POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

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

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Jessica Mustali, 10843734

Advisor:
Prof. Alfonso Gautieri

Co-advisors:

Academic year:
2022-2023

Abstract: Molecular dynamics (MD) simulations, which are central to drug discovery, offer detailed insights into protein-ligand interactions. However, analyzing large MD datasets remains a challenge. Current machine-learning solutions are predominantly supervised and are limited by data labelling and standardisation issues. In this study, we adopted an unsupervised deep-learning framework, previously benchmarked for rigid proteins, to study the more flexible SARS-CoV-2 main protease (M^{pro}). We ran MD simulations of M^{pro} with various ligands and refined the data by focusing on binding-site residues and time frames in stable protein conformations. The optimal descriptor chosen was the distance between the residues and the center of the binding pocket. Using this approach, a local dynamic ensemble was generated and fed into our neural network to compute Wasserstein distances across system pairs, revealing ligand-induced conformational differences in M^{pro}. Dimensionality reduction yielded an embedding map that correlated ligand-induced dynamics and binding affinity. Notably, the high-affinity compounds showed pronounced effects on the protein's conformations. We also identified the key residues that contributed to these differences. Our findings emphasize the potential of combining unsupervised deep learning with MD simulations to extract valuable information about protein-ligand molecular mechanisms and accelerate drug discovery, thereby setting the stage for rapid and refined therapeutic exploration.

Key-words: MD simulations, drug discovery, unsupervised deep learning

Introduction

The landscape of drug discovery has been traditionally characterized by profound challenges, such as escalating costs and protracted timelines. At present, the costs associated with drug development have escalated to exceed US$2.8 billion, and the process requires an average of 14 years to reach fruition [1–3]. To overcome these hurdles, computational methods have become increasingly prevalent in pipelines for expediting drug-discovery processes [4–6]. Among these methods, molecular dynamics (MD) simulations have pushed the confines of computationally driven drug discovery and design over the past decades, owing to the increasing availability of computational power and suitable software [7, 8]. Offering a dynamic, atomistic view of protein-ligand interactions, MD simulations represent a powerful tool in biophysics research.

The successful discovery and design of therapeutic agents significantly depends on the depth of our understanding of protein-ligand interactions [9]. The profound influence of these interactions on the pharmacodynamics and pharmacokinetics of drugs provides a rationale for the major emphasis laid on their study in the field of drug discovery and design [10, 11]. Comprehensive characterization of the protein-ligand interaction landscape can guide the optimization of lead compounds, facilitate predictions of drug responses, and help avoid undesirable off-target effects [12, 13]. However, decoding the intricate dynamics of protein-ligand interactions poses a formidable challenge owing to their inherent complexity and multifaceted nature [14, 15]. The classic static view of protein-ligand interactions, based primarily on the structures obtained from X-ray crystallography and NMR spectroscopy, does not capture the conformational dynamics and energetic nuances of protein-ligand crosstalk. MD simulations can provide perspectives beyond this static view, explore the dynamic behavior of protein-ligand systems in atomistic detail, and capture their temporal evolution [16]. These simulations can unravel the thermodynamic and kinetic properties of protein-ligand interactions by incorporating structural flexibility and entropic effects. Thus, they provide insights into both enthalpic and entropic contributions to the binding free energy [17–19]. However, the inherent complexity of the data generated by MD simulations and the high computational cost of long-duration simulations remain substantial challenges. [16, 20].

The synergy of machine learning (ML), particularly deep learning, with MD simulations represents a promising frontier in molecular system research. The applications of ML and deep-learning methods in MD simulations are diverse and growing. They range from deriving classical potential energy surfaces from quantum mechanical calculations [21–27] to enhancing MD sampling by learning bias potentials [28–32], and even include generating samples from the equilibrium distribution of a molecular system without performing MD altogether, as exemplified by Boltzmann generators [33, 34]. Recently emerged graph neural network (GNN)-based machine learned potentials (MLPs) have demonstrated excellent accuracy in predicting forces directly from atomic structures of biomolecules as well as small molecules [35–38]. ML algorithms that perform tasks such as dimensionality reduction, clustering, regression, and classification have also been proven to be conceptually potent tools for analyzing the large datasets obtained from MD simulations [8, 39–41].

While these applications enshrine the potential of ML and deep learning in this field, their specific application to the analysis of MD simulation data in the context of protein-ligand interactions has emerged only recently. Significant progress has been made in this direction using supervised training. Supervised machine-learning algorithms were successfully applied to the classification of ligand-determined GPCR conformational properties by Plante et al. [42]. This and other studies [43–45] have outlined the potential of ML to extract valuable functional information from MD simulation trajectories of protein-ligand complexes, setting the stage for future advances in this field. Despite this promise, the lack of labeled data represents a major limitation in the implementation of supervised deep-learning approaches [46, 47], and other issues that may affect the prediction quality of supervised deep neural networks include the dependence on the dataset and thus on experimental conditions. Thus, the need for data standardization and curation precedes the construction of robust predictive models [48]. Consequently, the implementation of unsupervised techniques to circumvent these concerns offers distinct advantages [49, 50]. Deep neural networks (DNNs) within unsupervised frameworks can learn the hierarchical representations of data and identify complex patterns in unlabelled high-dimensional MD data. This enables the capture of intricate protein-ligand interaction dynamics, which are often challenging to identify through traditional means. By producing a compact, lower-dimensional representation of MD data, these models facilitate in-depth exploration of system dynamics. Furthermore, the application of deep-learning models can reveal relationships between protein conformational dynamics and ligand-binding affinities, which would otherwise be difficult to identify. Considering this potential, a novel approach using unsupervised DNNs to extract features from the MD trajectory data of protein-ligand complexes was introduced in a previous paper [51]. The study showed that differences in protein dynamics induced by ligands are indicative of binding energy. However, the benchmarks in that study were limited to bromodomain 4, a rigid protein with diverse ligand structures, and protein tyrosine phosphatase 1B, a flexible loop-containing protein with a similar ligand structure. Therefore, these methods have not yet been validated against flexible proteins with various ligand structures.

In this study, we demonstrate the potency of this approach for the analysis of more complex flexible protein systems through a case study of the SARS-CoV-2 main protease (M^{pro}). M^{pro} is a key target for drug design against SARS-CoV-2 because of its critical role in mediating viral replication and transcription, high sequence conservation with other coronaviruses, and lack of human homologs [52]. An oral drug named Paxlovid (nirmatrelvir and ritonavir) has been approved for the inhibition of M^{pro} [53, 54], but its application is limited because of drug-drug interactions [55] and rebound effects [56, 57]. Understanding the dynamics of M^{pro}, both in its ligand-free form and when bound to potential inhibitors, is of significant interest in the ongoing efforts to develop extended and alternative treatments against SARS-CoV-2.

With the research presented here, we aim to offer valuable insights into the complex interplay between dynamic protein conformations and ligand binding by utilizing an advanced analytical framework that employs unsupervised deep learning for MD simulations. We believe that the innovative approach presented in this study holds significant potential for transforming the current landscape of protein-ligand complex analyses and drug

discovery.

Materials and Methods

In this study, we analyzed the structural and dynamic patterns induced by 11 different ligands on the SARS-CoV-2 Main Protease (M^{pro} or $3CL^{pro}$). The success of our machine learning-driven analysis relies on a substantial dataset obtained through extensive MD simulations, providing rich temporal information on the protein-ligand interactions. Our deep learning model calculated Wasserstein distance between different simulation data via unsupervised learning. Here, MD simulation data is used to train the deep learning model, and Wasserstein distance is calculated for the same dataset by iteratively training the model [58]. We performed three independent simulations, each spanning one microsecond, for each of the 11 protein-ligand systems, which we believe produce a sufficient amount of data. These simulations captured a diverse range of conformational states and ligand-induced dynamics (Fig. 1). The molecular dynamics simulations exhibited a performance rate of 310 ns/day, resulting in an approximate runtime of 77 hours for each simulation. Subsequently, the ML-driven analysis of the MD trajectories was completed within a single day. It's worth noting that while our approach may entail a relatively longer processing time compared to supervised learning methods, it doesn't rely on the availability of labelled data. In this section, we present methods for MD simulations, extracting features, refining MD data, and briefly introduce the ML approach [51].

Molecular dynamics simulations

We performed MD simulations of M^{pro} in the apo- and ligand-binding forms (1a). M^{pro} is a homodimeric cysteine protease composed of 306 amino acids per monomer. Each monomer contains three subdomains; domains I and II (residues 8-101 and 102–184, respectively) are characterized mainly by β-barrel motifs, whereas domain III (residues 201–306) primarily consists of α-helices. [59–61]. The substrate-binding region is located at the interface of domains I and II and consists of the key active-site residues Met49, Gly143, His163, His164, Glu166, Pro168, and Gln189, as well as Tyr54, Gly143, His163, which form an oxyanion loop. In addition, the M^{pro} active site cleaves peptide bonds using a catalytic dyad formed by a cysteine residue (Cys145) and a histidine residue (His41).
The structures of the apo- and holo-SARS-CoV-2 main protease (M^{pro}) were obtained from the Protein Data Bank [62] (PDB ID: 6M03, 6M2N, 6XMK, 6Y2F, 7JU7, 7K6D, 7K40, 7JYC, 6LZE, 6M0K, and 6WTK [7, 59, 61, 63–70]). The inhibitors of M^{pro} we considered in this study have various molecular weights, ranging from 270.24 g/mol to 709.98 g/mol, and a broad spectrum of IC50 values, ranging from 0.04 µM to 10.7 µM (table 1). Missing atoms from these structures were added using the homology model module of the Molecular Operating Environment (MOE) software [71]. The protonation state of the amino acids in the 6M03 system (apo structure) was set to pH 7 using protonate 3D in MOE. The protonation states of the amino acids of other protein structures were fitted to those of the 6M03 system, and the number of amino acids was set to 712 (homodimer of 306). These initially modeled protein structures were referred to as the initial structures, and their pocket conformations were nearly identical to those of the corresponding X-ray crystallographic (PDB) structures. Each ligand was responsible for recognizing the N-terminal fragments of the substrate peptide non-covalently occupying the active-site cleft of each Mpro monomer. The total charge of each ligand was set to neutral. To compare the effects of ligand charge on Mpro, we also prepared a 7JU7 system with a positively charged (pos) ligand in addition to a neutral one. The force fields of Mpro and each ligand were amber14SB [72] and the general AMBER force field (GAFF) [73], respectively. The partial charges for each ligand were calculated at the RHF/6-31G** level using Gaussian 16 software [74] and fitted by restrained electrostatic potential (RESP) charge fitting in the antechamber on AmberTools18 [75].
The following steps for MD simulations were performed using GROMACS 2023 [76]. The apo-protein and protein-ligand complexes were solvated in a periodic cubic water box of 10 nm, with TIP3P [77] water molecules used as the solvent model. The systems were then neutralized with the addition of Cl^- ions and Na^+ ions, with the ionic strength set to 0.15 M, resulting in a total of $\approx$ 100 000 atoms. Preliminary system energy minimisations were performed using the steepest descent algorithm for 10,000 steps, until the maximum force was reduced to less than 10.0 kJ mol^{-1}. Subsequently, the systems were equilibrated in the NVT ensemble for 200 ps at 310 K using the velocity-rescaling method [78], followed by NPT equilibration at 1 bar for 200 ps using the Berendsen barostat [79]. The heavy atoms of the protein were restrained in equilibrium processes with a spring constant of 1000 kJ per mol nm^2. Nonbonding interactions were computed using a cutoff value for the neighbor list at 1 nm, and the potential-shift-Verlet approach with a cutoff of 1.0 nm was used to handle van der Waals interactions, whereas the particle–mesh Ewald method was applied to describe electrostatic interactions. The LINCS algorithm was used to constrain the h-bonds, thus allowing a time step of 2 f s. The production phase consisted of three independent MD replicates for each system with a random initial velocity.

Table 1: Summary of the inhibitors of the SARS-CoV-2 considered in this study. The PDB structures, the molecular weights (MWs) in g/mol, and the experimental binding-affinity values (IC50) in µM are reported.

Cpd	PDB	Ligand	MW (g/mol)	IC50(µM)	Reference
1	6M0K	FJC	464.49	0.04	[61]
2	6LZE	FHR	452.55	0.053	[61]
3	6WTK	UED	405.49	0.4	[70]
4	6XMK	QYS	527.58	0.48	[83]
5	6Y2F	O6K	595.69	0.67	[59]
6	6M2N	3WL	270.24	0.94	[84]
7	7JU7	G65	498.64	2.5	[85]
8	7K40	U5G	521.69	4.13	[86]
9	7JYC	NNA	709.98	5.73	[86]
10	7K6D	SV6	681.87	10.7	[87]

Each simulation had a duration of 1 µs, and was performed using the velocity-rescaling method for temperature control and a Parinello-Rahman barostat [80].
The stability of the system was assessed by monitoring the convergence of the root mean square deviation (RMSD) of the protein. To determine the structural elasticity and residual fluctuation, the root mean square fluctuation (RMSF) profiles of the Cαatoms in the MD-simulated ensembles were calculated. To monitor the movement of the ligands relative to the binding pocket, we used the RMSD of the heavy atoms of the ligand after superimposing the backbone-binding site residues onto the reference structure.
The figures in the article were generated using *VMD* [81] and UCSF Chimera [82], while the analysis was performed via scripts written in Python 3.11 using the *matplotlib* libraries for plotting, and pandas, *numpy*, and *scipy* for data handling and statistics.

Descriptors of molecular systems from MD data

After obtaining trajectory data from MD simulations, a pivotal step is the selection of an appropriate trajectory-derived descriptor. This descriptor adequately represents the systems of interest for subsequent analyses using the DNN (Fig. 1b). We focused our analysis on the binding-site residues of the target protein. The rationale behind this choice relies on the ability to capture the difference in protein behavior in the context of ligand binding while dramatically reducing the dimensionality and computational cost considering the trajectories of all particles in the system. In a previous study [51], the protein fluctuation of binding-site residues in Cartesian coordinates was selected. However, in contrast to the relatively rigid proteins considered in that study, M^{pro} is highly flexible because its binding pocket consists of flexible loops [88–92], meaning that the fitting of structures may cause biases, and significant conformational changes can occur. Therefore, our descriptor was required to overcome two challenges: (1) the descriptor should avoid dependency on coordinate changes, and (2) conformations associated with the dynamics should be considered. After testing both coordinates and distance, we selected the distance between the center of mass of the binding-pocket residues (binding-site residue determination is discussed below) and the center of geometry of the binding-pocket. The selected distance conveys relevant information regarding M^{pro} structural and dynamic differences, providing a robust description of the thermodynamic and kinetic properties of the systems [11, 93, 94]. In addition, the distance representation of the trajectory is not affected by mixing the overall rotation and internal motion, which are issues that affect Cartesian coordinates [95]. Comparisons of different types of descriptors are presented in the results section.

