Water regulates the residence time of Benzamidine in Trypsin

Narjes Ansari
Italian Institute of Technology
Valerio Rizzi
Italian Institute of Technology https://orcid.org/0000-0001-5126-8996
Michele Parrinello (Michele.parrinello@iit.it)
Italian Institute of technology

Article

Keywords:

Posted Date: April 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1546274/v1

Version of Record: A version of this preprint was published at Nature Communications on September 16th, 2022. See the published version at https://doi.org/10.1038/s41467-022-33104-3.

Water regulates the residence time of Benzamidine in Trypsin

Narjes Ansari[1,], Valerio Rizzi[1], and Michele Parrinello[1,*]

[1]Italian Institute of Technology, Via E. Melen 83, 16152, Genova, Italy
[*]michele.parrinello@iit.it

April 11, 2022

Abstract

We simulate with state-of-the-art enhanced sampling techniques the binding of Benzamidine to Trypsin which is a much studied and paradigmatic ligand-protein system. We use machine learning methods and in particular Time-lagged Independent Component Analysis to determine efficient collective coordinates. These coordinates are used to perform On-the-fly Probability Enhanced Sampling simulations, which we adapt to calculate also the ligand residence time. Our results, both static and dynamic, are in good agreement with experiments. We underline the role of water in the unbinding process and find that the presence of a water molecule located at the bottom of the binding pocket allows via a network of hydrogen bonds the ligand to be released into the solution. On a finer scale, even when unbinding is allowed, another water molecule further modulates the exit time.

Introduction

The process of ligand-protein unbinding is crucial in biophysics and its full understanding would enhance not only our knowledge but also benefit the design of new drugs. In particular, two quantities are of great relevance, the ligand binding free energy and the inverse of the residence time k_{off} [1]. Being able to compute reliably these two quantities would be of great help. Here we describe a strategy that makes it possible to calculate accurately both quantities. We demonstrate this assertion with a state-of-the-art simulation of the Trypsin-Benzamidine system [2, 3, 4, 5, 6] and we find that the unbinding process can be more complex than what one could have anticipated. Although Trypsin-Benzamidine is one of the simplest cases of protein-ligand systems, high-resolution crystallographic experiments have recently demonstrated the presence of an extensive water structure in the binding cavity [7]. Our study reveals that water has not only a structural role but a dynamical one, since it regulates the unbinding process via a complex rearrangement of hydrogen bonds (HBs). In particular, the presence of a water molecule at a specific position in the binding cavity allows the unbinding process to take place. As the ligand begins to leave its binding pose, the number of water molecules in the binding cavity increases, leading to a finer regulation in which water determines two possible escape pathways with a k_{off} differing by one order of magnitude.

Several technical advances have made this study possible. Ligand unbinding is a rare event that takes place on a timescale of milliseconds and thus its study requires the use of an enhanced sampling method. Here, we profit from the flexibility and efficiency of the On-the-fly Probability Enhanced Sampling (OPES) method [8], that is the latest evolution of Metadynamics [9, 10]. Like umbrella sampling [11] and other enhanced sampling methods [12], OPES is based on the use of a set of collective variables (CVs) that are functions $s(R)$ of the atomic coordinates R and describe the slow modes of the system. A good choice of $s(R)$ is essential to obtain converged results in an affordable time. To determine good CVs we use

two machine-learning-based tools that we recently developed, Deep Local Discriminant Analysis (Deep-LDA) [13] and Deep Time-lagged Independent Component Analysis (Deep-TICA) [14]. In building these CVs, we shall pay great attention to the role of water and make use of the experience gained in Ref. 15.

Standard OPES is very efficient in calculating static properties such as binding free energies, but, in doing so, it alters the natural dynamics of the system so that a sensitive quantity like k_{off} cannot be easily extracted. Nevertheless, it has been pointed out that if an enhanced sampling method can be engineered such that no bias is added to the transition region, the value of k_{off} can still be computed [16, 17, 18, 19, 20]. Here we show that by an appropriate setting of the input parameters, OPES can be made to satisfy this condition. We call this approach OPES flooding ($OPES_f$) since it is inspired by the flooding approach [19] and use it to calculate the ligand residence time.

