



# UNIVERSITAT DE BARCELONA

Final Degree Project

**Biomedical Engineering Degree**

**“ Molecular dynamics of a subset of  
central nervous system proteins “**

Barcelona, 07 June, 2023

Author: Isabel Martín Valle

Director/s: Prof. Modesto Orozco

Tutor: Dra. Roser Sala



UNIVERSITAT DE  
BARCELONA

**Universitat de Barcelona**

Facultat de Medicina



INSTITUTE  
FOR RESEARCH  
IN BIOMEDICINE

**Institute for Research in Biomedicine**

Molecular Modeling and Bioinformatics Research Group

# Acknowledgements

I would like to express my deep gratitude to Prof. Modesto Orozco for offering me the valuable opportunity of doing this project within the Molecular Modeling and Bioinformatics research group at the Institute of Research in Biomedicine and Dr. Adam Hospital for his support, guidance and confidence in my work. I am also extremely grateful for the professional opportunities extended to me that exceeded my highest expectations, such as travelling to the University of Oxford and participating in several European projects.

It is inevitable that I follow my thanks with a mention of Daniel Beltrán. This project would not have been possible without his guidance. For the patience and the time he has dedicated to teaching me and solving my doubts, for transmitting me his passion for his work and for the support he has been during the realization of this project. For all this and much more, I am infinitely thankful.

I would like to extend my gratitude and recognition to my tutor Dra. Roser Sala for her mentorship and the knowledge and insights she has shared with me.

To all the researchers in the Molecular Modeling and Bioinformatics for treating me like another colleague and sharing their knowledge. I truly appreciate the opportunity to learn from all of you.

# Abstract

The human brain is organized in a hierarchical structure of distinct but closely connected levels. Although our knowledge about each individual level is extensive, what remains unknown is a comprehensive understanding of how events in low levels propagate through higher levels. This is a challenge for neuroscience.

This project focuses on low-level protein events. Its aim is to simulate the dynamics of a subset of central nervous system proteins as their mobility and flexibility are essential for neurological signal transduction. In particular, it is centred on the subset of CNS proteins related to Parkinson's disease. By reproducing their dynamics, the process of signal transduction in the brain can be better understood, as well as helping in the design of neuro-active drugs for Parkinson's disease.

This document provides a detailed description of the development of the project to obtain the dynamics of the proteins of study. The scope of the project goes from generating a list of proteins of interest, running and validating their dynamics and uploading the data to an open access database. A total of 49 simulations were successfully obtained and uploaded.

These simulations can be further studied to understand protein functions in neurological pathways as well as study possible drug binding interactions.

# Contents

<b>Acknowledgements .....</b>	<b>3</b>
<b>List of Figures .....</b>	<b>8</b>
<b>List of Tables.....</b>	<b>10</b>
<b>Abbreviations .....</b>	<b>12</b>
<b>1. Introduction.....</b>	<b>14</b>
1.1. <i>Motivation and topic presentation .....</i>	14
1.2. <i>Project objectives.....</i>	15
1.3. <i>Scope and limitations.....</i>	15
1.4. <i>Project justification.....</i>	16
1.5. <i>Research groups involved in this project.....</i>	16
1.6. <i>Trending practices performed during this project.....</i>	17
1.7. <i>Personal involvement.....</i>	17
<b>2. Background.....</b>	<b>19</b>
2.1. <i>Theoretical background .....</i>	19
2.1.1. <i>Molecular dynamics simulations .....</i>	19
2.1.2. <i>Force fields .....</i>	20
2.1.3. <i>Water systems .....</i>	20
2.1.4. <i>Periodic boundary conditions .....</i>	21
2.1.5. <i>Molecular dynamics algorithms.....</i>	22
2.1.6. <i>Type of files in molecular dynamics field.....</i>	22
2.2. <i>Applications of molecular simulations in proteins field .....</i>	23
2.3. <i>State of the art of molecular simulation.....</i>	24
<b>3. Market analysis .....</b>	<b>25</b>
3.1. <i>Current market for molecular simulations.....</i>	25
3.1.1. <i>Trends and prospects .....</i>	25
3.1.2. <i>Main competitors and collaborators .....</i>	25
3.2. <i>Potential of molecular simulations in the scientific market .....</i>	26

3.2.1.	Areas of opportunity .....	26
3.2.2.	Challenges and barriers .....	26
<b>4.</b>	<b>Conceptual Engineering.....</b>	<b>27</b>
4.1.	<i>Selection of the molecular system to study</i> .....	27
4.1.1.	Establishment of the selection criteria .....	28
4.2.	<i>Selection of databases and repositories to obtain molecular systems to study</i> .....	30
4.2.1.	Select central nervous system proteins.....	32
4.2.2.	Selection of resources to filter and reduce dataset .....	34
4.3.	<i>Software resources</i> .....	38
4.3.1.	Simulation software.....	38
4.3.2.	Force fields .....	39
4.3.3.	Software for visualizing results of simulations .....	39
4.4.	<i>Computing resources</i> .....	40
4.5.	<i>MDposit platform</i> .....	40
<b>5.</b>	<b>Detailed Engineering .....</b>	<b>42</b>
5.1.	<i>Selection of proteins of interest</i> .....	43
5.1.1.	Selection of central nervous system proteins.....	43
5.1.2.	Filters applied to reduce the dataset .....	51
5.2.	<i>Running the simulations</i> .....	58
5.2.1.	Obtaining the files to run the simulations .....	58
5.2.2.	Preparation of the systems for simulation .....	59
5.2.3.	Running the simulations.....	59
5.3.	<i>Results of the simulations</i> .....	60
5.4.	<i>Processing the simulations</i> .....	61
5.4.1.	Imaging and fitting.....	61
5.4.2.	Basic analysis of the simulations .....	62
5.4.3.	Specialized analyses of the simulations.....	63
5.5.	<i>Inclusion of the results in the MDposit platform</i> .....	63
<b>6.</b>	<b>Execution chronogram .....</b>	<b>65</b>
6.1.	<i>Work Breakdown Structure</i> .....	65
6.2.	<i>Tasks specifications</i> .....	66
6.3.	<i>GANTT diagram</i> .....	67
<b>7.</b>	<b>Technical viability .....</b>	<b>69</b>

<b>8. Economic viability .....</b>	<b>71</b>
<b>9. Legal aspects .....</b>	<b>73</b>
<b>10. Conclusions and future work.....</b>	<b>74</b>
<b>11. Bibliography .....</b>	<b>75</b>
<b>Appendix .....</b>	<b>83</b>

# List of Figures

Figure 1. Different protein structure representations: one atomistic model (left) and two coarse-grained models (center and right) of different resolution <sup>[12]</sup> .....	19
Figure 2. Protein embedded in a water box <sup>[18]</sup> .....	21
Figure 3. Diagram of PBC that show unit cell surrounded by its replicate <sup>[16]</sup> .....	21
Figure 4. Publication trend evolution for molecular dynamics simulation from 1977 to 2023 with a peak of 5,470 publications in 2021 <sup>[26]</sup> .....	24
Figure 5. Molecular modeling market evolution until 2030, reaching a market value of 7,8 Bn <sup>[28]</sup> . .....	25
Figure 6. Scheme of how proteins are involved in human brain signal transmission <sup>[32]</sup> . ....	28
Figure 7. Structure of the human dopamine D3 receptor in complex with eticlopride (PDB ID: 3PBL, UniProt ID: P35462) <sup>[43]</sup> . ....	31
Figure 8. For protein Adenosine receptor A2a (UniProt ID P29274), several experimentally determined structures are found, each corresponding to one PDB ID. Both displayed structures represent the same part of the protein (chain A, position 1-316), although with different conformations <sup>[45]</sup> .....	32
Figure 9. Coarse-grained (left) and atomistic models (right) of a membrane protein obtained from MemProtMD workflow <sup>[52]</sup> .....	35
Figure 10. Workflow followed of the execution part of the project (Detailed Engineering).....	42
Figure 11. Bgee dataset organization. ....	44
Figure 12. Tree hierarchy for central nervous system Uberon ID <sup>[72]</sup> .....	45
Figure 13. TSPN6 gene can be classified as expressed in components of the CNS but not classified as expressed in CNS itself. ....	46
Figure 14. CNS genes from Bgee (55,878) mapped to UniProt proteins (19,349). ....	47
Figure 15. Human Protein Atlas dataset organization. ....	47
Figure 16. 5.1% of proteins belonging to the raw HPA dataset are filtered out. ....	48
Figure 17. TISSUES database organization: gene identifier, gene name, tissue identifier, tissue name and integration of all confidence scores. ....	49



Figure 18. CNS genes from TISSUES mapped to UniProt. ....	50
Figure 19. Pharos database organization. ....	51
Figure 20. Comparison between the number of entries of each target classification in Pharos. ..	52
Figure 21. Open Targets - Target dataset organization showing a few columns. ....	52
Figure 22. Frequency histogram of which score values for PD are more recurrent and how they are distributed. ....	54
Figure 23. Atomistic-system.PDB file of a G Protein Coupled Receptor (PDB ID: 4GBR) where the structure of the system can be observed. A) Structure of all the system: water molecules, membrane and protein. B) Structure of the membrane and protein. C) Structure of the protein. ....	59
Figure 24. Resulting trajectory of the simulation. Mathematically in principle is correct but when visualizing it the periodic boundary conditions effects can be appreciated. Left image shows the entire system and right one shows the protein. ....	61
Figure 25. Simulation uploaded of Arginase-1 (PDB ID: 6QAF, UniProt ID: P22974). Overview section. ....	64
Figure 26. Simulation uploaded of Arginase-1 (PDB ID: 6QAF, UniProt ID: P22974). Trajectory section. ....	64
Figure 27. Work Breakdown Structure for this project. ....	65
Figure 28. GANTT diagram. ....	68
Figure 29. SWOT analysis of the resulting protein dynamics and their upload to MDposit platform. ....	69

# List of Tables

Table 1. Summary of the filters applied in order to select those CNS proteins of interest.....	29
Table 2. External resources used by <i>PDB</i> to identify membrane proteins. ....	34
Table 3. Databases considered as possible databases to select targetable proteins. ....	36
Table 4. Common software in MD field. ....	38
Table 5. Downloaded databases for selecting the proteins of interest. ....	43
Table 6. How data in Bgee database downloaded is structured. ....	44
Table 7. Columns of interest in HPA dataset. ....	48
Table 8. Columns of interest in Tissues database. ....	49
Table 9. Matrix of resulting <i>UniProt</i> IDs from each overlap .....	51
Table 10. Columns of interest for data manipulation of Pharos.....	51
Table 11. Columns of interest in Open Targets.....	53
Table 12. How data in Open Targets - Associations - direct (overall score) is structured. ....	53
Table 13. Summary of all filters applied and their output for selecting the proteins of interest. ....	57
Table 14. Files needed for running the simulations.....	58
Table 15. Table of tasks and their due date and dependencies.....	65
Table 16. Subtasks for task 1.1.....	66
Table 17. Subtasks for task 1.2.....	66
Table 18. Subtasks for task 1.3.....	66
Table 19. Subtasks for task 2.1.....	66
Table 20. Subtasks for task 2.2.....	67
Table 21. Subtasks for task 3.1.....	67
Table 22. Subtasks for task 3.2.....	67
Table 23. Subtasks for task 4.1.....	67
Table 24. Economic viability of the project. ....	71

Table 25. Resulting trajectories, each one from one <i>PDB</i> , which have been already uploaded to <i>MDposit</i> .....	83
---	----

# Abbreviations

Amino acid		Future and Emerging Technology	
(aa).....	27	(FET) .....	16
antibody		G-protein-coupled receptors	
(AB).....	53	(GPCRs) .....	26
Barcelona Supercomputing Center		High-Performance Computing	
(BSC-CNS) .....	17	(HPC) .....	17
Brenda Tissue Ontology		Human Protein Atlas	
(BTO).....	49	(HPA).....	33
Central nervous system		Institute for Research in Biomedicine	
(CNS).....	14	(IRB) .....	16
Centre Européen de Calcul Atomique et Moléculaire		Membrane Proteins of Known Structures	
(CECAM) .....	17	(mpstruc) .....	34
Druggable Genome		Molecular dynamics	
(IDG) .....	37	(MD).....	15
EMBL's European Bioinformatics Institute		Molecular Dynamics Data Bank	
(EMBL-EBI's) .....	32	(MDDB) .....	16
Energy Minimization		National Institutes of Health	
(EM).....	22	(NIH).....	37
European Bioinformatics Institute		Nuclear magnetic resonance	
(EBI) .....	30	(NMR).....	31
Food and Drug Administration		Orientations of Proteins in Membrane database	
(FDA).....	28	(OPM).....	34

Other clinical modalities	(RMSD).....	62
(OC).....		53
Parkinson's disease	(SM).....	28
(PD).....		15
Periodic boundary conditions	(INB).....	17
(PBC).....		21
Pressure Relaxation	(SIB).....	30
(PR).....		22
Protein Data Bank	(TCRD).....	37
(PDB).....		30
Protein Data Bank of Transmembrane Proteins	(TDL).....	37
(PDBTM).....		34
Protein Information Resource	UniProt Knowledgebase	
(PIR).....	(UniProtKB).....	30
Protolysis Targeting Chimeras	University of Oxford	
(PR).....	(UOXF).....	17
Root mean square distance	Work Breakdown Structure	
	(WBS).....	65

# 1.

# Introduction

## 1.1. Motivation and topic presentation

The human brain is an extremely complex system for processing information. It is organized in a hierarchical structure of distinct but closely connected levels, ranging from genes, proteins and cells to microcircuits, brain regions and the entire brain [1]. Currently, our knowledge about each individual level is extensive. However, what remains unknown is a comprehensive understanding of the causal relationships that enable events at the lowest level of the hierarchy to propagate through the various levels, ultimately giving rise to human cognition and behaviour [1]. Achieving this kind of understanding represents a challenge for neuroscience. By comprehending the intricacies of the brain, one could potentially prevent or find cures for neurological disorders or pave the way for the development of novel computing technologies that mimic the brain's capabilities [1].

This project focuses on the protein level, specifically on the central nervous system (CNS) proteins and their dynamics. Central nervous system proteins are involved in the neurological signal transduction and their flexibility and mobility are essential in this process [2]. To better understand signal transduction in the brain one approach is to study the dynamics of the CNS proteins. It is of high interest to focus on proteins with an impact on the treatment of pathologies related to the CNS for gaining insight about their mechanisms and understanding the way drugs bind to proteins. The simulation of the dynamics of these proteins constitutes the base for understanding higher levels of function.

Studying the dynamics of these molecules has always been a challenge. The atoms that constitute these molecules are in constant motion and their interactions depend on the complex dynamics involved. Obtaining a static snapshot of these molecules is not enough. To truly understand how they function, scientists need to be able to see these biomolecules in action, to perturb them at the atomic level and see how they respond [3]. Imagine a person who is trying to understand how a bicycle works by looking at a static picture. It would be challenging to understand the dynamic interactions between the different parts of the bicycle. Similarly, understanding the behaviour of biomolecules such as proteins requires a dynamic approach that considers the constant motion of the atoms and the interactions between them [3].

## 1.2. Project objectives

The goal of this work is to simulate the molecular dynamics (MD) of a group of interest of central nervous system protein and upload the resulting data in an online platform. To accomplish it, intermediate steps have been required, ranging from the acquisition of the necessary data and resources to the application of specific methodologies. Specifically, these intermediate steps, which are at the same time sub-objectives of this project, are the following:

**Aim 1.** To identify in databases which proteins are expressed in the central nervous system and fulfil some biological requirements as being membrane and Parkinson's disease (PD) targetable proteins.

**Aim 2.** To obtain the molecular dynamics of the identified proteins of interest.

**Aim 3.** To validate the simulations by processing the obtained results and using systematic analyses.

**Aim 4.** To contribute to scientific community by uploading these simulations in an open platform.

The tasks to carry out this project have been distributed around these four goals.

## 1.3. Scope and limitations

The scope of this project includes the obtention of a dataset of proteins of interest and the obtention of their dynamic molecular simulations. However, it excludes the biological conclusions of the resulting data, which in future work they may provide significant findings. It does also not contemplate the mathematical principles of molecular dynamics, but these are used for obtaining the resulting simulation.

Although efforts have been made to address the challenges encountered, from the beginning, the project had limitations. These include:

- Time and duration of the simulations. Molecular simulations require a lot of computational power and time, especially for simulating systems with a large number of atoms, as this project aims to do.
- Precision of the resulting data. Resulting simulations can be influenced by the quality of the models and approximations used as molecular models are simplifications of reality.
- Storage space. Molecular simulation results contain a lot of information and therefore the generated data is very heavy. Computers may not have enough space reserved to store them.

## 1.4. Project justification

This project has been carried out within the Molecular Modeling and Bioinformatics research group of the Institute for Research in Biomedicine (IRB), which actively participates in numerous research projects, including two large-scale European projects: Human Brain Project (HBP) and the Molecular Dynamics Data Bank (MDDDB) project – of which this research group is also the coordinator–.

Human Brain Project is one of the three Flagship projects under the Future and Emerging Technology (FET) program. Literally citing <sup>[1]</sup> *“the goal of the project is summarized as follows: The Human Brain Project should lay the technical foundations for a new model of ICT-based brain research, driving integration between data and knowledge from different disciplines, and catalysing a community effort to achieve a new understanding of the brain, new treatments for brain disease and new brain-like computing technologies”*.

On the other hand, reproducing exactly <sup>[4]</sup> *“MDDDB projects intends to design a European-scale repository of MD simulation (and associated analysis tools), which will: i) optimize computational resources; ii) favour the analysis (and meta-analysis) of trajectories for many different perspectives and fields; iii) guarantee a fast and efficient interchange of information between groups; and iv) facilitate the integration of the MD simulation field into neighbouring communities*.

The present work is part of these two initiatives as it includes the simulation of a specific group of central nervous system proteins, which are useful for the Human Brain Project; and the uploading of these simulations to a platform for the Molecular Dynamics Data Bank project, as these simulations will be presented as a use case.

## 1.5. Research groups involved in this project

This section concisely presents the research groups which have been involved in the development of the project, highlighting their areas of expertise and their contribution to the scope and objectives.

The main research group where everything about this project has been developed, as commented, is the Molecular Modeling and Bioinformatics led by Prof. Modesto Orozco from IRB. Their research revolved around investigating the molecular recognition processes biologically significant <sup>[5]</sup>. They are focused on studying the dynamics of nucleic acids, protein flexibility and interactions, and dynamic properties of macromolecules, as well as tuning the parametrization of molecular dynamics simulations and data mining <sup>[5]</sup>.

Furthermore, the project received assistance in its initial stage –goal one - from the Structural Bioinformatics and Network Biology research group led by Dr. Patrick Aloy from IRB. Their main



scientific research line is in the field of structural bioinformatics, specially studying how cell network and macromolecular complexes operate [6].

Additionally, we have been in contact with the Biggin Group led by Prof. Phil Biggin from University of Oxford (UOXF) to share the project progress and determine technical specifications of the MDDB project. The main focus of this group is to study conformational changes and properties of ligand-binding in receptor proteins, particularly in the brain, by using computational methods to design better therapies against neurological diseases [7].

### 1.6. Trending practices performed during this project

The purpose of this section is to highlight the utilization of current trending practices during the project. The first one is data analysis. Data analysis can be briefly described as the process of working with data to extract useful information [8,9]. This practice was carried out to achieve goal one, as multiple databases were examined and manipulated in order to select which proteins were of interest for this project. The second current trending practice is the use of High-Performance Computing (HPC) in the Barcelona Supercomputing Center (BSC-CNS). HPC is the capacity to quickly process large amounts of data and perform complex calculations [9]. This was used to accomplish goal two, as calculating molecular dynamics simulations involves doing a high number of calculations over a large number of particles and intermolecular forces.

### 1.7. Personal involvement

The realization of this project in the Molecular Modeling and Bioinformatics research group has offered the author the opportunity to travel to the University of Oxford and to participate in the kick-off of the MDDB European project.

The visit to the University of Oxford was made to hold a meeting with Prof. Phil Biggin and Prof. Philip J. Stansfeld, both experts in the molecular dynamics field. At this meeting, the selected proteins of interest were shared and commented; and the next steps for uploading the simulations were discussed as the UOXF also participates in the MDDB project.

Moreover, the kick-off of the MDDB project consisted in establishing common goals and summarizing the different work packages that all the participants should carry out during the following years. This kick off organized by the IRB and led by Prof. Modesto Orozco was attended by renowned people in the field of protein dynamics simulations such as the team leader of the Protein Databank in Europe (*PDBe*) Sameer Velankar, the Professor and *GROMACS* software project leader Erik Lindahl, the director of the Centre Européen de Calcul Atomique et Moléculaire (CECAM) Andrea Cavalli and the Computational Group Manager of Spanish National Bioinformatics Institute (INB) in BSC-CNS Prof. Josep Lluís Gelpí.

## 1.8. Structure and methodology

The present work begins with an introduction to molecular dynamics. Following this, a brief market analysis of the field is presented. The next section, Conceptual Engineering, discusses which proteins are of interest to simulate, and explores the best resources to select them and run their simulations. Next, the detailed Engineering section provides an explanation of how the selection of proteins of interest was carried out, the process of running their dynamics and analyzing the results; and ultimately uploading them to a platform. The last sections of this project correspond to the organization to carry out this project, its technical, economic viability and legal aspects. Finally, conclusions and future work are explained.

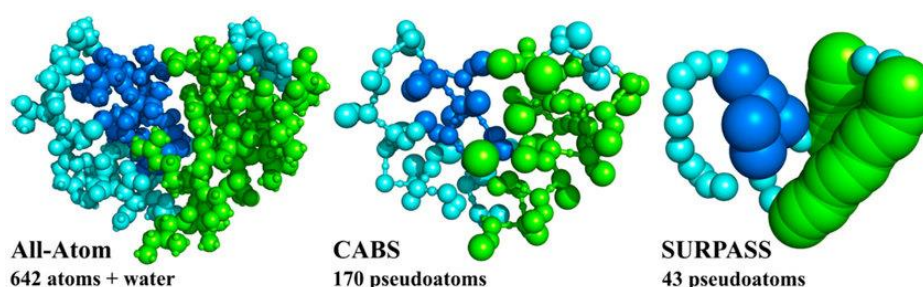
# 2.