Selection of the binding-pocket residues

The binding-site residues were selected by an analysis using the AmberTools *CPPTRAJ nativecontacts* module combined with the GROMACS *distance* module and VMD for visual inspection. For this purpose, trajectories spanning the last 200 ns were considered to determine protein-ligand atom pairs closer than 4.5 Å using CPPTRAJ native contacts. This distance cutoff value was demonstrated to be optimal, with performance equivalent to that of more sophisticated methods relying on residue–residue interaction energies [96]. The output file was processed using Python 3.11 scripts. This enabled the isolation of atom pairs engaged in protein-ligand hydrogen bonds (namely, oxygen–oxygen, nitrogen–oxygen, sulfur-oxygen, sulfur-nitrogen, and nitrogen–nitrogen)

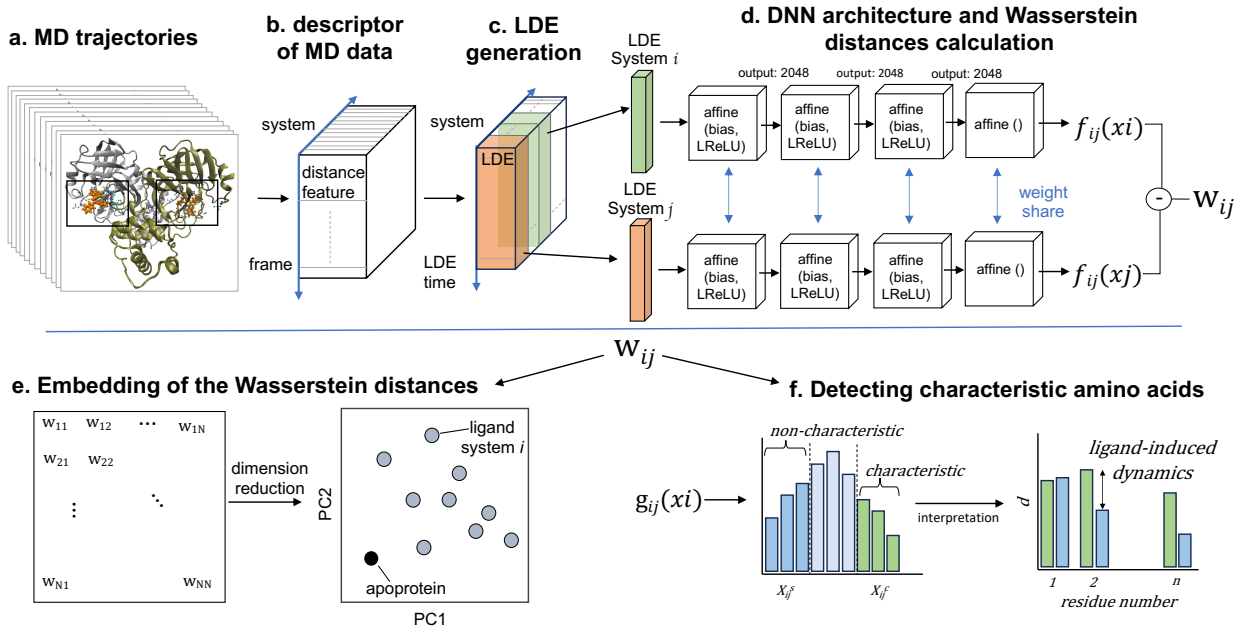


Figure 1: **a** MD trajectories for ligand-free (apo-protein) and ligand-bound (holo-protein) systems. **b** The distance between the center of mass of each binding-pocket residue and the center of geometry of the binding pocket is calculated over the trajectories. **c** Ligand-induced protein dynamics 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.

and those present in over 75% of the examined 200 ns timeframe. In-depth contact analysis of the identified atom pairs was performed using the GROMACS distance module supplemented by VMD visual validation. The M^{pro} residues that manifested from the contact analysis in both monomers of the dimeric M^{pro} across any simulated system were designated as binding-pocket residues. From this comprehensive analysis, 36 residues were identified for the dimeric M^{pro} (18 residues for each binding site) (Fig. 2). For subsequent analyses, the trajectories of the centers of mass of the binding-site residues were extracted from the total MD trajectory after fitting to the Cα of the binding-site residues of the apo-protein reference structure.

Selection of MD trajectory time windows for local dynamics ensemble generation

Flexible proteins adopt various atomistic conformations, some of which lead to ligand disassociation. To harness this effectively and extract highly stable conformations around the stable protein-ligand complex, MD trajectory should be refined by strategically selecting time windows. This selection process aims to represent nuanced local changes in the protein-ligand system, with an emphasis on periods where ligand interactions occur within the binding pocket, particularly during the most stable conformations of the system. To streamline this selection, we applied principal component analysis (PCA) to the distance features. In this context, the utility of PCA relies on its ability to distinguish relevant temporal patterns and transitions from extensive MD trajectory data [97, 98]. By mapping the MD frames onto these principal components, we could discern the specific structural variations. More importantly, this allowed us to pinpoint the time windows characterized by the adoption of stable conformations by the system. Using the insights derived from PCA, we chose distinct 300-ns intervals. These selected intervals became the foundation for constructing the local dynamics ensemble (LDE) trajectories. By narrowing our focus to these intervals, we enhanced the ability of LDE to encapsulate relevant conformational changes associated with ligand binding and bolstered the efficiency of the ensuing machine-learning analyses.

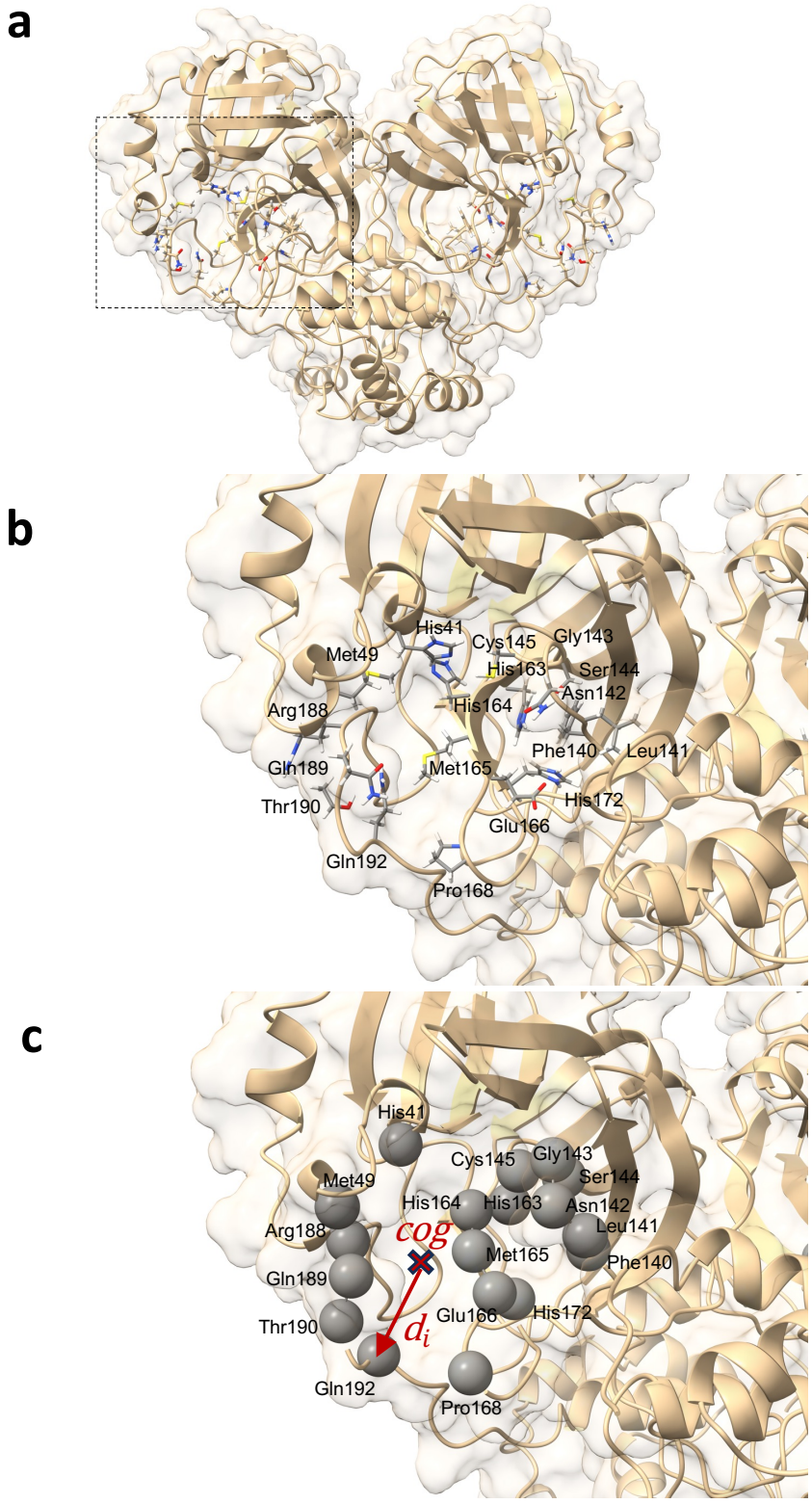


Figure 2: **a** Three-dimensional structure of SARS-CoV-2 M^{pro} dimer. **b** Binding site of M^{pro}. Selected binding-pocket residues are labeled and visualized in a licorice representation. **c** Binding-pocket residues are represented as spheres. The distance between the center of mass of each selected residue and the center of geometry (*cog*) of the binding pocket is calculated through the trajectory.

Analysis of protein conformation dynamics using ML

Here, we briefly introduce the machine-learning methods. The LDE, which is defined as an ensemble of short-term trajectories related to a descriptor of interest, was generated from the distance in the previous section. 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 system over time. Mathematically, the LDE from the starting time step t_0 is represented as a time series of configurations:

$$\boldsymbol{x} = [d(t_0 + \Delta), \ldots, d(t_0 + \delta)] \tag{1}$$

or when using time series of the displacement

$$\boldsymbol{x} = [d(t_0 + \Delta) - d(t_0), \ldots, d(t_0 + \delta) - d(t_0)] \tag{2}$$

In these equations, x denotes the LDE, $d(t)$ represents the distance between the center of mass of the binding-pocket residues and the center of geometry of the binding pocket at time t, Δ is the duration over which the LDE is defined, and δ is the time interval of the MD output selected to generate the LDE (in our case, 300 ns). In this study, the time window δ was 300 ns and the LDE time Δ was 64 ps. Each trajectory within the LDE captures the evolving change in the distance descriptor over a specified time window, thereby providing a dynamic snapshot of the system behavior.

Upon computing the LDE for every particle present in the binding site, a high-dimensional matrix is obtained (Fig. 1c). This matrix is a comprehensive representation of the structural and dynamic behavior of the system. The rows represent distinct particles within the binding site, with each row encapsulating the temporal evolution of the distance descriptor for that specific particle and providing a trajectory of its behavior over the course of the simulation. The matrix columns correspond to specific time points in the MD simulation and offer a cross-sectional overview of the structural configuration of the system at each time point. In essence, the matrix obtained by calculating the LDE for all the particles of the binding site encapsulated the temporal evolution of each particle and the state of the entire system at each time point.

For pairs of LDEs across all systems, the differences in LDEs were computed on the basis of the Wasserstein distance. Originating from optimal transport theory, the Wasserstein distance serves as a prominent metric to assess the difference between two probability distributions [99]. Opting for the Wasserstein distance over other metrics offers three notable benefits [58]:

1. It is suited for high-dimensional data, ensuring cost-effective computation via DNN.
2. It boasts mathematical properties inherent to a distance, which do not hold to divergence.
3. It eliminates the need for preliminary assumptions about the distribution.

Mathematically, the Wasserstein distance between two LDEs y_i and y_j, is expressed as:

$$W_{ij} = \sup_{|f_{ij}| \leq 1} \mathbb{E}_{\boldsymbol{x}_i \sim \boldsymbol{y}_i}[f_{ij}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}_j \sim \boldsymbol{y}_j}[f_{ij}(\boldsymbol{x})] \tag{3}$$

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. 3 is approximated by the network (Fig. 1d) with the 1-Lipschitz constraint. Conceptually, this function represents the optimal mapping function f_{ij} that transforms one system's LDE into another's. The expectations $\mathbb{E}_{x_i - y_i}[f_{ij}(x)]$ and $\mathbb{E}_{x_j - y_j}[f_{ij}(x)]$ are calculated over the probability distributions of the LDEs of system i and system j, respectively. The difference between these expectations yields the Wasserstein distance, providing a metric of the dissimilarity of the two systems. The DNNs consisted of multilayer perceptron used in a previous study [51]. Short-term trajectories x are flattened and used as input for the DNN. This DNN boasts three fully connected hidden layers, each with 2048 output nodes, and employs the leaky rectified linear unit (LReLU) as its activation function. The output layer has one node without bias and activation function. The initial values of parameters were sampled from uniform distributions (mean = 0, deviation = $\frac{1}{k}$), where k is the number of the input features of each layer. The networks were implemented in PyTorch [100] (Fig. 1d). In the optimization process, the loss function with gradient penalty was minimized [101]:

$$L = \mathbb{E}_{\boldsymbol{s} \sim \boldsymbol{y}_i}[f(\boldsymbol{s})] - \mathbb{E}_{\boldsymbol{t} \sim \boldsymbol{y}_j}[f(\boldsymbol{t})] - \mathbb{E}_{\boldsymbol{r} \sim \boldsymbol{R}}[f(\boldsymbol{r})(||\nabla_t f(\boldsymbol{r})|| - 1)^2] \tag{4}$$

where s and t are short-term trajectories, y in the probability distribution of local dynamics ensemble for system i and j, r is the interpolation between s and t and R is the probability distribution of r. For each learning iteration, short-term trajectories were selected randomly by deciding the initial step of the time sections. Model parameters were updated using the Adam optimizer [102] (learning rate = 1×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.9$). The size of the minibatch was 64. 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.

By computing the Wasserstein distance for all pairs of N systems, a distance matrix of (N,N) was obtained (Fig. 1e). This matrix provides a comprehensive view of the differences in protein dynamics owing to the presence of different ligands. The subsequent nonlinear dimensionality reduction and PCA resulted in the creation of an embedding map, thereby simplifying the visualization. The embedding process is guided by the minimization of the following expression:

$$\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n = \underset{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n}{\arg\min} \sum_{i<j} (W_{ij} - \|\boldsymbol{p}_i - \boldsymbol{p}_j\|)^2 \tag{5}$$

Here, $\boldsymbol{p}_i$ represents a three-dimensional vector corresponding to system i (embedded point of system i), where W_{ij} denotes the Wasserstein distance between systems i and j. Embedding optimization employs a two-pronged approach using simulated annealing for global-minimum exploration, followed by gradient descent for swift convergence [51]. This embedding cycle was iterated multiple times, and the most favorable result—having the least distance loss—was selected. Finally, PCA was performed on the set of embeddings; hence, the embedded vectors were used to represent the systems using principal components 1 and 2. This provides a compact and insightful representation of the complex high-dimensional dynamics inherent in the protein-ligand interactions. By facilitating the extraction of simple features, this embedding can deepen our understanding of global differences in systems.