Results

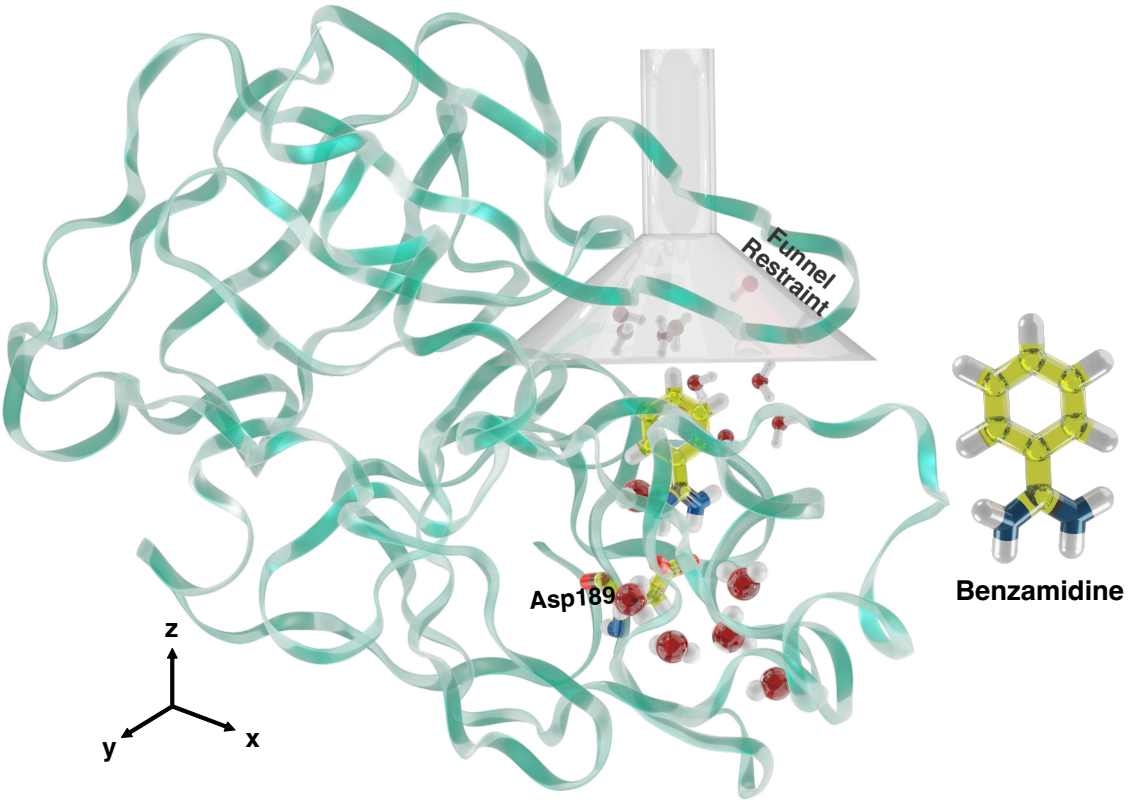

Figure 1: **The Trypsin-Benzamidine system.** A cartoon representation of Trypsin structure with the ligand Benzamidine and the Funnel restraint geometry.

Enhanced sampling technique for estimating free energies: OPES

To accelerate the occurrence of binding and unbinding events, we use OPES [8]. This method allows the system to overcome kinetic barriers by transforming the original CV probability distribution $P(s)$ into a smoother and thus simpler to sample one $P^{tg}(s)$. In the variant of OPES that we use here, we transform the s probability distribution $P(s) = \frac{\int \delta(s-s(R))e^{-\beta V(R)}dR}{Z}$, where $\beta = 1/k_B T$ is the inverse temperature, $V(R)$

the interaction potential and $Z = \int e^{-\beta V(R)} \mathrm{d}R$ the partition function, into the smoother well-tempered distribution $P^{\mathrm{tg}}(s) \propto [P(s)]^{1/\gamma}$ [10], where the parameter $\gamma > 1$ regulates the broadening of the target distribution. In standard metadynamics this transformation is achieved by iteratively building a bias potential $V(s)$, while in OPES one instead reconstructs the probability distribution $P(s)$ on-the-fly. At iteration step n, the probability distribution $P_n(s)$ is estimated as

$$P_n(s) = \frac{\sum_k^n w_k G(s, s_k)}{\sum_k^n w_k} \tag{1}$$

where $G(s, s_k)$ are multivariate Gaussian kernels evaluated at every step k, the weights $w_k = \mathrm{e}^{\beta V_{k-1}(s_k)}$ are computed from the bias at step $k-1$. In turn, the bias is

$$V_n(s) = \left(1 - \frac{1}{\gamma}\right) \frac{1}{\beta} \log\left(\frac{P_n(s)}{Z_n} + \epsilon\right) \tag{2}$$

where Z_n is a factor that measures the configuration space thus far explored and ϵ is a very important parameter that controls the maximum bias that can be deposited in the system.

Machine Learning-based CVs: Deep-LDA and Deep-TICA

As in our previous work [15], we use the machine learning-based Deep-LDA method [13] to design effective CVs to be used in conjunction with OPES. The method is based on the time-honored Linear Discriminant Analysis technique [21] employed in classification applications. Our aim is to build a CV that is able to distinguish between two sets of data. In our case, data come from unbiased simulations of the system in the bound (B) and unbound (U) states.

In standard LDA, one optimizes Fisher's ratio $\frac{\boldsymbol{w}^T \boldsymbol{S}_\mathrm{b} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_\mathrm{w} \boldsymbol{w}}$ to obtain the linear combination $s(R) = \boldsymbol{w}^T \mathbf{d}(\mathbf{R})$ of descriptors $\mathbf{d}(\mathbf{R})$ that best separates the two states. Fisher's ratio is written in terms of the scatter matrix $\boldsymbol{S}_\mathrm{b} = (\boldsymbol{\mu}_\mathrm{B} - \boldsymbol{\mu}_\mathrm{U})(\boldsymbol{\mu}_\mathrm{B} - \boldsymbol{\mu}_\mathrm{U})^T$ and the within matrix $\boldsymbol{S}_\mathrm{w} = \boldsymbol{S}_\mathrm{B} + \boldsymbol{S}_\mathrm{U}$, where $\boldsymbol{\mu}_\mathrm{B}, \boldsymbol{\mu}_\mathrm{U}$ indicate the average descriptors values and $\boldsymbol{S}_\mathrm{B}$, $\boldsymbol{S}_\mathrm{U}$ the descriptors variance matrices in the two states. The vector $\boldsymbol{w}$ that maximizes this ratio is the direction that optimally discriminates the states and provides the best-separated projection of the data in the one-dimensional s space [22].

In Deep-LDA, the set of N_d descriptors $\mathbf{d}$ is fed into a neural network (NN) that is trained by applying the LDA criterion to the last hidden layer $\mathbf{h}$ of the network. In analogy with LDA, a projection of the N_h components of the last hidden layer produces the Deep-LDA CV $s = \boldsymbol{w}^T \boldsymbol{h}$. This CV, being by construction a non-linear combination of the original input descriptors $\mathbf{d}$, is more expressive than a simple linear combination and has been successfully applied to solve a number of problems [15, 23, 24]. As s tends to produce sharp distributions, we apply the cubic transformation $s_w = s + s^3$ [25, 15] to make the CV more easily applicable to enhanced simulations.