## Background

### 2.1. Theoretical background

#### 2.1.1. Molecular dynamics simulations

Molecular dynamics is a computational simulation method that determines how the position of an atom changes over time, enabling one to understand the dynamic behaviour of the system [10]. Since its inception in the late 1970s, MD simulations have made significant progress, evolving from simulating a few hundred atoms to systems with biological relevance such as proteins in solution with explicit representations of the solvent, membrane-embedded proteins and large macromolecular complexes [10]. This is due to the improvement of high-performance computing and the evolution of MD algorithms [10]. MD is widely used in biochemistry and biophysics to facilitate the modeling and comprehension of various phenomena such as protein folding, drug-receptor interactions and conformational changes across different conditions [10]. The system of study can be represented at different levels of detail. Atomistic representation of a system consists of all the constituent atoms that make up the system. On the other hand, coarse-grained models are a simplified model of the atomistic one, as it only considers the most relevant constituents and therefore has lower resolution [11]. **Figure 1** shows the comparison between atomistic models and coarse-grained models of different resolution of a protein structure.



**Figure 1.** Different protein structure representations: one atomistic model (left) and two coarse-grained models (center and right) of different resolution [12].

The core of classical MD simulation relies on the classical equation of motion, specifically on the numerical integration of Newton's equation of motion. This results in the description of the system

as it calculates the positions and velocities of the atoms from the force acting on each one of them [3]. These forces are obtained by using complex equations called force fields [10].

### 2.1.2. Force fields

Force fields are mathematical models that describe the interactions between atoms and molecules in a system. These models use equations and parameters to calculate the forces and energies associated with the interactions between particles. In a typical force field, various terms are included to account for bonded interactions, such as bond stretching (2-body), bond angle (3-body), and dihedral angle (4-body) [13], which are approximated by spring-like terms [3]; and non-bonded interactions which include electrostatic and Van Der Waals interactions.

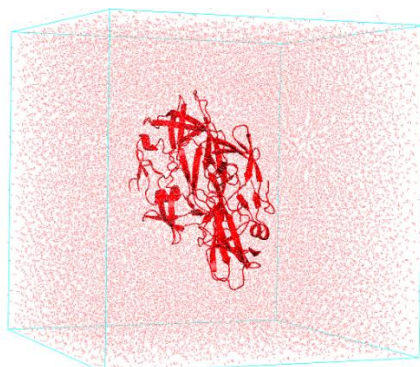
There are different force fields as there is no unique way of defining mathematical functions which describe the potential energy of the system [14]. This means that the interactions between atoms and molecules cannot be described precisely so different force fields use different approximations. Popular force fields are CHARMM, AMBER, GROMOS, OPLS, COMPASS and ParmBSC0/ParmBSC1 [15].

### 2.1.3. Water systems

Some biomolecules such as proteins are found in an aqueous environment together with ions in cell environment. The interaction between these biomolecules and water plays a crucial role in shaping their thermodynamics and conformational properties. Realistic water environment is essential for precise simulation of proteins and other biological molecules as it tries to mimic the environment where processes happen.

Water can be treated as a continuous medium or explicit individual molecule [16]. Systems with a continuum water model require less computational resources but key interaction characteristics are lost. On the other hand, although an explicit water model is more accurate, it demands a high computational cost for simulating as over 80% of the particles in the simulation will be water molecules and therefore there will be numerous water-water interactions [17].

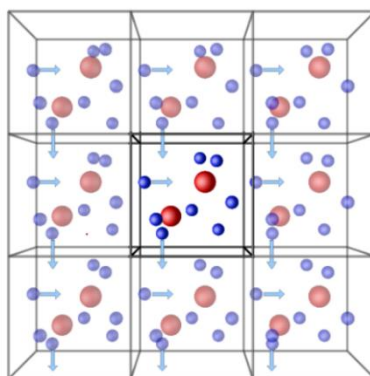
The system is usually set in a virtual box called simulation box which defines the spatial boundaries within which the molecular system is contained. In this simulation box the system of interest is found embedded with water molecules if it is a water system [18], as it is shown in **Figure 2**, where a protein embedded with water molecules inside a simulation box can be seen.



**Figure 2.** Protein embedded in a water box <sup>[18]</sup>.

#### 2.1.4. Periodic boundary conditions

Periodic boundary conditions (PBC) are a fundamental concept in computational simulations. These conditions make possible that the limits of the simulation box where the system is are not fixed <sup>[19]</sup>. This means that it tries to mimic an infinite system. For doing this, the simulation box is surrounded by itself in the three dimensions of space so as to have a continuous system (**Figure 3**).



**Figure 3.** Diagram of PBC that show unit cell surrounded by its replicate <sup>[16]</sup>

When one molecule diffuses across the boundary of the simulation box it reappears on the opposite side. In other words, each molecule always interacts with its neighbours despite being on opposite sides of the simulation box. All this is achieved with PBC parameters which also preserve thermodynamic properties of the simulated unit cell <sup>[20]</sup>.

Without using this method, the system will interact with vacuum; therefore, PBC method substitutes the surface anomalies originated from the interaction between the isolated system and vacuum with periodicity effect, which are generally much less severe <sup>[16,20]</sup>.

### 2.1.5. Molecular dynamics algorithms

The core of MD simulation relies on Newton's second law, also known as the equation of motion. This principle asserts that the movement of individual atoms within the system can be determined by the forces acting upon them. By integrating these equations of motion, a trajectory is generated, which shows the variations in position, velocity and acceleration of each particle as time progresses<sup>[16]</sup>. The resulting trajectory essentially is a movie of structural conformations the protein adopts over time <sup>[3]</sup>.

To ensure the trajectory stability, during numerical integration of movement, a short time step must be used <sup>[3]</sup>. The workflow of the molecular dynamics algorithms is as follows:

- Initialization. The initial parameters of the system are defined: configuration of particles (initial positions and velocities), interactions between particles, temperature, etc.
- Time Step. A discrete time increment, known as step, is established.
- Force Calculation. Using the current positions of the particles, the forces acting on each particle are calculated using force-field equations.
- Integration of Equations of Motion. Numerical methods are employed to integrate the equations of motion for the particles. These equations describe how forces affect the acceleration, velocity, and position of each particle.
- Update of Positions and Velocities. Using the results of the numerical integration, the positions and velocities of the particles are updated for the next time step.
- Repetition of Steps 3-5. Steps 3-5 are repeated in a loop to advance in time and simulate the system's evolution. In each iteration, forces are calculated, equations of motion are integrated, and positions and velocities are updated.

Before the initialization of the system, the simulation box with the system embedded with water must be designed (**Figure 2**). Moreover, when utilizing periodic boundary conditions, the simulation box interacts with numerous replicated images of itself. Consequently, if the simulation system carries a charge, the electrostatic energy would theoretically approach infinity <sup>[21]</sup>. To address this problem, it becomes necessary to introduce counter-ions that neutralize the system. After this step, it is also common to do an equilibration process for bringing the system to equilibrium, which is assessed by having stable values of various parameters <sup>[18,22]</sup>. This can be an Energy Minimization (EM) and a Pressure Relaxation (PR). The first one enables the particles to have a minimum energy configuration, while the second one aims to have a uniform pressure distribution in the system.

### 2.1.6. Type of files in molecular dynamics field

Throughout this project, terms such as trajectory, topology and structure files will appear. This section aims to describe briefly what information is contained in each one of them.

When running MD simulations, the obtained dynamics is stored in the trajectory file. This file can store the movement of the system in function of time. On the other hand, files of structure or topology are needed to calculate this trajectory. Structure files describe the position of each atom in the system while topology files contain the same information as structure ones but also the bonds between atoms and the charges of the atoms, among other information.

## 2.2. Applications of molecular simulations in proteins field

For the investigation of molecular properties and drug discovery, molecular dynamics is very helpful [23]. Simulations have also the ability to generate hypotheses that can guide new experimental investigations.

Assessing the mobility or flexibility of a system is perhaps the most straightforward application for a better understanding of the function of the system or the molecule such as the protein [3]. By using MD simulation, biomolecular processes such as ligand binding, protein folding or membrane transport can be seen in motion to better understand their mechanism and determine the basis of events that are challenging to address experimentally.

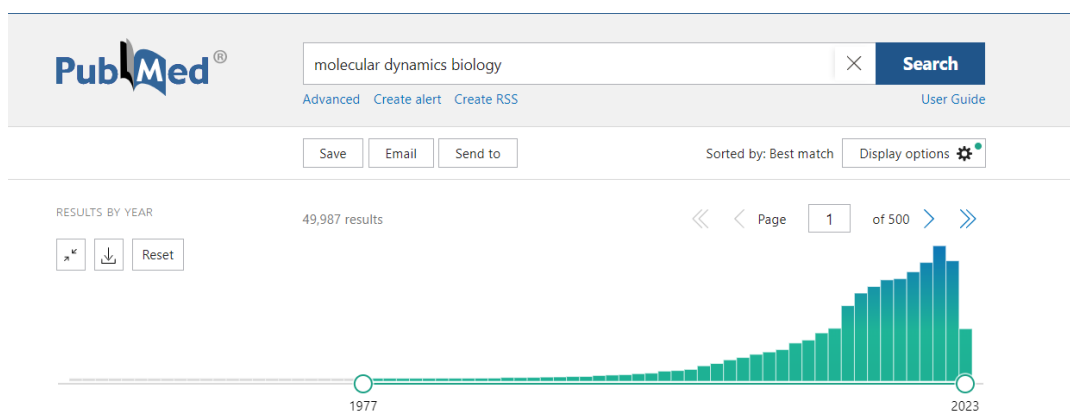
Moreover, a relevant application [3] of MD simulation is predicting how a biomolecular system will react to a perturbation by comparing the modified system with the unmodified one. Examples of these perturbations are mutations of some amino acid residues in a protein and its associated functional effect, changing the environment of a protein system such as the composition of lipids in a membrane or phosphorylate an amino acid, among others.

MD simulations have also proven to be useful to refine structural models [3] which have been obtained imprecisely by using experimental techniques such as X-ray crystallography. A more accurate structure configuration is obtained by guiding its dynamics and interactions.

Finally, and one of the most important and emerging applications of MD is drug discovery and its design-related processes [23]. Contemporary approaches to drug discovery often begin by identifying and validating a biologically significant target that can be modulated with drug molecules. A targetable molecule is defined as a molecule which its function is controlled by compounds such as a therapeutic drug [24]. These drug targets typically involve proteins such as receptors or enzymes although DNA and RNA molecules are also known to be targets. Protein flexibility plays a crucial role in determining the compatibility between the drug molecule and the protein's binding site as slight alterations of its conformation can impact how a drug binds. Therefore, MD simulations provide valuable insights into the target's dynamic behaviour with regard to drug design and they can be used to experimentally test these predictions of drug interactions.

### 2.3. State of the art of molecular simulation

Molecular dynamics simulations are not a novel concept [3]. In 1950s, the first MD simulations was performed about simple gasses [25]. When this first computer simulation was performed, it took some time for the scientific community to accept simulations as a tool for studying a system [25]. However, in 1971, a molecular dynamic study of a liquid water model provided significant data on its structure and diffusion, highlighting the potential of simulations to contribute to scientific progress. It was not until 1977 when the first MD of a protein was performed. These two achievements were recognized by the Nobel Prize in Chemistry [3]. During the 1980s and 1990s growing computer power as well as better optimization of software and algorithms made molecular simulations more common and widely used [25], especially in the field of biology [3]. The increasing number of publications in molecular dynamics reflects this trend as it is shown in **Figure 4**.



**Figure 4.** Publication trend evolution for molecular dynamics simulation from 1977 to 2023 with a peak of 5,470 publications in 2021 [26].

This trend is particular noticeable in the field of neuroscience [3]. The growing focus on MD simulations on this topic can be attributed to at least two underlying factors [3]. On one hand, there has been a breakthrough in determining experimental structures of molecules critical in neuroscience and medication targets – ion channels, transporters, neurotransmitters ... –, which are mainly membrane proteins. These structures serve as a basis for MD simulations. On the other hand, MD simulations have become more accessible due to the use of much more powerful computing resources as they require supercomputers to be run faster. Simulations have been used to study [3] proteins essential for neuronal signalling, to aid in the development of drugs that target the nervous system and to identify mechanisms underlying protein aggregation linked to neurodegenerative diseases.

Nowadays, molecular simulation has established itself as a valuable scientific tool, helpful in result interpretation, suggesting new experiments and even predicting outcomes [25,27].



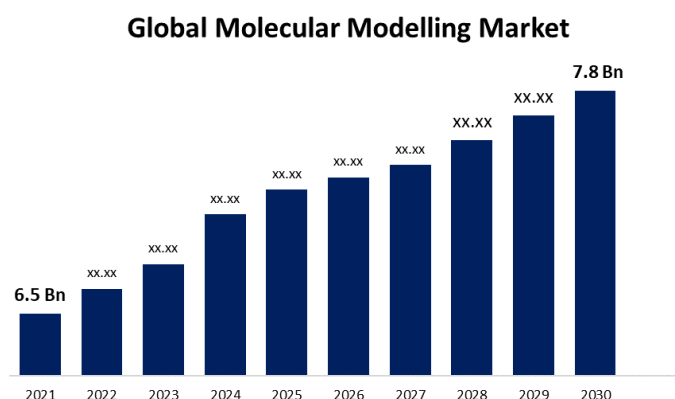
# 3.

## Market analysis

### 3.1. Current market for molecular simulations

#### 3.1.1. Trends and prospects

The molecular modeling market is projected to experience a growth rate of 14.43 % in the period of 2021 to 2028, reaching a market value of USD 7,8 billion by 2030 (**Figure 5**) [28]. Given the historical evolution and the rise of publications about the topic over the years, this prospect is not surprising. In fact, the increasing incidences of diseases such as cancer, cardiovascular and infections have led pharmaceutical and biotechnology companies make significant investments in the molecular modeling market [28]. These companies are the main drivers behind the growth of this market as they have recognised the need to incorporate structural biology and modeling techniques to develop a new generation of drugs. Furthermore, technological advancement will popularize molecular dynamics [28]



**Figure 5.** Molecular modeling market evolution until 2030, reaching a market value of 7,8 Bn [28].

#### 3.1.2. Main competitors and collaborators

Several research groups have made remarkable advances in MD simulations in various fields and, in particular, within the field of neuroscience, as mentioned before. Although this may be seen as a competition to see who can achieve the greatest number of simulations in the shortest time, it is not appropriate to do so. In this field, collaboration between different entities working in molecular dynamics is fundamental. These dynamics calculated by different groups can be uploaded all in

one online platform so collaboration and sharing of the data are ensured. This is what the MDDB project explained in the *Introduction* section intends to do.

However, on the other hand, from the perspective of creating a database, it becomes interesting to examine the competitiveness that there may be as two platforms which offer the same can be detrimental to the unification of information.

MDDB tries to unify efforts for the first time in a Europe-wide project, therefore it was not found another initiative with the same purpose. However, GPCR MD <sup>[29]</sup> is a database which stores molecular dynamics of G-protein-coupled receptors (GPCRs) and COVID-19 proteins. Their infrastructure may be similar to the one MDDB aims to create, but the scope of both databases is not comparable and therefore it is not considered as a competitor.

## 3.2. Potential of molecular simulations in the scientific market

### 3.2.1. Areas of opportunity

Several key components are found in the molecular modeling market, all of which are related to trends and prospects explained above. The main ones are increasing investments in drug research and development by biotechnology and pharmaceutical companies, technological advances in drug design and the growing prevalence of chronic diseases requiring new treatments. Therefore, the main areas of opportunities are found in the biological field <sup>[28]</sup>.

### 3.2.2. Challenges and barriers

The main challenges in using and developing MD simulations are the high cost of computing resources and a lack of skilled professionals in the field <sup>[28]</sup>. This is believed to hold back the growth of the market.

# 4.

## Conceptual Engineering

Before performing MD simulations, some choices must be taken. As this project is part of the Human Brain Project, central nervous system proteins will be simulated. However, it has to be decided a limited subset to focus on due to time limitations of this project. Additionally, other choices involve how to set up the molecular system for running their dynamics, as well as which computing tools and force fields will be employed.

### 4.1. Selection of the molecular system to study

Simulations are essential for bridging the gap between experimental and theoretical techniques in scientific studies <sup>[30]</sup>. This is especially true when studying dynamic and complex systems like the brain, which differ significantly between individuals and has a multi-scale design. Perception, cognition, learning, and memory are ultimately reliant on intricate chemical mechanisms. Simulating phenomena that take place on a variety of spatial and temporal scales is necessary for modeling these processes.

This project focuses on simulating central nervous system proteins as understanding their dynamics and flexibility can provide insights for better understanding their function in important neurological pathways or processes.

However, this project does not simulate all CNS proteins due to the limited time of execution, but those which satisfy some structure and functionality conditions. As commented in the *Background* section, two of the main applications of molecular dynamics are 1) understanding the function of proteins and therefore processes in which they are involved and 2) design / discover new drugs. To be centred in proteins with relevance in these two aspects, the following conditions – or filters – were chosen:

- Being membrane proteins.
- Being targetable proteins.

Moreover, to further delimit the subset these structure conditions were also considered:

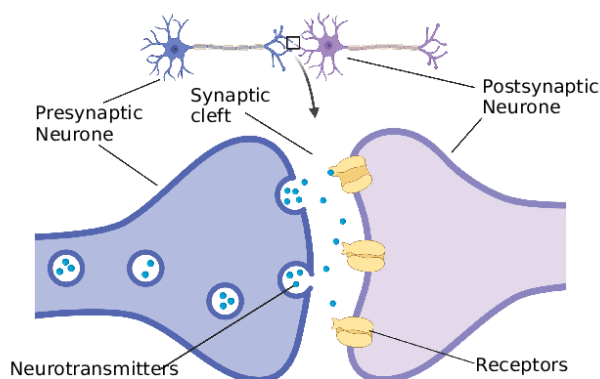
- Having a minimum amino acid (aa) sequence length.
- Having a structure experimentally determined.

A detailed explanation of this selection criteria is found in the next subsection, and it is summarized in **Table 1**.

#### 4.1.1. Establishment of the selection criteria

This section establishes the selection criteria for the choice of proteins to be included in the molecular simulations of the central nervous system. To make this selection, several relevant aspects have been considered, such as their presence in the plasma membrane, their length, their potential as therapeutic targets and their structure.

Regarding their presence in the plasma membrane, it has been decided to opt for membrane CNS proteins due to their indispensable role in the structure and function of cells, and therefore synapse communication process for transmitting information<sup>[30]</sup> as they act as molecular transporters, signal receptors and ion channels<sup>[31]</sup> (**Figure 6**).



**Figure 6.** Scheme of how proteins are involved in human brain signal transmission<sup>[32]</sup>.

Based on how they associate with the membrane, membrane proteins can be transmembrane – cross the membrane layer at least once – or monotopic/peripheral – they are attached to a single side of the bilayer --. This project will focus on those that are transmembrane. Understanding the properties of membrane-associated and membrane-traversing sections of membrane proteins may help to better comprehend their stability and activities, which also can be useful to design new drugs as many membrane proteins can be targetable<sup>[31]</sup>.

Furthermore, it has been decided to select only those proteins that are drug targetable for small molecules and Parkinson's disease. According to available data, there are approximately 3,000 genes that encode proteins which can be targeted by drugs. However, only a small fraction (10%) of these proteins has an approved drug by the Food and Drug Administration (FDA)<sup>[33]</sup>.

For nearly a century, small-molecule drugs have been the mainstay of the pharmaceutical industry<sup>[34]</sup>. These drugs are organic compounds with low molecular weight and offer unique benefits as

therapeutic agents. They can be administered orally and can readily cross cell membranes to reach intracellular targets. In addition, they can be designed to interact with biological targets through various mechanisms [34]. For these reasons, it has been decided to work with small molecular compound targets.

In addition, it has been chosen to work also with targets which are specific to Parkinson's disease. The study of Parkinson's disease is relevant due to the high prevalence and burden that this disease represents worldwide. Parkinson's is the second most common neurodegenerative disease after Alzheimer's [35]. As the world's population ages, the number of Parkinson's cases is expected to increase significantly, which in turn will have an impact on the quality of life of patients, their caregivers, and communities at large [36]. Research in this area is crucial to develop more effective treatments and improving patients' quality of life suffering from this disease. It is also important to study Parkinson's to better understand its aetiology and the underlying mechanisms that cause it, which may lead to new therapeutic approaches and better prevention.

Regarding the length of proteins, this study has taken into account those that may be of interest at a dynamic level. Although this project focuses on proteins which have a transmembrane domain, it is also interesting that they have an extramembrane domain too so that surface interactions can also be studied. Short sequences do not ensure this surface interactions. By visual inspection of the structure, it has been decided that those proteins with a length of less than 100 aa, with exceptions, are not interesting in this aspect.

Another condition to satisfy is that proteins of study must have a determined experimental structure. This project does not contemplate the study of proteins with a predicted structure. Due to the limited time of execution, a not very large dataset is studied and therefore when deciding, it is better to choose experimental structures – which are known to be real – rather than predictions, however accurate they may be nowadays.

**Table 1.** Summary of the filters applied in order to select those CNS proteins of interest.

<b>Filter</b>	<b>Detail</b>	<b>Justification</b>
<i>Membrane</i>	Transmembrane	Membrane CNS proteins are essential components of cell structure and function. They are a crucial part of the synapse communication process (electrical and chemical) [37]. Thus, they are interesting targets for the design of new drugs.
<i>Target</i>	Surely or potentially drug targetable for small molecules and Parkinson's Disease	Small molecule drugs are the pillar of pharmaceuticals. Moreover, PD is a prevalent disease which requires more investigation.

<i>Minimum aa sequence length</i>	Protein must have a length above 100 aa	By visual inspection, it was determined that those whose length is shorter are not that much interesting in terms of molecular dynamics as there are fewer surface interactions.
<i>Structure</i>	Experimental	This project does not contemplate predicted structures.

