an information-guided, flexible docking platform designed for modeling biomolecular complexes. It supports a wide range of interactions, including protein-protein, protein-nucleic acid, and protein-ligand complexes. Chemical modifications commonly occur across various biomolecules—such as nucleic acids, sugars, lipids, and proteins—and serve as key regulators of biological function. Currently, HADDOCK supports modeling of only standard nucleic acids.

In order to expand the capabilities of the software as integrative modelling framework we will work on **three main tasks**, namely:

- ***Adding support and parameters for the modeling nucleic acids with modified bases***. By incorporating parameters for modified RNA nucleotides we will enhance our capability of modelling protein-RNA interactions and provide deeper insights into the functional impact of RNA modifications on cellular processes. The India team will mainly work on the parametrisation of the modified nucleic acids, using various schemes/approaches and provide a translation from AMBER format to CNS format (CNS is the computational engine used by HADDOCK). The EU team will incorporate those new parameters into HADDOCK, testing and resolving any incompatibilities. In the first instance, the EU team will benchmark the normal nucleic acid parameters coming from AMBER to make sure they give an equal or better performance to the currently implemented nucleic acid parameters in HADDOCK. Later in the project (see timeline), a benchmark of protein - modified nucleic acids will be defined by the India team, and used by the EU team to benchmark the performance of the various parametrisation schemes for the modelling of those complexes. This task will result in a new release of the HADDOCK software with extended force field choices and capabilities in modelling modified nucleic acids.

- ***Developing improved AI models for protein-protein interactions.*** Protein-protein interactions (PPIs) are indeed fundamental to many biological functions. While Position-Specific Scoring Matrices (PSSMs) are widely used to capture evolutionary conservation at interaction sites and identify functionally

important residues, generating reliable PSSMs can be computationally demanding, particularly for large alignments. An alternative approach could involve replacing or augmenting PSSMs with contact maps, which depict spatial proximity between residues and offer insights into potential interaction sites. These maps, especially when derived from protein language models, could significantly reduce computational costs and improve efficiency in PPI prediction. The new AI models using contact maps will be developed and trained by the India team, extending the capabilities and information used by the DeepRank-GNN-Esm software. This process will happen in continuous concertation with the EU team. On the EU side, a new scoring module for DeepRank will be developed and implemented into HADDOCK. It will be updated later in the project to incorporate the new DeepRank version once the improved AI model has been finalized. The EU team will take care of benchmarking this new AI model for the scoring of protein-protein complexes.

- Broadening the accessibility and adoption of HADDOCK in India by ***building an efficient execution machinery for HADDOCK on Indian resources*** (especially ICE-CLOUD). Various container solutions will be investigated (e.g. Docker, aptainers) to find what works best on the Indian ICE-CLOUD resources (EU team), in concert with the Indian team which will work on defining, implementing and optimising the required microservices on the Indian cloud resources. A user interface will be developed, deployed and tested on ICE-CLOUD (India team) with input from the EU team for optimisation, user-friendliness and the development of training material (both teams) to facilitate the use of those resources by the large HADDOCK community in India.

2.2.2 Timeline

The following gantt chart presents the timeline of activities in sub-project2, both on the EU and India side. Task numbers correspond to the respective numbering in the EU and India proposals. Since the Indian project only started two and half months after the start of the EU one (EU project starting date is February 1st, official approval of the Indian proposal is April 17th), the timeline runs up to month 38 as the EU project has already been extended by two months to match the activities on the Indian side. The Indian due dates are reported in the 1 to 38 months scheme (which means 2 should be subtracted to match the due date in the Indian proposal).