In addition, the characteristic dynamics were extracted using the function $g(\boldsymbol{x}_i)$ (Fig. 1f). This function quantifies the contribution of a single short-term trajectory (the trajectory of the distance descriptor of one specific binding pocket residue) to the overall differences between the two systems. When juxtaposing the LDE trajectory of system i with that of reference system j, the function is represented as

$$g_{ij}(\boldsymbol{x}_i) = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{y}_i} \left[f_{ij}(\boldsymbol{x}_i) - f_{ij}(\boldsymbol{x}) \right] \tag{6}$$

Here lies the utility of $g(\boldsymbol{x})$: it quantitatively evaluates the uniqueness of a given short-term trajectory in comparison to the average trajectory of another system. For instance, a small $g(x)$ value for a trajectory in system i relative to system j suggests that system i's trajectory closely mirrors the general behavior observed in system j and vice versa. Building on this, because $g_{ij}(\boldsymbol{x}_i)$ encompasses short-term trajectories that span numerous residues, we can derive the residues that significantly affect the Wasserstein distance between systems, effectively shedding light on the contrasting protein differences (Fig. 1f). According to $g_{ij}(x_i)$, the short-term trajectories of system i are classified into three distinct groups: system$i$-characteristic, denoted as X_{ij}^C; system$j$-similar, denoted as X_{ij}^S; and middle X_{ij}^M:

$$\boldsymbol{x}_i \in \begin{cases} X_{ij}^C & \text{if } g_{ij}^C \leq g_{ij}(\boldsymbol{x}_i) \\ X_{ij}^S & \text{if } g_{ij}(\boldsymbol{x}_i) \leq g_{ij}^S \\ X_{ij}^M & \text{if } g_{ij}^S < g_{ij}(\boldsymbol{x}_i) < g_{ij}^C \end{cases} \tag{7}$$

The higher and lower thresholds g_{ij}^C and g_{ij}^S are determined by the top and bottom deciles of all sampled values of $g_{ij}(\boldsymbol{x}_i)$. In contrast to a previous study that utilized only fluctuation [51], we focused on fluctuating conformations represented by residue–residue distances, taking distance-based interpretations. If the average distance between the center of mass of residue k and the center of the geometry of the binding pocket is very different between groups X_{ij}^C and X_{ij}^S, the Wasserstein distance W_{ij} is highly influenced by residue k. Through this analysis, we identified the residues whose dynamics were highly 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 serve as inputs. Sampling of $g(x)$ was conducted at 64 ps intervals throughout the Molecular Dynamics (MD) trajectories.

Results and discussion

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

To compare the local conformational dynamics of M^{pro} in the presence and absence of 11 inhibitors, we conducted simulations of dimeric M^{pro} in inhibitor-unbound (apo state) and inhibitor-bound (holo state) states. We performed three MD simulations of 1 µs for each of the 12 systems (apo-protein system and 11 protein-ligand systems). To monitor the structural stability of M^{pro} during the simulations, we measured the RMSD of the C_{alpha} atoms from the starting crystallographic coordinates. As shown in the Supplementary Information (Supplementary Fig. 9), the plotted RMSD for MD run 1 provides evidence that all the simulated systems have reached convergence. The residue-based RMSF through the trajectory was calculated to assess the flexibility

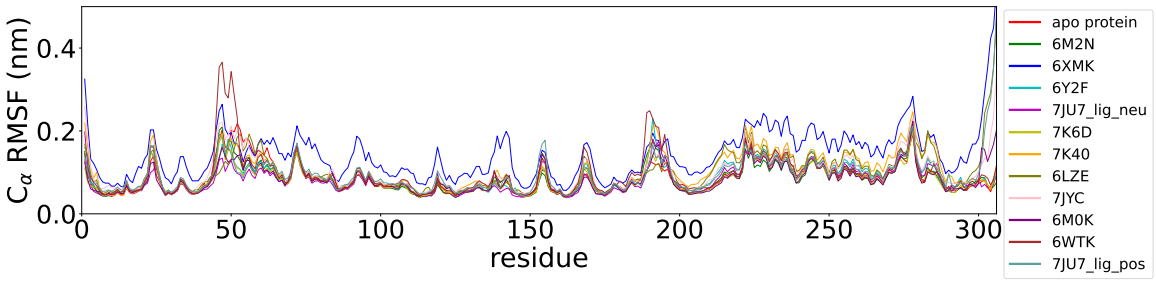


Figure 3: Residue-based root mean squared fluctuation (RMSF) of the protein backbone averaged between monomer A and monomer B in the first 1 µs MD simulation for the 12 systems.

of the residues (RMSF plot of MD run 1 in Fig.3). We computed the RMSF for each chain of the dimeric M^{pro} in every system and calculated the mean RMSF values between the two monomers. The overall RMSF analysis of the systems confirmed the structural flexibility of M^{pro}. The conformational flexibility of M^{pro} was experimentally assessed by Kneller et al. [88]. The structural heterogeneity of M^{pro} has also been highlighted in other studies using computational methods [89–92]. The flexibility of the protein structure plays a significant role in determining the thermodynamic properties of drug binding. This underscores the importance of considering intrinsic conformational flexibility and conformational selection when studying protein-ligand interactions [103, 104]. The RMSF data showed that the region from residue 45 to 53 and the region from residue 185 to 200 of the two protomers had a high RMSF. The largest differences in fluctuations between the systems were associated with these regions. Our findings find support in the study by Gorgulla et al. [91], which revealed the differences in the conformation and position of the Gln189-containing loop and the short Ser46-containing a-helix between three apo structures and five structures in the complex with inhibitors. These regions correspond to the two loops that enclose the catalytic pocket and physically occlude the path toward the catalytic site. The ligand-free system showed higher fluctuations than some protein-ligand systems and lower fluctuations than others. In conjunction with the RMSF data, this finding suggests that ligand binding cannot be simply correlated with the higher/lower induced fluctuation of M^{pro} residues. The approach proposed in this paper can overcome these challenges. Unsupervised deep learning can elucidate complex dynamic properties by detecting hidden patterns in MD data that conventional analysis methods such as RMSF cannot uncover. The RMSF plots for MD runs 2 and 3 are presented in the Supplementary Material (Supplementary Fig. 12 and 13).

Selection of binding-site residues and MD trajectory time windows

For effective analysis of the MD trajectory data, three core parameters must be determined: binding-site residues, appropriate time windows, and input type (descriptor and definition of LDE). The selection of the binding-site residues and time windows is rooted in MD analyses, while the input types require testing because of their dependence on the nature of the protein.

The binding-pocket residues of M^{pro} were determined through contact analysis, as detailed in the Methods section. The selected residues were as follows: His41, Met49, Phe40, Leu141, Asn142, Gly143, Ser144, Cys145, His163, His164, Met165, Glu166, Pro168, His172, Arg188, Gln189, Thr190, and Gln192. A comprehensive list of the amino acids in contact with each ligand over the three simulations is presented in Table 2.

The next core step involved the selection of frame windows guided by PCA. In the context of our MD simulations, PCA helped distinguish stable molecular conformations from fluctuations, ensuring that the chosen time intervals accurately represented the local changes induced by ligand binding. First, we visually inspected the MD trajectories using the VMD tool, supplemented by ligand RMSD plots available in the Supplementary Information (Supplementary Fig. 15). These plots were instrumental in monitoring the ligand movement relative to M^{pro}. A noteworthy observation was made for the system 6M2N: the ligands, initially situated at the binding site of the two monomers, migrated out of the pocket in all three simulations. One potential contributor to this behavior may be the lower molecular weight of the ligand in the system 6M2N. Consequently, it was not possible to identify the pertinent period of ligand interaction within the binding pocket of one of the protomers, and this system was excluded from further analysis. Our primary objective for using PCA was to identify the time windows that embodied stable conformations during pivotal ligand interaction events. Through this analysis, we identified a window of 300 ns demonstrating the enhanced structural stability of the protein-ligand complex. A visual illustration of our PCA results is provided in figure 4 for apo-protein system and a protein-ligand system as an example. The PCA plots of the selected frame windows, representative of stable conformations, for all the simulated systems are displayed in the Supplementary Information (Supplementary Fig. 14). In conclusion, contact analysis and PCA-based selection of time windows were central to guiding the relevance and efficiency

of the LDE trajectories, ensuring that our analysis captured the most relevant and stable interactions between ligands and proteins.

Table 2: Summary of the residues identified by the contact analysis conducted for each protein-ligand system over three MD simulations. The residues selected as binding-pocket residues are His41, Met49, Phe40, Leu141, Asn142, Gly143, Ser144, Cys145, His163, His164, Met165, Glu166, Pro168, His172, Arg188, Gln189, Thr190, Gln192.

Cpd	System	Residues
1	6M0K	[41,49,140-145,163,164,165,166,172,188,189]
2	6LZE	[41,140-145,163,164,165,166,172,188]
3	6WTK	[140-145,163,164,166,172]
4	6XMK	[140-144,163,164,165,166,172,188-190]
5	6Y2F	[41,140-145,163,164,165,166,172,189,190]
6	6M2N	[140-142,144,145,163,164,166,172]
7	7JU7neu	[41,49,140,141,144,145,163-166,172,189]
7	7JU7pos	[41,49,140,141,144,145,163-166,172,189,190]
8	7K40	[41,142,165,166,167,188-190,192]
9	7JYC	[41,140,142,143,145,164-166,189,190,192]
10	7K6D	[41,49,141-145,163-168,188-192]

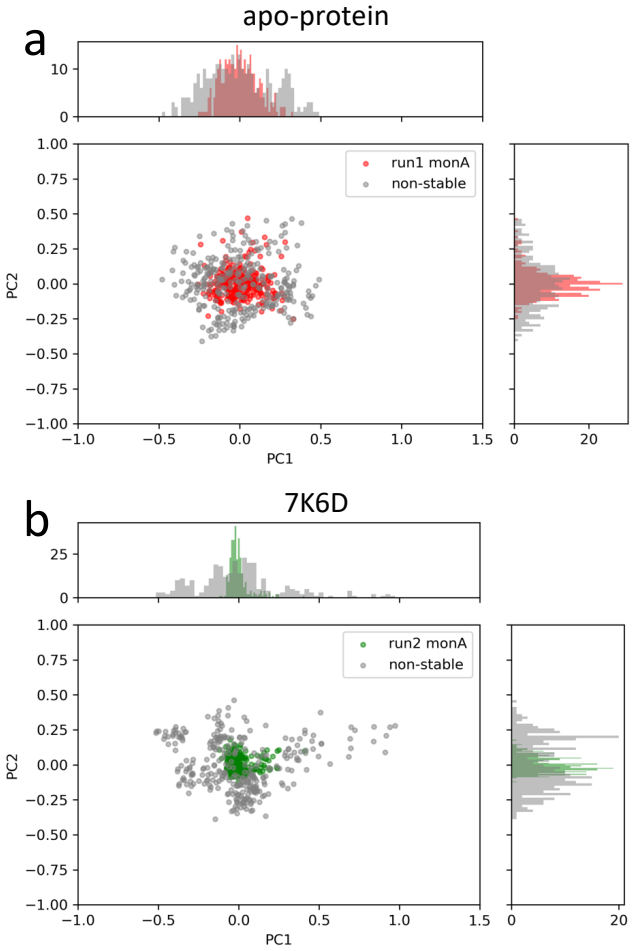


Figure 4: PCA plots of the stable-structure data selected for the generation of the LDEs of **a** apo-protein and **b** system 7K6D. In grey, PCA plots of non-stable-structure data for comparison.

Unsupervised deep learning-based insights into protein-ligand dynamics

Unsupervised deep learning offers the advantages of discovering hidden patterns and providing insights into complex datasets without prior labeling or categorization. Leveraging this approach, we sought to uncover the subtle protein-ligand interaction dynamics in the studied systems. Regarding the two other parameters for LDE, we selected the residue-pocket center distance and time series distances. For this selection, we assumed that both structure and fluctuation are important for representing flexible proteins, and fitting coordinates may result in a large bias for larger deformations.

Central to our methodology is the Wasserstein distance matrix derived from LDE. This 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 (Figure5a). This observation suggests that system 7JYC exhibits unique trajectories that were captured and highlighted by our unsupervised deep-learning methodology. Because we considered the time series of the distance to generate the LDE, the Wasserstein distance compares the probability distributions of the two LDEs, quantifying the differences in the conformations of the systems. While RMSD considers only the average difference between conformations, the Wasserstein distance also considers protein flexibility and is therefore more suitable for conveying a comprehensive view of fluctuating structures. Using the Wasserstein distance matrix, we constructed an embedding map that spatially arranges the systems. In this map, each system was represented as a point and its color corresponded to the experimental binding-affinity values ($pIC50$). A meaningful pattern emerged: systems with lower affinity values were situated closer to the apo-protein, indicating structural and dynamic behavior similar to that of the ligand-free state. Conversely, high-affinity systems were positioned further along PC2, indicating distinct ligand-influenced structures and dynamics (Figure 5b). We also noticed that the two systems with higher affinities, 6M0K and 6LZE, showed great similarities in the chemical structures of the ligands and were characterized by the same PC2 values. To reinforce the insights drawn from the embedding map, we correlated the experimental binding-affinity values (pIC50) with PC2 values from the embedding map. We observed a Pearson's correlation coefficient of 0.7 and a Spearman's correlation coefficient of 0.4. While the separation between high- (blue) and low-affinity (red) systems based on PC2 is evident, the classification of systems with moderate affinity seems complicated. This observation explains the differences between the two correlation metrics. The significant correlation between PC2 and IC50 for high- and low-affinity ligands, reflected by a Pearson's correlation of 0.7, indicates the potential of our deep-learning approach in highlighting the subtle shifts in ligand-induced trajectories within M^{pro} (Figure 6).

In addition to the time-series distance used as the descriptor of the MD, we investigated three other types of inputs: (1) time-series displacements of residue–pocket center distance, (2) time-series residue–pocket center xyz displacement, and (3) time-series displacements of residue–pocket center xyz displacement. The use of time-series displacements has demonstrated success in the previous study on rigid proteins [51], exhibiting a notable correlation with binding affinities. Time-series displacements primarily consider fluctuations, whereas time-series distances consider conformations. In the case of (1), although apo-proteins could not be distinguished from high-affinity ligands, low-affinity ligands were separated from high-affinity ligands (Fig. S9a). In case (2), high-affinity ligands were separated from low-affinity ligands within the embedding map (Fig. S9b). In case (3), 7K6D overlapped with 6M0K (Fig. S9c). These results indicate that the fluctuations themselves are insufficient to estimate binding-affinity-related features, and conformation is also important in the context of flexible proteins. When employing Cartesian coordinates, careful consideration must be given to the selection of fitting parameters. For example, if fitting encompasses terminal regions or involves the other monomer of the dimer, it can affect the Cartesian coordinates of the binding site regardless of the conformations of the binding pocket. In cases where the binding site undergoes significant changes, it is better to include the global protein conformations. Note that in contrast to the normal mode analysis, this embedding map does not have specific physical meanings in the PC axis.

In contrast to the correspondence between PC1 and binding affinity observed in this study, a previous study indicated a correlation between binding affinity and PC1 for rigid proteins [51]. Because PCs are determined to find the axes with the largest deviation, they depend on the number and nature of the systems. Hence, the ability to differentiate systems should be determined based on embedding maps to compare complex systems, although a single axis may be useful, as shown in Fig. 6. In addition, flexible proteins have a high degree of freedom, which may lead to a situation where a single Wasserstein distance is not sufficient to represent the various differences among the systems. Generally, if two high-dimensional manifolds are significantly different, the meaning of the distance becomes vague.

