POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Unsupervised Deep Learning for Molecular Dynamics Simulations: A Novel Analysis of Protein-Ligand Interactions in SARS-CoV-2 Mpro†

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Jessica Mustali

Advisor: Prof. Alfonso Gautieri

Academic year: 2022-2023

Introduction

Protein-ligand interactions are pivotal in modern drug discovery. While traditional methods offer structural snapshots, they often fall short in capturing the dynamic nature of these interactions. Molecular Dynamics (MD) simulations have emerged as a powerful tool, providing detailed insights into the temporal evolution of protein-ligand systems at the atomic level. However, the analysis of the extensive datasets generated by MD simulations poses significant challenges. While the integration of machine learning with MD shows promise, current methodologies primarily rely on supervised machine learning models, which face challenges with data labeling and standardization. To address this, an unsupervised deep learning framework tailored for efficient and meaningful pattern extraction from high-dimensional MD data was introduced and benchmarked in a previous study on relatively rigid proteins [1]. In the present study, we have further tested and adapted this framework for flexible proteins, with a case study on the SARS-CoV2 Main Protease (M^{pro}). Through this research, our aim is to provide valuable insights into the intricate interplay between dynamic protein conformations and ligand binding. We believe that the advanced analytical framework presented in this study has significant potential to revolutionise the analysis of protein-ligand complex MD data, thereby potentially accelerating the drug discovery process.

Materials and Methods

Molecular dynamics simulations

In this study, we conducted MD simulations of M^{pro} in both apo- and ligand-binding forms to analyze the structural and dynamic patterns induced by 11 different ligands. M^{pro} is a homodimeric cysteine protease comprising 306 amino acids per monomer, organized into three subdomains and the substrate-binding region located at the interface of domains I and II [2]. The inhibitors we considered in this study varied in molecular weight (ranging from 270.24 g/mol to 709.98 g/mol) and displayed a broad spectrum of IC50 values (ranging from 0.04 µM to 10.7 µM). Production simulations, each lasting 1 μs, were performed in triplicate for each system, providing rich temporal information on the protein-ligand interactions. The simulations exhibited a performance rate of 310 ns/day, resulting in an approximate runtime of 77 hours each. System stability was assessed by monitoring the convergence of protein root mean square deviation (RMSD). Ligand movement relative to the