While often successful in driving a system forth and back between different states, a Deep-LDA CV does not encode any information on the transition state. This information could have been obtained if one had access to a long dynamical trajectory in which many state-to-state transitions did occur. In this case, a way of building a good CV, would have been to use the Time-lagged Independent Component Analysis [26, 27], in which one looks for the most slowly decorrelating modes. These slow modes can be found using the variational principle [28]. If the modes are expressed as a linear combination of descriptors, the variational principle leads to a generalized eigenvalue equation. As in Deep-LDA, one can apply TICA not only to a linear combination of descriptors but also to the last hidden layer of a NN. This greatly improves the variational flexibility of the solution and thus its quality. In its original formulation, this approach was meant to be applied to unbiased trajectories. However, McCarty and Parrinello [29] using a linear approach and later Bonati et al. [14] using a non-linear one (Deep-TICA) have shown how to extract useful CVs from biased simulations. Here, we are going to apply Deep-TICA to the set of N_d descriptors on a converged OPES trajectory where the Deep-LDA CV was biased.

Enhanced sampling technique for estimating residence times: OPES flooding

As discussed in the introduction, to calculate rates from a biased simulations no bias must be deposited in the transition region. In such a case, the physical residence time t is related to the physical simulation time t_{MD} by [16, 17]

$$t = \left\langle e^{\beta V(s)} \right\rangle_V t_{\mathrm{MD}} \tag{3}$$

where the acceleration factor $\langle e^{\beta V(s)} \rangle_V$ is computed as an average along the simulation. To ensure that the condition under which Eq. 3 is satisfied, we introduce a variation of OPES that we call OPES$_{\mathrm{f}}$ that is inspired by the variational flooding method [19]. The condition for OPES$_{\mathrm{f}}$ to allow calculating physical rates are easily satisfied. The maximum amount of bias deposited can be controlled by the choice of the ϵ parameter in Eq. 2, making sure that it is lower than the free energy barrier. Furthermore via the parameter `EXCLUDED_REGION`, one can prevent OPES$_{\mathrm{f}}$ from depositing bias in a preassigned region of configuration space. Since, from an initial free energy surface (FES) estimate one can roughly estimate both the height and the location of the transition, the setup of a subsequent OPES$_{\mathrm{f}}$ rate calculation is relatively simple and straightforward.

Designing water CVs

In our OPES simulations we shall use as CVs the distance z from the binding site and a CV that encodes water behavior. Water is known to play a non-trivial role in ligand binding [30, 31, 32, 33, 34, 35, 36, 37, 38], therefore we want to use the power of machine learning techniques to capture and encode its behavior in a CV, generalizing the strategy that we have proposed for a smaller host-guest system in Ref. 15. The choice of an effective set of descriptors $\mathbf{d}$ is critical as it must capture the solvation in different situations that are relevant to the binding-unbinding process, i.e. the ligand itself, the binding position, the binding pocket and the binding path.

Some of these descriptors can be identified following physical intuition. For instance, to describe ligand's solvation we focus on its charged tail and use the coordination number of water around the Carbon atom of the amidine group ($\{\mathrm{G}\}$) (see Fig. 2c). Regarding the binding position, we analogously evaluate the coordination number between water and the Carbon atom of the carboxylate group in residue Asp189 ($\{\mathrm{H}\}$). These choices are similar in spirit to other solvation variables that have been devised in the past [39, 40, 41, 42, 15].

However, we follow a different approach to describe water behavior around the binding pocket and along the binding path. The S1 binding pocket of apo Trypsin form (see Fig. 1) is known from high resolution experiments [7, 43] to be characterized by the presence of a set of deeply buried water molecules. These water molecules have a long residence time and form what has been called in Ref. 7 the reservoir. To encode their role in the descriptor set, we propose a general method that can be applied to describe trapped water molecules in biological systems. This method consists of four steps: 1) selecting the α-Carbon atoms of the relevant part of the protein that encloses water (label 1 of Fig. 2a), 2) building the convex hull surface with the coordinates from step 1 (label 2 of Fig. 2a), 3) collecting the position of the water molecules that lie within the convex hull for longer than a pre-defined lifetime, and 4) clustering the collected data to determine areas of high water density and their centers (label 4 of Fig. 2a). We call those points hydration spots ($\{\mathrm{V}_i\}$) and use them as centers around which we calculate water coordination. The method is implemented in a Python script [44].

In Trypsin, we restrict the analysis to the region around the S1 binding pose where we build the convex hull. That area encloses both the deeply buried water molecules and the binding path of the ligand (see Fig. 2b). From a number of unbiased trajectories (both in states B and U) we collect the positions of the water molecules that have a residence time inside the convex hull longer than 100 ps. Then, using the clustering method introduced in step 4, we identify 16 Vs (see Fig. 2e) position. The number of clustering centers is chosen so that the relative distance between the Vs lies in a range of 2-3 Å. We find that there is a correspondence between the $\mathrm{V}_5 - \mathrm{V}_{12}$ centers and the position of the reservoir water molecules reported in Ref. 7. Furthermore, V_3 lies at the position of the long-lived water molecule that stabilizes binding