Interpretation of the contribution of residues to ligand-induced dynamics

The unsupervised deep-learning approach employed in this study enabled the extraction of significant features from the protein-ligand systems. Notably, the correlation between the PC2 component of the embedding map

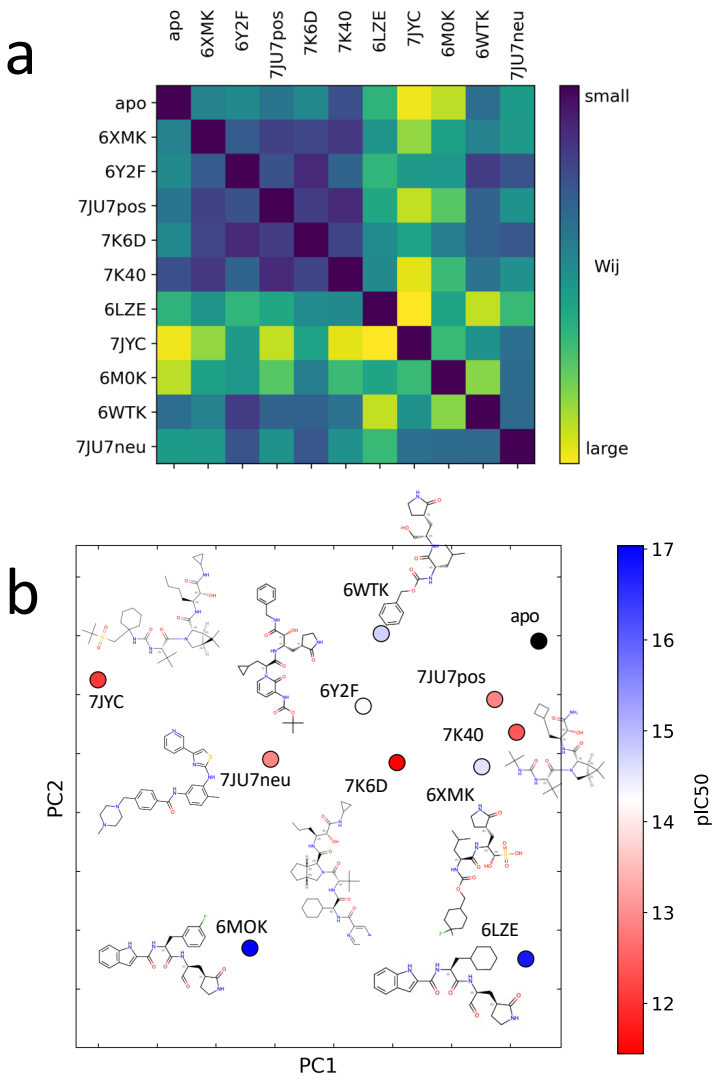


Figure 5: **a** Distance matrix of Wasserstein distances between the probability distributions of the LDEs for system pairs. A large Wasserstein distance (yellow) corresponds to a large difference in the protein structure and dynamics. **b** Embedded points of the distance matrix and chemical structure of the corresponding system. The points are colored according to the experimental binding-affinity values (pIC50). pIC50 corresponds to $-\log(IC50)$. IC50 values can be found in the table1.

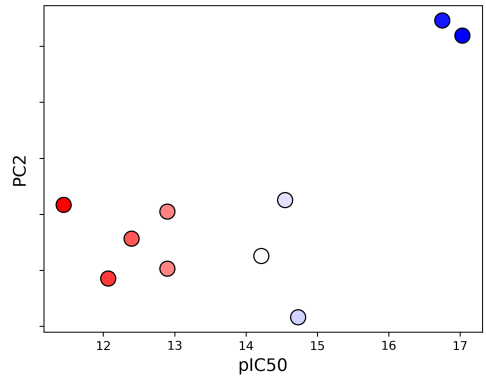


Figure 6: Correlation between PC2 and experimental binding-affinity data (pIC50). The correlation, quantified using Pearson Coefficient, is 0.7

and pIC50 indicates PC2's role in capturing conformational differences related 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 $g(x)$ (detailed in the methods section, see eq. 6), we examined the characteristic behavior differences between high- and low-affinity systems. This function allowed us to discern the characteristic dynamics of each system and to identify the residues that exhibited the most significant variations. First, according to the metrics derived from function $g_{ij}(x)$, the short-term trajectories of the LDE of system i are classified into three $g(x)$ groups, X_{ij}^{C} (high, characteristic of system i), X_{ij}^{S} (low, similar to system j) and X_{ij}^{M} (mid, non-characteristic of system i neither similar to system j). Then, the average value of the distance descriptor was calculated for each residue included in the LDE trajectories of the system i for each of the three LDE groups X_{ij}^{C}, X_{ij}^{M}, X_{ij}^{S}. Figure 7 shows the average distance from the center of the pocket for each LDE-residue of system 6M0K (with high binding affinity) when compared to system 7JYC (low binding-affinity system). The characteristic behavior X_{ij}^{C} in system 6M0K exhibited large movements in residues Met49 and Arg188-Gln189-Thr190. The largest differences between groups X_{ij}^{C} and X_{ij}^{S} corresponded to residues Arg188-Gln189-Thr190 and Met49. We also compared the characteristic trajectories of system 6M0K and system 6LZE (Figure 7b). In this case, large differences in the residues Arg188-Gln189-Thr190 between the characteristic (high) and similar (low) groups were absent, whereas the distinctions in residue Met49 persisted. The interpretation of this result in combination with visual inspection of the embedding map led us to conclude that (1) residues Met49 and Arg188-Gln189-Thr190 are highly influenced by the ligand-binding M^{pro}; (2) the conformation of residues Arg188-Gln189-Thr190, which is highly different between high- and low-affinity ligands, is predominantly represented in the PC2 feature; and (3) the conformation of residue Met49 is captured in PC1.

To support our findings, we referred to studies that offer complementary insights. MacDonald et al. [105] described how changes in substrate accommodation can cause significant alterations in catalytic efficiency. A widened active-site cleft between the M^{pro} residues Met49 and Asn142 led to decreased catalytic efficiency for the nsp8/9 substrate. This observation underscores Met49's critical role in ligand recognition and binding dynamics, consistent with our findings. Through binding free-energy decomposition analysis, Hamed et al. [106] highlighted the pivotal role of specific residues in ligand interactions. In addition to identifying Asp187 and Asp48 as essential for β-blocker agents, this study also highlighted the important roles of Met49 and Thr190, further validating our observations. A comprehensive analysis conducted by Amamuddy et al. [107] identified heightened mobility in residues such as Met49 and Tyr54, supporting our findings concerning Met49's significant movements. Furthermore, the identification of residues Asp187, Arg188, Gln189, Thr190, and Ala191 as flexible in slower modes indicates the importance of these residues in functional motion, which is consistent with our conclusions. Investigating M^{pro} mutations, Yang [108] et al. highlighted residues such as Met49 and Arg188-Gln189-Thr190 as pivotal for protein-ligand interactions. Their analysis of mutations affecting nirmatrelvir binding, particularly at Gln189 and Arg188, resonated with our findings, emphasizing the importance of these residues in ligand interaction and potential drug resistance. The shared insights across these independent studies bolster the robustness of our conclusions, contributing to a more comprehensive understanding of the dynamics governing protein-ligand interactions in M^{pro}.

Conclusions

In modern drug discovery, protein-ligand interactions play a crucial role in determining the efficacy and specificity of potential therapeutic agents. Traditional methods such as X-ray crystallography and NMR spectroscopy provide structural snapshots but often lack the capability to capture the dynamic nature of these interactions. MD simulations have emerged as a valuable tool in the drug-discovery process by providing a detailed characterization of the temporal evolution of protein-ligand systems at the atomic level. However, analyzing the vast datasets generated by MD simulations remains challenging. In this context, the integration of deep-learning techniques with MD simulations is a promising approach. Unsupervised deep-learning approaches can efficiently handle high-dimensional data and extract meaningful patterns and relationships. In this study, we adopted an unsupervised deep-learning framework specifically tailored for the analysis of MD simulation data of flexible protein-ligand complexes. We assessed the ability of our ML approach to capture patterns in MD trajectories induced by 11 different ligands of the SARS-CoV-2 main protease. To enhance both the relevance and efficiency of MD data analysis using ML, we focused on selected binding-pocket residues and time windows in stable protein conformations. The third core parameter to be determined for an effective analysis is the type of input: the descriptor (distance or coordinates) and the definition of LDE (time series or time displacements). After testing different types of inputs, we selected the time series of the distance between the centers of mass of the binding-site residues and the center of geometry of the binding pocket. As discussed in the previous section, Cartesian coordinates exhibited sub-optimal performance due to susceptibility to fitting selection and subsequent issues with coordinate rotations, which can compromise the representation of protein conformational

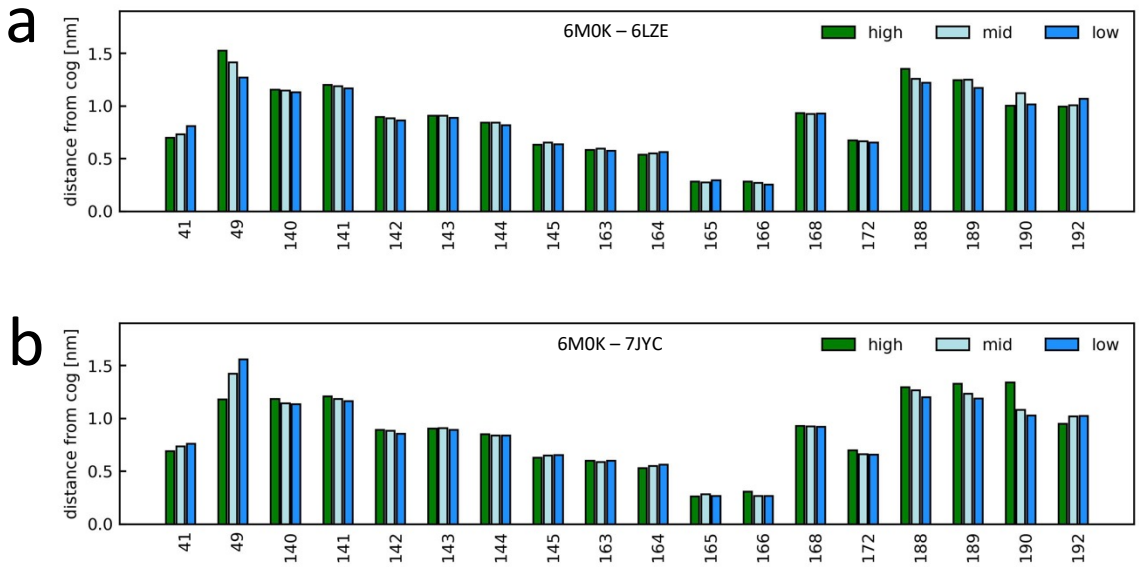


Figure 7: Characteristic dynamics were compared for selected system pairs, and the contributions of the binding-site residues were interpreted. The short-term trajectories of system i were classified into characteristic (high, characteristic of system i), non-characteristic (low, similar to system j), and others (mid), and the average value of the distance from the pocket center was calculated for each binding-site residue. **a** Characteristic dynamics analysis for system 6M0K (high-affinity system) compared to system 7JYC (low-affinity system). **b** Characteristic dynamics analysis for system 6M0K compared to system 6LZE (both high-affinity systems).

landscapes. In future investigations, it would be intriguing to incorporate bond angles and assess the performance of our method using this equivariant model as input [36, 37, 109]. Other types of features of protein complexes, such as surface volume, and features of ligands, such as molecular weights, have been used in previous works [8, 45, 110, 111]. In a supervised learning framework, these features are useful after normalization and the determination of optimal weight parameters. However, within our unsupervised learning framework, seamlessly incorporating such information into the coordinates of the protein is a complex task. We hypothesize that a self-supervised learning scheme might offer a viable avenue for achieving this, and we view it as a promising direction for our future research. Subsequently, Wasserstein distances between the LDE trajectories of the residue–pocket center distance were calculated using DNNs across all system pairs. Dimensionality-reduction techniques were employed to extract relevant variables. The distances between the systems in the embedding map were interpreted and related to the experimental binding affinities. Systems with lower affinity values were located closer to the apo-protein, whereas high-affinity systems were positioned further along PC2. We found a significant Pearson's correlation coefficient (0.7 between the ligand-induced dynamics reflected in PC2 and the experimental binding-affinity data). This finding implies that the most active compounds have the maximum impact on the local structure and dynamics of the target protein, resulting in them being further distanced from the ligand-free system. Moreover, we determined the binding-site residues that contributed the most to the ligand-induced changes in M^{pro}. These findings are consistent with the latest literature on this topic.
In a previous study, the DNN approach was employed for relatively rigid proteins [51], while in this study, it was tested and adopted for M^{pro}, a protein known for its high degree of flexibility (as discussed in the Results section). A recent study by Gu et al. [112] demonstrated that the application of classical machine-learning algorithms to MD trajectory-derived descriptors significantly enhanced the prediction performance of binding affinities for protein targets exhibiting considerable structural flexibility. The importance of using MD-generated descriptors instead of static 3D structural data of protein-ligand complexes as inputs has also been demonstrated by Ash and Fourches [113]. Additionally, complementary studies [45, 114, 115] further resonated with our approach, where a combination of DNNs with MD was deployed to capture the complex, nonlinear relationships in high-dimensional MD simulation data to leverage the intricate dynamics induced by the ligand. In the domain of methodologies that leverage deep learning for trajectory analysis, a noteworthy mention goes to the VAMPNet framework [49]. VAMPNet framework has indeed made significant contributions to the field by employing the variational approach for Markov processes (VAMP) to acquire a kinetic model from MD data. However, we would like to emphasize that there are notable distinctions with our approach that reflect their

specific applications and capabilities. VAMPNets excel at extracting metastable structures and determining the rate of conformational transition within a single system. In contrast, our method specifies the time scale of dynamics, and dynamical conformations are analyzed among multiple systems. This allows us to examine how dynamics vary across a set of systems. It's worth noting that we envision a potential synergy between our approach and VAMPNets, where the use of VAMPNets to extract input features for our method holds promise for enhancing the analysis of variances in conformational transitions. This underscores the complementary nature of our approach within the landscape of trajectory analysis methodologies. While our results are promising, it is important to acknowledge possible limitations and the directions for future work. Firstly, it is noteworthy that our approach is dependent on the initial conditions, specifically the initial structure of the protein and the chosen input feature. If the crystal structure of the protein-ligand complex is unavailable, docking plays a critical role in ensuring the effectiveness of the analysis. In addition, the sampling of MD simulations and the associated computational time are also recognized as intrinsic limitations. Efforts to optimize sampling strategies and potentially employ more efficient simulation techniques, such as metadynamics, are areas for future consideration [116, 117]. The integration of advanced simulation techniques to improve sampling could also be beneficial to reduce the dependence on the quality of the docking. Moreover, our study focused on a set of 11 ligands. Expanding this dataset to encompass a wider range of ligands will be crucial for a more comprehensive understanding of the method's capabilities. Looking ahead, we believe that the unsupervised deep-learning framework utilized in this study will be highly valuable in the early stages of drug discovery. When binding-affinity data are not yet available, this method may help identify the most promising compounds to prioritize for further analysis. The versatility of our approach offers potential extensions also to diverse protein-ligand interactions, including allosteric events, and holds promise for lead optimization. Using our approach, the effects of different variants of the same ligand can be analyzed to gain insights into the influence of ligand modifications on the dynamics of the target protein. Future work will also focus on extending our method to other datasets, and on leveraging the power of deep learning for feature selection. Integrating feature selection directly into the automated machine-learning component of our model will not only enhance the model's adaptability but also align it more closely with the objective of achieving a truly unsupervised approach. By harnessing the strengths of deep learning and MD simulations, we envision that our novel methodology will not only accelerate drug discovery but will also contribute to a deeper understanding of molecular mechanisms, thus paving the way for more targeted and efficient therapeutic interventions.

Bibliography and citations

References

[1] Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853, 2020.

[2] Michael Schlander, Karla Hernandez-Villafuerte, Chih-Yuan Cheng, Jorge Mestre-Ferrandiz, and Michael Baumann. How much does it cost to research and develop a new drug? a systematic review and assessment. *PharmacoEconomics*, 39:1243–1269, 2021.

[3] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[4] William L Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004.

[5] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe. Computational methods in drug discovery. *Pharmacological reviews*, 66(1):334–395, 2014.

[6] Aravindhan Ganesan, Michelle L Coote, and Khaled Barakat. Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug discovery today*, 22(2):249–269, 2017.