Task number	Activity	Project months (EU started at M1, India at M3)
		1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
	Extend the capabilities of HADDOCK in modeling nucleic acids with modified bases	
T3.2.1 (India)	Definition of force field topologies and parameters for modified RNA	
T3.2.1 (India)	Definition of a modified protein-RNA complexes benchmark set	
T3.2.1 (India)	Protein-modified RNA complexes benchmarking	
T3.5 (EU)	Implementation and benchmarking of Amber-derived force field for standard nucleic acids in HADDOCK	
T3.5 (EU)	Implementation and benchmarking of Amber-derived force field for modified nucleic acids in HADDOCK	
	improved AI models for protein-protein interactions	
T3.2.2 (India)	Understanding of language model for proteins and generation of embedding and training	
T3.2.2 (India)	Definition of new language model based features and implementation in DeepRankGNN-esm and training	
T3.2.2 (India)	Testing and benchmarking and training of machine-based scoring function for HADDOCK	
T3.6 (EU)	Implementation of a DeepRank-GNN-Esm module into HADDOCK	
T3.6 (EU)	Implementation of the new extended version of DeepRank-GNN-Esm from the Indian partner into HADDOCK	
T3.6 (EU)	Benchmarking of the new DeepRank module for the scoring of protein-protein complexes	
	Efficient execution machinery for HADDOCK on Indian resources	
T3.2.3 (India)	Architecture design and development, development of backend microservices, database design, and unit testing	
T3.2.3 (India)	Development of user interface, integration backend of microservices with user interface, deployment, testing, enhancing and bug fixing	
T3.2.3 (India)	Final Deployment and Release of microservice based HADDOCK on Cloud	
T3.7 (EU)	Containarisation of HADDOCK and instructions for its execution	
T3.7(EU)	Optimisation of HADDOCK execution of Indian resources	
T3.7(EU)	Development of training material for the use of the Indian resources	

Caption: The color code represents which team is responsible for the activity: light orange for the Indian Team and green for the EU team.

2.2.3 Associated Milestones

Milestones, which are checkpoints throughout the project demonstrating that the work is on track (only defined in the EU proposal), are listed in the following table together with the associated task number (both EU and India). In case a Milestone's due date has been changed because of the alignment of activities between EU and India, the original due date is indicated between brackets.

Milestones			
Associated Task	**Name (Milestone EU project number)**	**Means of verification**	**Due date (Month)**
T3.2.1 (India) T3.5 (EU)	HADDOCK3 release with support for modified RNA bases (**M7**)	HADDOCK3 release on GitHub and registered with DOI in Zenodo	**M18** (was M12)
T3.2.1 (India) T3.5 (EU)	Protein-modified RNA complexes benchmark available (**M8**)	Benchmark set available online (e.g. on GitHub)	**M18** (was M12)
T3.2.3 (India) T3.7 (EU)	Demonstration of HADDOCK3 execution on Indian infrastructure (**M9**)	Instructions available online to run HADDOCK3 on the Indian ICE-CUBE infrastructure	**M20**
T3.2.2 (India) T3.6 (EU)	New version of DeepRank-GNN-esm for the scoring of protein-protein complexes (**M10**)	Code available on GitHub and registered with DOI in Zenodo	**M30**

2.2.4 Associated deliverables

The following table lists deliverables (EU) or technical reports (India) and their respective due dates. The due dates slightly differ with the India reports being due toward the end of the project. Their content will however be aligned with the final Indian report which will be also presenting the activities in the last phase of the project, including dissemination and uptake in the Indian research community. The Indian due dates are reported in the 1 to 38 months scheme (which means 2 should be subtracted to match the due date in the Indian proposal). In case a Deliverable's due date has been changed because of the alignment of activities between EU and India, the original due date is indicated between brackets.

Deliverables			
Associated Task	**Deliverable Name (number)**	**Due date (India)**	**Due date (EU)**
Task 3.2.1 (India) Task 3.5 (EU)	India: Support for modified nucleic acid bases in HADDOCK EU: Integrative Modelling with modified nucleic acids (**D3.4**)	**M38**	**M26** (was M18)
Task 3.2.2 (India) Task 3.6 (EU)	India: Extension of DeepRankGNN by ESM Language model contact Maps EU: Improved AI models for protein-protein interactions - (**D3.5**)	**M38**	**M33**
Task 3.2.3 (India) Task 3.7 (EU)	Development of a HADDOCK execution machinery based on ICE-CUBE / ICE-FLAKE	**M38**	-

2.3 Project3: Framework for AI-driven Immunogenic Peptide Prediction

2.3.1 Description of work

This sub-project aims to develop an AI-driven framework capable of immunogenic peptide prediction. Of foremost importance for this task is the recognition mechanism by which the T-cell receptor (TCR) recognises a short peptide presented by the major histocompatibility complex (MHC) of the cell. Recognition of foreign antigens and discrimination of self relies on the highly specific binding of the TCR to the peptide-MHC (pMHC) complex. Despite its centrality to adaptive immunity, structural characterization of pMHC–TCR complexes remains sparse.