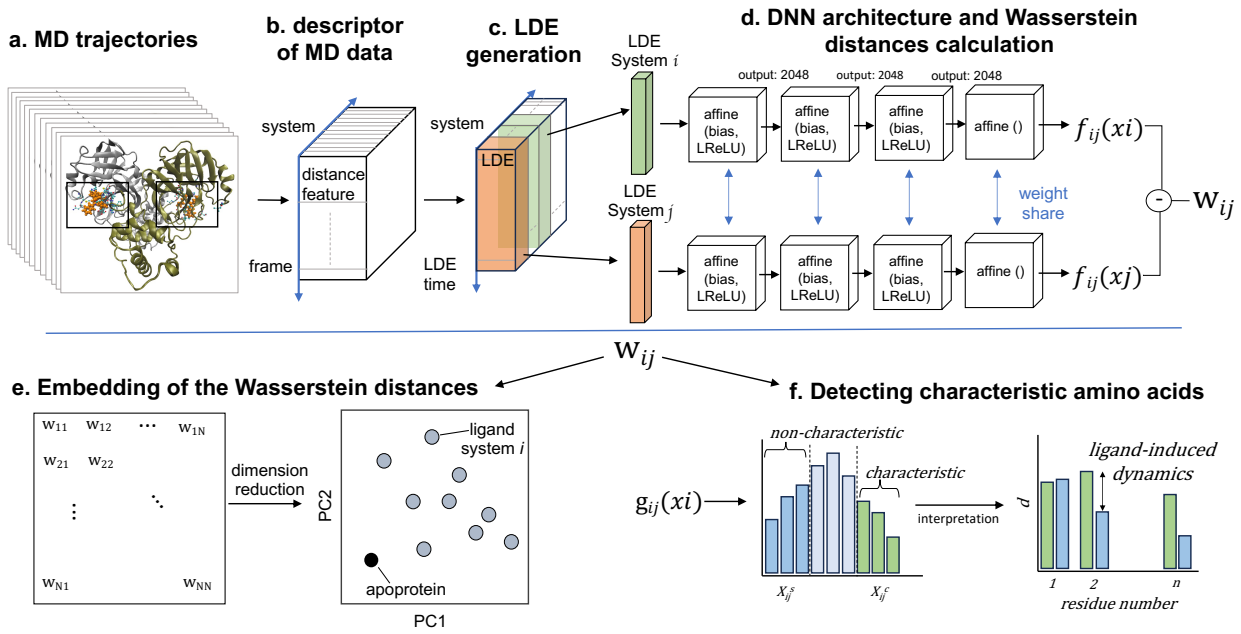

Figure 1: **a** MD trajectories for ligand-free (apo-protein) and ligand-bound (holo-protein) systems. **b** The distance between the center of mass of each binding-pocket residue and the center of geometry of the binding pocket is calculated over the trajectories. **c** Ligand-induced protein dynamics is represented by the local dynamics ensemble (LDE), which is an ensemble of short-term trajectories of the distance descriptor. **d** The difference between the LDEs of pairs of systems is calculated on the basis of the Wasserstein distance W_{ij} using the function f_{ij} approximated by deep neural networks (DNNs). **e** The Wasserstein distance matrix is embedded into points in a lower-dimensional space, and principal component analysis is performed to the embedded points. **f** The function $g_{ij}(x_i)$ helps interpret how specific residues contribute to the difference between the LDEs of system pairs, as determined by the DNNs. For both characteristic and non-characteristic trajectories, we computed the average value of the distance descriptor d_i for each residue. Notably, when there is a relevant difference in d_i values between characteristic and non-characteristic trajectories, the residues are highly influenced by the ligand.

binding pocket was monitored through the ligand heavy atoms RMSD.

Descriptors of molecular systems from MD data

In the process of analyzing MD trajectory data, input types (descriptor and LDE definition), binding-site residues, and appropriate time windows selection is crucial for subsequent Deep Neural Network (DNN) analyses (Fig. 1). Our focus lies on the binding-site residues, aiming to capture differences in protein behavior upon ligand binding while reducing computational complexity. Dealing with proteins with high-flexibility, it's necessary to carefully choose a descriptor that overcomes challenges related to coordinate dependency (e.g structure fitting issues arising from the combination of overall rotation and internal motion) and considers conformational dynamics. After careful testing, we opted for the distance between the center of mass of the binding-pocket residues and the center of geometry of the binding-pocket. This distance effectively encapsulates relevant information about the structural and dynamic differences of M^{pro}, offering a robust representation of the thermodynamic and kinetic properties of the systems.

Selection of the binding-pocket residues and MD trajectory time windows

In order to determine the binding-site residues, trajectories from the final 200 ns were considered to identify protein-ligand atom pairs involved in hydrogen bonds, within 4.5 Å in over 75% of the time frame. A total of 36 residues were iden-

tified (18 for each binding site). Trajectories of the centers of mass of these binding-site residues were extracted and the distance descriptor was calculated throughout the MD trajectories. We then strategically refined the MD trajectories by selecting specific time windows. PCA helped to distinguish stable molecular conformations from fluctuations, ensuring that the selected time intervals accurately represented the local changes induced by ligand binding. Leveraging the insights from PCA, we chose distinct 300-ns intervals characterized by enhanced stability. This targeted approach enhances the ability of subsequent machine-learning analyses to capture relevant conformational changes associated with ligand binding.

Analysis of protein conformation dynamics using ML

Here, we briefly introduce the machine-learning methods. The input of the DNNs is the Local Dynamics Ensemble (LDE), which is defined an ensemble of short-term trajectories of the distance descriptor. Derived from the MD simulation data, the LDE portrays the temporal evolution of this descriptor, thereby offering a snapshot of localized changes in the protein-ligand systems over time. Upon computing the LDE for every particle present in the binding site, a high-dimensional matrix is obtained (Fig. 1c), offering a comprehensive representation of the system's structural and dynamic behavior. To quantify differences between LDEs, the Wasserstein distance is employed. This metric, rooted in optimal transport theory, effectively assesses dissimilarities between two probability distributions [3]. Mathematically, the Wasserstein distance between two LDEs y_i and y_j, is expressed as:

$$W_{ij} = \sup_{|f_{ij}|\leq 1} \mathbb{E}_{\boldsymbol{x}_i\sim\boldsymbol{y}_i}[f_{ij}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}_j\sim\boldsymbol{y}_j}[f_{ij}(\boldsymbol{x})] \quad (1)$$