[7] Yao Zhao, Yan Zhu, Xiang Liu, Zhenming Jin, Yinkai Duan, Qi Zhang, Chengyao Wu, Lu Feng, Xiaoyu Du, Jinyi Zhao, et al. Structural basis for replicase polyprotein cleavage and substrate specificity of main protease from sars-cov-2. *Proceedings of the National Academy of Sciences*, 119(16):e2117142119, 2022.

[8] Yihang Wang, Joao Marcelo Lamim Ribeiro, and Pratyush Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current opinion in structural biology*, 61:139–145, 2020.

[9] Jacob D Durrant and J Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9(1):1–9, 2011.

[10] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184–10189, 2011.

[11] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling. *Nature chemistry*, 9(10):1005–1011, 2017.

[12] Robert A Copeland, David L Pompliano, and Thomas D Meek. Drug–target residence time and its implications for lead optimization. *Nature reviews Drug discovery*, 5(9):730–739, 2006.

[13] Clarisse G Ricci, Janice S Chen, Yinglong Miao, Martin Jinek, Jennifer A Doudna, J Andrew McCammon, and Giulia Palermo. Deciphering off-target effects in crispr-cas9 through accelerated molecular dynamics. *ACS central science*, 5(4):651–662, 2019.

[14] Xing Du, Yi Li, Yuan-Ling Xia, Shi-Meng Ai, Jing Liang, Peng Sang, Xing-Lai Ji, and Shu-Qun Liu. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144, 2016.

[15] Meir Wilchek, Edward A Bayer, and Oded Livnah. Essentials of biorecognition: The (strept) avidin–biotin system as a model for protein–protein and protein–ligand interaction. *Immunology letters*, 103(1):27–32, 2006.

[16] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[17] Pietro Cozzini, Glen E Kellogg, Francesca Spyrakis, Donald J Abraham, Gabriele Costantino, Andrew Emerson, Francesca Fanelli, Holger Gohlke, Leslie A Kuhn, Garrett M Morris, et al. Target flexibility: an emerging consideration in drug discovery and design. *Journal of medicinal chemistry*, 51(20):6237–6255, 2008.

[18] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, 59(9):4035–4061, 2016.

[19] Moon-Hyeong Seo, Jeongbin Park, Eunkyung Kim, Sungchul Hohng, and Hak-Sung Kim. Protein conformational dynamics dictate the binding affinity for a ligand. *Nature communications*, 5(1):3724, 2014.

[20] Ron O Dror, Robert M Dirks, JP Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–452, 2012.

[21] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[22] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[23] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13):136403, 2010.

[24] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9(1):3887, 2018.

[25] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

[26] Felix Brockherde, Leslie Vogt, Li Li, Mark E Tuckerman, Kieron Burke, and Klaus-Robert Müller. Bypassing the kohn-sham equations with machine learning. *Nature communications*, 8(1):872, 2017.

[27] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1):13890, 2017.

[28] Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. *Physical review letters*, 113(9):090601, 2014.

[29] Luigi Bonati, Yue-Yu Zhang, and Michele Parrinello. Neural networks-based variationally enhanced sampling. *Proceedings of the National Academy of Sciences*, 116(36):17641–17647, 2019.

[30] Jun Zhang, Yi Isaac Yang, and Frank Noé. Targeted adversarial learning optimized sampling. *The journal of physical chemistry letters*, 10(19):5791–5797, 2019.

[31] James McCarty and Michele Parrinello. A variational conformational dynamics approach to the selection of collective variables in metadynamics. *The Journal of chemical physics*, 147(20):–, 2017.

[32] Mohammad M. Sultan and Vijay S Pande. tica-metadynamics: accelerating metadynamics by using kinetically selected collective variables. *Journal of chemical theory and computation*, 13(6):2440–2447, 2017.

[33] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[34] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual review of physical chemistry*, 71:361–390, 2020.

[35] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

[36] Albert Musaelian, Anders Johansson, Simon Batzner, and Boris Kozinsky. Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. *arXiv preprint arXiv:2304.10061*, 2023.

[37] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[38] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, pages 1–11, 2023.

[39] Aldo Glielmo, Brooke E Husic, Alex Rodriguez, Cecilia Clementi, Frank Noé, and Alessandro Laio. Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, 121(16):9722–9758, 2021.

[40] Frank Noé. Machine learning for molecular dynamics on long timescales. *Machine learning meets quantum physics*, pages 331–372, 2020.

[41] Shreyas Kaptan and Ilpo Vattulainen. Machine learning in the analysis of biomolecular simulations. *Advances in physics: X*, 7(1):2006080, 2022.

[42] Ambrose Plante, Derek M Shore, Giulia Morra, George Khelashvili, and Harel Weinstein. A machine learning approach for the discovery of ligand-specific functional mechanisms of gpcrs. *Molecules*, 24(11):2097, 2019.

[43] Mariarosaria Ferraro, Elisabetta Moroni, Emiliano Ippoliti, Silvia Rinaldi, Carlos Sanchez-Martin, Andrea Rasola, Luca F Pavarino, and Giorgio Colombo. Machine learning of allosteric effects: the analysis of ligand-induced dynamics to predict functional effects in trap1. *The Journal of Physical Chemistry B*, 125(1):101–114, 2020.

[44] Filippo Marchetti, Elisabetta Moroni, Alessandro Pandini, and Giorgio Colombo. Machine learning prediction of allosteric drug activity from molecular dynamics. *The journal of physical chemistry letters*, 12(15):3724–3732, 2021.

[45] Salma Jamal, Abhinav Grover, and Sonam Grover. Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against alzheimer's disease. *Frontiers in pharmacology*, 10:780, 2019.

[46] Diogo M Camacho, Katherine M Collins, Rani K Powers, James C Costello, and James J Collins. Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592, 2018.

[47] Jintae Kim, Sera Park, Dongbo Min, and Wankyu Kim. Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, 22(18):9983, 2021.

[48] Maha A Thafar, Rawan S Olayan, Haitham Ashoor, Somayah Albaradei, Vladimir B Bajic, Xin Gao, Takashi Gojobori, and Magbubah Essack. Dtigems+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, 12(1):1–17, 2020.

[49] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):5, 2018.

[50] Tian Xie, Arthur France-Lanord, Yanming Wang, Yang Shao-Horn, and Jeffrey C Grossman. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nature communications*, 10(1):2667, 2019.

[51] 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.

[52] Sven Ullrich and Christoph Nitsche. The sars-cov-2 main protease as drug target. *Bioorganic & medicinal chemistry letters*, 30(17):127377, 2020.

[53] Dafydd R Owen, Charlotte MN Allerton, Annaliesa S Anderson, Lisa Aschenbrenner, Melissa Avery, Simon Berritt, Britton Boras, Rhonda D Cardin, Anthony Carlo, Karen J Coffman, et al. An oral sars-cov-2 mpro inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593, 2021.

[54] Jennifer Hammond, Heidi Leister-Tebbe, Annie Gardner, Paula Abreu, Weihang Bao, Wayne Wisemandle, MaryLynn Baniecki, Victoria M Hendrick, Bharat Damle, Abraham Simón-Campos, et al. Oral nirmatrelvir for high-risk, nonhospitalized adults with covid-19. *New England Journal of Medicine*, 386(15):1397–1408, 2022.

[55] Catia Marzolini, Daniel R Kuritzkes, Fiona Marra, Alison Boyle, Sara Gibbons, Charles Flexner, Anton Pozniak, Marta Boffito, Laura Waters, David Burger, et al. Recommendations for the management of drug–drug interactions between the covid-19 antiviral nirmatrelvir/ritonavir (paxlovid) and comedications. *Clinical Pharmacology & Therapeutics*, 112(6):1191–1200, 2022.

[56] Lindsey Wang, Nathan A Berger, Pamela B Davis, David C Kaelber, Nora D Volkow, and Rong Xu. Covid-19 rebound after paxlovid and molnupiravir during january-june 2022. *MedRxiv*, 2022.

[57] Yu Wang, Xubo Chen, Wenying Xiao, Danyang Zhao, and Liuliu Feng. Rapid covid-19 rebound in a severe covid-19 patient during 20-day course of paxlovid. *Journal of Infection*, 85(5):e134–e136, 2022.

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

[59] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of sars-cov-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, 368(6489):409–412, 2020.

[60] 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.

[61] Wenhao Dai, Bing Zhang, Xia-Ming Jiang, Haixia Su, Jian Li, Yao Zhao, Xiong Xie, Zhenming Jin, Jingjing Peng, Fengjiang Liu, et al. Structure-based design of antiviral drug candidates targeting the sars-cov-2 main protease. *Science*, 368(6497):1331–1335, 2020.

[62] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[63] B Zhang, Y Zhao, Z Jin, X Liu, H Yang, and Z Rao. The crystal structure of covid-19 main protease in apo form, publ, 2020.

[64] HX Su, WF Zhao, MJ Li, H Xie, and YC Xu. Sars-cov-2 3cl protease (3cl pro) in complex with a novel inhibitor. *PDB Protein Data Bank*, 2020.

[65] Hai-xia Su, Sheng Yao, Wen-feng Zhao, Min-jun Li, Jia Liu, Wei-juan Shang, Hang Xie, Chang-qiang Ke, Hang-chen Hu, Mei-na Gao, et al. Anti-sars-cov-2 activities in vitro of shuanghuanglian preparations and bioactive ingredients. *Acta Pharmacologica Sinica*, 41(9):1167–1177, 2020.

[66] Athri D Rathnayake, Jian Zheng, Yunjeong Kim, Krishani Dinali Perera, Samantha Mackin, David K Meyerholz, Maithri M Kashipathy, Kevin P Battaile, Scott Lovell, Stanley Perlman, et al. 3c-like protease inhibitors block coronavirus replication in vitro and improve survival in mers-cov–infected mice. *Science translational medicine*, 12(557):eabc5332, 2020.

[67] K Tan, NI Maltseva, LF Welk, RP Jedrzejczak, and A Joachimiak. The crystal structure of sars-cov-2 main protease in complex with masitinib, 2020.

[68] Nir Drayman, Jennifer K DeMarco, Krysten A Jones, Saara-Anne Azizi, Heather M Froggatt, Kemin Tan, Natalia Ivanovna Maltseva, Siquan Chen, Vlad Nicolaescu, Steve Dvorkin, et al. Masitinib is a broad coronavirus 3cl inhibitor that blocks replication of sars-cov-2. *Science*, 373(6557):931–936, 2021.

[69] Babak Andi, Desigan Kumaran, Dale F Kreitler, Alexei S Soares, Jantana Keereetaweep, Jean Jakoncic, Edwin O Lazo, Wuxian Shi, Martin R Fuchs, Robert M Sweet, et al. Hepatitis c virus ns3/4a inhibitors and other drug-like compounds as covalent binders of sars-cov-2 main protease. *Scientific reports*, 12(1):12197, 2022.

[70] Wayne Vuong, Muhammad Bashir Khan, Conrad Fischer, Elena Arutyunova, Tess Lamer, Justin Shields, Holly A Saffran, Ryan T McKay, Marco J van Belkum, Michael A Joyce, et al. Feline coronavirus drug inhibits the main protease of sars-cov-2 and blocks virus replication. *Nature communications*, 11(1):4282, 2020.

[71] Chemical Computing Group Inc. Molecular operating environment (moe), 2016.

[72] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

[73] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[74] MJ ea Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, VPGA Barone, GA Petersson, HJRA Nakatsuji, et al. Gaussian 16, revision c. 01, 2016.

[75] DA Case, IY Ben-Shalom, SR Brozell, DS Cerutti, TE Cheatham III, VWD Cruzeiro, TA Darden, RE Duke, D Ghoreishi, MK Gilson, et al. Amber 2018; 2018. *University of California, San Francisco*, 2018.

[76] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[77] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

[78] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):–, 2007.

[79] Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[80] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[81] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[82] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[83] Athri D Rathnayake, Jian Zheng, Yunjeong Kim, Krishani Dinali Perera, Samantha Mackin, David K Meyerholz, Maithri M Kashipathy, Kevin P Battaile, Scott Lovell, Stanley Perlman, et al. 3c-like protease inhibitors block coronavirus replication in vitro and improve survival in mers-cov–infected mice. *Science translational medicine*, 12(557):eabc5332, 2020.

[84] Hai-xia Su, Sheng Yao, Wen-feng Zhao, Min-jun Li, Jia Liu, Wei-juan Shang, Hang Xie, Chang-qiang Ke, Hang-chen Hu, Mei-na Gao, et al. Anti-sars-cov-2 activities in vitro of shuanghuanglian preparations and bioactive ingredients. *Acta Pharmacologica Sinica*, 41(9):1167–1177, 2020.

[85] Nir Drayman, Krysten A Jones, Saara-Anne Azizi, Heather M Froggatt, Kemin Tan, Natalia Ivanovna Maltseva, Siquan Chen, Vlad Nicolaescu, Steve Dvorkin, Kevin Furlong, et al. Drug repurposing screen identifies masitinib as a 3clpro inhibitor that blocks replication of sars-cov-2 in vitro. *BioRxiv*, pages 2020–08, 2020.

[86] Chunlong Ma, Michael Dominic Sacco, Brett Hurst, Julia Alma Townsend, Yanmei Hu, Tommy Szeto, Xiujun Zhang, Bart Tarbet, Michael Thomas Marty, Yu Chen, et al. Boceprevir, gc-376, and calpain inhibitors ii, xii inhibit sars-cov-2 viral replication by targeting the viral main protease. *Cell research*, 30(8):678–692, 2020.

[87] Brandon J Anson, Mackenzie E Chapman, Emma K Lendy, Sergii Pshenychnyi, TD Richard, Karla JF Satchell, and Andrew D Mesecar. Broad-spectrum inhibition of coronavirus main and papain-like proteases by hcv drugs. 2020.

[88] Daniel W Kneller, Gwyndalyn Phillips, Hugh M O'Neill, Robert Jedrzejczak, Lucy Stols, Paul Langan, Andrzej Joachimiak, Leighton Coates, and Andrey Kovalevsky. Structural plasticity of sars-cov-2 3cl mpro active site cavity revealed by room temperature x-ray crystallography. *Nature communications*, 11(1):3202, 2020.

[89] Budheswar Dehury, Sarbani Mishra, and Sanghamitra Pati. Structural insights into sars-cov-2 main protease conformational plasticity. *Journal of Cellular Biochemistry*, 2023.

[90] Maria Bzówka, Karolina Mitusińska, Agata Raczyńska, Aleksandra Samol, Jack A Tuszyński, and Artur Góra. Structural and evolutionary analysis indicate that the sars-cov-2 mpro is a challenging target for small-molecule inhibitor design. *International Journal of Molecular Sciences*, 21(9):3099, 2020.

[91] Christoph Gorgulla, Krishna M Padmanabha Das, Kendra E Leigh, Marco Cespugli, Patrick D Fischer, Zi-Fu Wang, Guilhem Tesseyre, Shreya Pandita, Alec Shnapir, Anthony Calderaio, et al. A multi-pronged approach targeting sars-cov-2 proteins using ultra-large virtual screening. *Iscience*, 24(2):–, 2021.

[92] Terra Sztain, Rommie Amaro, and J Andrew McCammon. Elucidation of cryptic and allosteric pockets within the sars-cov-2 main protease. *Journal of chemical information and modeling*, 61(7):3495–3501, 2021.

[93] Georg Diez, Daniel Nagel, and Gerhard Stock. Correlation-based feature selection to identify functional dynamics in proteins. *Journal of Chemical Theory and Computation*, 18(8):5079–5088, 2022.

