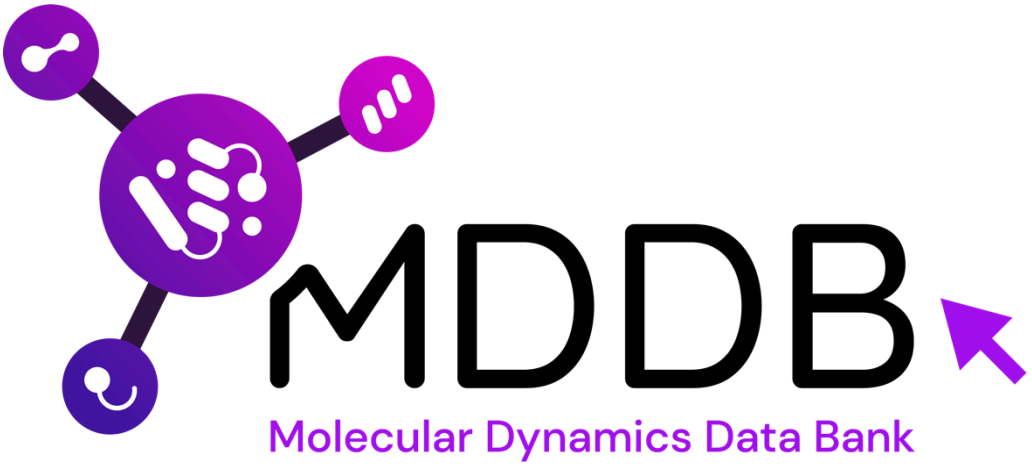
MDDB
Molecular Dynamics Data Bank

Introduction: Protein structure determination has been a revolution in science, with numerous Nobel Prizes awarded both for methods, applications and - in 2024 – prediction methods. However, an equally important challenge is to understand how proteins move, e.g. between alternative states, and there has been similar tremendous progress on computational methods that make it possible to use computer simulations to study atomic motions in molecules.

The single most common such technique is Molecular Dynamics (MD) simulations that first emerged 50 years ago, and is now used by tens of thousands of research groups worldwide and one of the primary users of High-Performance Computers (HPC). The output of these simulations are sequences of conformations of molecules represented as coordinates, and a typical so-called trajectory can contain millions of frames for millions of atoms. Analyzing these collected coordinates enables us to determine the dynamic properties of a molecular system, fueling research projects in biophysics, protein design, de novo antibody and vaccine designs, therapeutic nucleic acids, drug design and even cell or synthetic biology. As an example of the impact, high-impact publications describing the structure of a new protein today are often expected to also include simulations showing functional motions.

By definition, the larger the trajectories and the simulated system, the richer the information provided by MD simulations. However, this also increases the computational cost. In the seventies, slow computers limited simulations to the picosecond range and system sizes to a few hundred atoms. Today, High-Performance Computers (HPC) enable simulations involving systems with multi-million to billion atoms, extending into the microsecond and millisecond regimes.

Consequently, the output from individual simulations has grown from a few kilobytes in the 70s-90s to megabytes in the 00s, gigabytes a decade ago, and now the terabyte range. In the new era of open science, the traditional approach from the 70s of running the simulation, performing a hypothesis-driven analysis, deriving conclusions, writing the paper, and removing the trajectory, is no longer advisable. Modern simulations: i) consume large amounts of compute resources, ii) contain far more information than can be uncovered through hypothesis-driven analysis alone, and iii) may have hidden insights that can only be accessed through meta-analysis by combining multiple simulations. The lack of FAIR access to MD data is also impeding the development of artificial intelligence (AI) techniques for extracting insights into dynamic properties of biological systems and processes. The “flagship” AlphaFold2 technology has revolutionized the field of structural biology in just a few years leading to recognition with the Nobel Prize in a very short time, but the development was only made possible because of the wealth of structural data collected in the Protein Data Bank to learn from.

Further development of AI techniques, able to tackle the prediction of dynamic properties, deal with single-stranded nucleic acids, intrinsically disordered proteins, drug-macromolecule binding, phase transitions, or complex protein machines will only be possible with similar access to training data describing dynamics – and presently only molecular dynamics (MD) simulations can provide this. Thousands of research groups worldwide are collecting MD trajectories, tackling larger systems over longer time scales with increasingly accurate force fields, but the data is often stored locally or on Zenodo without following any standards, making it inaccessible and ineffective for AI development.

In a position paper (Hospital et al., Nature Methods, 2024) signed by more than 150 scientists, the community **highlighted the need to integrate MD-derived information into an**

international data ecosystem to ensure that MD data can contribute to the advancement of Science. **Europe is in a unique position to lead this integration worldwide, but this requires us to act immediately**.