This sub-project will develop an AI-based model to recognise immunogenic responses involving these protein complexes. To achieve this, automated workflows combining hybrid model development, protein complexes modelling, and molecular dynamics to extract structural features will be exploited. The previously described HADDOCK and GROMACS softwares will be used within the workflows. The project is structured into **four main tasks**, namely:

- ***Genomic-based identification of peptides and TCR loops***. The Indian team will compile sets of TCR-peptide and peptide-MHC from curated datasets such as VDJdb (curated database of T-cell receptor (TCR) sequences with known antigen specificities), IEDB (Immune epitope database), McPAS-TCR (pathology associated T-cell receptor database), MIRA (A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2). This will result in input training data for specificity prediction. Concurrently, various embedding techniques for TCR-peptide and peptide-MHC will be explored and feature engineering will be performed to optimize and incorporate relevant information. This is a preparatory work for the following Tasks that will start shortly afterwards (see timeline). Conventional supervised classification methods as well as deep learning approaches will be explored to create a sequence-based classifier of pMHC-TCR binding affinity. Validation on unseen datasets will be performed to check model accuracy. The main part of the work will be carried on by the Indian partners with inputs from EU collaborators on benchmark data and format for the following tasks.
- ***Workflows for 3D modelling of pMHC-TCR complexes.*** Following the identification of the binding and non-binding datasets (see the timeline), the EU partners will design several modelling protocols and build corresponding workflow implementations for the modelling of peptide-MHC and pMHC-TCR complexes. The implementations will exploit existing software such a PANDORA,

AI-based structural prediction tool and the docking/refinement machinery offered by HADDOCK. To tackle the conformation challenge of modelling the TCR loops, the EU team will explore an approach combining a generative AI model based on Flow Matching with HADDOCK in the context of antibody-antigen. This task will result in three workflows: for peptide-MHC complexes modelling, for pMHC-TCR complexes modelling and a final one harvesting the prediction from the ML hybrid model developed under Task 3.11. The latter will incorporate contact predictions and backbone conformational restraints, provided by the Indian team, to guide modelling of pMHC-TCR.

- ***Construct automation workflows to simulate the conformational dynamics of the model complexes*** generated by the previous step (T3.10). The EU team will design a protocol to perform first equilibrium simulations, followed by an enhanced sampling technique and finally the extraction of structural and binding parameters to be used as inputs for the downstream ML. Work on the simulation protocol started with the first peptide-HMC and TCR sequences with a corresponding experimental structure identified by the Indian Team. In this task, the EU team will concentrate on technology scouting to select an appropriate Workflow Management System (WMS) software to implement automated, fault-tolerant, and reproducible workflows (e.g. COMPSs, Pegasus etc.). The software stack will be easily deployable on Indian infrastructure, ensuring software integration with HPC systems. The pipeline will use the GROMACS package as the back-end simulation engine and will manage all inputs and outputs at each of the above stages. This involves preparation of the input complex topologies, creation of parameter files defining the conditions for equilibrium simulations, extraction of the representative complex conformation for enhanced sampling, definition of protocols and reaction coordinates for binding free energy estimations. Finally, it will include scripts to derive observables from simulations and export them in csv format for use in the downstream ML.