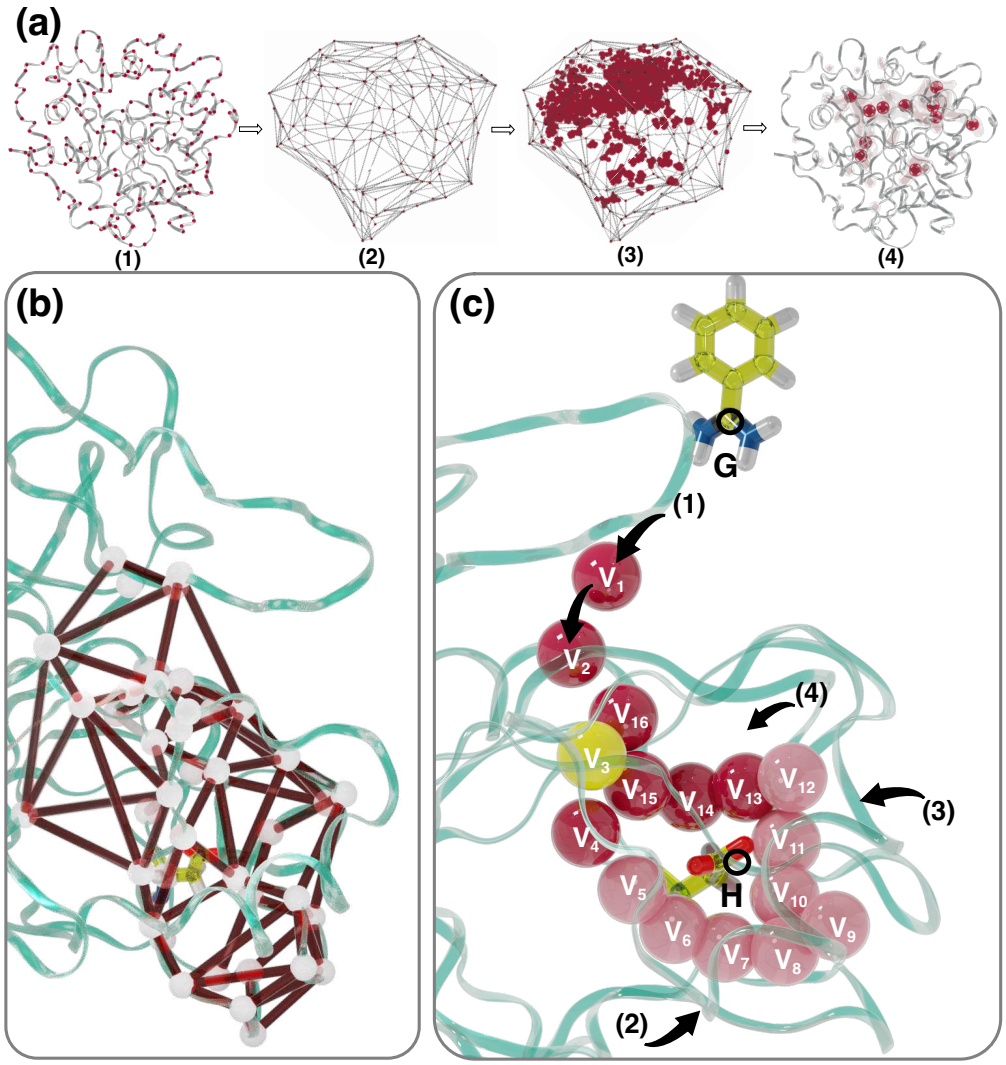

Figure 2: **Identification of long-lived water molecules.** (a) Graphical representation of the four steps strategy used to identify the long-lived hydration spots. (1) α-Carbon atoms of the residues located on a protein's outer surface, represented by red dots. (2) Convex hull surface built from the α-Carbon atoms shown in (1). (3) Distribution of the long-lived water molecules inside the surface. (4) Centers of the water distribution from step (3) obtained with a clustering algorithm. (b) Convex hull surface built around the binding pose of Trypsin. (c) Position of the 16 hydration spots V_i, shown by spheres. The hydration spots $V_5 - V_{12}$ coincide with the position of the reservoir water molecules of Ref. 7 and the dark red spheres lie on the binding path. The yellow sphere on V_3 shows the position of the key W1 water molecule [7] that forms a hydrogen bond with the ligand. Black arrows indicate the four possible paths of entrance/exit for water molecules in the binding pocket.

and that in Ref. 7 is called W1. We use as descriptors all the water coordination on ({G}, {H}, {V_i}) to generate Deep-LDA s_w and Deep-TICA s_t water CVs.