#### 4.2. Selection of databases and repositories to obtain molecular systems to study

The aim of this section is to select the appropriate resources to identify those CNS proteins that meet the selection criteria established in the previous section. For this, it is necessary to consult and use several databases and repositories that contain different types of information about proteins. Several databases will be employed to verify the information and establish a consensus among them. These databases will be chosen depending on the condition that it is wanted to satisfy.

The article *Protein Databases on the Internet* <sup>[38]</sup> provides a starting point for exploring the potential of protein databases found on the Internet. [UniProt](#) is the most commonly used database for sequences, while the [Protein Data Bank](#) (*PDB*) is the most widely used resource for structural information, according to the information provided in this article. Due to their relevance in the field, these two databases will be used to validate the information obtained in the other databases.

##### **UniProt**

*UniProt*, according literally from its documentation, is a comprehensive resource for protein sequence and annotation data <sup>[39]</sup>. It is originated from an initiative of the *UniProt* Consortium groups, composed of the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). *UniProt* provides detailed and up-to-date information on proteins including amino acid sequences, three-dimensional structures (if available), functional annotations, gene expression information and more, from a wide variety of organisms.

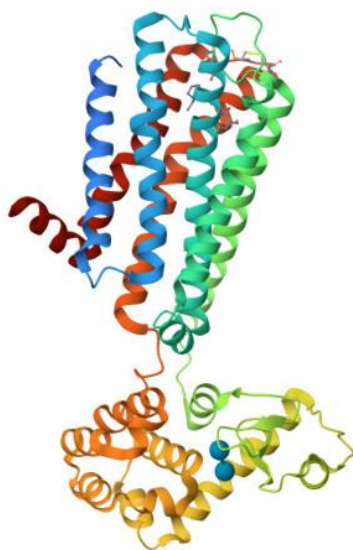
The *UniProt* database is organized into three different database layers <sup>[40]</sup>. The main layer and the one that will be used in this project is *UniProt Knowledgebase* (*UniProtKB*). It contains protein sequences and annotations obtained from scientific literature and protein sequencing resources.

*UniProtKB* is composed of two sections: *UniProtKB/Swiss-Prot* and *UniProtKB/TrEMBL*. *UniProtKB/TrEMBL* (unreviewed) contains protein sequences that come with computationally generated annotations and extensive functional characterization. On the other hand,

*UniProtKB/Swiss-Prot* (reviewed) [41] is a superior protein sequence database characterized by meticulous manual annotation, ensuring high quality and non-redundancy. *UniProtKB/Swiss-Prot* will be used to ensure high quality entries. It is noteworthy to mention that this particular database demonstrates minimal redundancy by consolidating all protein sequences encoded by a single gene into a unified *UniProtKB/Swiss-Prot* entry. Each protein entry is unequivocally identified by a permanent and unique identifier (primary key) and, therefore this identifier is useful for connecting information between different databases.

### Protein Data Bank

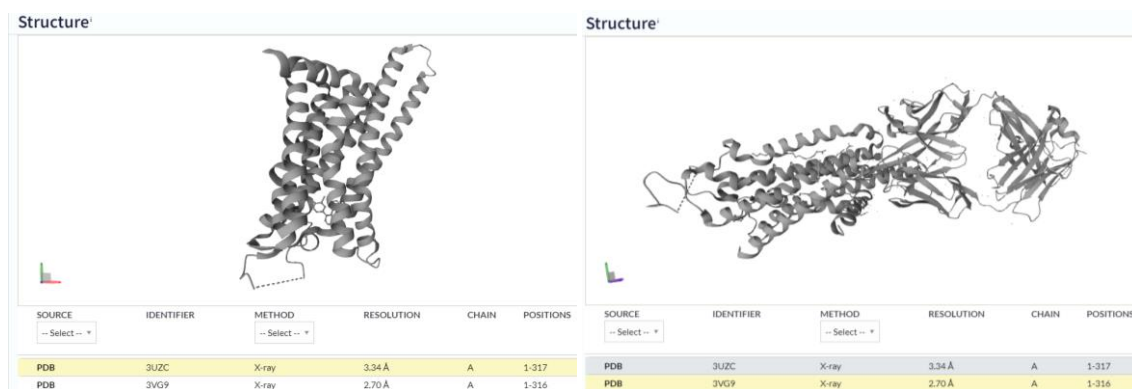
The *PDB* [42] is a repository that stores information on the three-dimensional structures of proteins (**Figure 7**) and other biological macromolecules, such as nucleic acids and protein complexes. The *PDB* collects experimental data on structures solved using techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), cryo-electron microscopy and other methods. These three-dimensional structures provide detailed information about the spatial arrangement of the atoms that make up a molecule, which is critical to understand its biological function, its interaction with other molecules and to run the protein simulation.



**Figure 7.** Structure of the human dopamine D3 receptor in complex with eticlopride (PDB ID: 3PBL, UniProt ID: P35462) [43].

Each structure and all its data in the *PDB* are identified by a unique *PDB* identifier which consists in a 4-character alphanumeric ID. Regarding this, there is a close relationship between *UniProt* and *PDB*. A particular *UniProt* code may have several *PDB* identifiers associated with it (**Figure 8**). This is because a protein can have various three-dimensional structures determined through different experimental conditions, post-translational modifications, protein isoforms, and genetic variations. Additionally, proteins may have distinct structures in their active and inactive forms, or when bound to different ligands or cofactors [44]. Moreover, another factor contributing to multiple *PDB* structures

is the experimental determination of a small portion of the protein each time. This results in various *PDBs* from the same protein but from different parts of it. This may cause redundancies because some *PDBs* will include parts of others *PDBs*. Knowing this information is important when deciding which structures one wants to simulate. This will be taken into account in the *Detailed Engineering* section.



**Figure 8.** For protein Adenosine receptor A2a (UniProt ID P29274), several experimentally determined structures are found, each corresponding to one *PDB* ID. Both displayed structures represent the same part of the protein (chain A, position 1-316), although with different conformations [45].

#### 4.2.1. Select central nervous system proteins

To filter by expression in components of the human body, specifically by central nervous system, a search was performed in the databases that *UniProt* uses to indicate expression and two of them are considered in this discussion section as at a glance it is easier to access to their data:

- [ExpressionAtlas](#), *Differential and Baseline Expression*
- [Bgee dataBase](#) for *Gene Expression Evolution*

*ExpressionAtlas* is an EMBL's European Bioinformatics Institute (EMBL-EBI's) database for gene and protein expression. It gathers information on the quantity and distribution of RNA and proteins. The website allows for gene searches within or between species to reveal tissues and cell types where a gene is expressed [46]. Taking a first look at the website, genes can be searched by species and biological conditions. Genes of interest can be filtered by selecting Homo Sapiens and central nervous system.

On the other hand, *Bgee*, taking the definition written on its website, is defined as a multi-species animal gene expression comparison and retrieval database that indicates where a gene is expressed [47]. *Bgee* provides four databases with different information on each of the 52 species it includes. The "Gene expression call" one provides information on which genes are expressed in each anatomical entity (present/absent expression calls) or which are expressed according to a combination of different parameters such as anatomical entity, developmental and life stage, sex and strain or ethnicity [47]. This database makes use of the EMBL-EBI Ontology definition [48] to



describe what is considered an anatomical entity. Thus, *Bgee* considers that an anatomical unit is a biological element that can be an individual member belonging to a given species or its structural organization of that organism. It is to be expected from this definition that the central nervous system is considered as an anatomical entity.

It should be noted that an ontology is a way of relating different concepts and categories that represent the same subject. It provides an ID for each entry which not only identifies uniquely the entry but its relationships with other entries.

Considering all this information, *Gene expression call* dataset of *Homo Sapiens* is a potent candidate to be used in this project.

One advantage of this project is the possibility of being in direct contact with the research groups mentioned in the *Introduction* section. A brief meeting with the research group of Structural Bioinformatics and Network Biology in IRB was done and it was recommended to use [Human Protein Atlas](#) (HPA) and [TISSUES](#).

*Human Protein Atlas* is a resource that aims to map all human proteins in cells, tissues and organs by using omics technologies [49]. *HPA* classifies proteins into 12 possible sections: Tissue, Brain, Single cell Type, Tissue Cell type, Pathology, Disease, Immune cell, Blood protein, Subcellular, Cell line, Structure and Metabolic; although a free search for a particular gene or protein can also be performed.

*TISSUES* [50] is a database that classifies genes according in which tissue they are expressed. It integrates data collected from several sources and it is publicly accessible through a web interface.

### **Selection**

On one hand, *Bgee dataBase* and *ExpressionAtlas* can both be of use as they allow the user to filter by component. However, not all entries in *ExpressionAtlas* are classified by organism part, making it difficult to have a good approach of which genes are expressed in the CNS. Therefore, *Bgee* was selected; specifically, the *Homo Sapiens - Gene Expression Call - Simple File database*, which enables to identify proteins expressed in the CNS and it summarizes information of the classification for human.

On the other hand, regarding the recommended databases, both of them, *HPA* and *TISSUES*, were selected as they are validated by experts in the field.

It should be noted that, *HPA* is also used by *UniProt* as a cross-reference for classifying organisms, confirming its robustness. Its brain-specific section makes it useful for the classification of CNS proteins. For this work, the section of interest is of *HPA* database is *Brain*. This contains the distribution and expression of proteins in the different areas of the mammalian brain.

Moreover, regarding *TISSUES* database, its documentation is published in the *Database* journal in Oxford Academic, *Volume 2018 (2018)*. Although it has a small number of citations (33), it is considered a robust database because of how data has been treated and *analysed*, as it is explained in its documentation [50].

As a summary, three databases were chosen in order to select and collect proteins expressed in central nervous system:

- *Bgee* database.
- *Human Protein Atlas*.
- *TISSUES*.

#### 4.2.2. Selection of resources to filter and reduce dataset

##### *Select membrane proteins*

To select membrane proteins, it was decided to initially use *UniProt* to do a first filter as this information is already contained in *UniProt*. To specifically classify those that are transmembrane, the external resources used by *PDB* to identify membrane proteins were examined (**Table 2**).

**Table 2.** External resources used by *PDB* to identify membrane proteins.

Resources used by <i>PDB</i>	Definition	Number of citations (by Scopus)
<a href="#">Orientations of Proteins in Membrane database</a> ( <i>OPM</i> )	Database that assesses the location of the lipid bilayer by a transfer energy function.	899
<a href="#">Protein Data Bank of Trans-membrane Proteins</a> ( <i>PDBTM</i> )	Updated database that makes a trans-membrane protein selection of the <i>PDBs</i> .	194
<a href="#">MemProtMD</a>	Database that uses an automatic annotation pipeline to determine $\alpha$ -helical and $\beta$ -barrel domains and then uses molecular dynamics to establish the protein-lipid interactions.	196
<a href="#">Membrane Proteins of Known Structures</a> ( <i>mpstruc</i> )	Curated database of membrane proteins of known 3D structure.	221

##### **Selection**

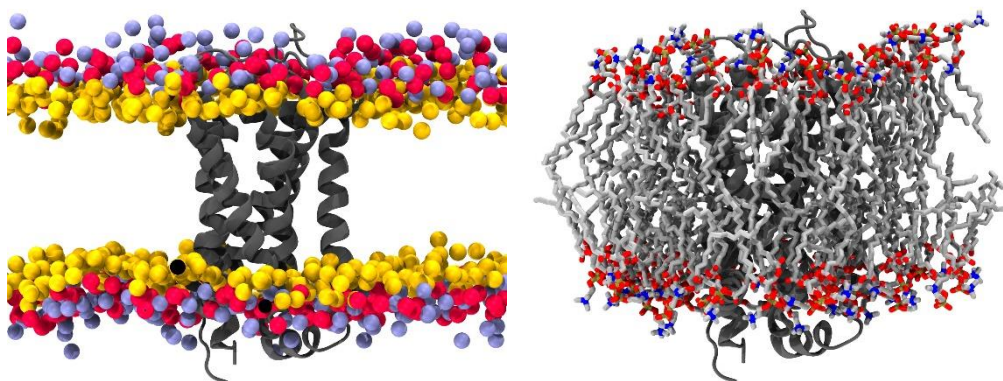
All the previous databases aim to classify membrane proteins, although they differ from each other in proteins they record as membrane ones [51]. *OPM* contains the greatest number of membrane proteins, even though the majority are peripheral membrane proteins. The *mpstruc* database has

a lot of transmembrane proteins, in particular those which have a beta barrel domain. Moreover, *PDBTM* and *MemProtMD* have also a large number of transmembrane proteins detected.

As this project is focused on selecting membrane proteins with a transmembrane domain, *OPM* is discarded as it does not give as much information as the other databases. Additionally, it is wanted a database with a robust tool for identifying these transmembrane domains. This is offered by the *MemProtMD* database, a resource from the Department of Biochemistry of the University of Oxford.

*MemProtMD*<sup>[52]</sup> presents an automatic pipeline to detect  $\alpha$ -helical and  $\beta$ -barrel which also establishes the protein-lipid interactions. This project had the opportunity of being in contact with the research group which developed this tool to be more informed of how it works.

The automatic pipeline of *MemProtMD* not only detects membrane proteins but it reinserts the protein into their membrane environment. Firstly, it detects  $\alpha$ -helical and  $\beta$ -barrel domains using algorithms (Octopus) and structure analysis to detect hydrophobicity, accessibility to the membrane and protein length. Next, to embed the protein in a solvated bilayer, coarse-grained simulations are used. This results in well-defined interactions between the different components of the system: protein, bilayer, and solvent. Moreover, from this coarse-grained system, an atomistic system is derived. **Figure 9** shows an image of the coarse-grained and atomistic models obtained from this pipeline.



**Figure 9.** Coarse-grained (left) and atomistic models (right) of a membrane protein obtained from *MemProtMD* workflow<sup>[52]</sup>.

*MemProtMD* makes available the resulting files that contain information about the positions of the atoms, the types of atoms present, the interactions between them and other relevant features to automatically run the simulations, of both atomistic and coarse-grained model of each protein. The use of atomistic models in this molecular simulation project has been chosen, but it is important to keep in mind that creating these systems from scratch can be a time-consuming and labor-intensive process. This involves determining the appropriate parameters for each type of molecule, the

inclusion of solvents such as water, and the precise setting of initial conditions; as well as building missing atoms to the structure. Since the generation of complete atomistic systems can be laborious, it was chosen to use atomistic systems that have already been prepared and are available for use such as the ones present in *MemProtMD*.

Regarding membrane information again, although *MemProtMD* was chosen to be used in order to select transmembrane proteins, the *UniProt* database was chosen to previously be used to initially filter out those proteins without any membrane interaction and work with a little dataset. The one used of *UniProt* to filter out was the one with the following values *Status: Reviewed*, *Popular organisms: Human*, *Protein existence: Protein level* and *Subcellular location: Intramembrane and Transmembrane* – although Intramembrane were Transmembrane proteins -.

### Select targetable proteins

In order to identify which proteins are targets, we have searched which databases or repositories accessible through the Internet contain this information. The following articles were found in the literature: *Drug–target interaction prediction: databases, web servers and computational models* [53] and *A Review of Target Identification Strategies for Drug Discovery: from Database to Machine-Based Methods* [54], both summarizing databases involved in drug-target identification. From these databases described in the article, three of them were studied as they had the largest number of citations by Scopus (**Table 3**).

Moreover, two other databases were studied as possible candidates to be used for selecting pharmacological target proteins: *Open Targets* and *Pharos* database (**Table 3**). As this project had the opportunity of being in contact with different research groups, these two databases were recommended by the Structural Bioinformatics and Network Biology research group; in particular by Dr. Adrià Fernández, a PhD specialized in drug discovery.

**Table 3.** Databases considered as possible databases to select targetable proteins.

<b>Database</b>	<b>Brief definition</b>	<b>Number of citations (by Scopus)</b>
<a href="#">DrugBank</a>	It combines comprehensive target information (such as sequencing, structure, and pharmaceutical data) with detailed drug data (such as chemical, pharmacological, and pharmaceutical data) [54]. The database is regularly updated.	3,675
<a href="#">ChEMBL</a>	In the disciplines of drug discovery and medicinal chemistry research, ChEMBL is currently a well-known resource. The ChEMBL database collects and archives data on bioactivity, molecules, targets, and medications that have been standardised and taken from a variety of sources [55].	2,526
<a href="#">BindingDB</a>	The BindingDB is a binding database that contains small molecule ligands and protein targets, along with experimentally verified protein-ligand binding	1,291

	affinities.	
<a href="#">Open Targets</a>	Free accessible informatic platform that identifies targets using data publicly available <sup>[56]</sup> .	254
<a href="#">Pharos</a>	Online platform that presents information derived from the Target Central Resource Database (TCRD) which classifies molecules depending on their Target Development Level (TDL) <sup>[33]</sup>	149

### **Selection**

From the several databases displayed in the above table, it was chosen to work with *Open Targets* and *Pharos* platforms as they were highly recommended by experts in the field.

*Open Targets Platform* serves as a knowledge repository that can be used to describe targets, diseases, and drugs, in the context of drug discovery, while also highlighting the relationships between these entities, with a particular emphasis on the associations between targets and diseases <sup>[56]</sup>. Additionally, the platform is equipped with an in-house scoring system that assesses the likelihood of a particular target being linked to a specific disease, and the resulting evidence is integrated from various sources to generate a list of ranked gene-disease associations.

On the other hand, *Pharos* is an online platform that presents information derived from the *Target Central Resource Database*, which is one of the components of the Illuminating the Druggable Genome (IDG) program, an initiative launched by the National Institutes of Health (NIH) to identify proteins that can potentially be influenced by small molecules or biologics <sup>[57,58]</sup>. *Pharos* classifies each target by its Target Development Level. According to publications, tool compounds and other characteristics, TDL describes how much the targets are or are not explored. Those targets that have not been studied yet as possible targets are labelled as *Tdark*, those that have approved drugs as *Tclin*, those that have small molecule activity in *ChEMBL* as *Tchem* and finally those that are not associated with small molecule or drug activities as *Tbio* <sup>[57,58]</sup>.

Both platforms make use of multiple other reliable sources, which ensure the accuracy and reliability of the data, as any inconsistencies or errors are more likely to be identified and corrected. Although both platforms chosen have fewer citations than the ones commented in **Table 3**, both of them have been published in the *Nucleic Acids Research* magazine, a reputable scientific journal in the field of molecular biology and genetics which ensures their consistency.

#### *Select amino acid sequence length*

To consider the length of the proteins, *UniProt* was used as it displays the canonical length sequence of amino acids for each protein and as this database was previously validated.

### Select experimental structure

Moreover, *UniProt* is a database well mapped with *PDB* database and therefore it was used in order to know which proteins have a *PDB* structure. As mentioned in the previous *section 1.2*, it is possible for each *UniProt* entry to have multiple *PDB* identifiers, corresponding to different the experimental structures determined for that particular protein. This will have to be taken into account as we are interested in running unique amino acid sequences. This is because calculating simulations takes a lot of time and we have to ensure that this time is well spent in studying different proteins.

## 4.3. Software resources

### 4.3.1. Simulation software

Nowadays, there are numerous software packages available with diverse functionalities to choose in the field of molecular dynamics. While most of these packages try to cover a broad range of capabilities, each of them possesses distinctive features or advantages. A significant proportion of the popular MD packages can use force field, structure, and trajectory file formats that were initially introduced in other packages. This allows for validation and facilitates replication of published results, even in the absence of the original software <sup>[59]</sup>. In this project, one will be chosen to run the simulations.

With the purpose of deciding which software to use, common softwares <sup>[59,60]</sup> were studied (**Table 4**).

**Table 4.** Common software in MD field.

Software	Definition
<a href="#">GROMACS</a>	This modeling package is principally designed to carry out molecular dynamics simulations on biochemical molecules, such as proteins and lipids <sup>[59]</sup> . Moreover, it includes essential dynamics analysis and numerous utilities for analysing trajectories <sup>[61]</sup> .
<a href="#">AMBER</a>	It is one of the most widely used simulation software package <sup>[62]</sup> . <i>AMBER</i> consists of a variety of applications that cooperate to set up, run and <i>analyse</i> MD simulations. It also includes classical molecular mechanics force fields that were essentially designed for the simulation of biomolecules, such as amino acid, phospholipids, nucleic acids and carbohydrates, among others <sup>[62,63]</sup> .
<a href="#">CHARMM</a>	This software targets biological systems such as those found in solution, crystals and membrane settings <sup>[63]</sup> . These include peptides, proteins, small molecule ligands, lipids, nucleic acids, and others. It also has applications for inorganic materials. Moreover, it has a complete set of tools for analysing the simulations and can achieve high-performance.

**NAMD**

It is an open-source software used to simulate biomolecules such as proteins, nucleic acids, lipids and carbohydrates. It is designed to work in parallel which means that it takes advantage of the processing power of computer clusters, allowing the users to do high-performance simulation of large biomolecular systems <sup>[64]</sup>.

**Selection**

Although all these simulation softwares can simulate membrane proteins, particular attention was given to the *GROMACS* software. According to the official website, *GROMACS* stands out due to some distinctive features. Its most significant one is that it offers exceptionally high performance compared to other programs when working with CPUs and MPI, due to its numerous optimizations in the code <sup>[64]</sup>, which makes it useful when working with supercomputers that only have CPUs such as StarLife, which is explained in 4.4. *Computing resources* section. Moreover, *GROMACS* is a user-friendly and easy-to-navigate software, with topologies and parameter files written in clear text format; it includes a broad range of flexible tools for analysing trajectories; and it can be run in parallel enhancing its performance capabilities <sup>[64]</sup>. *GROMACS* was the chosen software not only because of these advantages but also because of its simplicity of use and the fact that is a free distributed software <sup>[59]</sup>. The version of *GROMACS* used is 2022.3 as from various tests it was seen that this was the most compatible one with the files to be run.

**4.3.2. Force fields**

The selection of the force field is a crucial step in a MD project. Force fields are in continuous evolution; nevertheless, the following four force fields are the most popular for protein simulations: *AMBER99SB-ILDN*, *CHARMM36*, *GROMOS 53a6* and *OPLS-AA/M* <sup>[65]</sup>.