- ***Structural and sequence based machine learning***. In this task, the Indian partners will focus on predicting TCR's binding specificities with a hybrid model based on sequential data from Task 3.8 and on the structural model of pMHC-TCR complexes generated in Task 3.10. The workflow developed in T3.10 will allow to find the binding conformations of the peptides in MHC grooves and the binding poses of TCRs interacting with peptide-MHC complexes. The Indian partners will develop a framework for extracting the features from the interface of pMHC-TCR including the interacting residue pairs, which are crucial for accurate binding affinity assessments. The peptide backbone conformation will allow to finetune the TCR specificity, providing deeper insight into structural dynamics of pMHC-TCR interface. The features extracted from the binding interface will be used to train the AI model and retrain the models developed

in Task 3.9. Finally, EU and Indian teams will collaborate to rigorously test and validate the hybrid model to ensure its robustness and reliability for predicting TCR-binding peptides and their MHC restriction.

2.3.2 Timeline

The following gantt chart presents the timeline of activities in sub-project2, both on the EU and India side. Task numbers correspond to the respective numbering in the EU and India proposals. Since the Indian project only started two and half months after the start of the EU one (EU project starting date is February 1st, official approval of the Indian proposal is April 17th), the timeline runs up to month 38 as the EU project has already been extended by two months to match the activities on the Indian side. The Indian due dates are reported in the 1 to 38 months scheme (which means 2 should be subtracted to match the due date in the Indian proposal).

		Project months (EU started at M1, India at M3)																																				
Task number	**Activity**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Genomic based identification of peptides and TCR loops																																							
T3.3.1 (India) T3.8 (EU)	Preprocessing of peptide-TCR and peptide-MHC datasets to compile binding and non-binding sets																																						
T3.3.2 (India) T3.8 (EU)	Feature embedding of TCR-peptide interactions																																						
T3.3.3 (India) T3.8 (EU)	Ensemble/Deep learning methods to implement a supervised classifier using sequence-derived features																																						
T3.3.3 (India) T3.8 (EU)	Benchmark the developed model against unseen datasets																																						
Workflow and 3D modeling of peptide MHC/TCR complexes																																							
T3.9 (EU)	Development of a structural modeling workflow for peptide-MHC complexes																																						
T3.9 (EU)	Development of a structural modeling workflow for pMHC-TCR complexes																																						
T3.9 (EU)	ML-restrained modeling of pMHC-TCR complexes using prediction from model developed in T3.11																																						
Construct automation workflows																																							
T3.10 (EU)	Protocol design and workflow implementation for MD simulations of the selected system using GROMACS as backend																																						
T3.10 (EU)	Protocol design and workflow implementation for MD-based free energy, including binding affinity predictions of the selected system using GROMACS as backend																																						
T3.10 (EU)	Features generation for downstream AI-based hybrid model																																						
Structural and sequence based machine learning																																							
T3.3.4 (India)	Analysis of structural features from pMHC-TCR interface, e.g. residue pairs, and of peptide backbone conformations to tune TCR specificity																																						
T3.3.5 (India)	Exploration and training of AI-based hybrid model																																						
T3.3.6 (India)	Benchmark the developed hybrid model by testing and validating on experimental datasets																																						
T3.11 (EU)	Benchmark the developed hybrid model by testing and validating on experimental datasets																																						

Caption: The color code represents which team is responsible for the activity: light orange for the Indian Team and green for the EU team. In the Task number column, the main contributor to the task is written in black, while the corresponding task for the other team is written in light grey.

2.3.3 Associated Milestones

Milestones, which are checkpoints throughout the project demonstrating that the work is on track (only defined in the EU proposal), are listed in the following table together with the associated task number (both EU and India). In case a Milestone's due date has been changed because of the alignment of activities between EU and India, the original due date is indicated between brackets.

Milestones			
Associated Task	**Name (EU project number)**	**Means of verification**	**Due date (Month)**
T3.3.6 (India) T3.11 (EU)	Workflow for AI-driven Immunogenic Peptide Prediction (**M15**)	Documentation available on-line	**M17**

2.3.4 Associated deliverables

The following table lists deliverables (EU) or technical reports (India) and their respective due dates. The due dates slightly differ with the India reports being due toward the end of the project. Their content will however be aligned with the final Indian report which will be also presenting the activities in the last phase of the project, including dissemination and uptake in the Indian research community. The Indian due dates are reported in the 1 to 38 months scheme (which means 2 should be subtracted to match the due date in the Indian proposal). In case a Deliverable's due date has been changed because of the alignment of activities between EU and India, the original due date is indicated between brackets.