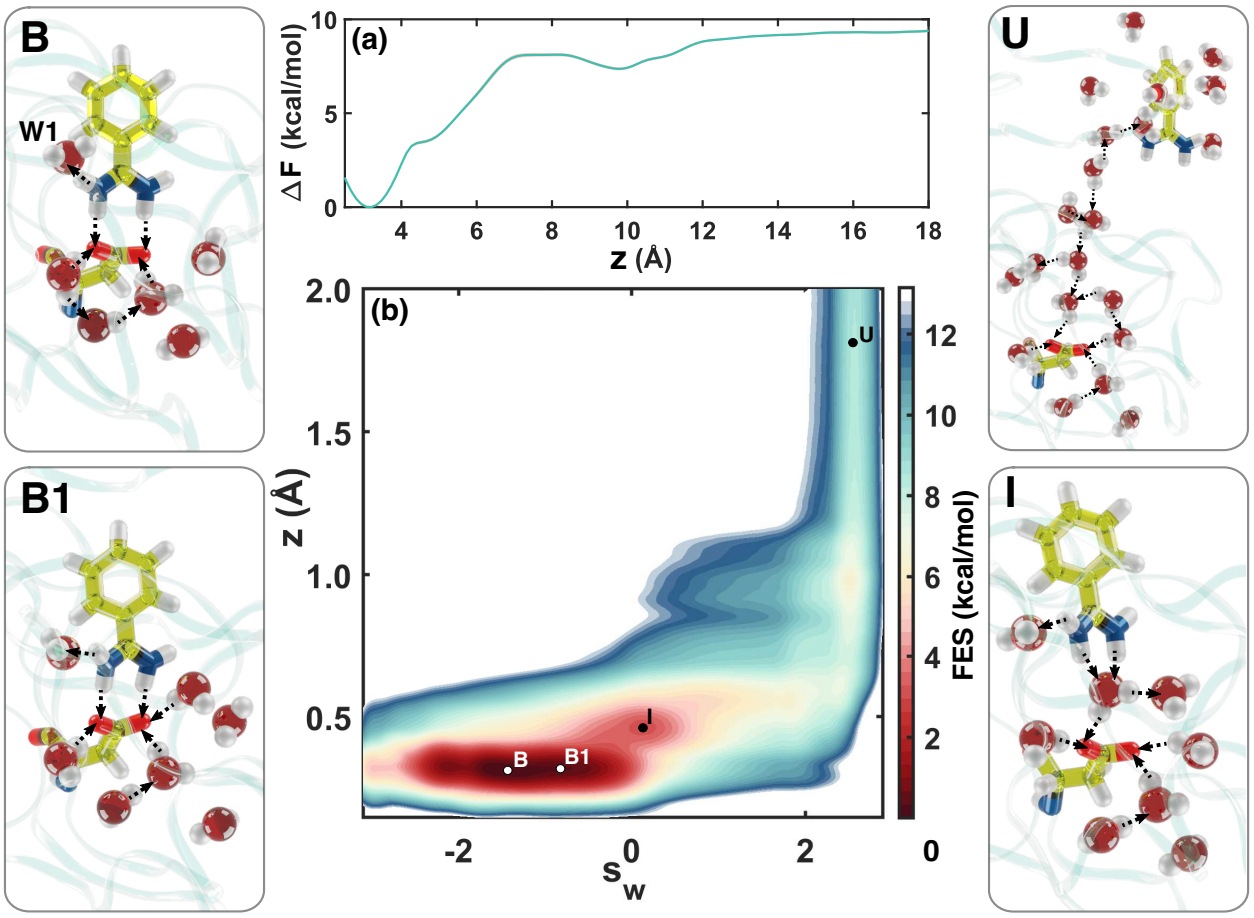

Figure 3: **Mechanistic interpretation of the binding process**. (a) 1D free energy surface of Trypsin-Benzamidine reconstructed using reweighting along z with the statistical uncertainty within the line width. (b) 2D free energy landscape along z and s_w. In the panels, representative configurations of the relevant states are shown with a focus on the solvation pattern around the ligand and the binding pose.

Table 1: Our simulation estimates of the free energy, enthalpy and entropy of binding at T=300 K and corresponding experimental data [45, 46].

	ΔF	ΔU	-TΔS
This work	6.36 ± 0.07	4.12 ± 1.56	2.23 ± 1.56
EXP.	6.36	4.52	1.84

Static properties: binding free energy, enthalpy and entropy

To calculate the binding free energy, we perform an OPES simulation using 32 walkers, biasing z and the Deep-LDA water CV s_w, for a total simulation time of 3.2 μs. More details about the simulation are provided in the Supporting Information (SI). Convergence is achieved as several binding and unbinding events occur in a quasi-static regime of the bias (see Fig. S-1). In Fig. 3a, we show the FES projection on z, calculated as a block average with an error bar that is within the line width of the plot. We calculate the free energy difference between the B and U states using the funnel correction in Eq. 4 obtaining a value of 6.36 ± 0.07 kcal/mol, in an embarrassing and certainly fortuitous agreement with the experimental value

of 6.36 kcal/mol [45, 46].

The high level of accuracy that the combination of OPES and good quality CVs, makes it possible to estimate separately enthalpy ΔU and entropy ΔS of binding. This is achieved by converging binding free energy calculations in a range of temperatures and using the relationship $\Delta F = \Delta U - T\Delta S$. In Tab. 1 we report our estimate for these thermodynamic quantities and observe that they are also in good agreement with experiments. Further details about these simulations are provided in the SI.

In Fig. 3b, we show the two-dimensional FES of binding projected on z and s_w, along with a number of representative snapshots of the different states and their typical water arrangement. State B is the global minimum of the FES and corresponds to the protein-ligand crystallographic structure (e.g. PDB 3atl [47]). In line with experiments [7], the W1 water molecule connects the ligand to Trypsin residues Tyr228, and Ser190, forming a total of 3 HBs. Furthermore, in the reservoir region below residue Asp189 we observe on average the presence of 5 water molecules, in agreement with the X-ray structure [7]. State B1 is a less stable binding pose where the ligand is in the same configuration as B, but the reservoir contains on average one more water molecule. In state I, the reservoir region contains another extra water molecule that tends to bridge the amidine group of the ligand and the binding site of Asp189, thus weakening binding. In the fully dissociated state U, the binding cavity returns to its apo form in which the 5 water of the reservoir are in the experimental structure of the holo form. The space previously occupied by the ligand is replaced by about 4-5 water molecules. The fact that in the apo form the reservoir structure of the water is preserved underlines once more the relevance of this structure (see Fig. 3).

Note that, as the ligand progresses towards the U state, the number of water molecules in the binding site increases, underlining the role of water throughout the whole unbinding process. In spite of the limit of being trained on data coming exclusively from B and U, the Deep-LDA CV in combination with OPES is able to capture this non-trivial water behavior and converge the FES by virtue of the ability of OPES to deal with non-optimal CVs [48]. Nevertheless, s_w is not able to resolve the diverse water arrangements in states such as B and B1, as we shall see below. In order to capture the finer details of the role of water in the binding pose, the intermediate states and the path to unbinding, we use Deep-TICA to determine a new water CV that we call s_t. Since Deep-TICA is trained on trajectories coming from biased simulations, the resulting CV is informed on the water modes that Deep-LDA is not able to resolve.

Using the data generated in the Deep-LDA-driven simulation, we project in Fig. 4 (a,b) the FES along different pairs of CVs: z, s_t and s_w, s_t, respectively. We find that s_t resolves the original B state into states B and B′, with state B being about 12 kJ/mol more stable than B′. From these two states, two different unbinding pathways depart. We denote the states belonging to the less stable branch with a prime. Fig. 4 (d,e) shows typical configurations of the ligand and the surrounding water molecules in state B and B′. The number of water molecules in the reservoir is the same, but the water arrangement is different. In B, water molecules are part of an extended HB network that includes residue Asp189, while in state B′ this network is less structured.

