Cancer and cardiovascular diseases are the leading causes of death in the EU, highlighting the urgent need for more effective therapeutics for these diseases. Unfortunately, drug discovery is a slow process, taking 16 to 20 years from target identification to drug approval. To accelerate drug discovery, pharmaceutical companies use computational tools. Among the most successful of these are physical molecular simulations, but these are limited by the availability of experimental structural data. Recent advances in machine learning-based structure prediction, such as AlphaFold, offer the potential to supply structural data suitable for physics-based modeling without the need to experimentally solve structures. However, these tools produce 3D structures missing key physical details, which are vital for accurate molecular modeling. A critical physical detail is the assignment of relevant protein protonation states, where misprediction results in large errors in drug binding affinity predictions, slowing down drug discovery.

To address this limitation, we propose to build PROTONMD, the first open, kinome-wide database of constant-pH molecular dynamics (CpHMD) simulations. PROTONMD will provide high-quality reference data on equilibrium protonation states and their coupling to protein conformational dynamics, becoming a foundational database for future ML forcefields and structure prediction tools that incorporate protonation state effects.
In this project, we will perform CpHMD simulations for 240 representative human kinases, with three independent replicas per system, yielding a total of 720 trajectories of 500 ns each. Kinases are an ideal initial target due to their central role in cancer and cardiovascular disease, their well-characterized conformational landscape, and the known coupling between protonation states and functional conformations. The resulting database will include CpHMD trajectories, equilibrated protonation state ensembles, and associated metadata, providing a unique and reusable resource for the community. PROTONMD will serve as the foundational dataset to enable accurate protein protonation state prediction, and ultimately improving the reliability of computational drug discovery pipelines.