where x_i and x_j are short-term trajectories of systems i and j, respectively. The function $f_{ij}(\boldsymbol{x})$ that solves the maximization problem in Eq. 1 is approximated by the network (Fig. 1d) with the 1-Lipschitz constraint. The DNNs consisted of multilayer perceptron used in a previous study [1]. Short-term trajectories x of the LDEs are flattened and used as input for the DNN. In the optimization process, the loss function with gradient penalty was minimized. The optimization process was performed for up to 500,000 steps per model, when the moving averages of DNN output over 10,000 steps converged. The mean value of the last 10,000 steps was used as the Wasserstein distance. Nonlinear dimensionality reduction and Principal Component Analysis (PCA) yield an embedding map, thus providing a compact and insightful representation of the complex high-dimensional dynamics inherent to the presence of different ligands.
Additionally, we conducted an analysis to extract the characteristic dynamics using the function $g(\boldsymbol{x}_i)$ (Fig. 1f). This function evaluates the contribution of a single short-term trajectory to the overall differences between the two systems. A small $g(x)$ value for a trajectory in system i compared to system j indicates that system i's trajectory closely mirrors the general behavior observed in system j and vice versa. Since $g_{ij}(\boldsymbol{x}_i)$ includes short-term trajectories over many residues, we can identify the residues that significantly influence the Wasserstein distance between systems, and are therefore strongly affected by ligand binding. The computation of $g(x_i)$ was executed using the optimized Deep Neural Networks (DNNs). For this, the specific short-term trajectory of system i and the average local trajectory of the other system j served as inputs. The ML-driven analysis of the MD trajectories was completed within a single day.

Results and discussion

Flexibility of SARS-CoV-2 M^{pro}

The Root Mean Square Fluctuations (RMSF) of the residues were computed throughout the trajectory, providing valuable insights into their dynamic flexibility. The results of the RMSF analysis provided evidence for the intrinsic flexibility of M^{pro}, a feature that was corroborated by a number of experimental and computational investigations. Understanding protein flexibility is crucial in the context of drug binding thermodynamics and underscores the importance of considering conformational dynamics in protein-ligand interaction studies. The ligand-free system showed higher fluctuations than some protein-ligand systems and lower fluctuations than others. This suggests that ligand binding cannot simply be correlated with the higher/lower induced fluctuations of M^{pro} residues. Unsupervised deep learning can over-

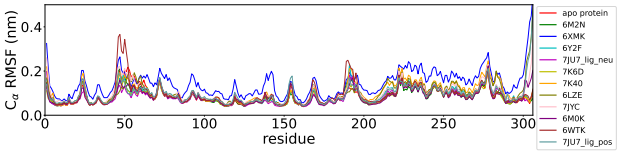

Figure 2: Residue-based root mean squared fluctuation (RMSF) of the protein backbone averaged between monomer A and monomer B in the first 1 µs MD simulation for the 12 systems.

come these challenges and reveal complex dynamic properties by detecting hidden patterns in MD data that conventional analysis methods such as RMSF cannot uncover.

Unsupervised deep learning-based insights into protein-ligand dynamics

The Wasserstein distance matrix provides a quantitative measure of ligand-induced changes across systems. The color-coded representation of this matrix shows the relative distances between the systems, with system 7JYC distinctly separated from the other systems (Figure3a). This observation suggests that system 7JYC exhibits unique trajectories that were captured and highlighted by our unsupervised deep-learning methodology. Starting from the Wasserstein distance matrix, we constructed an embedding map that spatially arranges the systems. Here, each system is represented as a point, with color indicating the experimental binding-affinity values ($pIC50$). A meaningful pattern emerged: systems with lower affinity are found closer to the apo-protein, reflecting structural and dynamic similarities to the ligand-free state. Conversely, high-affinity systems occupy positions further along PC2, denoting distinct ligand-influenced structures and dynamics (Figure 3b). Furthermore, we observed that systems 6M0K and 6LZE, exhibiting higher affinities, displayed notable similarities in ligand chemical structures and shared identical PC2 values. To reinforce the information obtained from the embedding map, we correlated the experimental binding affinity values (pIC50) with PC2 values. The Pearson's correlation coefficient of 0.7 affirm the substantial correlation between PC2 and IC50 for high- and low-affinity ligands, underscoring the potential of our deep-learning approach to detect subtle shifts in ligand-induced trajectories within M^{pro} (Figure 4).