As *MemProtMD* files will be used to run the simulations, the preparation of the system as well as the parameters such as the force field, among others, are already determined. *MemProtMD* uses *GROMOS53a6* or *CHARMM36* forcefields, depending on the simulation. The *GROMOS* force fields are united atom force fields, i.e., without explicit aliphatic (non-polar) hydrogens; while *CHARMM36* it does consider hydrogens (all-atom force field).

**4.3.3. Software for visualizing results of simulations**

Visualizing dynamic molecular processes allows quick and easy exploration of the dynamic transitions between the states of the system of study <sup>[66]</sup>. Some programs that prove helpful in visualizing either a trajectory file, a coordinate file, or both <sup>[67]</sup> are the following: *VMD*, *PyMol* and *ChimeraX* <sup>[68]</sup>.

*VMD* is the most commonly employed among the molecular graphics tools that support molecular dynamics. It is capable of producing 'movies', examining characteristics like atomic fluctuations,



and providing flexible integration with both other computational tools and user's personal scripts [66]. Moreover, *VMD* is compatible with the trajectory formats generated by *GROMACS*, the software selected in this project, and is the main software used in the research group where this project was carried out. For all this, it was chosen as the software for visualizing the resulting simulations.

#### 4.4. Computing resources

Simulations in MD demand short time steps, often only a few femtoseconds each, as it was previously commented in the *Background* section. A typical simulation encompasses millions or even billions of time steps, along with the assessment of millions of interatomic interactions within a single time step [3]. This results in the fact that simulations are extremely computationally demanding and therefore they must be performed in supercomputers.

The Molecular Modeling and Bioinformatics research group, where this project has been developed, has at their disposal the possibility of running the dynamics at the BSC-CNS (Barcelona Supercomputing Center – Centro Nacional de Supercomputación) due to their direct contact with Prof. Josep Lluís Gelpí, the group manager of the INB in the BSC-CNS. Specifically, this project will be using the StarLife infrastructure as it has already installed the *GROMACS* version 2022.3, version needed for running the simulations of *MemProtMD*.

StarLife is a distinctive infrastructure designed to enhance the competitiveness of Barcelona's biomedical cluster, through the collaboration between Centro de Regulación Genómica, the IRB and the BSC-CNS, with the backing of La Caixa, la Generalitat de Catalunya and Fondo Europeo de Desarrollo Regional. StarLife offers 138,2 Teraflops from a total of 54 nodes (2160 cores).

To access to StarLife a ssh connection will have to be established. A ssh is a network protocol used from the terminal that enables one to establish a connection between different remote hosts. Moreover, these simulations will be run directly in StarLife terminal using bash. Bash is an interpreter and programming language that is used to interact with the operating system (Linux/Unix) using text commands [69].

#### 4.5. *MDposit* platform

This project aims to contribute to scientific community by sharing the protein dynamics and flexibility that will be obtained so other researcher can use them. A way of doing this is by using the platform *MDposit*, which has been created by the research group where this project has been carried out.

*MDposit* is a publicly accessible platform developed to enable web-based access to atomistic molecular dynamics simulations. The primary objective of this initiative is to facilitate and encourage the sharing of data among the global scientific community, with the ultimate goal of contributing to research efforts.

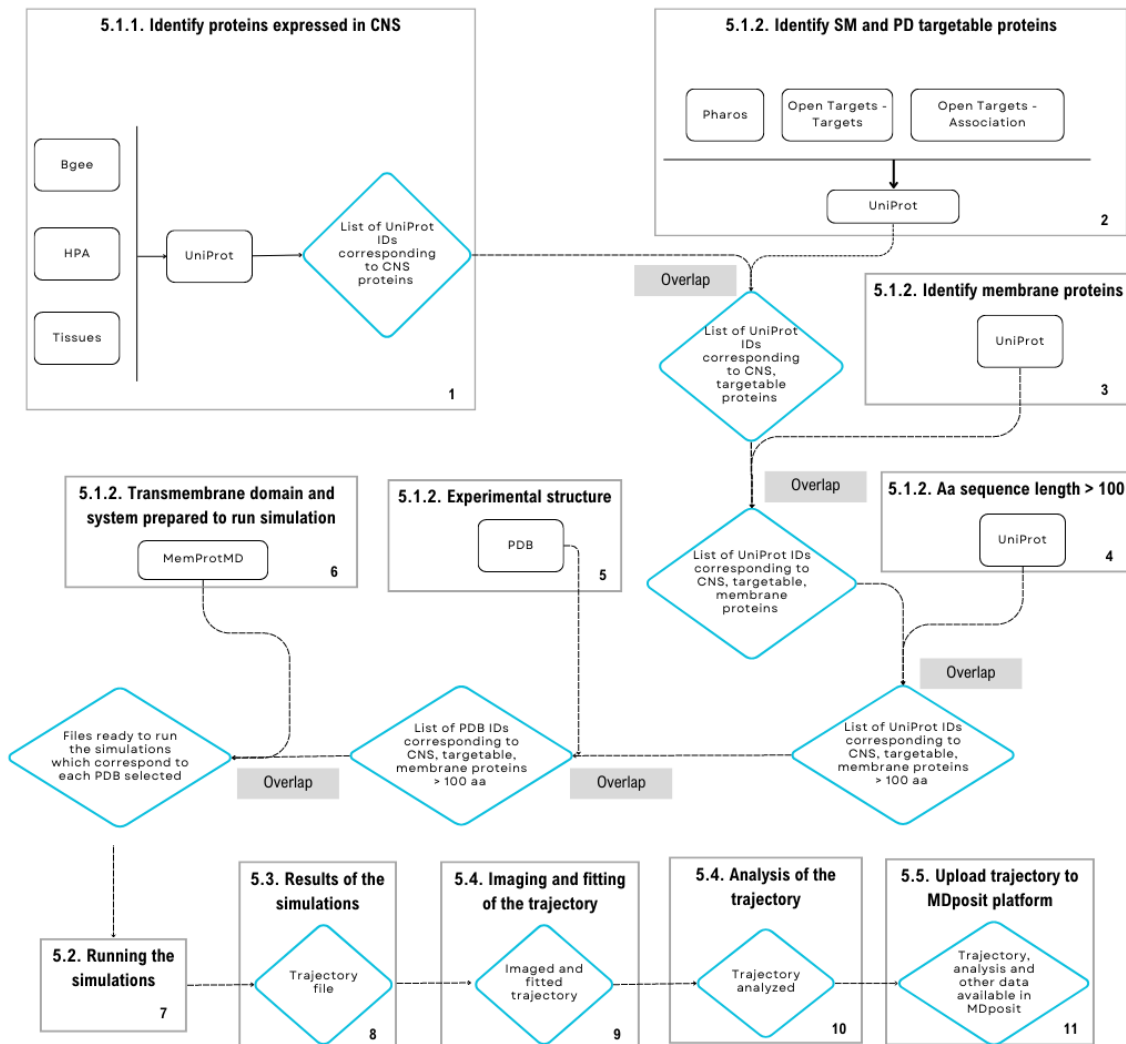


As this project does not intend to modify the platform, just upload some data to it, a brief summary of the sections that contains is done. This brief overview is considered enough for understanding how the data is organized and how it will be uploaded. The platform structure contains:

- An overview section which provides detailed information about the simulated *PDB* structure and the parameters used in the simulation such as software program and its version.
- A trajectory section, where the user can access the resulting trajectory from the simulation. The trajectory can be visualized in various ways using the configuration panel. The configuration panel allows the user to select the biomolecule of interest, customize the viewing options, adjust the trajectory speed and specify the number of frames to display. Additionally, this section includes protein functional analysis.
- An analysis section which offers several quality control tests, including RMSDs (Root Mean Square Deviation), RMSD per residue, RMSD pairwise, radius of gyration, fluctuation, PCA (Principal Component Analysis) and solvent accessible surface. It may also include interaction analysis such as distance per residue, electrostatic potential surface, hydrogen bonds, energies and pocket analysis.
- A download section which makes available the download of these trajectories.
- A REST-API section which allows the user to access to all the information in the database programmatically. The user can also download coordinates for specific frames and structure domains.

# 5. Detailed Engineering

The Detailed Engineering section includes several steps. First, a list of CNS proteins that satisfies the conditions established in the previous section must be obtained. Once these proteins are known and the files needed to run the simulation are downloaded, the molecular dynamics simulations are run. This will result in the trajectory of the protein, which among other data, will be published in the *MDposit* database. **Figure 10** shows the workflow which will be followed in this section in relation to the selection taken in the *Conceptual Engineering* section.



**Figure 10.** Workflow followed of the execution part of the project (Detailed Engineering).

## 5.1. Selection of proteins of interest

To select the proteins of interest relying on databases, a Jupyter Notebook file was programmed in Python language. This code can be found in the following [GitLab](https://mmb.irbbarcelona.org/gitlab/imartinv/model_cns_isabel) link ([https://mmb.irbbarcelona.org/gitlab/imartinv/model\\_cns\\_isabel](https://mmb.irbbarcelona.org/gitlab/imartinv/model_cns_isabel)) . With the appropriate environment, packages and disk space, this code will select automatically the proteins of interest of this project and download the needed files for each one to run their simulation. As databases are always being updated, little differences in the final list of proteins may be found every time the code is executed. For this project, the workflow was run between February and April of 2023.

A summary of the databases which will be downloaded and manipulated in this section, which were previously selected in the *Conceptual Engineering* section, are displayed in **Table 5**.

**Table 5.** Downloaded databases for selecting the proteins of interest.

Databases downloaded name	Format	Source	Interest
<a href="#">Homo Sapiens simple expression</a>	Compressed tsv in gz	<i>Bgee</i>	Select CNS proteins
<a href="#">Human Brain Proteome</a>	tsv	<i>Human Protein Atlas</i>	Select CNS proteins
<a href="#">Human All channels integrated</a>	tsv	<i>TISSUES</i>	Select CNS proteins
<a href="#">Targets</a>	json	<i>Open Targets</i>	Select target proteins
<a href="#">Association</a>	json	<i>Open Targets</i>	Select PD target proteins
<a href="#">Pharos</a>	tsv	<i>TCRD</i>	Select target proteins
<a href="#">Transmembrane and Intramembrane</a>	tsv	<i>UniProt</i>	Validate, mapping, to select membrane proteins and select minimum length proteins

When using all these databases and comparing information, it is essential to identify uniquely the proteins and be able to map them between the different databases. To map the proteins between the different databases, the *UniProt* ID will be used. This approach also serves the purpose of validating that the proteins identified in these three are also listed in *UniProt*, which is considered as a reference database, as it was previously commented. Therefore, all proteins selected from each database will be mapped to *UniProt* to obtain their ID and be able to overlap information (**Figure 10**).

### 5.1.1. Selection of central nervous system proteins

The goal of this section is to identify and select proteins specifically expressed in the central nervous system, which is a crucial requirement for this project. To achieve this, *Bgee*, *HPA*, and *TISSUES* databases will be used, as it was previously selected in the *Conceptual Engineering*

section. The aim is to combine the information from these databases and reach a consensus on which proteins are present in the CNS.

## **Bgee**

### *Data overview*

After downloading the *Bgee* database entitled *Present/Absent expression calls - Anatomical entities only- Simple file*, in .tsv format with a size of 162 MB, it was visualized how the data is organized, with the help of the relevant documentation. The database to be worked with has a dimension of 9,093,493 rows × 9 columns. The data is classified according to nine columns (**Figure 11**).

	Gene ID	Gene name	Anatomical entity ID	Anatomical entity name	Expression	Call quality	FDR	Expression score	Expression rank
18	ENSG00000000003	TSPAN6	UBERON:0000007	pituitary gland	present	gold quality	1.000000e-14	93.38	3090.0
38	ENSG00000000003	TSPAN6	UBERON:0000451	prefrontal cortex	present	gold quality	4.110953e-14	74.56	11900.0
47	ENSG00000000003	TSPAN6	UBERON:0000941	cranial nerve II	present	gold quality	1.000000e-02	91.36	4030.0
51	ENSG00000000003	TSPAN6	UBERON:0000955	brain	present	gold quality	1.000000e-14	75.86	11300.0
52	ENSG00000000003	TSPAN6	UBERON:0000956	cerebral cortex	present	gold quality	1.000000e-14	74.74	11800.0

**Figure 11.** *Bgee* dataset organization.

From these, the columns of interest to select protein expressed in the CNS are the ones displayed in **Table 6**.

**Table 6.** How data in *Bgee* database downloaded is structured.


Column name	Description <sup>[70]</sup>
Gene ID	Unique gene identifier from <a href="#">Ensembl</a> database.
Gene name	Gene name.
Anatomical entity ID	Unique and unequivocal identifier of the anatomical entity, according to Uberon Ontology.
Anatomical entity name	Name of the anatomical entity.
Expression	Expression value - present or absent - according to the chosen condition parameters, in this case, anatomical.

For doing this selection, the *Anatomical entity ID* column has to be used as it contains the identifiers of the anatomical entity of where the gene can be expressed. As previously mentioned in **Table 6**, these identifiers are IDs from Uberon Ontology. Uberon Ontology gives to the central nervous system entity the ID: *UBERON:0001017* <sup>[71]</sup>. This ID not only defines the central nervous system but all its components and how they are related as this ontology is classified as a tree hierarchy (**Figure 12**).

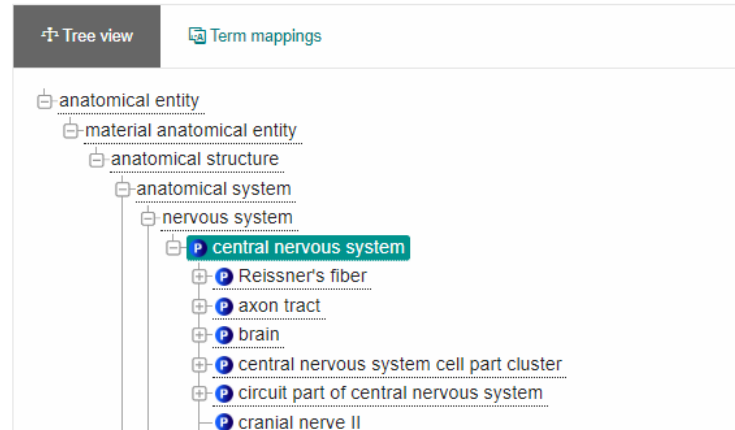
Ontologies are used in several databases such as *Bgee* and *TISSUES* to identify where a gene in a tissue or anatomical entity name is expressed, and know the relation of this with others.

OLS / Uber-anatomy ontology **UBERON** / UBERON:0001017 

# central nervous system

[http://purl.obolibrary.org/obo/UBERON\\_0001017](http://purl.obolibrary.org/obo/UBERON_0001017) 

The central nervous system is the core nervous system that serves an integrating and co-ordinating function in vertebrates and the spinal cord. In invertebrates it includes central ganglia plus nerve cord. [ <http://amigo.geneontology.org/amigo/term/UBERON:0001017> ] <https://sourceforge.net/p/geneontology/ontology-requests/11422/> <http://www.frontiersinzoology.com>



**Figure 12.** Tree hierarchy for central nervous system Uberon ID [72].

Moreover, the *Expression* column was used to know if the gene is expressed, meaning it results probably in a protein.

Finally, when genes expressed in the CNS were identified, these should be mapped to *UniProt* because of previous justifications. As this database does not contain a column for identifying these genes expressed with a *UniProt* ID, a mapping between the *Bgee* database and *UniProt* database has to be done. To do this mapping and obtain their *UniProt* IDs, the Gene ID column is essential. This column provides another type of identifier which is from *Ensembl* database; another resource known in this field. Therefore, this gene ID from *Ensembl* has to be map with *UniProt* in order to know the *UniProt* ID.

When doing this selection, it was not taken into account the expression level because although a gene has a low expression in the CNS, it may have an important function in it.

It was expected that each entry of each column contains a single value. However, the *Anatomical entity ID* column contained two identifiers in some cases (e.g. 'UBERON:0000473  $\cap$  CL:0000089') which can difficult the manipulation of the data if not separated.

### Data manipulation

Each gene in this database can be expressed in a multitude of anatomical entities. During the inspection of the Anatomical entity ID column, it was observed that when a gene has multiple entries for different tissues, if it already has a specific entry for one part of the central nervous system, an additional entry for the central nervous system term is not found. Therefore, when selecting genes

which may be in the central nervous system, it must be taken into account also the entries which are classified as components of central nervous system (e.g., medulla) as if not, information will be lost. To clarify this, **Figure 13** is displayed. It can be observed that the gene *TSPAN6* may be classified as expressed in pituitary gland and prefrontal cortex, both components of the CNS, but it is not classified as expressed in CNS itself. Therefore, if one wants to select the genes which may be expressed in the CNS it should also consider the genes expressed in components that make up the CNS.

	Gene ID	Gene name	Anatomical entity ID	Anatomical entity name	Expression	Call quality	FDR	Expression score	Expression rank
18	ENSG00000000003	TSPAN6	UBERON:0000007	pituitary gland	present	gold quality	1.000000e-14	93.38	3090.0
38	ENSG00000000003	TSPAN6	UBERON:0000451	prefrontal cortex	present	gold quality	4.110953e-14	74.56	11900.0

**Figure 13.** *TSPAN6* gene can be classified as expressed in components of the CNS but not classified as expressed in CNS itself.

To determine these CNS components, as Uberon Ontology provides a tree view of each Uberon ID, all descendants from CNS will be components that make it up. Therefore, it is of interest to get the Uberon ID of the CNS but also all of its descendants. This was done by using the Uberon Ontology API which provides this information for each ID. Once knowing all these IDs, the *Anatomical entity ID* column was filtered to only stay with those genes classified as CNS and its components.

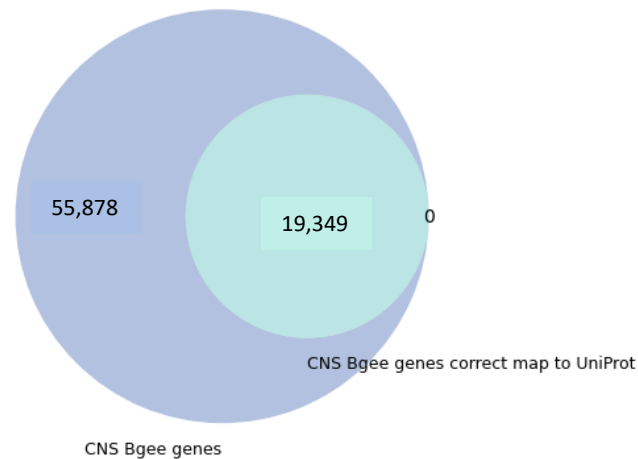
Second, to only stay with those genes which are expressed in these anatomical entities, the *Expression* column was filtered by *present* as it means that the gene is expressed in that anatomical entity name.

When all these steps were done, a list of non-repeated Gene IDs was obtained. These were genes expressed in the CNS. This does not mean that these genes cannot be expressed in another anatomical entity.

Next, to determine if these genes are protein coding, validate their existence and compare them with other databases, they will be mapped to the *UniProt* database. If a protein in *Bgee* database has a *UniProt* identifier, it means that *UniProt* database contains this protein.

To do this mapping, the mapping tool provided by *UniProt* was used. The tool was accessed via API and it requests from which-to-which database ID the mapping is going to be done. In this case, the *Bgee* Gene ID column contained *Ensembl* IDs (ENSG ID), and the desired *UniProt* ID was *UniProtKB*, as chosen during the Conceptual Engineering phase. Thus, the mapping involved converting from ENSG ID to *UniProtKB* ID. When doing this mapping, a number of 19,349 genes from *Bgee* were found in *UniProt* from a total of 55,878 CNS genes in *Bgee* (**Figure 14**).

Venn diagrams of mapping Bgee genes to UniProt proteins



**Figure 14.** CNS genes from Bgee (55,878) mapped to UniProt proteins (19,349).

However, it must be taken into consideration that only reviewed and human *UniProt* IDs were of interest, as this project aims to be the more accurate as possible for human proteins. When considering this a total of 18,459 *UniProt* IDs are acquired.

From these numbers it can be seen that a great number of genes in *Bgee* are not found in *UniProt*. This was observed that it is due to novel genes found in *Bgee*, inaccuracy of the mapping tool or non-coding-protein genes collected in *Bgee*, as *UniProt* only considers genes which codifies for proteins.

Overall, the final output of all these steps is a list of unique 18,459 *UniProt* IDs obtained from the *Bgee* dataset.

## Human Protein Atlas

### *Data overview*

The *HPA* database of the *Human Brain* section was downloaded in .tsv format of 25.1 MB in size and has a dimension of 16,465 rows x 4 columns. The *HPA* database used in this project classifies the data in 4 columns (**Figure 15**), although it can also be structured depending on the user's preference.

	Gene	Gene description	Uniprot	Evidence
0	A2M	Alpha-2-macroglobulin	P01023	Evidence at protein level
1	A2ML1	Alpha-2-macroglobulin like 1	A8K2U0	Evidence at protein level
2	A4GALT	Alpha 1,4-galactosyltransferase (P blood group)	Q9NPC4	Evidence at protein level
3	AAAS	Aladin WD repeat nucleoporin	Q9NRG9	Evidence at protein level
4	AACS	Acetoacetyl-CoA synthetase	Q86V21	Evidence at protein level

**Figure 15.** Human Protein Atlas dataset organization.

From these four columns, the columns of interest to work with are the ones displayed in **Table 7**: *UniProt* and *Evidence*. This is because *HPA* already gives evidence that these proteins are from the central nervous system and its components, and therefore few manipulations have to be done, only selecting those with protein evidence to be accurate and collect their *UniProt* IDs.