Deliverables			
Associated Task	**Deliverable Name (number)**	**Due date (India)**	**Due date (EU)**
Task 3.10 (EU)	EU: Automation workflow for immunogenic peptide prediction (**D3.3**)	-	**M17**
Task 3.3.1 (India) Task 3.11 (EU)	India: Delineation and incorporation of 3D structural features to develop a hybrid TCR-peptide prediction model EU: Framework for AI-driven Immunogenic Peptide Prediction (**D3.7**)	**M38**	**M38**

3 Communication and interaction structure

To promote interactions between the EU and Indian teams various measures and communication channels have been put into place:

1) Within the GANANA project ***Slack space***, a general WP3 life sciences and three specific sub-project channels have been defined which have been joined by the respective teams from the EU and Indian partners. This provides an easy and fast channel for day-to-day communication. Slack also allows researchers to send direct messages to individuals, or groups.
2) For more formal communication, and document sharing a ***Teams workspace***, GANANA-Global, has been defined, hosted by KTH, which brings together EU and Indian participants. This is also the location where joint documents (e.g. templates, drafts and completed milestones and deliverables are shared).
3) For software-related developments, next to the code-specific repositories (e.g. Gromacs and HADDOCK each have their own repositories), a **GANANA-EU-INDIA Git organization** has been created where project-specific developments (e.g. workflows, benchmarks, input files, …) will be shared, organized per work-package. For the data generated during the timeline of the project we refer to D1.3 – Data Management Plan.
4) Each sub-project is holding ***bi-weekly meetings*** to discuss and plan the work to be done. Here sub-project specific aspects are discussed. These meetings usually take place via Teams.

4 Sub-projects current status

4.1 Project1: Accelerating and Enhancing Molecular Dynamics Simulations

Fixed precision coordinates for improved performance of MD simulation **(EU-task 3.1 - India task 3.1.1)**. The objective of the task is to use 32-bit fixed-precision for particle coordinates to prevent rounding errors when handling large molecular systems.

India is in process of hiring a person for the India task. Meanwhile, exploration of various methods for converting double precision to fixed point precision is being explored. Study/exploration of source code of Gromacs is being done.

KTH is hiring a person for the EU-task starting for month 12. In the meantime, KTH has started exploratory work on the design and an initial prototype of fixed-precision

class. India partner started to familiarize with the GROMACS code using training materials and initiative connected with BioExcel CoE, like workshop "learn to code in GROMACS"

***Modernised communication algorithms for molecular dynamics simulations* (EU-task 3.2 - India Task 3.1.4).** The objective is to improve the scalability of MD simulations with GPU-centric algorithms. KTH worked on refining the support for GPU-resident parallelization and direct-GPU communication; this improvements will help both the redesigned algorithms as well as the current mature algorithms and implementation still using a CPU-initiated communication approach (background on the GROMACS heterogeneous parallelization: https://pubs.aip.org/aip/jcp/article/153/13/134110/199476).

KTH has worked on an initial NVSHMEM-based prototype for domain-decomposition halo exchange for improving multi-GPU scaling. The prototype is ready for in-depth correctness and performance testing on Indian software/hardware stacks. KTH has also done early small-scale testing that revealed promising performance behavior, but has also identified the need for further optimizations.
C-DAC will formally commence its work from Month 7 (Month 9 for EU counting) of the project and is currently in the process of recruiting a dedicated resource for this activity. In the meantime, C-DAC has initiated preliminary exploratory work focused on evaluating GROMACS communication patterns, starting with a comparison of the staged and new direct domain-decomposition scheme on CPUs.