One striking difference between B and B′ is that in state B there is one water molecule trapped deeply in the binding cavity, in an intermediate position between residues Tyr182, Lys219 and Gln219 which corresponds to descriptor V9 (see the semi transparent sphere in Fig. 4d). The FES projection along V9 and s_t in Fig. 4c, confirms that V9 discriminates well between B and B′. Further analysis reveals that state B presents a slightly larger volume of the reservoir region than B′ (see SI), which, in turn, possibly facilitates the formation of an extended HB network. A study of the relative importance of the water descriptors on the NN CVs can be also found in the SI.

The presence of a water molecule in V9 discriminates well between B and B′, as can be seen if we project the FES along V9 and s_t. Thus this molecule stabilizes the cavity HB network that is such an important feature of this protein.

Dynamic properties: ligand residence time

We compute the ligand-unbinding rate by employing the OPES flooding technique on simulations that start from state B. In this approach a choice of CVs that captures well the complexity of the path from state B to state U is essential. The Deep-LDA coordinate built only on the knowledge of the B and U states

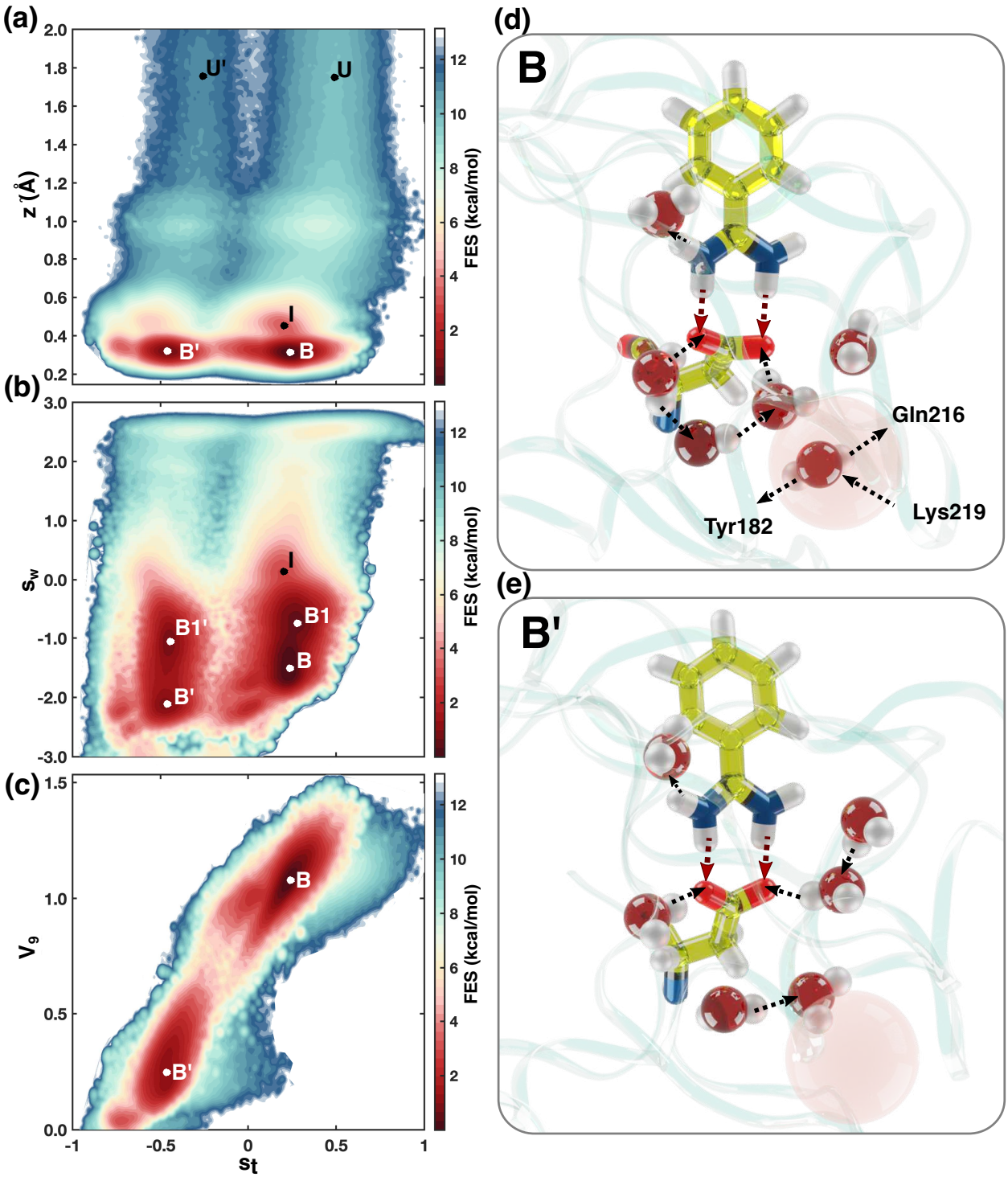

Figure 4: 2D FES along s_t and (a) z, (b) s_w, (c) V9. In (d) and (e), representative configurations of the ligand and its solvation pattern in states B and B′, respectively. The location of V9 is highlighted with a semi-transparent pink sphere.

is not adequate to the purpose. In fact, being able to fill all the different metastable bound states is crucial for promoting transitions to the U state without remaining stuck in intermediate states. For this reason besides z we use Deep-TICA CV s_t. By construction, s_t is able to extract the slow modes of the system and as a consequence to drive it out of deep basins towards the transition state. The use of OPES$_f$ for the calculation of rates requires that we define an excluded region to avoid depositing bias in the transition state region. An analysis of the FES in Fig. 4a suggests to prevent depositing bias in the region $z > 6$ Å. We run a total of 55 ligand unbinding simulations and, by using Eq. 3, we determine for each simulation a physical ligand residence time t.