The role of Europe: MD is one of the few fields Europe dominates. Many leading data producers are based in Europe, and the most widely used MD software is developed in Europe. The first initiatives to store MD data following the FAIR (Findable, Accessible, Interoperable and Reusable) principles were developed in Europe, and many of the life science core data resources, e.g. PDBe, the central repository of structural data, are based in Europe. Furthermore, the **Molecular Dynamics Data Bank (MDDB)**, a European-funded project, is a collaboration of the leading groups in the field (Institute for Research in Biomedicine in Barcelona –IRB; University of Oxford -UOXF-; Royal Institute of Technology in Stockholm –KTH-; Barcelona Supercomputing Center –BSC; EMBL-EBI –PDBe-; Nostrum Biodiscovery –NBD- and Centre Européen de Calcul Atomique et Moléculaire –CECAM), has positioned Europe at the forefront of developing an MD data ecosystem. Today, dozens of private and academic institutions, as well as HPC centers in the US, Japan, and even China, have approached us to extend this European initiative worldwide. **Europe cannot miss the opportunity to maintain its leadership. Strengthen European values of open science in the field, and exploit the possibilities it will bring for AI training Transforming the MDDB project into a stable infrastructure would be crucial to achieving this goal.** Very importantly, an MDDB European data resource will have Synergies with other life science core resources such as UniProt, ChEMBL, and PDB) It will collaborate with PRACE, EuroHPC-JU, EOSC, and Fenix infrastructures to design user-friendly interfaces and an optimal hardware framework for the sustainable storage and maintenance of large datasets. This collaboration will open exciting new scenarios for AI groups, enabling them to leverage the rich HPC European ecosystem and optimize the use of HPC centers. Joining efforts with ELIXIR will create unique links between simulation data and Core Data Resources, adhering to the guidelines set by ELIXIR Core Data Resources as a public Open Science repository. This comprehensive approach not only strengthens the infrastructure but also promotes a more integrated research environment in Europe.

The value of an international repository of MD trajectories? Integrating the fragmented MD trajectory data in a single repository would facilitate:

- General accessibility to trajectories in a “one-stop-shop” with professional management.
- Persistent and curated data beyond the typical 2-years of lifetime of a webserver.
- A detailed description of general principles of macromolecular interactions and dynamics.
- Training of a new generation of AI techniques specialized in interactions and dynamics.
- Definition of the general mechanism of drug-protein and drug-RNA interactions.
- Determination of complex Intrinsically Disordered Proteins (IDP) and single stranded nucleic acids dynamics.
- Training of coarse-grained and mesoscopic models for cellular-scale simulations.
- Training and refining of enhanced sampling strategies and MD-generative models.
- Improvement of force-fields by “active learning” protocols.
- Guaranteeing the reproducibility and robustness of simulations.

The global requirements of a MD repository:

- Coordinated local repositories able to store Petabyte-scale data and provide analysis capabilities to external users.
- Efficient and standardized graphical interfaces connected to the stored data including powerful analysis tools, tailored to the specific nature of the systems stored.
- A general coordinator for worldwide access and allowing meta-analysis from different repositories.
- A commonly accepted, standardized metadata allowing characterization of the nature of the system, the details of the simulations, their purpose, and any information the user might need to analyze a single trajectory or a set of them.
- Maintain data standards reducing the problems derived at the analysis level from the output of different MD codes.
- Some commonly accepted rules about what trajectories merit to be stored.

The Proposed infrastructure:

The MDDB infrastructure will adopt a distributed infrastructure strategy, with nodes (ideally thematic) installed in centres with sizeable data storage capabilities, such as HPC centres. The current nodes integrated in the prototype and accepting MD trajectories include BSC, CINECA, JSC and different Tier-1 computers around the world:

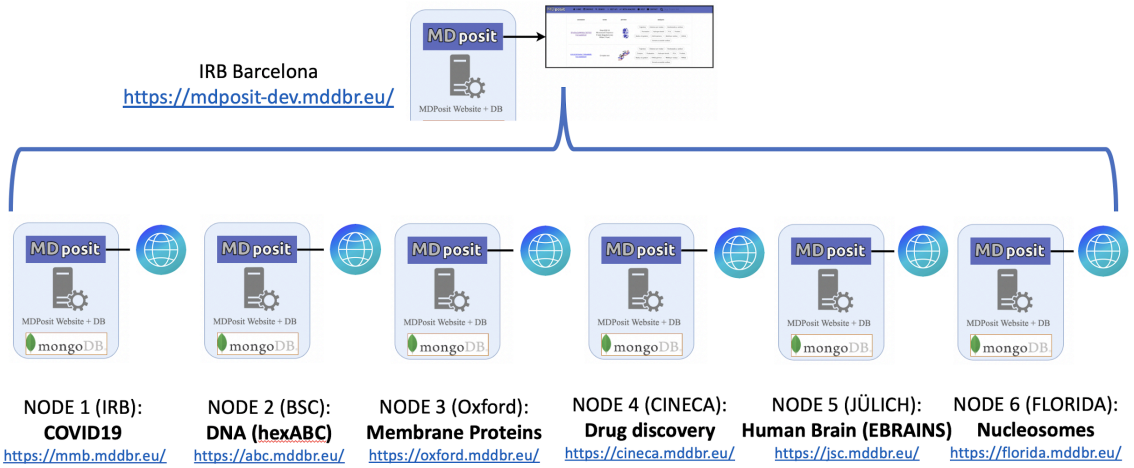

***Figure 1**. Distributed data infrastructure in development within the MDDB project.*

Several non-European centres, such US-National centers (Argonne and Los Alamos), Cloudbased consortia (such as Folding@Home), Rikken Institute in Japan and many centres in Europe have offered to host nodes of the infrastructure. There are close to 30 private and public institutions offering us to provide trajectories (just in drug-protein environment we are managing close to 105 trajectories, which are being deployed in CINECA).

The current MDDB consortium has developed a prototype for the distributed infrastructure, coordinated by a central database storing metadata of all the different integrated repositories: MDposit. MDposit can direct users to the desired trajectory regardless of its physical location and perform meta-analysis on different trajectories, even when stored in separate repositories. This system eliminates the need for data transfer across the internet and avoids redundant data replication at different sites (Figure 2, top).

The prototype also contains a reference implementation that can be used to deploy a new node. The deployment uses a containerized system that orchestrates, manages and scales containers across a cluster of physical or virtual machines. This system automatically deploys a database, a workflow to run the quality checks and analyses on the uploaded trajectories, a loader to upload the data into the database, and a standard web interface to facilitate access to this data, among other features (Figure 2, bottom). The code is easy to use and robust that installation in HPC centres such as CINECA or JULICH took only 1-2 days. Each of the nodes provides not only disk space, but also physical or virtual machines to support this reference implementation. In the next step, the system will be linked to High Performance Computation, a crucial element for many AI training procedures.

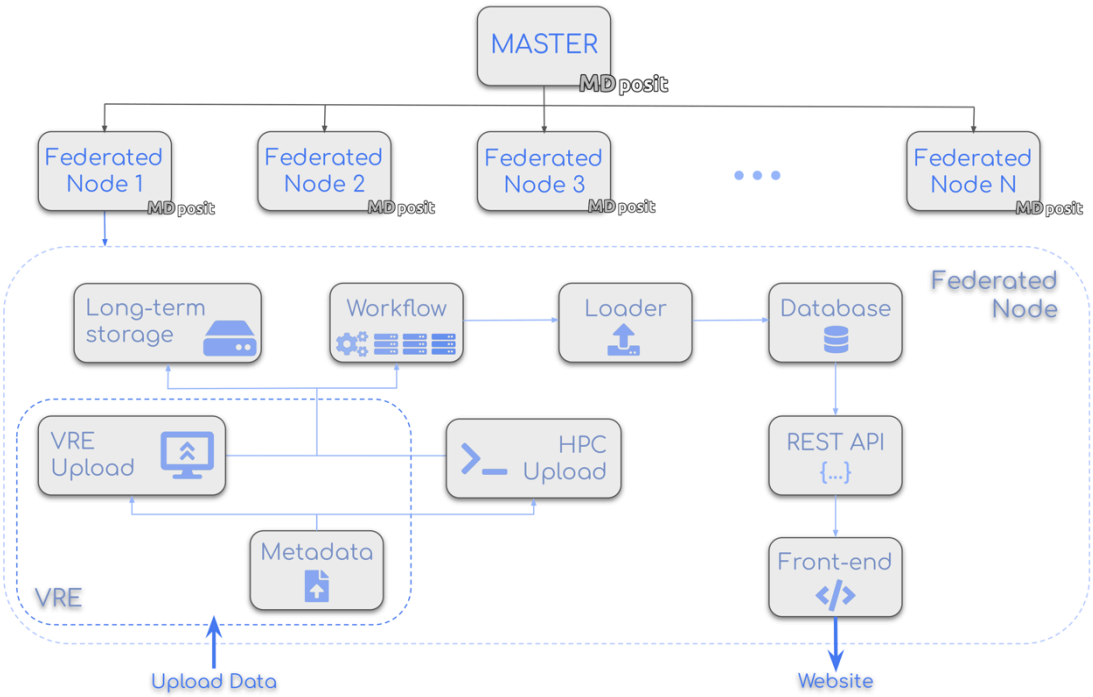

***Figure 2**. MDDB infrastructure prototype schema. Top: master interface, storing global metadata and linking all integrated repositories; Bottom: federated node reference implementation, including database, worker, loader, web interface and REST API programmatic access.*

After one and a half years of work, the MDDB consortium has developed technical solutions to enable the creation of a global MD-data infrastructure. The overwhelming commitment of the community and the enthusiastic support from computer centres in Europe and beyond is evident. The next challenge is transforming this 3-year initiative into a stable, enduring infrastructure.