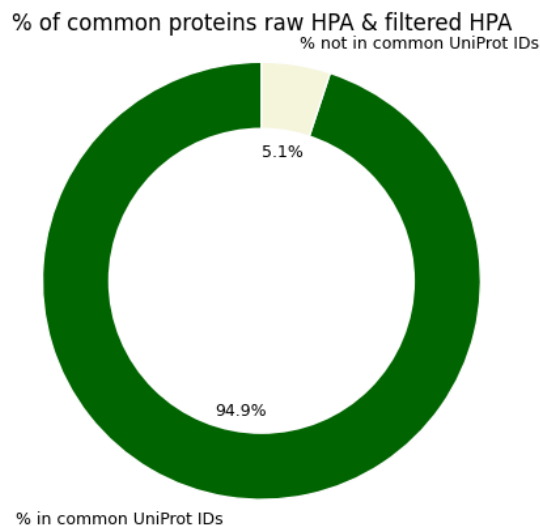
**Table 7.** Columns of interest in HPA dataset.

Column name	Description
<i>UniProt</i>	Unique <i>UniProt</i> ID
<i>Evidence</i>	Protein evidence scores generated from several sources

#### Data manipulation

First, those genes whose value in the *Evidence* column is not *Present at protein level* were discarded as if protein evidence is found, it is more reliable that a gene is clearly a coding gene. From this filter, 15,782 proteins (*UniProt* IDs) out of 16,465 proteins were obtained.

Then, the *UniProt* ID of the resulting proteins was selected. It should be noted that this column has null values since there are still genes that *HPA* has in its database, but *UniProt* does not yet include them and, therefore, they do not have a *UniProt* ID. Specifically, there are 100 genes that still do not have *UniProt* ID. Those genes without a *UniProt* ID are discarded. From this, a list of 15,631 proteins is obtained. **Figure 16** show the proportion of proteins discarded by these filters.



**Figure 16.** 5.1% of proteins belonging to the raw HPA dataset are filtered out.

Overall, this section provides a list of 15,631 *UniProt* IDs which come from HPA dataset.

## TISSUES

### Data overview



The *All channels integrated* database for human was downloaded in .tsv format and this file has a size of 496.2 MB. This database has a dimension of 8,938,525 rows (entries) x 5 columns. These five columns into which the database sorts the data are: gene identifier, gene name, tissue identifier, tissue name and integration of all confidence scores; and the columns take values such as the ones shown in **Figure 17**.

1	18S_rRNA	18S_rRNA	BTO:0001489	Whole body	3.422
2	18S_rRNA	18S_rRNA	BTO:0000284	Organism form	3.217
3	18S_rRNA	18S_rRNA	BTO:0001773	Oocyst	3.186
4	18S_rRNA	18S_rRNA	BTO:0001481	Plant	3.167

**Figure 17.** *TISSUES* database organization: gene identifier, gene name, tissue identifier, tissue name and integration of all confidence scores.

However, the columns of interest are the ones displayed in **Table 8** as they enable to filter by genes expressed in the CNS and obtain their gene ID to do the mapping with *UniProt* ID.

**Table 8.** Columns of interest in *TISSUES* database.

Column name	Description
Gene identifier	Unique ID for each gene
Tissue identifier	Tissue identifier from BTO ontology

### Data manipulation

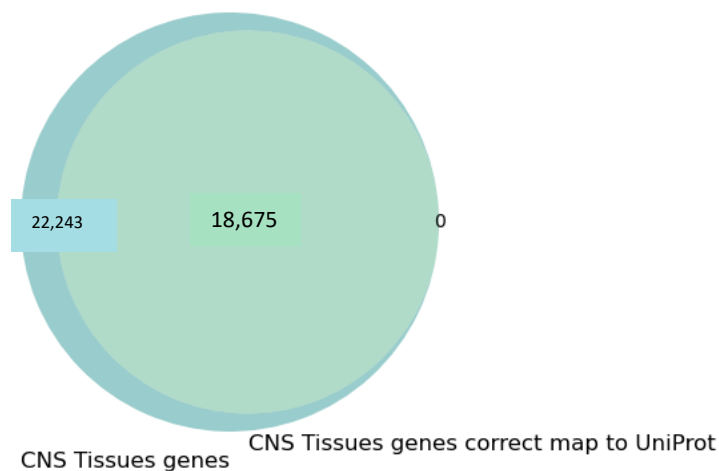
First, the genes expressed in CNS were obtained. To do this, the column Tissue identifier was used as it contains unique *Brenda Tissue Ontology* (BTO) identifiers for each tissue. This ontology is similar to *Uberon* but with different terms. The *Brenda Tissue Ontology* is a structured encyclopaedia of tissue terms containing more than 4,600 different anatomical structures, tissues, cell lines and cell types classified according to Gene Ontology Consortium principles. The BTO identifier to select CNS tissue is *BTO\_0000227* [73].

When doing this selection, it was seen that the same situation encountered in *Bgee* happened in this dataset as well: although genes can be classified as expressed in several tissues, when expressed in some component of the CNS it was not also classified as CNS itself. Therefore, it is necessary to filter by the term CNS and also by all its components. To find all the terms of the different CNS components, the EMBL-EBI BTO Ontology Search Service (OLS) API is used. This API is organized by tree hierarchy, so the children of the CNS term are the ones of interest. From this, a list of 22,343 gene IDs was obtained.

The next step, as *Bgee*, is to map from the gene ID proportionated by the *TISSUES* database to *UniProt* ID, as *TISSUES* database does not contain a column for identifying these genes with a

*UniProt* ID either. However, it identifies them with *STRING* identifiers, as it is written in the documentation of the database. To map the *STRING* identifier to the *UniProt* identifier, the API of the *UniProt* mapping tool was again used. As the mapping tool resulted in a list of few *UniProt* IDs - which was observed that can be due to very novel proteins with no evidence yet in *UniProt*, inaccuracy of the mapping tool or even insertion in *TISSUES* of non-coding genes – a mapping from *Ensembl* ID to *UniProt* ID was also done. This was because in *TISSUES* documentation is explained that *STRING* IDs were mostly obtained from *Ensembl* IDs, and therefore, maybe, some of them were not mapped from *Ensembl* to *STRING* and *Ensembl* IDs were also used. A total of 18,675 *UniProt* IDs were found in *UniProt* out of 22,343 gene IDs from *TISSUES* (**Figure 18**).

Venn diagrams of mapping Tissues genes to UniProt proteins



**Figure 18.** CNS genes from *TISSUES* mapped to *UniProt*.

It needs to be taken into account that this project is focused in reviewed *UniProt* IDs and for human. From this, the overall, the output of this section was a list of 18,429 *UniProt* IDs obtained out of 22,243 CNS genes from *TISSUES*.

### Intersection between the three databases

At this point, three lists of *UniProt* IDs were obtained, one for each database and validated with *UniProt*. Moreover, for further validation and consensus between these databases, an intersection of these three lists was done. It was of interest to only select those *UniProt* IDs that were in the three databases. The results of this intersection are shown in **Table 9**. The *UniProt* IDs common in the three databases will be selected and the ones not present in it will be discarded.

From the intersection numbers in **Table 9**, it can be seen that *Bgee* and *TISSUES* have a lot of proteins in common; and almost all HPA proteins are found in the other databases. Therefore, also a good approximation would have been also to only use HPA dataset although by using the intersection of the three it is more accurate.

A total of 15,317 central nervous system proteins (*UniProt* IDs) were obtained.

**Table 9.** Matrix of resulting UniProt IDs from each overlap

	Bgee	TISSUES	Human Protein Atlas
Bgee	18,459	17,659	15,542
TISSUES	17,659	18,429	15,361
Human Protein Atlas	15,542	15,361	15,631

Overlap between the three databases: **15,317 CNS proteins**

### 5.1.2. Filters applied to reduce the dataset

Once obtained the dataset of which proteins are expressed in the CNS, it was of interest to reduce this dataset by applying some filters which were commented in the *Conceptual Engineering* section. These filters aim to specifically choose those proteins which satisfy some requirements and therefore are of interest of this project.

#### *Drug Targetable. Druggable with SM activity and Parkinson Diseases*

As discussed in the previous section, three databases are used to perform this filtering: *Open Targets - Target*, *Open Targets - Associations - direct (overall score)* and *Pharos*.

The *Pharos* database, downloaded as a .csv file, has a size of 370.7 kB. This database consists of 20412 rows × 3 columns. It is structured according to **Figure 19**.

	UniProt_accession	Pharos_target	TDL
0	P32929	P32929	Tchem
1	A4D0Y5	A4D0Y5	Tdark
2	Q49A92	Q49A92	Tbio
3	Q9UFW8	Q9UFW8	Tbio
4	Q96K31	Q96K31	Tdark

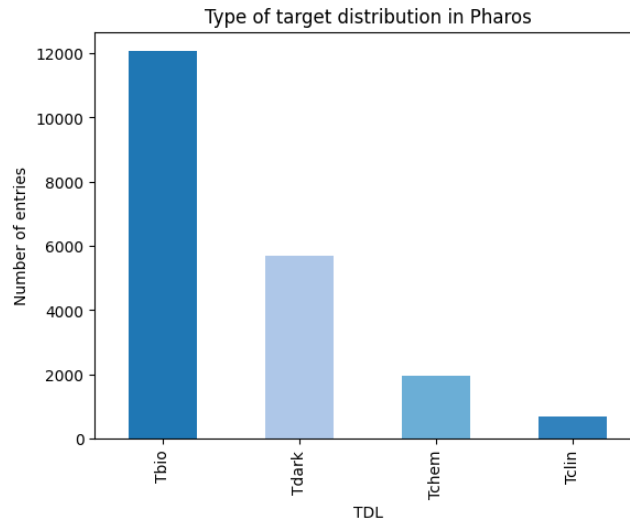
**Figure 19.** Pharos database organization.

From these columns, the ones of interest (**Table 10**) are *UniProt\_accession*, so as to get the protein *UniProt* ID and be able to compare this dataset with the previous ones obtained, and TDL, to select which target development level is chosen according to the criteria established in *Conceptual Engineering* section.

**Table 10.** Columns of interest for data manipulation of Pharos.

Columns	Description
<i>UniProt_accession</i>	Unique <i>UniProt</i> ID
TDL	Target Development Level (Tchem, Tclin, Tbio, Tdark)

The criteria established that the interesting *UniProt* IDs were the ones that were classified as Tchem and Tclin in the TDL column, as Tchem are proteins known to bind small molecules and Tclin are the ones with already approved drugs, as commented in Conceptual Engineering. These categories include fewer entries (**Figure 20**) in the dataset compared to Tbio and Tdark, which makes sense as there are a lot more of unstudied proteins than studied ones. These few entries can repunctuate in a decrease of number of *UniProt* IDs of the final wanted list, after all the filters.



**Figure 20.** Comparison between the number of entries of each target classification in Pharos.

Moreover, *Open Targets* database was used for filtering by druggable proteins. The *Open Targets - Target* database was downloaded in 200 files in .json format, with sizes between 2.3 MB and 5.6MB, which were joined into a single table with Python. This table - or database - has a size of 62,678 rows  $\times$  28 columns. Although the dataset is organized in 28 columns, **Figure 21** shows a small section of how this organization is.

	id	approvedSymbol	biotype	transcriptIds	canonicalTranscript	canonicalExons	genomicLocation	approvedName
0	ENSG00000020219	CCT8L1P	processed_pseudogene	[ENST00000465400]	{'id': 'ENST00000465400', 'chromosome': '7', 'start': 152445477, 'end': 152447150}	[152445477, 152447150]	{'chromosome': '7', 'start': 152445477, 'end': 152447150}	chaperonin containing TCP1 subunit 8 like 1, p...
1	ENSG00000059588	TARBP1	protein_coding	[ENST00000496673, ENST00000483404, ENST00000048...]	{'id': 'ENST0000040877', 'chromosome': '1', 'start': 234420812, 'end': 234425673}	[234420702, 234420812, 234425673, 234425793, 2...]	{'chromosome': '1', 'start': 234391313, 'end': 234425673}	TAR (HIV-1) RNA binding protein 1
2	ENSG00000070182	SPTB	protein_coding	[ENST00000553938, ENST00000389720, ENST00000055...]	{'id': 'ENST00000644917', 'chromosome': '14', 'start': 64774397, 'end': 64774527}	[64785537, 64785627, 64774397, 64774527, 64769...]	{'chromosome': '14', 'start': 64746283, 'end': 64774527}	spectrin beta, erythrocytic
3	ENSG00000070366	SMG6	protein_coding	[ENST00000354901, ENST00000570756, ENST00000026...]	{'id': 'ENST00000263073', 'chromosome': '17', 'start': 2292977, 'end': 2297863}	[2172658, 2172859, 2292871, 2292977, 2297863, ...]	{'chromosome': '17', 'start': 2059839, 'end': 2292977}	SMG6 nonsense mediated mRNA decay factor
4	ENSG00000072071	ADGRL1	protein_coding	[ENST00000361434, ENST00000589616, ENST00000059...]	{'id': 'ENST00000361434', 'chromosome': '19', 'start': 14159401, 'end': 14155...}	[14160112, 14160297, 14159401, 14159584, 14155...]	{'chromosome': '19', 'start': 14147743, 'end': 14159584}	adhesion G protein-coupled receptor L1

**Figure 21.** *Open Targets - Target* dataset organization showing a few columns.

However, from all these columns the ones of interest are described in **Table 11**.

**Table 11.** Columns of interest in Open Targets.

Column	Description
<i>Biotype</i>	Type of gene
<i>proteinIds</i>	Unique identifier protein ID for several sources, including <i>UniProt</i>
<i>Tractability</i>	Target key data for tractability assessments such as small molecule (SM), antibody (AB), Proteolysis Targeting Chimeras (PR) and other clinical modalities (OC). It includes common assessments from ChEMBL for all modalities: Approved Drug, Approved Clinical and Phase 1 Clinical, and specific ones for each modality.

First, the table was filtered according to the *protein\_coding* value of the *biotype* column since it is only wanted to consider those genes that are coding for proteins.

As it was previously mentioned in earlier sections, this project focuses on those proteins that are targetable for Small Molecules. Although, almost all human proteins in *Open Targets* are considered targets, according to the *Open Targets* definition of targetability, there are levels of targetability. To keep only those proteins that meet this condition, the tractability column was filtered to keep only those values that contain the modality of *SM* and have *True* value in *Approved Drug* or *Advanced Clinical* or *Phase 1 Clinical* or *High-Quality Pocket*. This selection therefore will give proteins which are targetable for Small Molecules and already have or potentially have a drug, as considering this selection it is assured that there is a high probability that the protein has been verified to be a real target of small molecules.

From this filtering, only those *UniProt* IDs from the *UniProt\_swissprot* source (*proteinIDs* column) were selected since it is the one used in this project.

The *UniProt* IDs finally obtained by *Pharos* and those obtained by *Open Target* are then intersected to obtain a list of proteins IDs that are druggable, or potentially targetable, by Small Molecules.

Afterwards, the *Open Targets - Associations - direct (overall score)* database, also downloaded in 200 .json files, ranging in size from 166.2 kB to 2.1 MB, were also merged into a single table. The use of this database was to select targetable proteins for a specific disease, in particular this project is focused in the Parkinson's disease. This database has dimensions of 214,6271 rows × 4 columns. These four columns in which the database is structured are displayed in **Table 12**.

**Table 12.** How data in Open Targets - Associations - direct (overall score) is structured.

Column	Description
--------	-------------

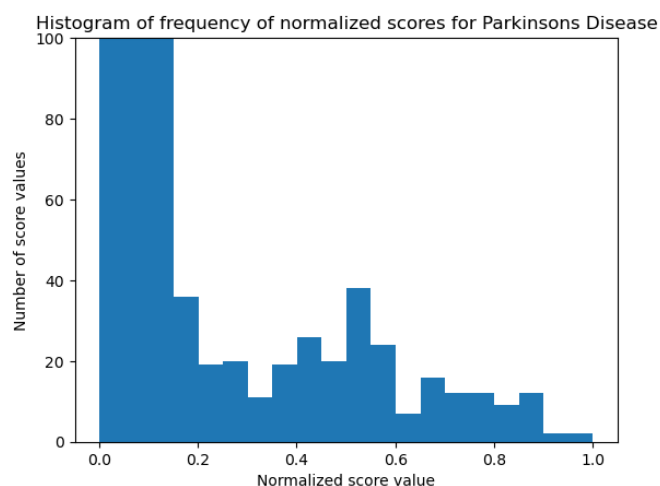
<i>diseaseId</i>	MONDO disease identifier
<i>targetId</i>	Unique identifier for target
<i>Score</i>	Overall association score that summarises evidence from several databases. It is calculated with a formula where each database has a weight and take values from 0 to 1 [74].
<i>evidenceCount</i>	Number of sources that give evidence of the information

From this dataset, the unique MONDO identifier for Parkinson's Disease (*MONDO\_0005180*) was used. MONDO is also an ontology but for diseases not for anatomical entities as BTO or UBERON, which were previously used.

By filtering the column *diseaseId* by this PD MONDO ID, only targets for PD are being considered. The number of entries when doing this is decreased to 2855 out of 2146271 total entries, meaning that there are 2855 proteins which are considered targets for PD.

Out of the 2855 entries available, it is important to determine which entries are more reliable as potential targets for Parkinson's disease as this dataset provides this information in the Scores column. The range of score values of these 2855 goes from 0.11 and 0.71 approximately. At higher score, more reliable and larger number of sources support that the target is for PD [75].

It was of interest to set a threshold value to have strong evidence that a specific target is for the Parkinson's Disease. To do this, first all values were normalized to have a score range from 0 to 1 so as to better assess the threshold value. **Figure 22** displays a frequency histogram of which score values are more recurrent and how they are distributed.



**Figure 22.** Frequency histogram of which score values for PD are more recurrent and how they are distributed.

From this plot it is seen that most of the normalized scores are between  $\sim 0$  and 0.2. This means that they have very little evidence because there are few sources that have studied this target or

the sources which have studied it were low reliable <sup>[75]</sup>. It is seen that from 0.5 upwards, there are a smaller number of scores with that values but with stronger reliability. It was considered that the normalized value threshold would be 0.5 as it is the half value of the maximum and it allows to discard very low scores but stay with enough proteins to study. With this threshold, a total of 132 protein target IDs were obtained.

To identify proteins that can be targeted by small molecules and are relevant to Parkinson's disease , an overlap of the two datasets obtained (druggable for SM and targetable for PD) was done. This overlapping process resulted in 85 *UniProt* IDs.

Next, these 85 *UniProt* IDs were compared to the previously obtained list of 15,317 *UniProt* IDs representing proteins expressed in the central nervous system. The aim was to narrow down the list to proteins that are both expressed in the CNS and potential targets for SM and PD.

From this comparison, a final intersection was obtained, yielding a total of 77 *UniProt* IDs. These 77 proteins represent the desired outcome of this section, as they are expressed in the CNS and can be targeted by SM for the treatment of Parkinson's disease.

### *Membrane proteins*

With *UniProt* DB, the previous list of 77 proteins was filtered to obtain only those which are membrane proteins. This was done to do a first initial filter to filter out those proteins that do not have any membrane interaction. However, as explained in the *Conceptual Engineering*, *UniProt* does not filter correctly by those proteins which only have a transmembrane domain as it also includes proteins with just few interactions on the membrane surface. Therefore, in further steps *MemProtMD* was also used to only select those membranes with a transmembrane domain, as this is what this project focus on doing; as it was previously mentioned.

The list of 77 proteins was intersected with this dataset of membrane proteins of *UniProt* and it resulted in a list of 75 proteins (*UniProt* IDs) which are from the central nervous system, targetable and membrane.

### *Amino acids sequence length*

Those proteins whose sequence length was less than 100 aa were not of interest by the criteria established in the *Conceptual Engineering*. Therefore, it was checked that the previous list of 75 proteins were all above this threshold. The minimum length of a protein of this list was 271 amino acids. Therefore, all the 75 proteins were considered as they had an interesting structure.

It must be considered that this section has taken into account the sequence length of *UniProt* which in fact indicates how many amino acids are in the canonical sequence. *UniProt* sequence length

and *PDB* sequence length are not the same as *PDB* may only consider some parts of the sequence. This can result in simulating shorter proteins than expected.

#### *Experimental structure: PDB*

From the above list of 75 proteins, it was of interest to select only those proteins which had at least a *PDB* ID, therefore an experimentally determined structure. The reason of this filter is explained in the *Conceptual Engineering* section; but in summary, predicted structures were not considered as this project focuses on a short dataset to simulate due to time limitations.

To know which proteins had at least one *PDB*, a file which contains the mapping between *UniProt* IDs and *PDB* IDs was used. This file can be found in the [GitLab](#) repository and it was provided by the Structural Bioinformatics and Network Biology research group. By using these relations between the two databases, a total of 55 proteins had at least one *PDB* structure out of the 75 proteins. From these 55 proteins, a total of 765 unique *PDBs* were obtained, meaning a total of 765 protein structures. These mapping between *UniProt* and *PDB* is necessary as to run the simulations the structure is needed.

It was considered interesting to collect technical information about these *PDBs* so as to further understand their structure and use it for the next step. This was done by using the *mmb.irbbarcelona* API by the Molecular Modeling and Bioinformatics research group, which stores all this information. It was collected the experimental method, the experimental resolution, information about the chains (type, sequence, fragments) and the chain ID of each *PDB*.

After this step, it was reviewed that all the 55 proteins were expressed in the CNS, as it is the more critical condition of project as it revolves around the Human Brain European Project.

#### *Transmembrane domain and files ready to run simulations*

From all the previous obtained *PDBs*, only those *PDBs* with a transmembrane domain were of interest, as it was previously explained in the *Conceptual Engineering*. To select those with this domain, it was previously selected that the best database to use was *MemProtMD*. This only stores ready-to-run-simulation files from proteins with this specific domain.

By using the API of *MemProtMD* it was obtained that it contains a total of 6352 *PDBs* which are transmembrane and ready to be simulated. From these, only 260 *PDBs* from the previous obtained list achieved from all the filtering conditions (765 *PDBs*) were in *MemProtMD*.

Once knowing this, the files from the 260 *PDBs* were attempted to be downloaded in order to already have the needed files to run the simulation of proteins which satisfies all the conditions. These files aim to be downloaded are atomistic models as it was accorded in the *Conceptual Engineering*. The downloading process was done by using the *MemProtMD* API. When trying to



download all the files from the 260 *PDBs*, files from 6 *PDBs* could not be correctly downloaded. Therefore, 254 *PDBs* could be simulated.