[94] Wei Deng, Curt Breneman, and Mark J Embrechts. Predicting protein- ligand binding affinities using novel geometrical descriptors and machine-learning methods. *Journal of chemical information and computer sciences*, 44(2):699–703, 2004.

[95] Florian Sittel, Abhinav Jain, and Gerhard Stock. Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates. *The Journal of Chemical Physics*, 141(1):–, 2014.

[96] Xin-Qiu Yao, Mohamed Momin, and Donald Hamelberg. Establishing a framework of using residue–residue interactions in protein difference network analysis. *Journal of chemical information and modeling*, 59(7):3222–3228, 2019.

[97] Charles C David and Donald J Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. *Protein dynamics: Methods and protocols*, pages 193–226, 2014.

[98] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[99] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[100] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[101] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[102] D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, page 6. San Diego, California;, 2015.

[103] Marta Amaral, DB Kokh, J Bomke, A Wegener, HP Buchstaller, HM Eggenweiler, P Matias, C Sirrenberg, RC Wade, and MJNC Frech. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nature communications*, 8(1):2276, 2017.

[104] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–796, 2009.

[105] Elizabeth A MacDonald, Gary Frey, Mark N Namchuk, Stephen C Harrison, Stephen M Hinshaw, and Ian W Windsor. Recognition of divergent viral substrates by the sars-cov-2 main protease. *ACS Infectious Diseases*, 7(9):2591–2595, 2021.

[106] Mohammed IA Hamed, Khaled M Darwish, Raya Soltane, Amani Chrouda, Ahmed Mostafa, Noura M Abo Shama, Sameh S Elhady, Hamada S Abulkhair, Ahmed E Khodir, Ayman Abo Elmaaty, et al. β-blockers bearing hydroxyethylamine and hydroxyethylene as potential sars-cov-2 mpro inhibitors: Rational based design, in silico, in vitro, and sar studies for lead optimization. *RSC advances*, 11(56):35536–35558, 2021.

[107] Olivier Sheik Amamuddy, Gennady M Verkhivker, and Ozlem Tastan Bishop. Impact of early pandemic stage mutations on molecular dynamics of sars-cov-2 mpro. *Journal of chemical information and modeling*, 60(10):5080–5102, 2020.

[108] Kai S Yang, Sunshine Z Leeuwon, Shiqing Xu, and Wenshe Ray Liu. Evolutionary and structural insights about potential sars-cov-2 evasion of nirmatrelvir. *Journal of Medicinal Chemistry*, 65(13):8686–8698, 2022.

[109] Zun Wang, Hongfei Wu, Lixin Sun, Xinheng He, Zhirong Liu, Bin Shao, Tong Wang, and Tie-Yan Liu. Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics. *The Journal of Chemical Physics*, 159(3), 2023.

[110] Debby D Wang, Le Ou-Yang, Haoran Xie, Mengxu Zhu, and Hong Yan. Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods. *Computational and structural biotechnology journal*, 18:439–454, 2020.

[111] Debby D Wang, Mengxu Zhu, and Hong Yan. Computationally predicting binding affinity in protein–ligand complexes: free energy-based simulations and machine learning-based scoring functions. *Briefings in bioinformatics*, 22(3):bbaa107, 2021.

[112] Shukai Gu, Chao Shen, Jiahui Yu, Hong Zhao, Huanxiang Liu, Liwei Liu, Rong Sheng, Lei Xu, Zhe Wang, Tingjun Hou, et al. Can molecular dynamics simulations improve predictions of protein-ligand binding affinity with machine learning? *Briefings in Bioinformatics*, 24(2):bbad008, 2023.

[113] Jeremy Ash and Denis Fourches. Characterizing the chemical space of erk2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *Journal of chemical information and modeling*, 57(6):1286–1299, 2017.

[114] Oleksandr Yakovenko and Steven JM Jones. Modern drug design: the implication of using artificial neuronal networks and multiple molecular dynamic simulations. *Journal of Computer-Aided Molecular Design*, 32:299–311, 2018.

[115] WF Drew Bennett, Stewart He, Camille L Bilodeau, Derek Jones, Delin Sun, Hyojin Kim, Jonathan E Allen, Felice C Lightstone, and Helgi I Ingólfsson. Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning. *Journal of Chemical Information and Modeling*, 60(11):5375–5381, 2020.

[116] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[117] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *The Journal of chemical physics*, 151(7), 2019.

Supplementary Information

Additional Figures

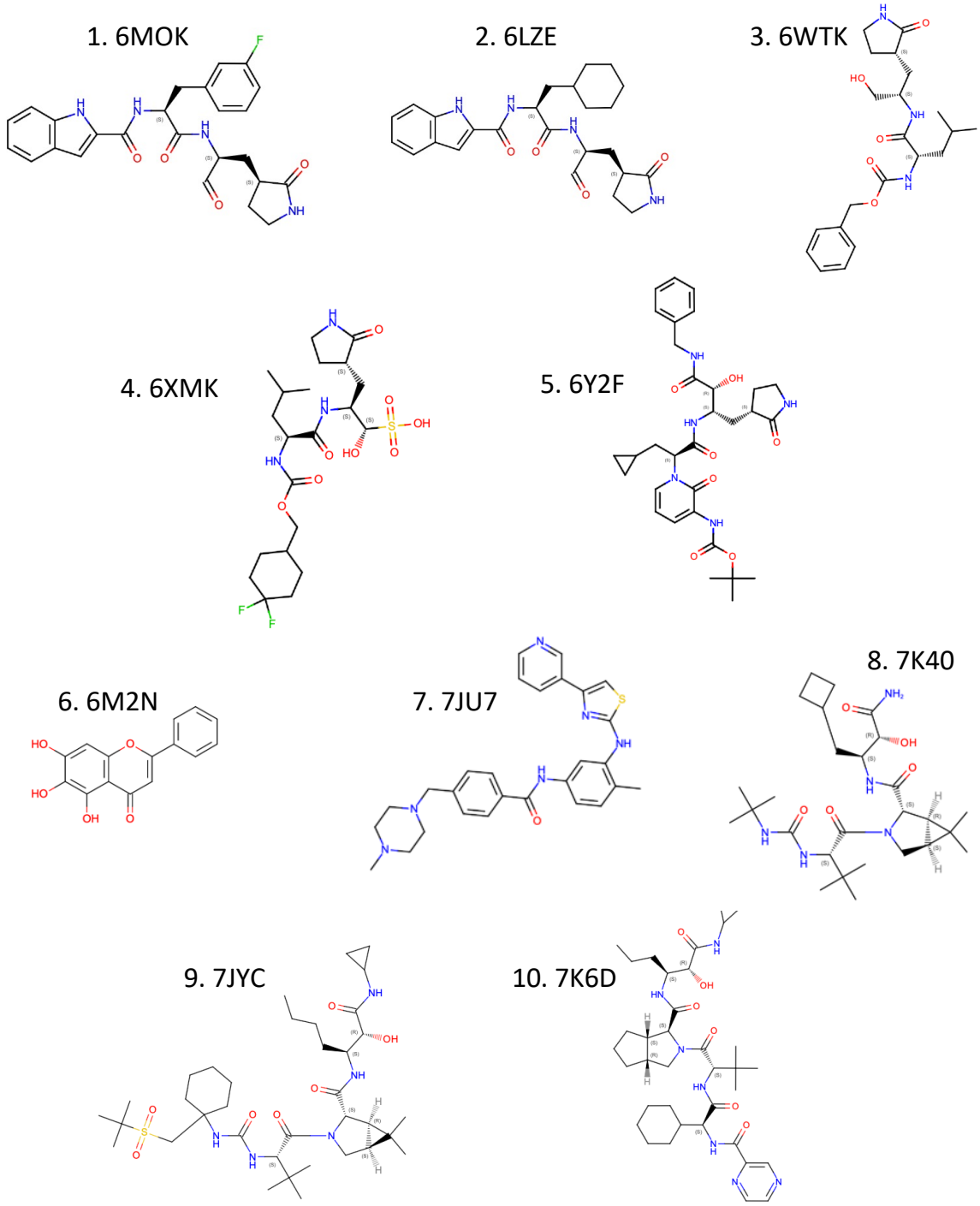


Figure 8: Chemical structure of the ligands. Compounds are labeled with Arabic numerals in descending order of affinity.

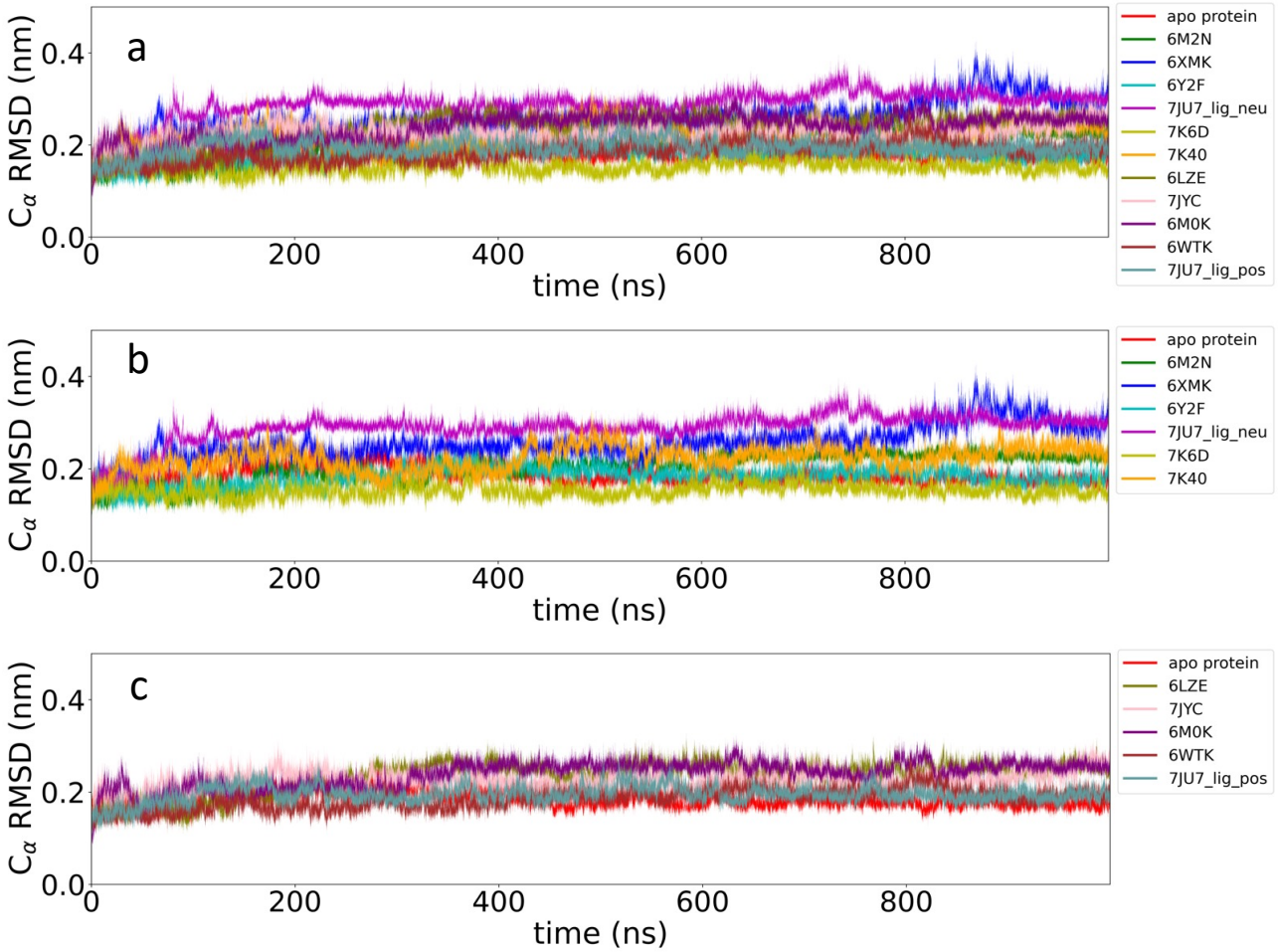


Figure 9: **a** Root mean squared deviation (RMSD) of the protein backbone in the first 1 µs molecular dynamics (MD) simulation for the 12 systems. In figure **b** and **c** the protein-ligand systems have been split for a more clear representation

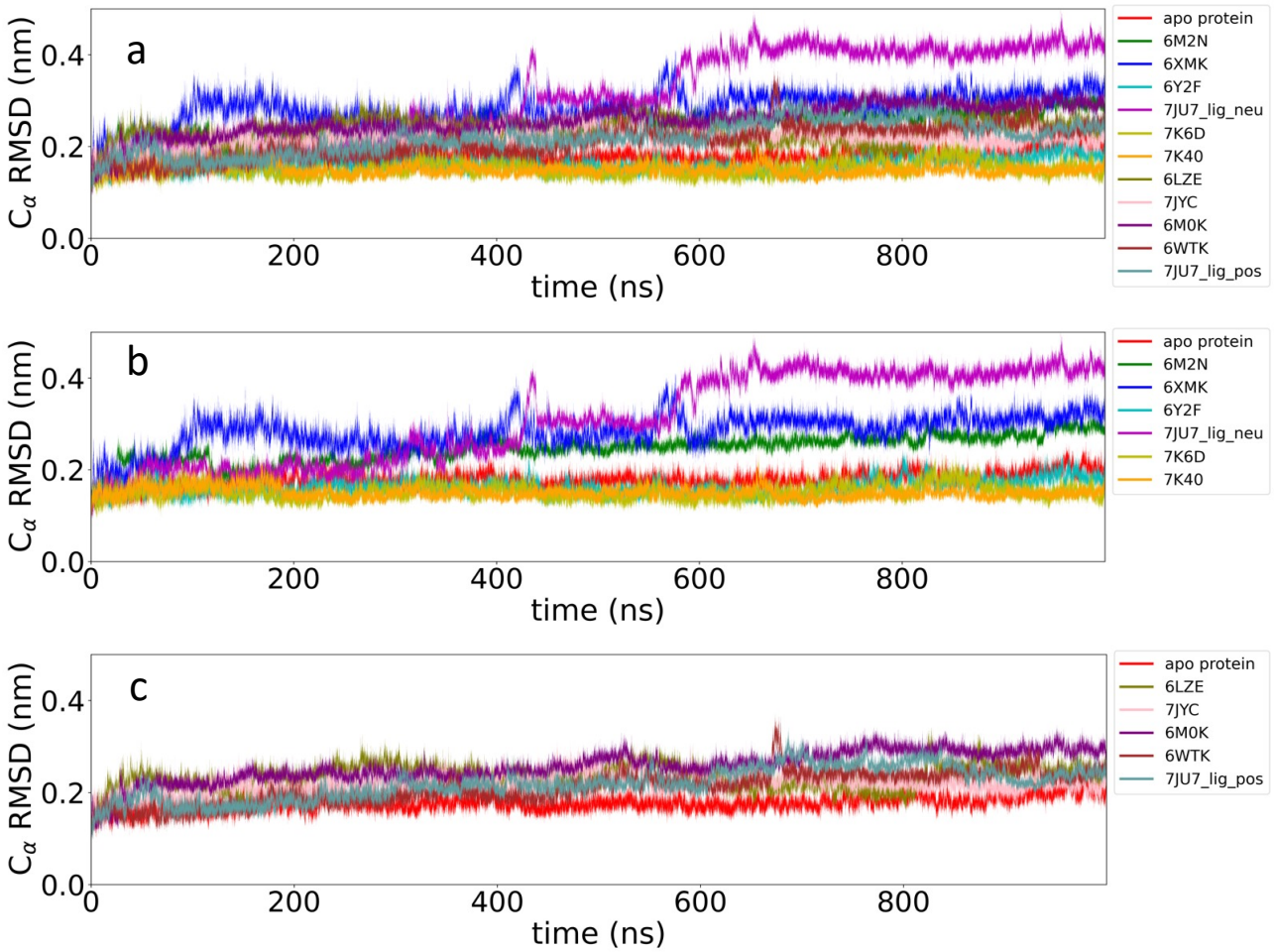


Figure 10: **a** Root mean squared deviation (RMSD) of the protein backbone in the second 1 µs molecular dynamics (MD) simulation for the 12 systems. In figure **b** and **c** the protein-ligand systems have been split for a more clear representation