The distribution of transition times of a rare event dominated by a single barrier is expected to be Poissonian $\frac{1}{\tau}e^{-t/\tau}$ where τ is the characteristic time of the associated homogeneous process [49]. We fit all our data to such a model and find that the quality of the fit is poor (see Fig. S-7), as indicated by its low p-value. It is known that complex biological processes present multiple timescales and are not expected to necessarily follow such a simple model [50, 1, 51]. This led us to further analyze the unbinding trajectories and to identify the presence of two possible unbinding mechanisms: a faster and a slower one. The key difference between the two is related in the water network arrangement in the binding pocket (see Fig. 5).

All unbinding events start with the water reservoir acquiring an extra molecule and the system reaching state B1 (see Fig. 5b). Then two possibilities occur. In the faster case, the presence of a water molecule in the vicinity of W1 weakens the bond between the ligand and W1 (see Fig. 5c), which leads to another water molecule to bind to residue Asp189 (see Fig. 5d) and finally brings about the ligand-unbinding event. In the slower mechanism, W1 is stable and, as the reservoir increases its water content (see Fig. 5e), eventually one water molecule bridges the ligand and Asp189 (see Fig. 5f), driving the system towards unbinding. As reported in Tab. 2, both mechanisms fit well the homogeneous model and present a discrepancy in the resulting τ of about one order of magnitude. In particular, the slower mechanism τ that will dominate the residence time has a $k_{off} = 1/\tau = 687\ s^{-1}$ that is in close agreement with the experimental value of $k_{off} = 600 \pm 300\ s^{-1}$ [52].

Table 2: **Ligand residence time.** τ is the characteristic time from an exponential fit [49], $k_{off} = 1/\tau$ and p-value measures the quality of the fit. μ and σ are the average and the standard deviation of the data, respectively. We show results from all our data and from data split into the two observed mechanisms. The experimental results are taken from Ref. 52.

	τ (10^{-3} s)	k_{off} (10^{2} s^{-1})	p-value	$\mu \pm \sigma$ (10^{-3} s)
All data	0.64	15.6	0.06	1.10 ± 1.49
Faster path	0.18	55.4	0.62	0.23 ± 0.29
Slower path	1.45	6.87	0.63	1.58 ± 1.59
EXP.	-	6.00 ± 3.00	-	-

Discussion

Our work shows clearly that water plays a major role in the activation of Trypsin and modulates the release process of Benzamidine. Through the use of water-focused machine-learned CVs, we estimate static and dynamic properties in excellent agreement with experiments and we analyze the results with a level of resolution that lets us identify the importance of each individual water molecules. We observe how the presence of water in key positions affects the binding stability and determines unbinding mechanisms with significantly different ligand residence times. In conclusion, one might say that the set of reservoir water molecules, present both in the apo and the holo form, are an essential part of the enzyme, its structure and its activity. We believe that this is not just specific to the Trypsin-Benzamidine system and we argue that water would play such a non-trivial role at least in all the cases where the ligand has to negotiate its way out of a water-rich enzymatic cavity.

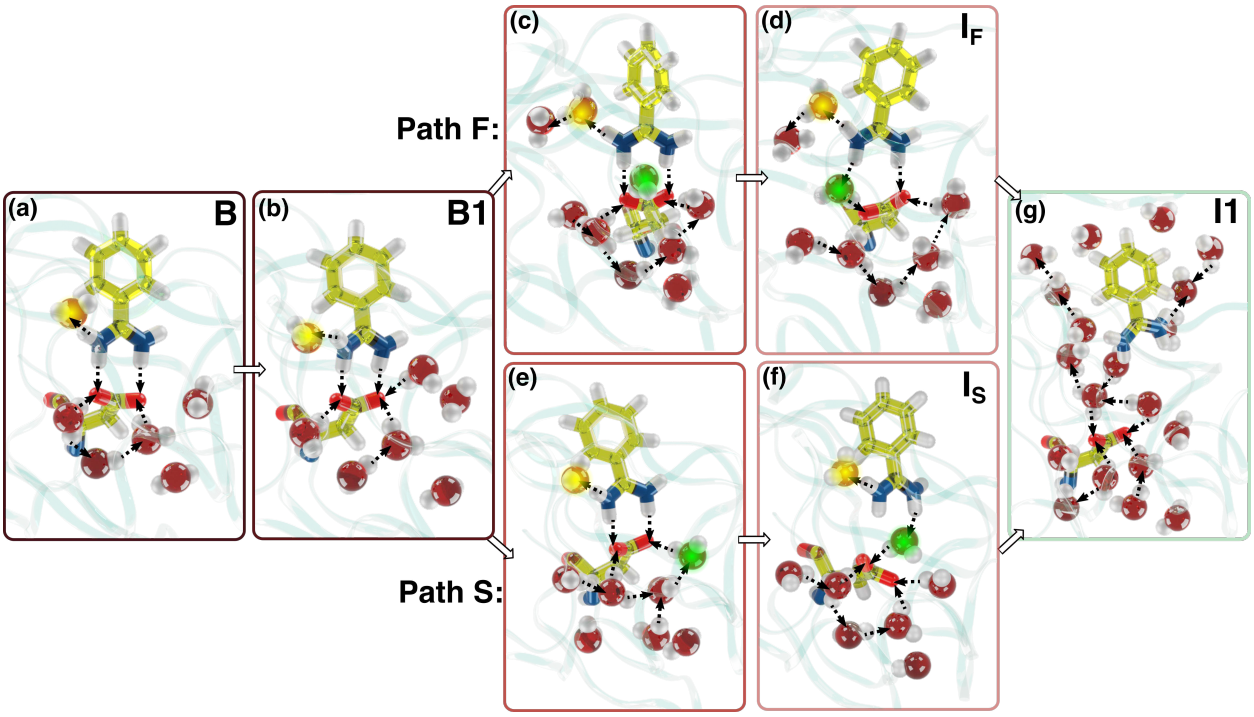

Figure 5: **Two ligand unbinding mechanisms.** Representative configurations of states (a) B and (b) B1. (c) and (e) the pre-intermediate step of fast and slow mechanisms, respectively. (d) and (f) the intermediate step of the faster (I_F) and the slower (I_S) unbinding mechanisms. (g) Intermediate I_1 that presents a number of water molecules between ligand and the binding pose. The yellow sphere highlights the location of the W1 water molecule. The green sphere indicates the position of the water molecule that bridges the ligand and the Asp189 residue. The protein is shown as a semi-transparent ribbon.