Furthermore, it was informed to the University of Oxford which CNS proteins were included (260 but 254 correctly downloaded) and which not (505) in the *MemProtMD* platform. This was done with the purpose of facilitating collaborative efforts - as the UOXF is part of the HBP and MDDDB projects - aimed at preparing simulations for the proteins that are currently not present in their database.

### *Selecting PDBs with unique amino acid sequence*

In the Conceptual Engineering section, it was explained that the *PDB* repository can contain multiple structure files with the same amino acid sequence for the same protein. This occurs because different experimental determinations capture the protein in different folding states, such as active or inactive forms. Consequently, there are several *PDB* IDs that essentially represent the same protein structure in terms of amino acid sequence.

Since this project focuses on protein dynamics, it is important to simulate unique amino acid sequences of proteins. Interactions and dynamics depend on these sequences. Simulating structures with the same amino acid sequence would result in redundant simulations, wasting computing time.

To avoid this redundancy and prevent simulating the same structure multiple times, a thorough comparison of the sequences was conducted for all 254 *PDBs*. When one or more *PDBs* had identical sequences, only one of them was selected to be simulated. Further detail of how this was done is found in the [GitLab](#) repository.

Out of the 254 *PDBs* set, only 157 *PDBs* were found to have a unique amino acid sequence. As a result, these 157 *PDBs* will be the ones selected for running their dynamics.

A summary of the filters applied and their output is shown in **Table 13**.

**Table 13.** Summary of all filters applied and their output for selecting the proteins of interest.

<i>Filter</i>	<i>Output</i>
<i>Central Nervous System - Bgee</i>	18459 proteins
<i>Central Nervous System - HPA</i>	15631 proteins
<i>Central Nervous System - TISSUES</i>	18429 proteins
<i>Central Nervous System – Bgee, HPA, TISSUES</i>	15317 proteins
<i>Target for SM and PD</i>	85 proteins
<i>CNS target for SM and PD</i>	77 proteins
<i>CNS – target -membrane</i>	75 proteins
<i>CNS – target -membrane -length</i>	75 proteins

<i>With PDB structure</i>	55 proteins
<i>PDBs in MemProtMD</i>	260 proteins
<i>PDBs in MemProtMD correct downloaded</i>	254 proteins
<i>PDBs without redundancy</i>	157 proteins
<b>Total of PDBs to study</b>	157 protein structures
<b>Total of proteins</b>	37 proteins

**Final result of section 5.1:** 157 atomistic files from *MemProtMD* which each one corresponds to a PDB ID. These PDBs are from 37 different proteins (*UniProt* IDs).

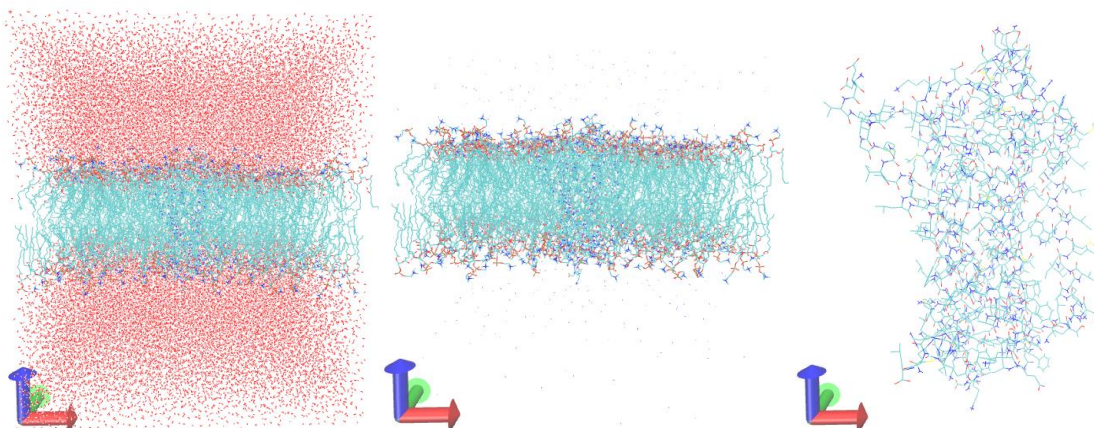
## 5.2. Running the simulations

### 5.2.1. Obtaining the files to run the simulations

The files to run the simulations were obtained from *MemProtMD* and are of the order of megabytes. These files are downloaded in a zip folder which contains the files and folders shown in **Table 14**.

**Table 14.** Files needed for running the simulations.

<i>File</i>	<i>Description</i>
<i>Readme text file</i>	Provides an overview of each relevant folder and file in the downloaded zipped folder. And the instructions for running the simulation
<i>ff folder</i>	Includes the files needed for the CHARMM36 force field.
<i>itp folder</i>	Contains topology files for membrane lipids and other components.
<i>mdp folder</i>	Contains .mdp files describing simulation parameters such as simulation duration, temperature, pressure, among others
<i>.ndx file</i>	Contains an index of the atoms in the membrane protein and is used to select specific groups of atoms during the simulation. In other words, it allows to identify and work with specific groups of atoms.
<i>.top file</i>	Contains the topology of the membrane protein (residues, atoms and interactions).
<i>.PDB file</i>	Stores the information about the three-dimensional structure protein. This file contains detailed information about the atoms, chemical bonds and spatial arrangement of the atoms that make up the molecule. It can be visualized with VMD software program ( <b>Figure 23</b> ).
<i>.itp files</i>	Include information on chemical bonds, bond angles, torsions, charges, etc.



**Figure 23.** Atomistic-system.PDB file of a G Protein Coupled Receptor (PDB ID: 4GBR) where the structure of the system can be observed. A) Structure of all the system: water molecules, membrane and protein. B) Structure of the membrane and protein. C) Structure of the protein.

### 5.2.2. Preparation of the systems for simulation

Preparing systems for simulation is a crucial step in performing accurate and detailed molecular simulations. Before the simulations can be run, it is necessary to prepare the molecular systems, which involves a number of steps including building the system structure, assigning the boundary conditions, defining the simulation parameters and generating the initial trajectories.

In this case, *MemProtMD* already makes available all the files required for the simulation as well as the parameter values and boundary conditions. The boundary conditions are important to ensure that the system is simulated in a suitable environment, an environment that resembles the real conditions.

### 5.2.3. Running the simulations

To obtain the trajectory of the protein, which is the result of interest when running these simulations, the process starts with an Energy Minimization to reduce the energy of the system and eliminate possible errors in the structure. Next, a Position Restrained Protein MD Simulation is performed to balance the atomistic system prior to the molecular dynamics simulation. Finally, a 100 ns MD simulation is performed to simulate the time evolution of the system. From this last step, the trajectory of the protein will be obtained.

Since many protein structures were obtained for running, the process has been automated so as not to launch them one by one. For doing this, one bash script was written for each step that composes the overall simulation: one for the EM step, another of the PR step and one for the 100 ns simulation step. Each script will correspond to one Slurm job and they will be run with dependencies, meaning that first, the EM step will be run, then the PR step and finally the 100 ns MD simulation step. Slurm is a task and cluster management system <sup>[76]</sup> which schedules a set of instructions that have to be performed called job <sup>[77]</sup>. It is useful as it allows distributing the work of

multiple users among the resources available in a supercomputer. When working with supercomputers it is essential to learn how Slurm works because it is a fundamental tool of supercomputers. From the personal experience of the author of this project, this may require time and help.

It should be noted that supercomputers have a time limit, 72 hours in the case of Starlife, for each job so as not to hijack resources for too long. This means that after 72 hours, the job is cancelled. Therefore, the 100 ns simulation step has to be resumed from a check point created by *GROMACS* every time this time limit passes. The simulation step will be executed for a total of 360 hours, enough for the 100 ns simulation to be run.

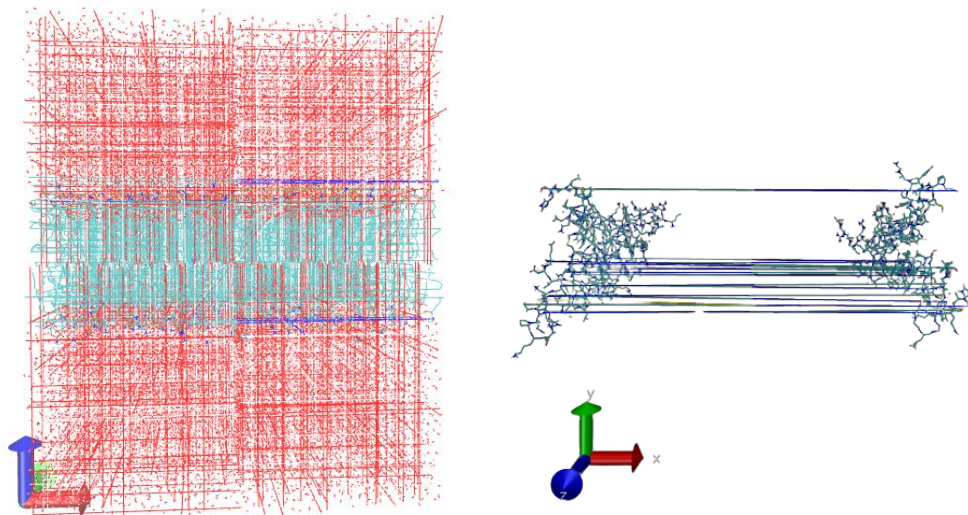
The main resulting file of interest is the one that stores the trajectory. Moreover, the structure file generated will be also useful for next steps.

It should be noted that when running the minimization of energy and the position restrained protein steps, some *GROMACS* warnings were raised, but none of significance, as the 100 ns simulation was successfully completed.

It has to be noted also that due to the time limit of this project out of the 157 *PDBs* wanted to run, 51 were actually simulated. However, this project will have continuity within the research group where it has been carried out. Therefore, the dynamics of these remaining *PDBs* will be run in future work as they will also be used as a use case for the MDDB project.

### 5.3. Results of the simulations

To visualize the resulting trajectories of the simulations, a structure file and the trajectory file are needed to be loaded into the visualization software VMD. *MemProtMD* puts at disposal a structure file. However, a resulting structure file was also obtained from the process and this will be the one used. The resulting trajectory is then loaded into this structure file and the movement of the system can be visualized. **Figure 24** shows one frame of the resulting trajectory. As the simulation is performed under Periodic Boundary Conditions, atoms which cross the limits appear on the other side; that is why frames lines in all the box are observed. These lines represent bonds between atoms of the same molecule as some of the atoms of the molecule may be on the other side of the box. If one wants to observe how the protein moves, it is impossible to do it with this bond lines. Therefore, a preprocess of this trajectory is needed in order to obtain a clear representation without crossing lines. This is the process known as Imaging or Removing Boundary Conditions.



**Figure 24.** Resulting trajectory of the simulation. Mathematically in principle is correct but when visualizing it the periodic boundary conditions effects can be appreciated. Left image shows the entire system and right one shows the protein.

## 5.4. Processing the simulations

### 5.4.1. Imaging and fitting

For a correct visualization and the subsequent analysis of the protein trajectory, it is required to do an imaging process.

As a result of using PBC, when observing a simulation, one may mistakenly think there is an error when notices that molecules deviate from the center of the box, diffuse out of the box, create holes, or appear broken or distorted (**Figure 24**). However, this is to be expected because any atoms leaving the simulation box from one side will enter from the opposite side due to PBC, as commented. In most cases, molecules are free to diffuse within the box and are not constrained to stay in one location.

To correct any visual problems observed during simulation, commands from *GROMACS* was used to process the trajectory files after simulation ([GitLab](#)).

To perform the imaging, an automated protocol provided and designed by Daniel Beltrán from the Molecular Modeling and Bioinformatics research group was used. This protocol is designed to be able to image different systems, which in the case of this project, are membrane protein systems. This protocol needs as input parameters a trajectory and a structure file. This will be the ones obtained from the results of the simulation, the trajectory and structure files, as well as a file which contains information about the trajectory; information useful for the analysis and information about the protein structure that is being simulated.

To ensure that the system (membrane + protein) is centered in the middle of the box and that atoms do not jump the borders of the box, several *GROMACS* commands were used. This can be found in the bash scripts of the [GitLab](#) repository. It is worth mentioning that these commands were used in comparison to others because they gave the best results in terms of imaging and subsequent analysis. When analysing the trajectories, it was seen that some of them did not pass the quality control due to a bad imaging. That is why several tests were done in order to find which imaging parameters were better.

From all the resulting files of this imaging process, the ones of interest to assess if the imaging process was well done are a structure file with no water, as it requires less computational power and a trajectory file which stores the imaged trajectory.

Moreover, to better visualize the protein trajectory, a process of fitting was done. The aim of this method is to rotate and translate the protein so the protein is always fixed and at the center of the box. This was done with a bash script which iterates for all finished imaged trajectories and selects for which molecule the fitting is wanted (protein), which will be the output system (all system) and does a rotation and translation of the protein. Doing the fitting is not always recommended to membrane protein systems because it leads to sudden membrane movements. However, with the systems provided with *MemProtMD* this does not happen and the result is optimum.

After the processing of the trajectory, it is important to analyse molecular dynamics trajectories to ensure that the results obtained are accurate. If they are not accurate, the results obtained from them can be unreliable, which can lead to erroneous conclusions. Two types of calculations can be done: basic analysis and specialized analysis.

#### 5.4.2. Basic analysis of the simulations

To assess trajectory quality and check whether the imaged trajectory is physically and visually coherent, three basic analyses were done with the same automated protocol used for imaging as it includes also these analyses:

- Coherent bonds: it checks that each atom has a number of expected and physically possible bonds, for example, no hydrogen atom is expected to have more than one covalent bond.
- Stable bonds: it finds which are the stable bonds of the trajectory and checks that the trajectory contains them.
- Trajectory integrity: it checks that there are no abnormally sharp jumps in the positions of the atoms at any point along the trajectory (as **Figure 24** shows). It is checked with the root mean square distance between two consecutive frames.

When running this analysis, it must be considered that trajectories whose structure file does not contemplate hydrogens in the membrane, as some in *MemProtMD*, will generate an error in the coherent bonds test, as some bonds are missing and it is considered as a 'type of coarse-grained' model. For this reason, if the membrane does not have hydrogens, the coherent bonds test is skipped.

Once the trajectory passes these analyses, it can be considered that it has been correctly calculated and imaged.

When doing these analyses for the 51 simulations, some of them did not pass them. Most of those that did not pass was due to a wrong imaging and this is why the parameters of imaging were changed as commented before, and these simulations were imaged again. Moreover, two failed due to internet connection, as the analyses requires it. Of these two, one was repaired but the other one, due to the time limit, was not.

#### 5.4.3. Specialized analyses of the simulations

To have further information about the trajectory which can be of help to extract biological conclusions, several specialized analyses were done with the same automated protocol. This include RMSD, RMSD per residue, RMSD pairwise, Radius of gyration, Fluctuation, Principal Component Analyses and Solvent accessible surface. These analyses will be available in the *MDposit* platform.

#### 5.5. Inclusion of the results in the *MDposit* platform

To upload the resulting imaged and fitted trajectory and its analyses to the *MDposit* platform, one must jump via ssh into a remote machine located in the BSC. This machine grants access to the platform, enabling the upload of all the data. Moreover, to be able to do this, a software proportionated by the research group where this project has been carried out was used.

The URLs for the resulting 50 simulations which this project calculated and uploaded to *MDposit* can be found in the **Appendix**. An example of one protein structure simulation uploaded is displayed in **Figure 25** and **Figure 26**. When the simulations are uploaded, a visual inspection is done to check that everything has been uploaded correctly. From this visual inspection, it was seen that the simulation of the 5JQH *PDB* was incorrect as a lipid floating in the environment was found. Therefore, this simulation was removed from the platform.

The screenshot shows the MDposit website interface. At the top, there is a navigation bar with 'MDposit' logo and links for HOME, BROWSE, SEARCH, HELP, CONTACT, and REST API. Below this, a secondary navigation bar includes OVERVIEW, TRAJECTORY, ANALYSES, and DOWNLOADS. The main content area is titled '645cb1559bc81a1b4e777253 - Overview' and includes a 'DATA IN THIS PAGE' link. The text describes a molecular dynamics simulation of the crystal structure of A2AAR-BRIL in complex with the antagonist ZM241385, produced from *Pichia pastoris*. It specifies the simulation type as 'Trajectory' and 'Classical MD'. The authors are listed as Isabel Martin, and the groups are Orozco Lab, IRB Barcelona. The program used is GROMACS, version 2022.3. A link to the structural data source is provided. Below this, the 'Adenosine receptor A2a' section lists the gene (ADORA2A), organism (Homo sapiens), and UniProt ID (P22974). The PDB accession is 6aqf, and the crystal structure is described as 'Crystal structure of A2AAR-BRIL in complex with the antagonist ZM241385 produced from Pichia pastoris'. The experimental method is x-ray. Organisms listed are Homo sapiens and Escherichia coli. The keyword is 'Membrane protein'. A 3D ribbon diagram of the protein structure is shown on the right.

**Figure 25.** Simulation uploaded of Arginase-1 (PDB ID: 6QAF, UniProt ID: P22974). Overview section.

The screenshot shows the MDposit website interface for the 'Trajectory' section of simulation 645cb1559bc81a1b4e777253. The navigation bar at the top is the same as in Figure 25. The main content area is titled '645cb1559bc81a1b4e777253 - Trajectory' and includes a 'DATA IN THIS PAGE' link. The 'Domains' section shows 'Overall' selected. The main visualization area displays a 3D ribbon diagram of the protein structure, colored in red and yellow, set against a grey mesh background representing the simulation box. Below the visualization is a video player interface with a progress bar, play/pause buttons, and a settings icon.

**Figure 26.** Simulation uploaded of Arginase-1 (PDB ID: 6QAF, UniProt ID: P22974). Trajectory section.

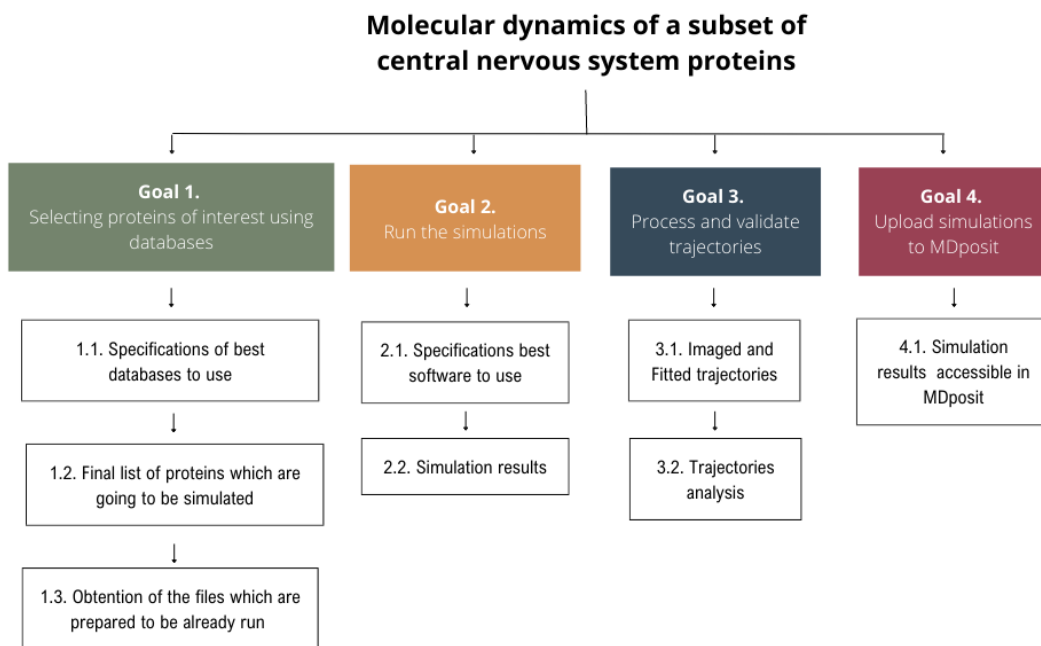


# 6.

## Execution chronogram

### 6.1. Work Breakdown Structure

The Work Breakdown Structure (WBS) is a hierarchical, itemized representation of the major components of a project. It decomposes the project by tasks around the goals initially set. **Figure 27** shows the different tasks for each goal that compose the project.



**Figure 27.** Work Breakdown Structure for this project.

Each task has a due date in order to be able to complete the project on time, in a total of 8 months. Moreover, the initiation of some of these tasks relies on the total / partially completion of others.

**Table 15** shows the due date and the dependencies of each task.

**Table 15.** Table of tasks and their due date and dependencies.

Task No	Task Name	Goal	Type	Due Date (month)	Previous task
1.1	Specifications of best databases to use	G1	Document – Conceptual Engineering	1	-

1.2	Final list of proteins which are going to be simulated	G1	UniProt ID and PDB ID lists	3	1.1
1.3	Obtention of the files prepared to be run	G1	Zip files	3	1.2
2.1	Specifications of best software to use	G2	Document – Conceptual Engineering	5	-
2.2	Simulation results	G2	Trajectory file	5	2.1, 1.3
3.1	Imaged and fitted trajectories	G3	Trajectory files	6	2.2
3.2	Trajectories analysis	G3	Json and trajectory files	6	3.1
4.1	Simulation results accessible in <i>MDposit</i>	G4	Web interface	7	3.2

## 6.2. Tasks specifications

The following tables display which subtasks have to be done for completing the tasks.