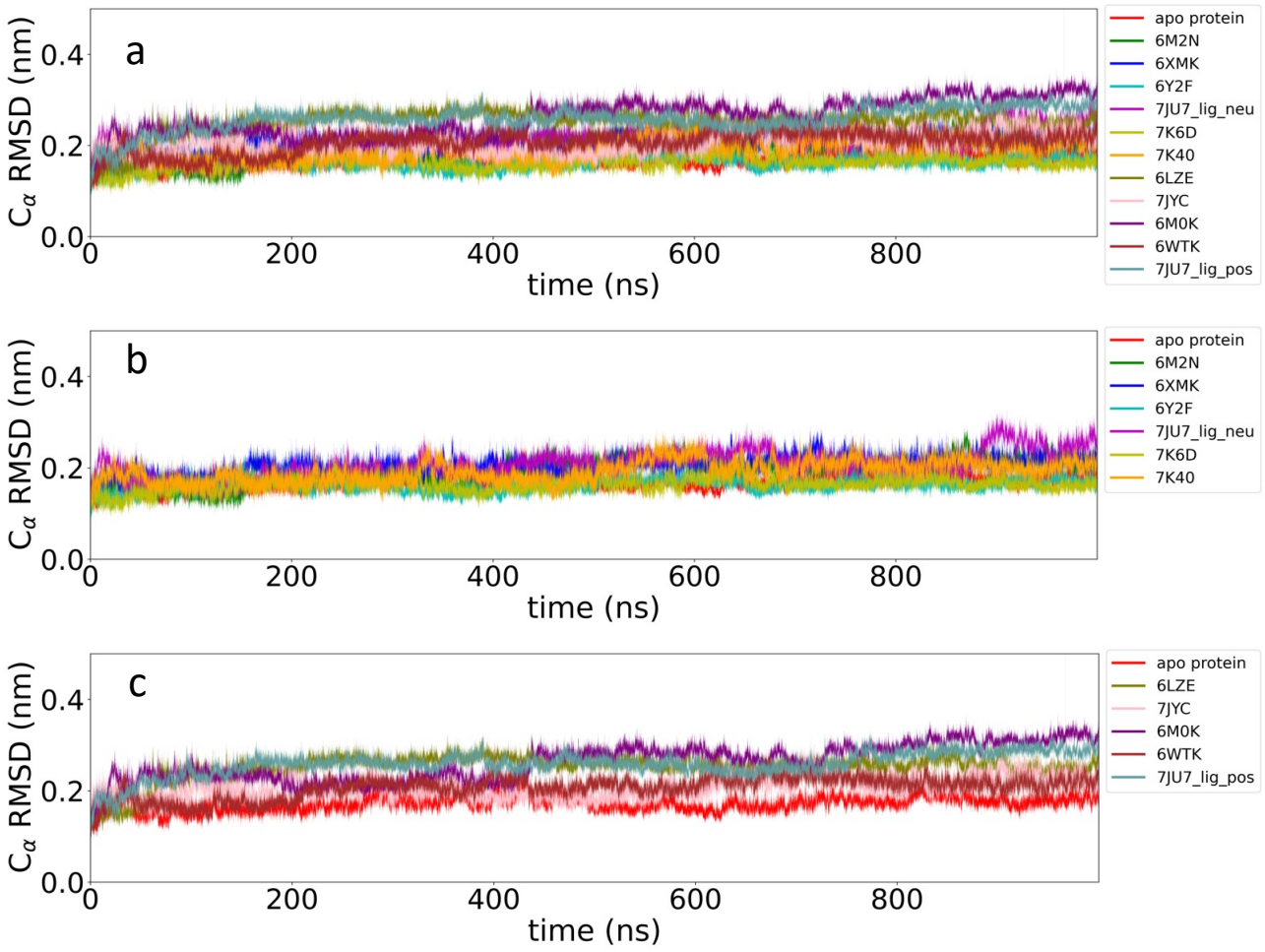


Figure 11: **a** Root mean squared deviation (RMSD) of the protein backbone in the third 1 µs molecular dynamics (MD) simulation for the 12 systems. In figure **b** and **c** the protein-ligand systems have been split for a more clear representation

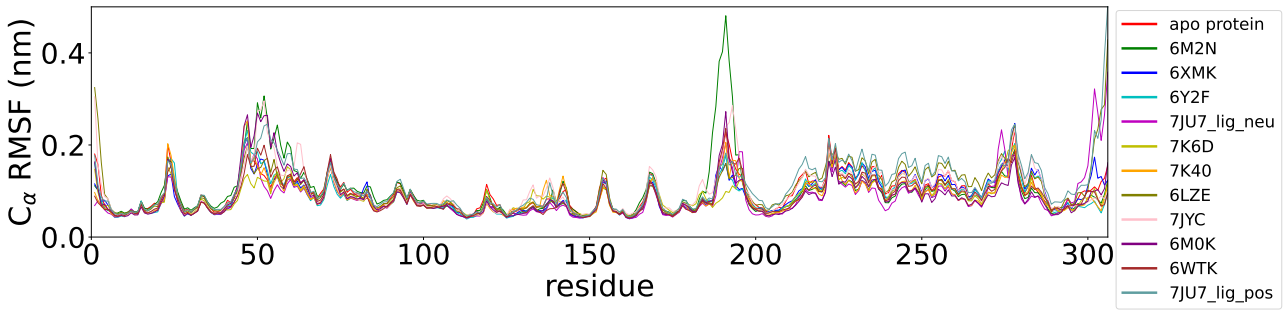


Figure 12: Residue-based root mean squared fluctuation (RMSF) of the protein backbone averaged between monomer A and monomer B in the second 1 µs MD simulation for the 12 systems.

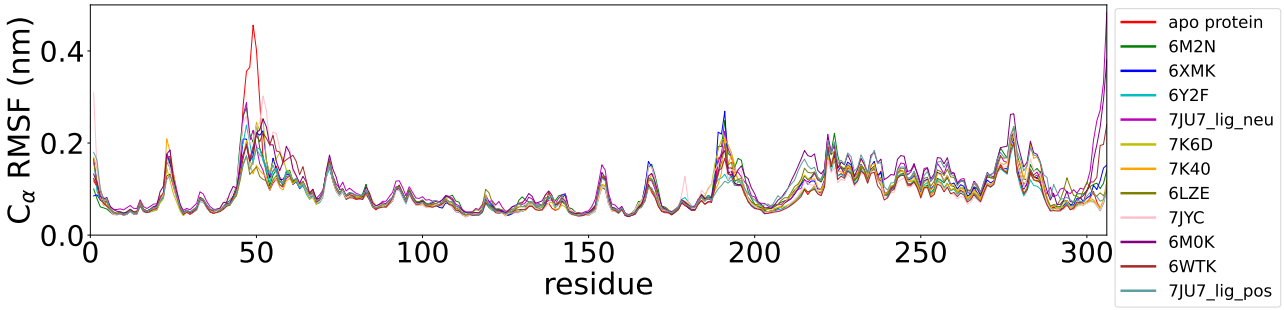


Figure 13: Residue-based root mean squared fluctuation (RMSF) of the protein backbone averaged between monomer A and monomer B in the third 1 µs MD simulation for the 12 systems.

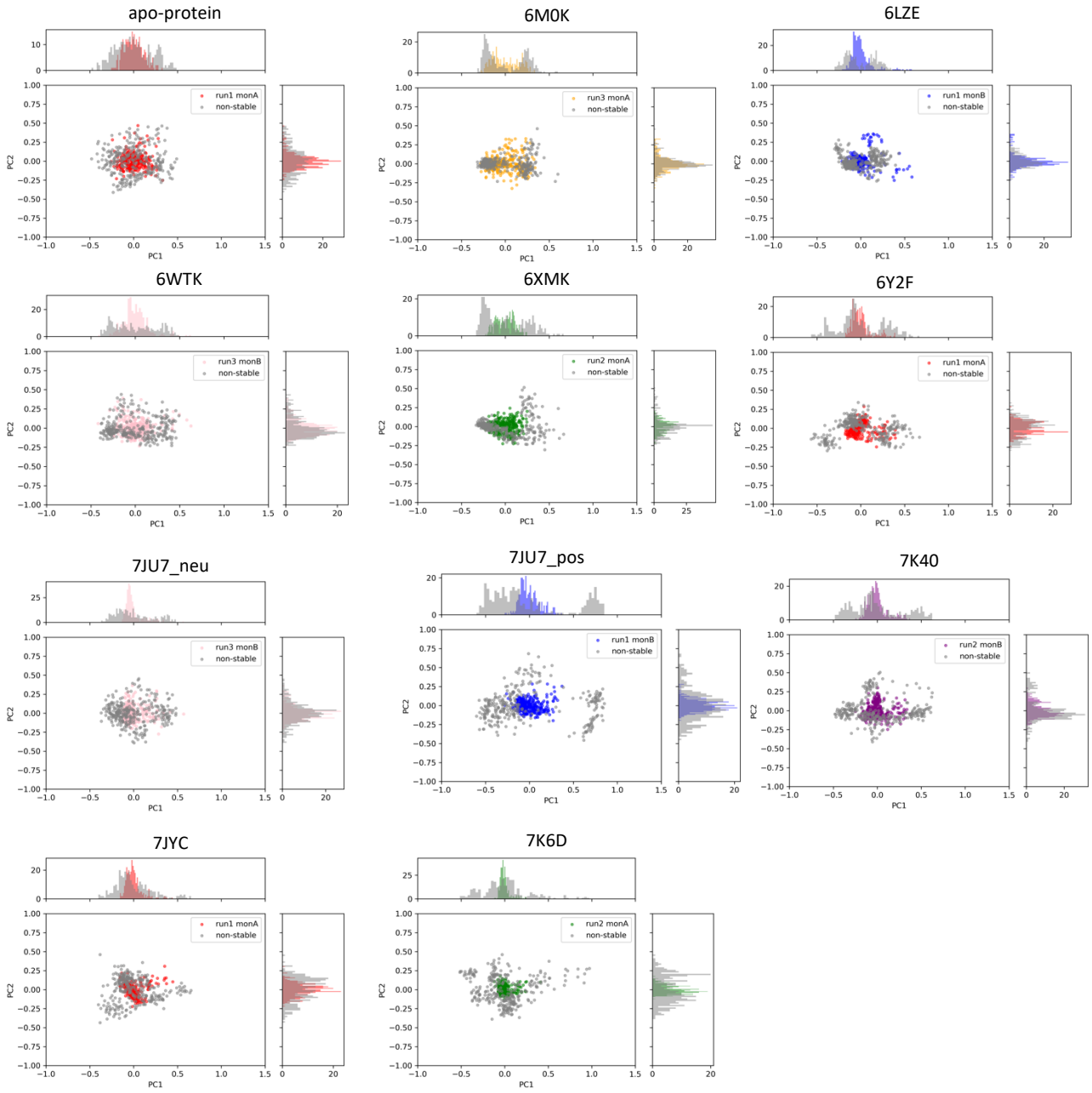
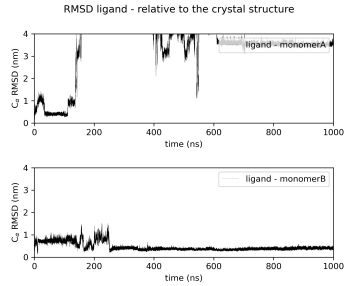
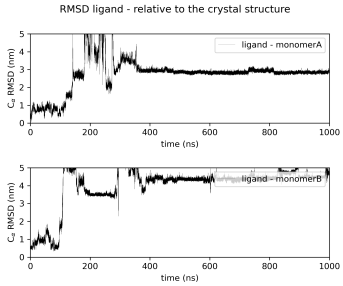


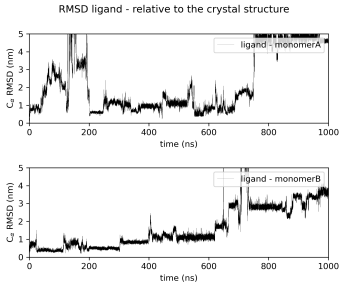
Figure 14: PCA plots of the stable-structure data selected for the generation of the LDEs. In grey, PCA plots of non-stable-structure data for comparison.