Co-design software/hardware for molecular dynamics simulations (EU-task 3.3 - India task 3.1.3). The objective of this task is to improve efficiency of HPC hardware / software stacks for molecular dynamics workloads via co-design. KTH has been working to ensure that GROMACS supports SYCL for performance-portability on AMD, Intel, and NVIDIA GPUs (e.g, SYCL backend general design and optimizations for AMD platforms: https://arxiv.org/pdf/2405.01420), including the support for GPU-aware communication algorithms. KTH will provide test-cases to India partners to evaluate and optimize the communication library and SYCL software stack for MD applications using GROMACS as a case study. The design of the test-case is based on the C-DAC inputs about hardware/software stack specifications.
C-DAC is delivering a GROMACS build compiled with the ParaS compiler, which is based on the SYCL 2020 specifications. The build is configured to support a range of architectures, including NVIDIA A100/H100 GPUs, AMD MI300x GPUs, as well as Intel Sapphire Rapids (SPR), AMD Genoa, and ARM Neoverse V2 CPUs. The relevant SYCL kernels within GROMACS have been identified. The build is progressing with these kernels, along with the required adaptations, specific to the ParaS compiler

***Interface for arbitrary reaction coordinates and unified interface for hybrid QM/MM* (EU-task - India task 3.1.2)**. The objective of the task is the introduction of a ML based reaction coordinate interface into the GROMACS pulling code and generalization of the ML/MM and QM/MM interfaces.

Exploration of existing Machine learning methods to generate reaction coordinates is done by C-DAC, India. The existing libraries like Colvar, which calculates collective variables for enhanced MD sampling, have been discussed with EU partners. Study and implementation of ML methods (which are not part of Colvar Library) like time lagged Independent Component Analysis (tICA) is discussed with EU partners and development of the same is taken up by C-DAC India.

JYU is evaluating in collaboration with India the best approach to introduce ML based reaction coordinate into GROMACS. Exploratory work has been done in evaluating the possibility to reuse available GROMACS capabilities/interface, such as the Colvars library, libtorch and transformation coordinates. KTH/JYU have previously contributed to the implementation of those interfaces. To ensure that evaluation targeted India needs, India partners will share details about the type of simulation data and analysis used to design ML reaction coordinates. Concerning the ML/MM/QM interface, one of the first steps is to select the QM software to be interfaced. The choice will be oriented toward a QM software, that has all the functionality requested by the India community, but also fulfills the EU requirement, such as OpenSource, as an example option we are currently investigating the ORCA package.

4.2 Project2: Framework for Integrative Modelling of Biomolecular Complexes

Adding support and parameters for the modeling nucleic acids with modified bases (India Task 3.2.1- EU Task 3.5). A first meeting took place in which the most common nucleic acid modifications were presented and prioritized for implementation. The Indian team will take care of converting parameters and topologies from AMBER to CNS format. As this means using a different nucleic acid force field as currently implemented into HADDOCK, the EU/UU team will first benchmark the force field impact on the docking performance for protein-DNA complexes without any modified bases.

Currently, five RNA-modified nucleotides have been selected based on their prominent occurrence in biological systems. We are following two approaches:

The first approach involves converting the parameters available in the modrna08 forcefield of AMBER for modified RNA nucleotides from the AMBER format to the CNS format, enabling their implementation in HADDOCK.

The second approach involves obtaining the structure of each modification, optimizing it, and performing quantum chemical calculations at the DFT level using Veloxchem to derive partial charges. These charges are then used in RESP fitting to

generate the required parameters. These parameters are also converted to the CNS format for HADDOCK implementation.
The implementation and benchmarking in HADDOCK are being carried out by the EU partner. Meanwhile, the Indian partner plans to run molecular dynamics (MD) simulations of protein–RNA complexes using both sets of parameters to evaluate their effectiveness in capturing the influence of modifications on the overall stability of the complex.
This comparative analysis will help determine which parameter set performs better within HADDOCK.