**Table 16.** Subtasks for task 1.1

<b>Task number</b>	1.1	<b>Goal number</b>	G1
<b>Task name</b>	Specifications of best databases to use.		
<b>Type</b>	Document – Conceptual Engineering	<b>Due month</b>	1
<b>List of subtasks</b>			
<ul style="list-style-type: none"> <li>- Read literature to be informed which are the most common ones.</li> <li>- Meeting with Structural Bioinformatics and Network Biology IRB research group.</li> </ul>			

**Table 17.** Subtasks for task 1.2

<b>Task number</b>	1.2	<b>Goal number</b>	G1
<b>Task name</b>	Final list of proteins which are going to be simulated.		
<b>Type</b>	UniProt ID and PDB ID lists	<b>Due month</b>	3
<b>List of subtasks</b>			
<ul style="list-style-type: none"> <li>- Data analysis and manipulation of different databases via Python language.</li> <li>- Understand data.</li> <li>- Meeting with University of Oxford to inform them about the process of obtention of the list, share data to help add new membrane protein simulations relative to CNS in <i>MemProtMD</i>.</li> </ul>			

**Table 18.** Subtasks for task 1.3.

<b>Task number</b>	1.3	<b>Goal number</b>	G1
<b>Task name</b>	Obtention of the files prepared to be run.		
<b>Type</b>	Zip files	<b>Due month</b>	3
<b>List of subtasks</b>			
<ul style="list-style-type: none"> <li>- Download prepared files from <i>MemProtMD</i>.</li> </ul>			

**Table 19.** Subtasks for task 2.1

<b>Task number</b>	2.1	<b>Goal number</b>	G1
<b>Task name</b>	Specifications of best software to use		
<b>Type</b>	Document – Conceptual Engineering	<b>Due month</b>	6

List of subtasks	
-	Know which software is better to use to run the obtained files.
-	Different tests to see which software version and HPC resource is better to use.

Table 20. Subtasks for task 2.2.

Task number	2.2	Goal number	G2
Task name	Specifications of best databases to use.		
Type	Trajectory file (.xtc)	Due month	6
List of subtasks			
-	Program environment.		
-	Program different bash scripts to automatically run the simulations.		
-	Wait for the simulations to finish running.		
-	Check if warnings were raised and fatal errors stopped the running of the simulation.		

Table 21. Subtasks for task 3.1.

Task number	3.1	Goal number	G3
Task name	Imaged and fitted trajectories		
Type	Trajectory files	Due month	7
List of subtasks			
-	Test which imaging parameter is better to use for these trajectories.		
-	Run workflow of imaging provided by Daniel Beltrán from Molecular Modeling and Bioinformatics IRB research group.		
-	Do the fitting of the previous obtained file.		
-	Visual inspection of the resulting processed trajectories.		

Table 22. Subtasks for task 3.2.

Task number	3.2	Goal number	G3
Task name	Trajectories analysis.		
Type	Json and trajectory files	Due month	7
List of subtasks			
-	Run workflow of analysis provided by Daniel Beltrán from Molecular Modeling and Bioinformatics IRB research group.		
-	Understand why some trajectories do not pass the analysis and tune the parameters of the workflow and imaged process.		

Table 23. Subtasks for task 4.1.

Task number	4.1	Goal number	G4
Task name	Simulation results accessible in <i>MDposit</i> .		
Type	Web interface	Due month	8
List of subtasks			
-	Upload one by one the obtained simulation results in <i>MDposit</i> .		
-	Visual inspection that everything has been correctly uploaded and repair if something has not.		

### 6.3. GANTT diagram

**Figure 28** shows the GANTT diagram for the project. It displays which is the time expected to be used to finish a task and when it has to be done. It should be noted that it is possible to start a task which depends on other although the prior task has not been fully completed. For instance, consider the imaging task, where simulations are needed for this task. It is not necessary for all simulations to be completed before initiating the imaging process. Instead, as soon as some of the simulations are finished, the imaging process for those completed simulations can commence. This approach allows for a more efficient workflow, as it enables parallel execution of tasks.

It also should be noted that although the due date for work package 1 is M3 because results have to be shown to UOXF, after the meeting some corrections are expected to be done.

Given that the GANTT diagram already illustrates the timelines and dependencies of each task, a CPM/PERT diagram was omitted because it could be redundant.

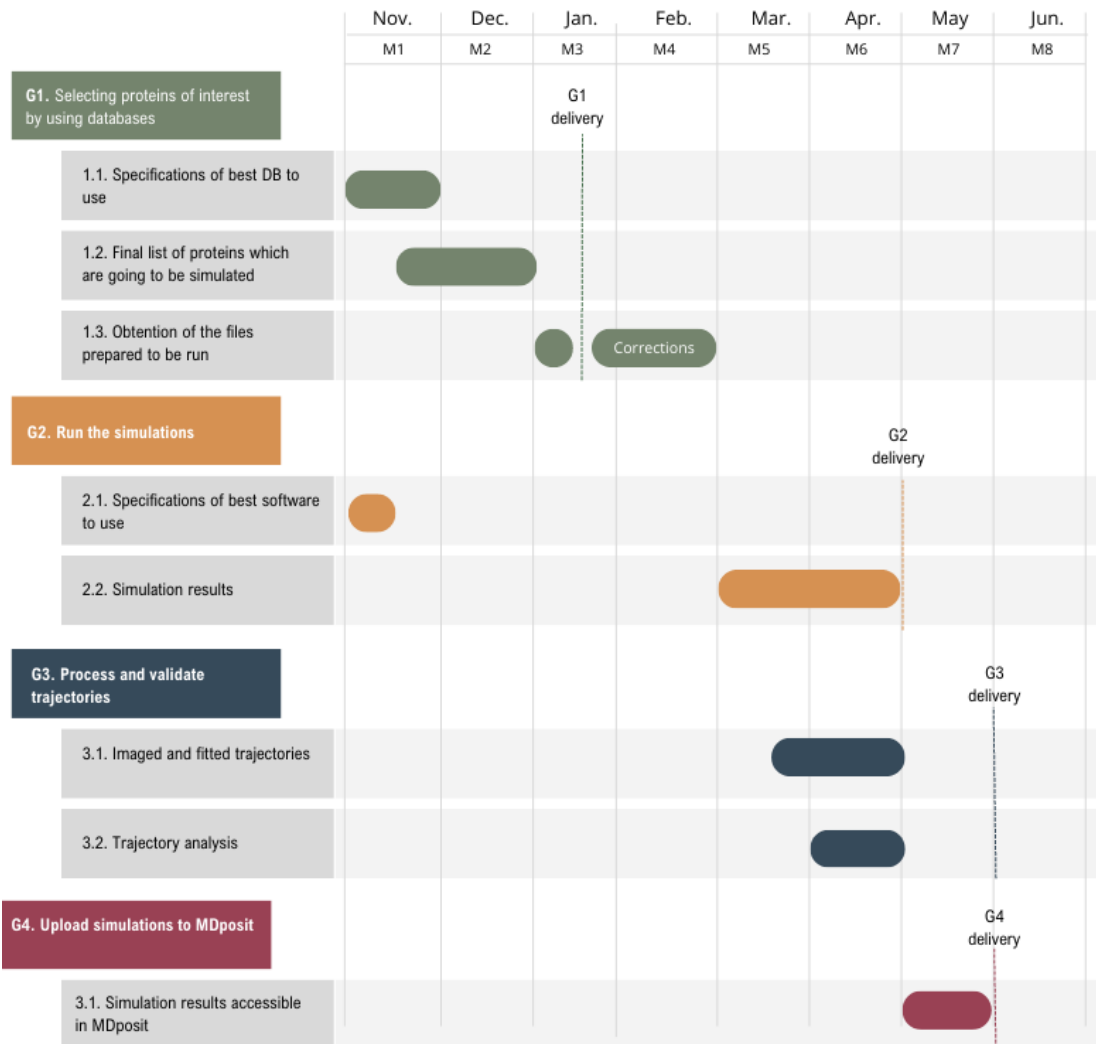
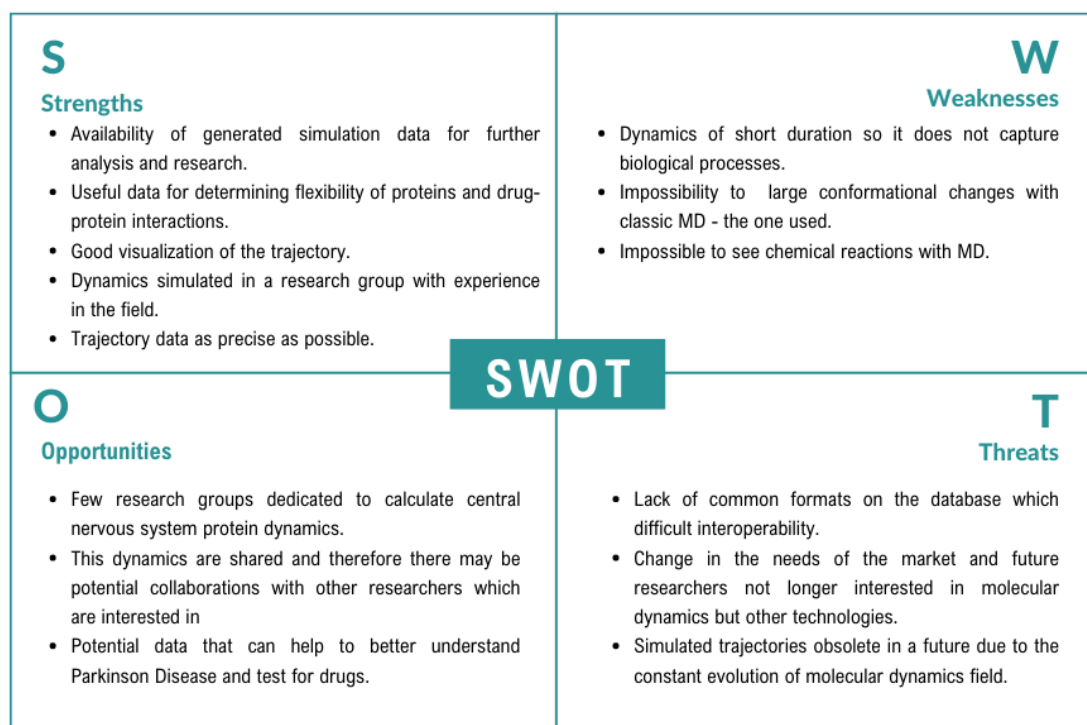


Figure 28. GANTT diagram.

# 7.

## Technical viability

This section deals with the analysis of the technical feasibility of carrying out a molecular dynamics project, in which some central nervous system protein dynamics simulations have been calculated and the results have been loaded into a database (**Figure 29**). In this context, SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis will be used as a strategic approach to evaluate the internal and external elements that can influence the project's technical feasibility and its resulting data.



**Figure 29.** SWOT analysis of the resulting protein dynamics and their upload to MDposit platform.

### Strengths

The dynamics simulated have been carried out in a research group which already has a lot of expertise in this field. This is a clear strength when talking about obtaining data as it mostly ensures that it is coherent as it is known which analysis must be made, what it needs to be observed and

what needs to be improved. Moreover, this central nervous system protein dynamics can be useful in other fields such as the study of drug-protein interactions. And this can be done by other research groups more specialized because all the generated data is published in the *MDposit* database, where it can be downloaded.

### **Weakness**

The dynamics simulated in this project are of short duration, of 100 ns. Certain biological processes such as protein folding occur on a much larger timescale, typically in the order of milliseconds. Therefore, the calculated dynamics would not be able to show these processes. Moreover, there is also an impossibility of seeing large conformational changes or allosteric processes with a single classic molecular dynamic. It would be necessary to move to more advanced methods such as enhanced sampling methods and take into account variations in structure such as the addition of possible ligand and ions. Finally, it is also impossible to see chemical reactions (e.g., proton transfer) with MD.

### **Opportunities**

These simulations can be well received in the computational / experimental scientific fields as they provide useful information and there are few research groups dedicated to obtaining by molecular dynamics the movement of central nervous system proteins which can be to understand Parkinson Disease and test for new drugs. As these dynamics are shared for future analysis, they may lead to future potential collaborations with different research groups.

### **Threats**

The external limitations or obstacles that can stop the future use of these calculated molecular dynamics simulations are, first a lack of interoperability between the *MDposit* database and other databases, a change in the needs in the scientific field which will make these dynamics unwanted; and becoming obsolete trajectories due to the constant evolution of molecular dynamics field.

# 8.

## Economic viability

The successful implementation of any project generally involves the allocation of financial resources to cover the associated costs. In the case of this particular project, an economic investment is also required to ensure its proper development.

A detailed description of the prices required to carry out all stages of the project from its inception is provided in the following lines. This involves identifying and calculating the different elements and activities required, such as the acquisition of materials, hiring of personnel, production costs, operating expenses, among others.

The purpose of this information is to provide a clear and transparent view of the economic resources required for the execution of the project from scratch. These economic resources are specified for each goal described in the GANTT diagram in section 6. *Execution chronogram* and are displayed in **Table 24**.

**Table 24.** *Economic viability of the project.*

	Resources needed	Price (€)	Total price (€)
<b>G1. Selecting proteins of interest by using databases</b>			
1.1. Specifications of best DB to use	Access to documentation	~ 100	2,100
	Computer	~ 2,000	
1.2. Final list of proteins which are going to be simulated	Computer with large storage disk to store DBs downloaded	Already considered	0
	Jupyter notebook	0	
1.3. Obtention of the files prepared to be run	Computer	Already considered	
<b>G2. Run the simulations</b>			
	Access to documentation	~ 100	100

2.1. Specifications of best software to use	Computer	Already considered	
2.2. Simulation results	Supercomputer StarLife	Total core-hours: 480,000 → Total of core-hours in StarLife for the realization of this project. Data proportioned by BSC-CNS support. Price/core-hour: 0.25 → Data proportioned by BSC-CNS support	120·10 <sup>3</sup>
	GROMACS	0	
<b>G3. Process and validate trajectories</b>			
3.1. Imaged and fitted trajectories	GROMACS	0	
	IRB workflow	Internal resource	
3.2. Trajectory analysis	GROMACS	0	
	IRB workflow	Internal resource	
<b>G4. Upload simulations to MDposit</b>			
4.1. Simulation results accessible in MDposit	Supercomputer StarLife	Already considered	133.33 (8 months)
	IRB software	Internal resource	
	Hosting and maintenance of the database	200 / year → Data proportioned by IRB.	
Human resources	27,725 / year → Average salary of a researcher in Spain <sup>[78]</sup> .		18,480 (8 months)
<b>TOTAL</b>			<b>140,813</b>

The above table presents the fundamental resources required to estimate the approximate amount of money needed to carry out this project from the beginning. Taking into account this project has lasted a total of 8 months, the total cost of the project from scratch is approximately 140,813 €.

The use of HPC can justify this high price. HPC are often expensive to acquire and maintain as this resource needs significant amount of inversion for hardware components such as numerous CPUs, cooling systems and expert people to correctly administrate these systems. Therefore, the high price of the project is considered feasible within the typical ranges associated with the use of HPC.



# 9.

## Legal aspects

In the context of legal aspects, it is important to note that the code developed in this project has an open-source license.

The specific license used for the code in this project is an Apache 2.0, which is available in the [GitLab](#) repository. An Apache 2.0 license ensures the security and availability of the developed open-source code <sup>[79]</sup>. This license enables software developers to access, modify, update or distribute the code. In this way, this project also contributes to future work as it provides the opportunity to modify the filtering conditions of the list of proteins to generate new datasets. Additionally, it offers the necessary files to easily execute their dynamics.

Likewise, it is important to mention that this project does not use data from patient at any stage of its development. The data employed in this project has been obtained from databases that collect information from various sources such as academic research or published studies. This data does not pertain to specific patients but rather encompasses generalizations.

# 10.

## Conclusions and future work

- This project has successfully obtained the molecular dynamics simulations pre-processed and validated of a subset of proteins of the central nervous system.
- For doing this, all the sub-objectives established in the Introduction section, such as using data analysis of databases to obtain the subset of study, have been accomplished.
- By uploading these simulations in the open platform *MDposit*, further analysis of this simulations can be done by other research groups in order to better understand their function or study possible drug binding on them as they are targetable proteins.
- Due to time limitations, this project could not simulate all the protein structures of the subset of interest but some of them. Future work will involve running the simulations of these remaining proteins. Therefore, this project has not only fulfilled its own objectives, but also set the basis for further developments.
- When running the simulations, some trajectories raise some errors in the analysis and visual inspection. Future work will also involve understanding and correcting them.
- Moreover, the simulations that have been run are of 100 ns. It would be of interest also to expand this simulation time to 1 us to have a better overview of the dynamics of the proteins.

# 11.

## Bibliography

1. *The Human Brain Project A Report to the European Commission*. (n.d.).
2. Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., Carrillo, O., Camps, J., Fenollosa, C., Repchevsky, D., Gelpí, J. L., & Orozco, M. (2010). MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure*, *18*(11), 1399–1409. <https://doi.org/10.1016/j.str.2010.07.013>
3. Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. In *Neuron* (Vol. 99, Issue 6, pp. 1129–1143). Cell Press. <https://doi.org/10.1016/j.neuron.2018.08.011>
4. *Molecular Dynamics Data Bank. The European Repository for Biosimulation Data | MDDB Project | Fact Sheet | HORIZON | CORDIS | European Commission*. (n.d.). Retrieved June 3, 2023, from <https://cordis.europa.eu/project/id/101094651>
5. *Molecular Modelling and Bioinformatics | IRB Barcelona*. (n.d.). Retrieved June 3, 2023, from <https://www.irbbarcelona.org/es/research/molecular-modelling-and-bioinformatics>
6. *Structural Bioinformatics and Network Biology | IRB Barcelona*. (n.d.). Retrieved June 4, 2023, from <https://www.irbbarcelona.org/es/research/structural-bioinformatics-and-network-biology>
7. *Prof Phil Biggin | Biochemistry*. (n.d.). Retrieved June 3, 2023, from <https://www.bioch.ox.ac.uk/research/biggin>
8. *What Is Data Analysis? (With Examples) | Coursera*. (n.d.). Retrieved June 3, 2023, from <https://www.coursera.org/articles/what-is-data-analysis-with-examples>
9. *What is High Performance Computing | NetApp*. (n.d.). Retrieved June 3, 2023, from <https://www.netapp.com/data-storage/high-performance-computing/what-is-hpc/>
10. Hospital, A., Goñi, J. R., Orozco, M., & Gelpí, J. L. (2015). Molecular dynamics simulations: Advances and applications. In *Advances and Applications in Bioinformatics and Chemistry* (Vol. 8, Issue 1, pp. 37–47). Dove Medical Press Ltd. <https://doi.org/10.2147/AABC.S70333>

11. *Atomistic models - Latest research and news | Nature*. (n.d.). Retrieved June 3, 2023, from <https://www.nature.com/subjects/atomistic-models>
12. Kmiecik, S., Kouza, M., Badaczewska-Dawid, A. E., Kloczkowski, A., & Kolinski, A. (2018). Modeling of protein structural flexibility and large-scale dynamics: Coarse-grained simulations and elastic network models. *International Journal of Molecular Sciences*, 19(11). <https://doi.org/10.3390/IJMS19113496>
13. *Bonded interactions - GROMACS 2023.1 documentation*. (n.d.). Retrieved June 3, 2023, from <https://manual.gromacs.org/current/reference-manual/functions/bonded-interactions.html>
14. *Force Fields and Interactions – Practical considerations for Molecular Dynamics*. (n.d.). Retrieved June 3, 2023, from [https://computecanada.github.io/molmodsim-md-theory-lesson-novice/01-Force\\_Fields\\_and\\_Interactions/index.html](https://computecanada.github.io/molmodsim-md-theory-lesson-novice/01-Force_Fields_and_Interactions/index.html)
15. Martin, M. G. (2006). Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor–liquid coexistence curves and liquid densities. *Fluid Phase Equilibria*, 248(1), 50–55. <https://doi.org/10.1016/J.FLUID.2006.07.014>
16. Topal, B. (n.d.). *Computational Study of the Structure and Dynamics of Androgen Receptor Polyglutamine Tract*. [www.tdx.cat](http://www.tdx.cat)
17. *Water models – Practical considerations for Molecular Dynamics*. (n.d.). Retrieved June 3, 2023, from [https://computecanada.github.io/molmodsim-md-theory-lesson-novice/09-water\\_models/index.html](https://computecanada.github.io/molmodsim-md-theory-lesson-novice/09-water_models/index.html)
18. *Setting up a Molecular Dynamics simulation - Compchems*. (n.d.). Retrieved June 3, 2023, from <https://www.compchems.com/setting-up-a-molecular-dynamics-simulation/#system-preparation>
19. *Molecular Simulation/Periodic Boundary Conditions - Wikibooks, open books for an open world*. (n.d.). Retrieved June 3, 2023, from [https://en.wikibooks.org/wiki/Molecular\\_Simulation/Periodic\\_Boundary\\_Conditions](https://en.wikibooks.org/wiki/Molecular_Simulation/Periodic_Boundary_Conditions)
20. *Periodic Boundary Conditions – Practical considerations for Molecular Dynamics*. (n.d.). Retrieved June 3, 2023, from [https://computecanada.github.io/molmodsim-md-theory-lesson-novice/04-Periodic\\_Boundary/index.html](https://computecanada.github.io/molmodsim-md-theory-lesson-novice/04-Periodic_Boundary/index.html)
21. *Solvating a System, Adding Ions and Generating Input Files – Running Molecular Dynamics on Alliance clusters with AMBER*. (n.d.). Retrieved June 7, 2023, from [https://computecanada.github.io/molmodsim-amber-md-lesson/12-Adding\\_Ions/index.html](https://computecanada.github.io/molmodsim-amber-md-lesson/12-Adding_Ions/index.html)
22. Ge, Y., Calabró, G., Bayly, C. I., & Mobley, D. L. (2021). *Sensitivity of Molecular Dynamics Simulations to Equilibration Scheme: A Case Study of Bromodomain*