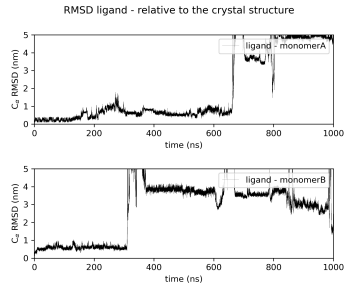
(a) RMSD ligand 6M2N run 1



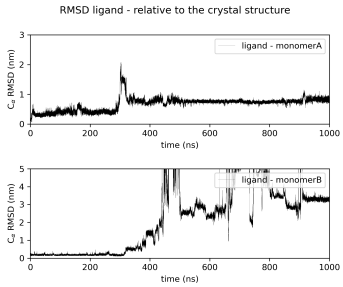
(b) RMSD ligand 6M2N run 2



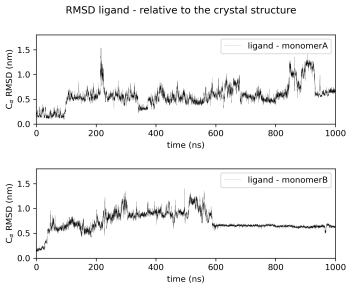
(c) RMSD ligand 6M2N run 3



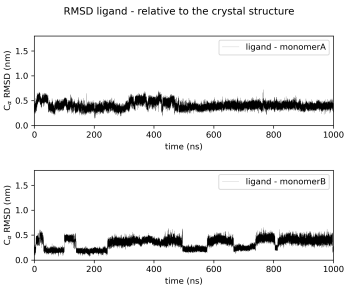
(d) RMSD ligand 6XMK run 1



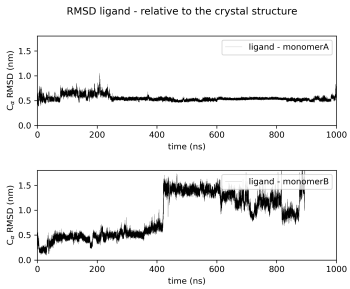
(e) RMSD ligand 6XMK run 2



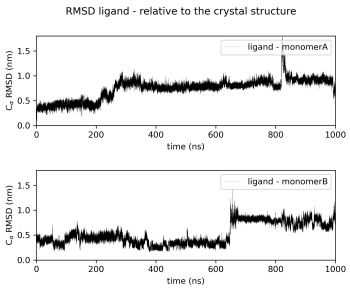
(f) RMSD ligand 6XMK run 3



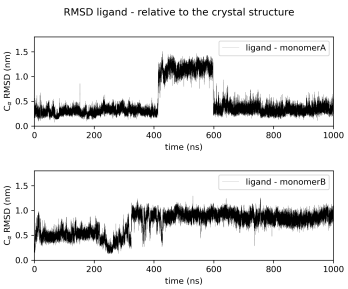
(g) RMSD ligand 6Y2F run 1



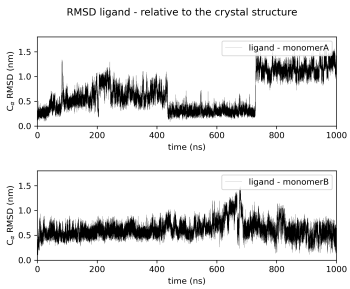
(h) RMSD ligand 6Y2F run 2



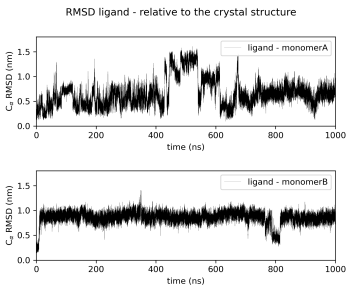
(i) RMSD ligand 6Y2F run 3



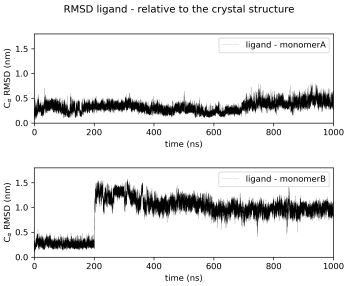
(j) RMSD ligand 7JU7pos run 1



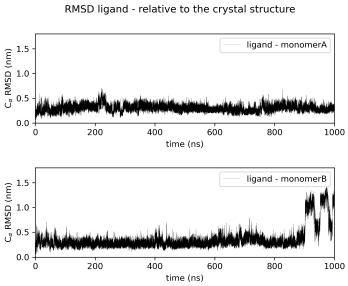
(k) RMSD ligand 7JU7pos run 2



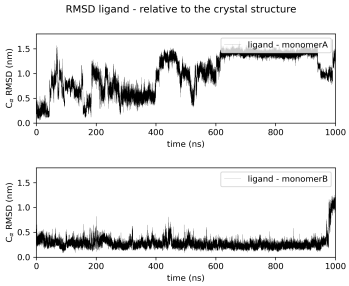
(l) RMSD ligand 7JU7pos run 3



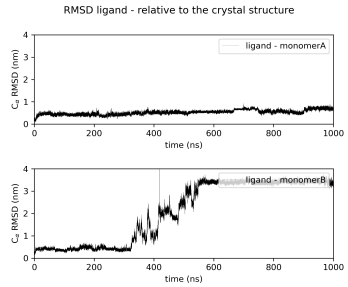
(a) RMSD ligand 7JU7neu run 1



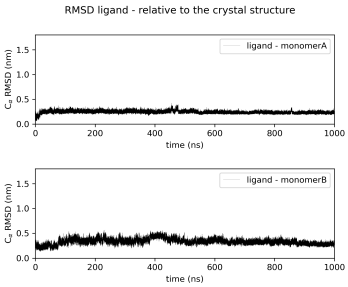
(b) RMSD ligand 7JU7neu run 2



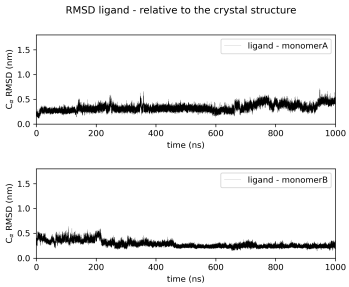
(c) RMSD ligand 7JU7neu run 3



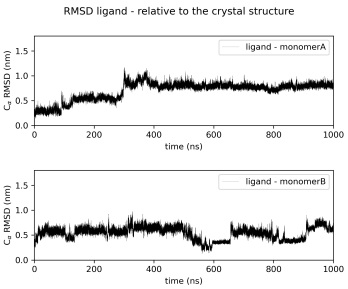
(d) RMSD ligand 7K6D run 1



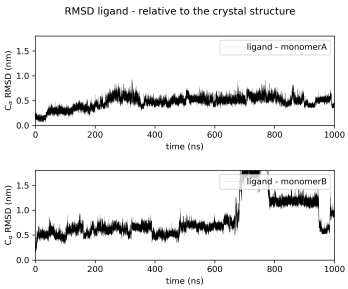
(e) RMSD ligand 7K6D run 2



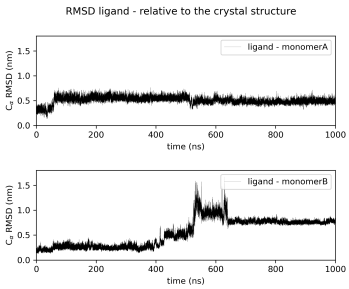
(f) RMSD ligand 7K6D run 3



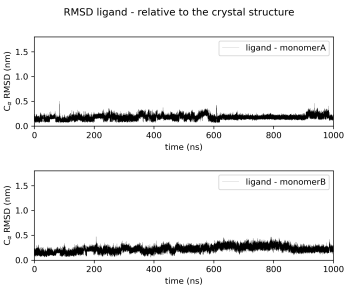
(g) RMSD ligand 7K40 run 1



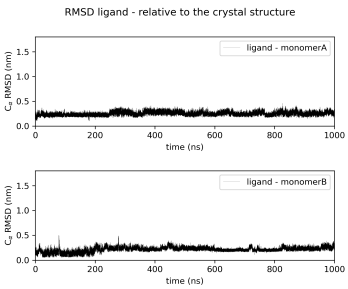
(h) RMSD ligand 7K40 run 2



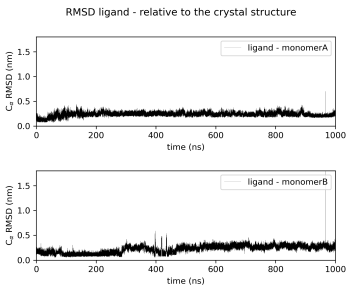
(i) RMSD ligand 7K40 run 3



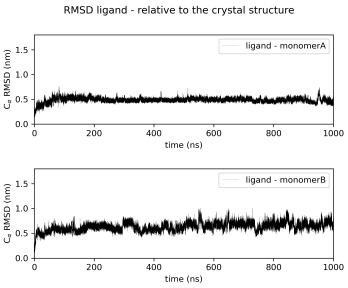
(j) RMSD ligand 6LZE run 1



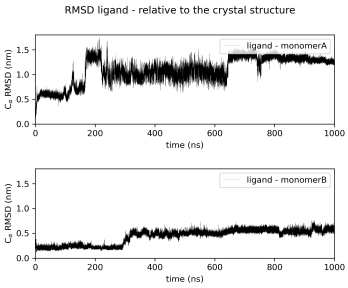
(k) RMSD ligand 6LZE run 2



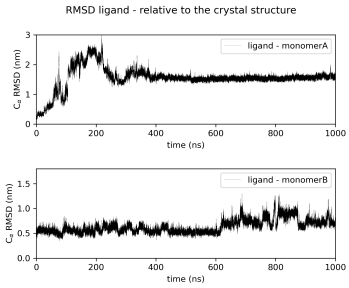
(l) RMSD ligand 6LZE run 3



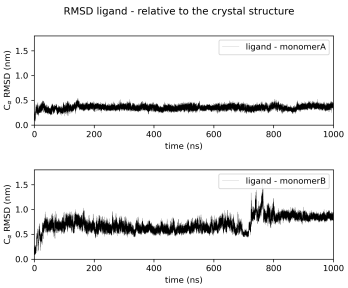
(a) RMSD ligand 7JYC run 1



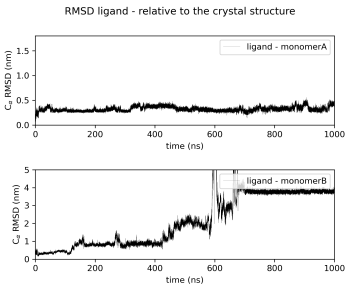
(b) RMSD ligand 7JYC run 2



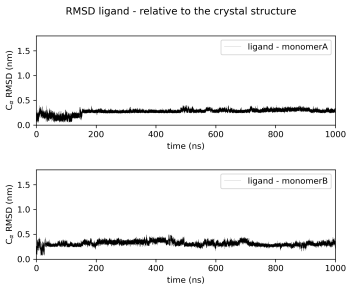
(c) RMSD ligand 7JYC run 3



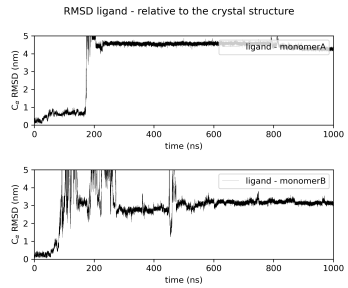
(d) RMSD ligand 6M0K run 1



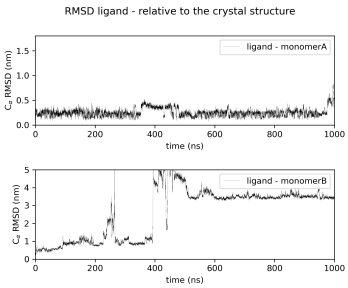
(e) RMSD ligand 6M0K run 2



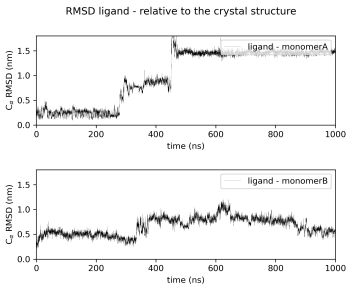
(f) RMSD ligand 6M0K run 3



(g) RMSD ligand 6WTK run 1



(h) RMSD ligand 6WTK run 2



(i) RMSD ligand 6WTK run 3

Figure 15: Ligand RMSD for the 11 systems (vertical) in the three MD simulations (horizontal): movement of the ligand relative to the main protein.

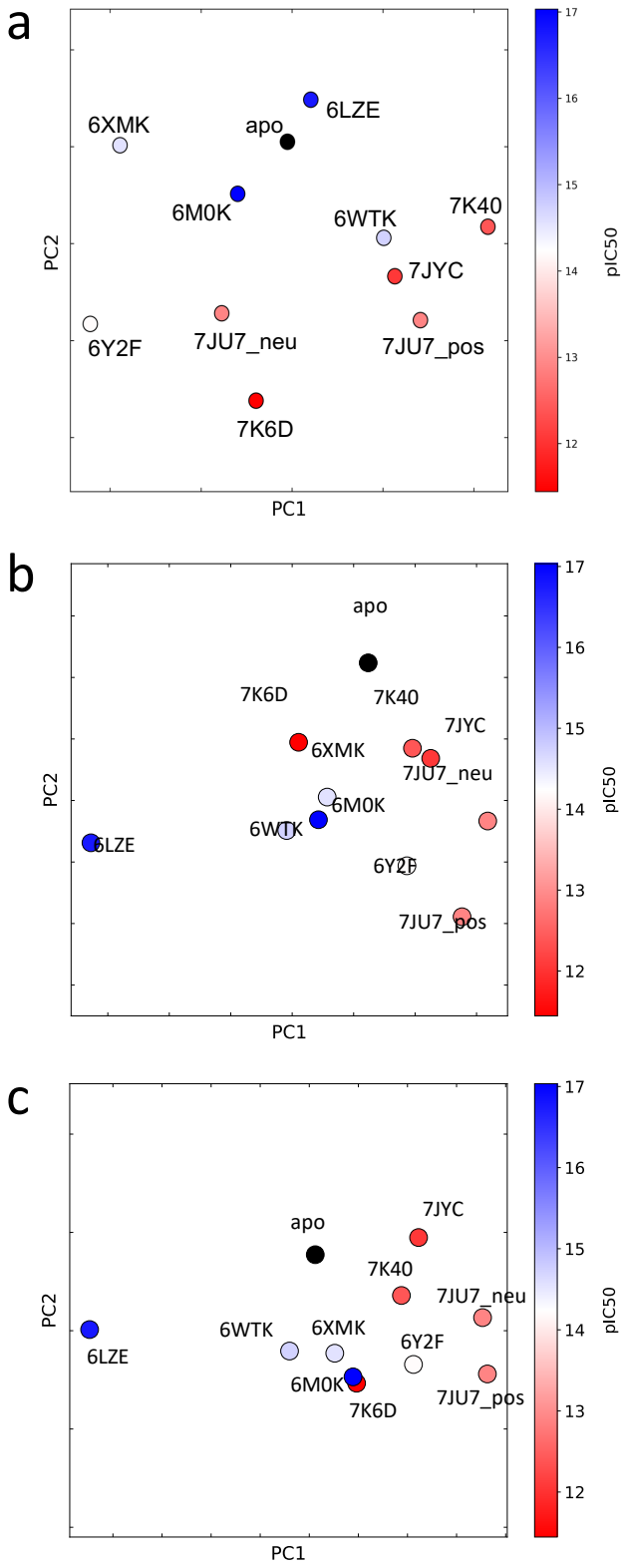


Figure 16: **a** Embedded points of the distance matrix using as input the time-series displacements of residue–pocket center distance **b** Embedded points of the distance matrix using as input the time-series residue–pocket center xyz displacement **c** Embedded points of the distance matrix using as input the time-series displacements of residue–pocket center xyz displacement

Abstract in lingua italiana

Le simulazioni di dinamica molecolare (MD) rivestono un ruolo centrale nella scoperta e nello sviluppo di farmaci, fornendo la possibilità di esplorare a livello atomico le interazioni proteina-ligando. Tuttavia, l'analisi di grandi volumi di dati MD rimane una sfida. Tra gli approcci di apprendimento automatico al problema si trovano principalmente modelli supervisionati, con limiti legati all'etichettatura e alla standardizzazione dei dati. In questo studio è stato adattato e utilizzato un framework di deep-learning (apprendimento profondo) non supervisionato, precedentemente testato su proteine relativamente rigide, per studiare proteine flessibili attraverso un indagine con oggetto la proteasi principale del SARS-CoV-2 (M^{pro}). Abbiamo eseguito simulazioni MD su M^{pro} con diversi ligandi e abbiamo elaborato i risultati concentrandoci sui residui situati nel sito di legame e su time frame caratterizzati dall'adozione di conformazioni stabili della proteina. Dopo aver testato diversi tipi di dati, abbiamo selezionato come descrittore ottimale dei dati MD la distanza tra i residui e il centro del sito di legame. Un insieme di traiettorie rappresentative della dinamica del descrittore (denominate local dynamc ensemble LDE) è stato generato e utilizzato come input della rete neurale per calcolare le distanze di Wasserstein tra le coppie di sistemi, rivelando differenze nella conformazione della proteina target M^{pro} dovuta ai ligandi. Utilizzando tecniche di riduzione della dimensionalità abbiamo prodotto una mappa che fornisce una semplice rappresentazione grafica della distanza relativa tra i sistemi. I risultati ottenuti mettono in relazione la dinamica indotta dai ligandi con le misure sperimentali di efficacia dei ligandi (IC50) con un coefficiente di Pearson di 0.7. Particolarmente evidenti sono stati gli effetti dei composti ad alta affinità sulla conformazione della proteina. Abbiamo anche condotto un'analisi per identificare i residui del sito di legame che hanno contribuito maggiormente alla differenze tra i sistemi, trovando conferma con altri risultati in letteratura. Il nostro metodo mette in evidenza come l'utilizzo di deep learning non supervisionato per l'analisi delle simulazioni MD abbia il potenziale di estrarre informazioni preziose sui meccanismi molecolari tra farmaco e target e accelerare la scoperta di farmaci, ponendo così le basi per un'esplorazione terapeutica rapida e efficace.

Parole chiave: Simulazioni di Dinamica Molecolare (MD), scoperta di farmaci, apprendimento non supervisionato