Developing improved AI models for protein-protein interactions (India Task 3.2.2- EU Task 3.6). The first meeting to discuss work in this task actually took place in January in Pune (i.e. before the start of the project) during the Accelerating Biology 2025 meeting in which Prof. Bonvin was speaking. During this meeting the general strategy to extend DeepRank with contact maps and the benchmark set of complexes to use for this purpose was discussed. The Indian team is extending the Deeprank by using Graph attention network. To facilitate the use of DeepRank and promote in the future the results of this task, the EU team (UU) has started building a web server for DeepRank (the prototype is already accessible at wenmr.science.uu.nl). In the future, under Task 3.7, this interface could be offered under the ICE-CLOUD infrastructure.

Efficient execution machinery for HADDOCK on Indian resources(India Task 3.2.3 - EU Task 3.7). First meetings took place to define the requirements to provide haddock3 under the ICE-CLOUD infrastructure. Currently ICE-CLOUD only supports Docker containers. For haddock3, instructions to build Docker containers have been published on the haddock3 user manual. The Indian team built a first haddock3 image which is now available on ICE-CLOUD for testing purposes.

4.3 Project3: Framework for AI-driven Immunogenic Peptide Prediction

Genomic based identification of peptides and TCR loops (India Task 3.3.1/3.3.2/3.3.3 - EU Task 3.8). The current main effort of the sub-project is devoted to this task which is propedeutic for the followings (see timeline in 2.3.3 and more details on the Description of Work in 2.3.1). Indian team from C-DAC, in collaboration with NII, started the exploration of genomic databases to compile two datasets of pMHC-TCR sequences: a binding set and a non-binding set. NII, during the recurrent meetings, gave a seminar on the topic, providing important insights regarding which features could be relevant to be extracted. Shortly after the first meeting of the sub-project (May 2025), few peptide-MHC sequences with binding and non-binding TCR sequences have been identified. These have been shared with KTH to perform a preliminary study on the simulation protocol to implement. Before the kick-off

meeting (May 2025), C-DAC conducted a literature review of the available state-of-the-art pMHC-TCR specificity prediction AI tools to evaluate possible approaches for the hybrid model.

Workflow and 3D modeling of peptideMHC-TCR complexes (EU Task 3.9). Before the start of the project, exploratory work has been carried out by UU to develop an AI-based protocol for modeling. This task will build on the work related to structural predictions by AlphaFold enhanced by a flow matching technique, studied in the context of antibodies (Bioinfo. Adv.). This approach might be useful to sample conformation of the CDR loops of the TCR. Peptide-MHC complexes can be built using different software (e.g. HADDOCK or PANDORA (Frontier. Imm.)). The modelled peptide-MHC complexes and TCR will be used as starting points for docking with HADDOCK to model the full pMHC-TCR complexes. Feedback from T3.11 on the pMHC-TCR interface will be used to guide docking in a second iteration of the modeling. UU hired a person from M7 that already took part in the recurrent meetings and will work on the topic. Regarding workflow development and its implementation on HPC infrastructures, CINECA started technology scouting after the kick-off meeting in May 2025 (see the following task for more details).

Construct automation workflows (EU Task 3.10). As presented in the first Task, KTH used some preliminary pMHC-TCR sequences with corresponding experimental structures to start the design of the simulation protocol. Discussions regarding the relative free energies of binding calculations highlight the possibility of exploiting single-residue mutations and continuous lambda protocol. CINECA explored possible Workflow Management System (WMS) softwares. Main features that the WMS software has to satisfy are: open source, actively supported, ease to deploy on HPC systems, fault tolerant, supporting Data Provenance, compliant with FAIR principles and with APIs to use HPC's job schedulers such as Slurm. A possible candidate could be the WSM developed by BSC called PyCOMPSs, for which tests are performed on the EuroHPC supercomputer Leonardo.

Structural and sequence-based machine learning (India Task 3.3.4/3.3.5/3.3.6 - EU Task 3.11). As described in the first Task, C-DAC conducted literature review of the available state-of-the-art pMHC-TCR specificity prediction AI tools in order to explore current sequence-based approaches. After the development of the sequence-based model in T3.8, features from the structure-based model will be incorporated to build an hybrid model. Therefore, the main effort for this task will take place when structural features from Task 3.10 will be available (see timeline in 2.3.3 and more details on the Description of Work in 2.3.1).