- Protein BRD4-Ligand Complex System*. <https://doi.org/10.26434/CHEMRXIV-2021-DGZRT>
23. Salo-Ahen, O. M. H., Alanko, I., Bhadane, R., Alexandre, A. M., Honorato, R. V., Hossain, S., Juffer, A. H., Kabedev, A., Lahtela-Kakkonen, M., Larsen, A. S., Lescrinier, E., Marimuthu, P., Mirza, M. U., Mustafa, G., Nunes-Alves, A., Pantsar, T., Saadabadi, A., Singaravelu, K., & Vanmeert, M. (2020). Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development. *Processes* 2021, Vol. 9, Page 71, 9(1), 71. <https://doi.org/10.3390/PR9010071>
  24. *Learn About Target Protein | Chegg.com*. (n.d.). Retrieved June 3, 2023, from <https://www.chegg.com/learn/topic/target-protein>
  25. Ciccotti, G., Dellago, C., Ferrario, M., Hernández, E. R., & Tuckerman, M. E. (2022). Molecular simulations: past, present, and future (a Topical Issue in EPJB). *The European Physical Journal B* 2021 95:1, 95(1), 1–12. <https://doi.org/10.1140/EPJB/S10051-021-00249-X>
  26. *molecular dynamics biology - Search Results - PubMed*. (n.d.). Retrieved June 3, 2023, from <https://pubmed.ncbi.nlm.nih.gov/?term=molecular+dynamics+biology>
  27. Lazim, R., Suh, D., & Choi, S. (2020). Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems. *International Journal of Molecular Sciences* 2020, Vol. 21, Page 6339, 21(17), 6339. <https://doi.org/10.3390/IJMS21176339>
  28. *Molecular Modelling Market Size To Worth \$ 7.8 Bn by 2030*. (n.d.). Retrieved June 3, 2023, from <https://www.sphericalinsights.com/press-release/molecular-modelling-market>
  29. Rodríguez-Espigares, I., Torrens-Fontanals, M., Tiemann, J. K. S., Aranda-García, D., Ramírez-Anguita, J. M., Stepniewski, T. M., Worp, N., Varela-Rial, A., Morales-Pastor, A., Medel-Lacruz, B., Pándy-Szekeres, G., Mayol, E., Giorgino, T., Carlsson, J., Deupi, X., Filipek, S., Filizola, M., Gómez-Tamayo, J. C., Gonzalez, A., ... Selent, J. (2020). GPCrmd uncovers the dynamics of the 3D-GPCRome. *Nature Methods*, 17(8), 777–787. <https://doi.org/10.1038/S41592-020-0884-Y>
  30. *Simulations*. (n.d.). Retrieved June 3, 2023, from <https://www.humanbrainproject.eu/en/science-development/focus-areas/simulations/>
  31. Bittrich, S., Rose, Y., Segura, J., Lowe, R., Westbrook, J. D., Duarte, J. M., & Burley, S. K. (2022). RCSB Protein Data Bank: Improved annotation, search and visualization of membrane protein structures archived in the PDB. *Bioinformatics*, 38(5), 1452–1454. <https://doi.org/10.1093/BIOINFORMATICS/BTAB813>
  32. *A-schematic-illustration-of-a-typical-chemical-synapse-created-with-BioRendercom.png (680x476)*. (n.d.). Retrieved June 7, 2023, from <https://www.researchgate.net/profile/Shuang-Gao>

- 12/publication/362568402/figure/fig3/AS:1187035071164428@1660022623803/A-schematic-illustration-of-a-typical-chemical-synapse-created-with-BioRendercom.png
33. Sheils, T., Mathias, S. L., Siramshetty, V. B., Bocci, G., Bologna, C. G., Yang, J. J., Waller, A., Southall, N., Nguyen, D. T., & Oprea, T. I. (2020). How to Illuminate the Druggable Genome using Pharos. *Current Protocols in Bioinformatics*, 69(1), e92. <https://doi.org/10.1002/CPBI.92>
  34. *Small molecules*. (n.d.). Retrieved June 3, 2023, from <https://www.astrazeneca.com/r-d/next-generation-therapeutics/small-molecule.html>
  35. Herrera, F. E. (2008). *Computational approaches to the investigation of proteins involved in Parkinson's Disease*. SISSA. <http://hdl.handle.net/20.500.11767/4656>
  36. *El párkinson, una pandemia en 2040 - Federación Española de Parkinson*. (n.d.). Retrieved June 3, 2023, from <https://www.esparkinson.es/parkinson-pandemia-2040/>
  37. *The synapse (article) | Human biology | Khan Academy*. (n.d.). Retrieved June 3, 2023, from <https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/the-synapse>
  38. Xu, D. (n.d.). *Protein Databases on the Internet*. <https://doi.org/10.1002/0471142727.mb1904s68>
  39. *About UniProt | UniProt help | UniProt*. (n.d.). Retrieved June 3, 2023, from <https://www.uniprot.org/help/about>
  40. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L.-S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(Database issue), D115. <https://doi.org/10.1093/NAR/GKH131>
  41. *UniProtKB | UniProt help | UniProt*. (n.d.). Retrieved June 3, 2023, from <https://www.uniprot.org/help/uniprotkb>
  42. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/NAR/28.1.235>
  43. *RCSB PDB - 3PBL: Structure of the human dopamine D3 receptor in complex with eticlopride*. (n.d.). Retrieved June 7, 2023, from <https://www.rcsb.org/structure/3pbl>

44. *Why do I find many cross-references to PDB in UniProtKB/Swiss-Prot? | UniProt help | UniProt.* (n.d.). Retrieved June 3, 2023, from [https://www.uniprot.org/help/multiple\\_pdb\\_xrefs](https://www.uniprot.org/help/multiple_pdb_xrefs)
45. *ADORA2A - Adenosine receptor A2a - Homo sapiens (Human) | UniProtKB | UniProt.* (n.d.). Retrieved June 7, 2023, from <https://www.uniprot.org/uniprotkb/P29274/entry#expression>
46. Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., ... Brazma, A. (2020). Expression Atlas update: from tissues to single cells. *Nucleic Acids Research*, *48*(D1), D77–D83. <https://doi.org/10.1093/NAR/GKZ947>
47. Bastian, F. B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S. S., de Farias, T. M., Moretti, S., Parmentier, G., de Laval, V. R., Rosikiewicz, M., Wollbrett, J., Echchiki, A., Escoriza, A., Gharib, W. H., Gonzales-Porta, M., Jarosz, Y., Laurency, B., Moret, P., Person, E., ... Robinson-Rechavi, M. (2021). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, *49*(D1), D831. <https://doi.org/10.1093/NAR/GKAA793>
48. *Ontology Lookup Service < EMBL-EBI.* (n.d.). Retrieved June 3, 2023, from <https://www.ebi.ac.uk/ols/index>
49. Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., Huang, J., Du, Y., Lin, L., Dong, Z., Yang, L., Liu, X., Jiang, H., Xu, X., Wang, J., ... Mulder, J. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science*, *367*(6482). <https://doi.org/10.1126/SCIENCE.AAY4106>
50. Palasca, O., Santos, A., Stolte, C., Gorodkin, J., & Jensen, L. J. (2018). TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database: The Journal of Biological Databases and Curation*, *2018*(2018). <https://doi.org/10.1093/DATABASE/BAY003>
51. Shimizu, K., Cao, W., Saad, G., Shoji, M., & Terada, T. (2018). Comparative analysis of membrane protein structure databases. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1860*(5), 1077–1091. <https://doi.org/10.1016/J.BBAMEM.2018.01.005>
52. Newport, T. D., Sansom, M. S. P., & Stansfeld, P. J. (2019). The MemProtMD database: A resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Research*, *47*(D1), D390–D397. <https://doi.org/10.1093/NAR/GKY1047>
53. Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., & Zhang, Y. (2016). Drug–target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics*, *17*(4), 696–712. <https://doi.org/10.1093/BIB/BBV066>

54. *A Review of Target Identification Strategies for Drug Discovery: from Database to Machine-Based Methods.* (n.d.). <https://doi.org/10.1088/1742-6596/1893/1/012013>
55. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. P. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, *43*(Web Server issue), W612. <https://doi.org/10.1093/NAR/GKV352>
56. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., Baker, J., Ferrer, J., Raies, A., Razuvayevskaya, O., Faulconbridge, A., Petsalaki, E., Mutowo, P., MacHlitt-Northen, S., Peat, G., ... McDonagh, E. M. (2021). Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research*, *49*(D1), D1302–D1310. <https://doi.org/10.1093/NAR/GKAA1027>
57. Sheils, T. K., Mathias, S. L., Kelleher, K. J., Siramshetty, V. B., Nguyen, D. T., Bologa, C. G., Jensen, L. J., Vidović, D., Koletić, A., Schürer, S. C., Waller, A., Yang, J. J., Holmes, J., Bocci, G., Southall, N., Dharkar, P., Mathé, E., Simeonov, A., & Oprea, T. I. (2021). TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Research*, *49*(D1), D1334–D1346. <https://doi.org/10.1093/NAR/GKAA993>
58. Nguyen, D. T., Mathias, S., Bologa, C., Brunak, S., Fernandez, N., Gaulton, A., Hersey, A., Holmes, J., Jensen, L. J., Karlsson, A., Liu, G., Ma'ayan, A., Mandava, G., Mani, S., Mehta, S., Overington, J., Patel, J., Rouillard, A. D., Schürer, S., ... Guha, R. (2017). Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Research*, *45*(D1), D995–D1002. <https://doi.org/10.1093/NAR/GKW1072>
59. Adcock, S. A., & McCammon, J. A. (2006). Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, *106*(5), 1589–1615. [https://doi.org/10.1021/CR040426M/ASSET/CR040426M.FP.PNG\\_V03](https://doi.org/10.1021/CR040426M/ASSET/CR040426M.FP.PNG_V03)
60. Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, *99*(6), 1129–1143. <https://doi.org/10.1016/J.NEURON.2018.08.011>
61. *Modeling and Simulation Software.* (n.d.). Retrieved June 3, 2023, from <https://www.rcsb.org/docs/additional-resources/modeling-and-simulation-software>
62. Salomon-Ferrer, R., Case, D. A., & Walker, R. C. (2012). An overview of the Amber biomolecular simulation package. *WIREs Comput Mol Sci*. <https://doi.org/10.1002/wcms.1121>
63. *CHARMM | CHARMM.* (n.d.). Retrieved June 3, 2023, from <https://academiccharmm.org/>
64. *Gromacs | Molecular Dynamics solutions | LinuxVixion.* (n.d.). Retrieved June 3, 2023, from <https://www.linuxvixion.com/gromacs/>



65. *Force fields in GROMACS — GROMACS 2022 documentation*. (n.d.). Retrieved June 3, 2023, from <https://manual.gromacs.org/2022/user-guide/force-fields.html>
66. O'Donoghue, S. I., Goodsell, D. S., Frangakis, A. S., Jossinet, F., Laskowski, R. A., Nilges, M., Saibil, H. R., Schafferhans, A., Wade, R. C., Westhof, E., & Olson, A. J. (2010). Visualization of macromolecular structures. *Nature Methods* 7:3, 7(3), S42–S55. <https://doi.org/10.1038/nmeth.1427>
67. *Visualization Software - GROMACS 2024-dev-20230601-042103d documentation*. (n.d.). Retrieved June 3, 2023, from <https://manual.gromacs.org/documentation/nightly/how-to/visualize.html>
68. *Molecular Graphics Software*. (n.d.). Retrieved June 3, 2023, from <https://www.rcsb.org/docs/additional-resources/molecular-graphics-software>
69. *Bash (Unix shell) - Wikipedia*. (n.d.). Retrieved June 3, 2023, from [https://en.wikipedia.org/wiki/Bash\\_\(Unix\\_shell\)](https://en.wikipedia.org/wiki/Bash_(Unix_shell))
70. *Gene expression calls*. (n.d.). Retrieved June 3, 2023, from <https://www.bgee.org/support/gene-expression-calls#generation-of-presentabsent-expression-calls-per-gene-and-condition>
71. *central nervous system*. (n.d.). Retrieved June 3, 2023, from [https://www.ebi.ac.uk/ols/ontologies/uberont/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FUBERON\\_0001017](https://www.ebi.ac.uk/ols/ontologies/uberont/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FUBERON_0001017)
72. *central nervous system*. (n.d.). Retrieved June 7, 2023, from [https://www.ebi.ac.uk/ols/ontologies/uberont/terms?obo\\_id=UBERON:0001017](https://www.ebi.ac.uk/ols/ontologies/uberont/terms?obo_id=UBERON:0001017)
73. *central nervous system*. (n.d.). Retrieved June 3, 2023, from [https://www.ebi.ac.uk/ols/ontologies/bto/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FBTO\\_0000227](https://www.ebi.ac.uk/ols/ontologies/bto/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FBTO_0000227)
74. *Target - disease associations - Open Targets Platform Documentation*. (n.d.). Retrieved June 3, 2023, from <https://platform-docs.opentargets.org/associations>
75. *Target - disease associations - Open Targets Platform Documentation*. (n.d.). Retrieved June 4, 2023, from <https://platform-docs.opentargets.org/associations>
76. *Slurm – PROTEUS*. (n.d.). Retrieved June 3, 2023, from <https://proteus.ugr.es/docs/slurm/>
77. *Job (computing) - Wikipedia*. (n.d.). Retrieved June 3, 2023, from [https://en.wikipedia.org/wiki/Job\\_\(computing\)](https://en.wikipedia.org/wiki/Job_(computing))
78. *Sueldo: Investigador en España en 2023 | Glassdoor*. (n.d.). Retrieved June 7, 2023, from [https://www.glassdoor.es/Sueldos/investigador-sueldo-SRCH\\_KOO,12.htm](https://www.glassdoor.es/Sueldos/investigador-sueldo-SRCH_KOO,12.htm)

79. *What Does "Apache 2.0 License" Mean? - Planet Crust.* (n.d.). Retrieved June 7, 2023, from [https://www.planetcrust.com/what-does-apache-2-0-license-mean?utm\\_campaign=blog](https://www.planetcrust.com/what-does-apache-2-0-license-mean?utm_campaign=blog)

# Appendix

**Table 25.** Resulting trajectories, each one from one PDB, which have been already uploaded to MDposit

<b>PDBs</b>	<b>Website</b>
1. <b>2K58</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/64414f46d8203151cefd371f/overview">https://MDposit-dev.mddbr.eu/#/id/64414f46d8203151cefd371f/overview</a>
2. <b>2KSR</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644650e4df1b306958ae526d/overview">https://MDposit-dev.mddbr.eu/#/id/644650e4df1b306958ae526d/overview</a>
3. <b>2R4S</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644651c375cb356a864857fe/overview">https://MDposit-dev.mddbr.eu/#/id/644651c375cb356a864857fe/overview</a>
4. <b>2YDO</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6447dbd90fba421040cbf1ef/overview">https://MDposit-dev.mddbr.eu/#/id/6447dbd90fba421040cbf1ef/overview</a>
5. <b>3EML</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6447dc8b388f6326e0d21435/overview">https://MDposit-dev.mddbr.eu/#/id/6447dc8b388f6326e0d21435/overview</a>
6. <b>3KJ6</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6447ddb161110c4a215c91f9/overview">https://MDposit-dev.mddbr.eu/#/id/6447ddb161110c4a215c91f9/overview</a>
7. <b>3NYA</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3a280b169a4e220d884f/overview">https://MDposit-dev.mddbr.eu/#/id/644a3a280b169a4e220d884f/overview</a>
8. <b>3P0G</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3af158458a67899273c9/overview">https://MDposit-dev.mddbr.eu/#/id/644a3af158458a67899273c9/overview</a>
9. <b>3PBL</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3b671c08a47680458b4c/overview">https://MDposit-dev.mddbr.eu/#/id/644a3b671c08a47680458b4c/overview</a>
10. <b>3PDS</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3be445f0eb06436c8b72/overview">https://MDposit-dev.mddbr.eu/#/id/644a3be445f0eb06436c8b72/overview</a>
11. <b>3QAK</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3c5db26b5c135c1edee1/overview">https://MDposit-dev.mddbr.eu/#/id/644a3c5db26b5c135c1edee1/overview</a>
12. <b>3RZE</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3cc87f18871ed442c565/overview">https://MDposit-dev.mddbr.eu/#/id/644a3cc87f18871ed442c565/overview</a>
13. <b>3VG9</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3d2bd2b43f298d1d3940/overview">https://MDposit-dev.mddbr.eu/#/id/644a3d2bd2b43f298d1d3940/overview</a>
14. <b>3VGA</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3d97ca09c034829ccd33/overview">https://MDposit-dev.mddbr.eu/#/id/644a3d97ca09c034829ccd33/overview</a>
15. <b>4COF</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458fc9a39473d49cf4972e9/overview">https://MDposit-dev.mddbr.eu/#/id/6458fc9a39473d49cf4972e9/overview</a>
16. <b>4E1Y</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a3e2d6bf8af4347a808cf/overview">https://MDposit-dev.mddbr.eu/#/id/644a3e2d6bf8af4347a808cf/overview</a>
17. <b>4GBR</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a63ec05c7f3693df9d870/overview">https://MDposit-dev.mddbr.eu/#/id/644a63ec05c7f3693df9d870/overview</a>
18. <b>4MQS</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a64621c907e7b16025d4f/overview">https://MDposit-dev.mddbr.eu/#/id/644a64621c907e7b16025d4f/overview</a>
19. <b>4MQT</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a662d9ebb3531bccaed24/overview">https://MDposit-dev.mddbr.eu/#/id/644a662d9ebb3531bccaed24/overview</a>
20. <b>4UG2</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a66935604bf3da7eee0f5/overview">https://MDposit-dev.mddbr.eu/#/id/644a66935604bf3da7eee0f5/overview</a>
21. <b>4UHR</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a66fde3a94049332ab54b/overview">https://MDposit-dev.mddbr.eu/#/id/644a66fde3a94049332ab54b/overview</a>
22. <b>5OJM</b>	Error in internet connection when doing the analysis.

23.	<b>5CXV</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6762b56e9853f4271cc2/overview">https://MDposit-dev.mddbr.eu/#/id/644a6762b56e9853f4271cc2/overview</a>
24.	<b>5D5B</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a67dee7db1660e391f913/overview">https://MDposit-dev.mddbr.eu/#/id/644a67dee7db1660e391f913/overview</a>
25.	<b>5IU8</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a68726757e26f7d16c337/overview">https://MDposit-dev.mddbr.eu/#/id/644a68726757e26f7d16c337/overview</a>
26.	<b>5IUA</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a69accd84850f07f31483/overview">https://MDposit-dev.mddbr.eu/#/id/644a69accd84850f07f31483/overview</a>
27.	<b>5JQH</b>	Removed. By visual inspection it was seen that the structure was wrong – a lipid was floating.
28.	<b>5JTB</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6a9a568b012676b22d7c/overview">https://MDposit-dev.mddbr.eu/#/id/644a6a9a568b012676b22d7c/overview</a>
29.	<b>5MZP</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6b18a0587f330bd54c66/overview">https://MDposit-dev.mddbr.eu/#/id/644a6b18a0587f330bd54c66/overview</a>
30.	<b>5N2R</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6bbe75f42cb915159/overview">https://MDposit-dev.mddbr.eu/#/id/644a6bbe75f42cb915159/overview</a>
31.	<b>5NLX</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6c680c45895292dd86cb/overview">https://MDposit-dev.mddbr.eu/#/id/644a6c680c45895292dd86cb/overview</a>
32.	<b>5OLO</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6d305358c665108aa97c/overview">https://MDposit-dev.mddbr.eu/#/id/644a6d305358c665108aa97c/overview</a>
33.	<b>5OM1</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6d889fe2986dc2627aa8/overview">https://MDposit-dev.mddbr.eu/#/id/644a6d889fe2986dc2627aa8/overview</a>
34.	<b>5UIG</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6debaedd72776308939e/overview">https://MDposit-dev.mddbr.eu/#/id/644a6debaedd72776308939e/overview</a>
35.	<b>5UVI</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6e4952c87f01bf091029/overview">https://MDposit-dev.mddbr.eu/#/id/644a6e4952c87f01bf091029/overview</a>
36.	<b>5VRA</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/644a6ef008f2ce1105bbab1c/overview">https://MDposit-dev.mddbr.eu/#/id/644a6ef008f2ce1105bbab1c/overview</a>
37.	<b>5WF5</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458be035d8cd640f57f489d/overview">https://MDposit-dev.mddbr.eu/#/id/6458be035d8cd640f57f489d/overview</a>
38.	<b>5WF6</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458be7a0b416a421f0ee152/overview">https://MDposit-dev.mddbr.eu/#/id/6458be7a0b416a421f0ee152/overview</a>
39.	<b>5WIU</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458bede362e6a4327ca1c54/overview">https://MDposit-dev.mddbr.eu/#/id/6458bede362e6a4327ca1c54/overview</a>
40.	<b>5WIV</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458bf913fd7b5443386745b/overview">https://MDposit-dev.mddbr.eu/#/id/6458bf913fd7b5443386745b/overview</a>
41.	<b>5X7D</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458bffaab2b6c4540781084/overview">https://MDposit-dev.mddbr.eu/#/id/6458bffaab2b6c4540781084/overview</a>
42.	<b>5YC8</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/6458e25c8b6c0646fca4807e/overview">https://MDposit-dev.mddbr.eu/#/id/6458e25c8b6c0646fca4807e/overview</a>
43.	<b>5ZK3</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645901330e87044ae5c09842/overview">https://MDposit-dev.mddbr.eu/#/id/645901330e87044ae5c09842/overview</a>
44.	<b>5ZK8</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645cb041ba067e16ecec3e59/overview">https://MDposit-dev.mddbr.eu/#/id/645cb041ba067e16ecec3e59/overview</a>
45.	<b>5ZKB</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645cb0aba835ae180666c357/overview">https://MDposit-dev.mddbr.eu/#/id/645cb0aba835ae180666c357/overview</a>
46.	<b>6A93</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645cb0e47d1e96191cbd6def/overview">https://MDposit-dev.mddbr.eu/#/id/645cb0e47d1e96191cbd6def/overview</a>
47.	<b>6A94</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645cb11591eec41a33e043dc/overview">https://MDposit-dev.mddbr.eu/#/id/645cb11591eec41a33e043dc/overview</a>
48.	<b>6AQF</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645cb1559bc81a1b4e777253/overview">https://MDposit-dev.mddbr.eu/#/id/645cb1559bc81a1b4e777253/overview</a>
49.	<b>6BQG</b>	<a href="https://MDposit-dev.mddbr.eu/#/id/645cb172297c021c5a68a227/overview">https://MDposit-dev.mddbr.eu/#/id/645cb172297c021c5a68a227/overview</a>

- 50. **6BQH** | <https://MDposit-dev.mddbr.eu/#/id/645cb197563b941d6bce49ee/overview>
- 51. **6CM4** | <https://MDposit-dev.mddbr.eu/#/id/645cb1b81917f61e7869e58c/overview>