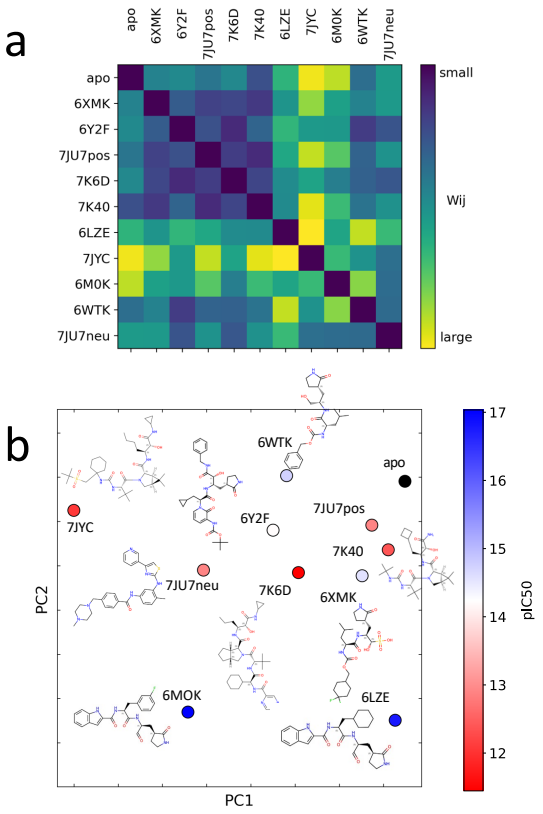

Figure 3: **a** Distance matrix of Wasserstein distances between the probability distributions of the LDEs for system pairs. A large Wasserstein distance (yellow) corresponds to a large difference in the protein structure and dynamics. **b** Embedded points of the distance matrix and chemical structure of the corresponding system. The points are colored according to the experimental binding-affinity values (pIC50). pIC50 corresponds to $-\log(IC50)$.

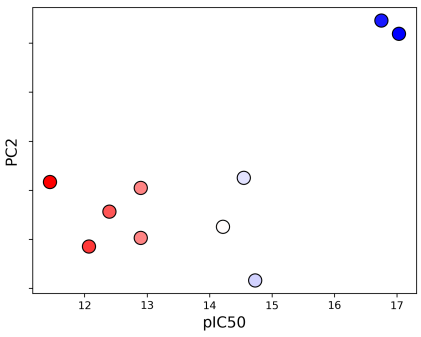

Figure 4: Correlation between PC2 and experimental binding-affinity data (pIC50). The correlation, quantified using Pearson Coefficient, is 0.7

Interpretation of the contribution of residues to ligand-induced dynamics

The correlation observed between the PC2 component of the embedding map and pIC50 values underscores PC2's significance in capturing conformational disparities linked to ligand-binding affinity. To delve deeper into the molecular underpinnings of this observation, we aimed to identify specific amino acids that showed prominent dynamic disparities between the highest and lowest binding-affinity systems using the function $g(x)$. Notably, Met49 and Arg188-Gln189-Thr190 emerge as pivotal residues in ligand binding. Conformational disparities in Arg188-Gln189-Thr190 are primarily captured in PC2, while Met49 is represented in PC1. The convergence of insights across various independent studies bolsters the robustness of our conclusions, providing a more comprehensive understanding of the dynamics governing protein-ligand interactions in M^{pro}.

Conclusions

Molecular dynamics (MD) simulations are central in the drug discovery process, offering atomistic insights into protein-ligand interactions crucial for therapeutic design. However, effectively analyzing extensive MD datasets remains a standing challenge. Throw a case study on SARS-CoV-2 M^{pro} we tested our unsupervised deep-learning framework for the analysis of MD simulation data of flexible protein-ligand complexes. First we conducted MD simulations of M^{pro} with various ligands. In the pursuit of effective MD trajectory analysis, the MD data were refined by focusing on binding site residues and time frames in stable protein conformations and the optimal input type was ascertained. We tested different input types and selected the distance between each residue and center of the binding pocket as descriptor and defined the Local Dynamic Ensemble LDE as the time series of the descriptor. We fed the Local Dynamic Ensemble (LDE) into our neural network to compute the Wasserstein distance across system pairs, revealing ligand-induced conformation differences in M^{pro}. Dimension reduction yielded an embedding map correlating ligand-induced dynamics and binding affinity with a Pearson coefficient of 0.7. This finding implies that the most active compounds had the maximum impact on the local structure and dynamics of the target protein, resulting in them being further distanced from the ligand-free system. We also identified the residues that contribute most to the difference between the systems, and the results are consistent with the latest literature on the subject. While our results are promising, we acknowledge potential limitations and avenues for future research. The effectiveness of our approach relies on initial conditions, specifically the initial structure of the protein and the chosen input feature. Expanding the dataset to encompass a broader range of ligands and optimizing simulation strategies are areas for further exploration. Looking ahead, we believe that the unsupervised deep-learning framework utilized in this study will be highly valuable for early-stage drug discovery, aiding in the prioritization of promising compounds. It also could be extended to diverse protein-ligand interactions, including allosteric events. In conclusion, this novel methodology, combining the strengths of deep learning and MD simulations, can help improve our understanding of molecular mechanisms and accelerate drug discovery, thereby setting the stage for rapid and refined therapeutic exploration.

1. Bibliography

References

[1] Ikki Yasuda, Katsuhiro Endo, Eiji Yamamoto, Yoshinori Hirano, and Kenji Yasuoka. Differences in ligand-induced protein dynamics extracted from an unsupervised deep learning approach correlate with protein–ligand binding affinities. *Communications biology*, 5(1):481, 2022.

[2] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, et al. Structure of mpro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293, 2020.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.