Methods

We perform all simulations with the molecular dynamics code GROMACS 2020.4 [53] patched with PLUMED 2.8 [54] and the Pytorch library 1.4 [55, 13]. We use TIP3P water, while the protein is described by the Amber-14SB force field and the ligand interactions are taken from the Amber GAFF library [56]. We employ Ewald summation [57] for long-range electrostatic interactions with a cutoff of 10 Å for both the Coulomb and the van der Waals interactions. All simulations are run in the NPT ensemble with a timestep of 2 fs. We use the Parrinello-Rahman barostat [58] and the stochastic velocity rescaling thermostat [59], both with a coupling constant of 1 ps.

We train a Deep-LDA CV by running unbiased simulations on states B and U for 60 ns and evaluating the water coordination on a descriptor set $\mathbf{d}$ made by the 18 components ($\{G\}$, $\{H\}$, $\{V_i\}$). For training, we take as state B the initial configuration used in Ref. 60. We use a NN made of 4 layers with 2.5×10^{-5} learning rate. To train the Deep-TICA CV, we take the converged OPES trajectory used for calculating the binding free energy and feed the descriptors set $\mathbf{d}$ to a NN made of 4 layers with 1.0×10^{-4} learning rate and 0.07 lag time. Further details can be found in the SI.

One of the difficulties in ligand-unbinding problem arises from the fact that once the ligand leaves the protein, it has to explore a large conformational space. To tackle this issue, we use the Funnel restraint proposed in Ref. 61. In this method, the space available to the ligand in the unbound state is limited by confining it to a cylindrical volume above the binding site (see Fig. 1) which introduces an entropic restraint in the U state. To calculate the absolute free energy of binding, one needs to apply a correction

to the apparent free energy $W(z)$ that comes from the simulation:

$$\Delta F = -\frac{1}{\beta} \log \left(C^0 \pi R_{\text{cyl}}^2 \int_{\text{B}} \text{d}z \, \exp\left(-\beta(W(z) - W_{\text{U}})\right) \right) \tag{4}$$

where $C^0 = 1/1660$ Å^{-3} is the standard concentration, z is the distance between the center of mass of the ligand and the binding site along the funnel's axis, $W(z)$ is the free energy along the funnel axis and W_{U} is its reference value in state U.

Data Availability

All the inputs and instructions to reproduce the results presented in this manuscript are soon to be uploaded to the PLUMED-NEST repository. A tutorial on Deep-LDA training can be found at this link while a tutorial on Deep-TICA is at this link.

References

[1] Pan, A. C., Borhani, D. W., Dror, R. O. & Shaw, D. E. Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today* **18**, 667–673 (2013). URL http://dx.doi.org/10.1016/j.drudis.2013.02.007https://linkinghub.elsevier.com/retrieve/pii/S1359644613000627.

[2] Gervasio, F. L., Laio, A. & Parrinello, M. Flexible Docking in Solution Using Metadynamics. *Journal of the American Chemical Society* **127**, 2600–2607 (2005). URL https://pubs.acs.org/doi/10.1021/ja0445950.

[3] Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences* **112** (2015).

[4] Wolf, S., Lickert, B., Bray, S. & Stock, G. Multisecond ligand dissociation dynamics from atomistic simulations. *Nature Communications* **11**, 2918 (2020).

[5] Brotzakis, Z. F., Limongelli, V. & Parrinello, M. Accelerating the calculation of protein–ligand binding free energy and residence times using dynamically optimized collective variables. *Journal of Chemical Theory and Computation* **15**, 743–750 (2019).

[6] Ray, D., Stone, S. E. & Andricioaei, I. Markovian weighted ensemble milestoning (m-wem): Long-time kinetics from short trajectories. *Journal of Chemical Theory and Computation* **18**, 79–95 (2022).

[7] Schiebel, J. *et al.* Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nature Communications* **9**, 3559 (2018). URL http://dx.doi.org/10.1038/s41467-018-05769-2http://www.nature.com/articles/s41467-018-05769-2.

[8] Invernizzi, M. & Parrinello, M. Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *The Journal of Physical Chemistry Letters* **11**, 2731–2736 (2020). URL https://link.aps.org/doi/10.1103/PhysRevLett.115.070601.

[9] Laio, A. & Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99**, 12562–12566 (2002). URL http://www.pnas.org/cgi/doi/10.1073/pnas.202427399.

[10] Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **100**, 020603 (2008). URL https://link.aps.org/doi/10.1103/PhysRevLett.100.020603.

[11] Torrie, G. & Valleau, J. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23**, 187–199 (1977).

[12] Valsson, O., Tiwary, P. & Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry* **67**, 159–184 (2016). URL http://www.annualreviews.org/doi/10.1146/annurev-physchem-040215-112229.

[13] Bonati, L., Rizzi, V. & Parrinello, M. Data-Driven Collective Variables for Enhanced Sampling. *The Journal of Physical Chemistry Letters* **11**, 2998–3004 (2020). URL http://arxiv.org/abs/2002.06562https://pubs.acs.org/doi/10.1021/acs.jpclett.0c00535.

[14] Bonati, L., Piccini, G. & Parrinello, M. Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences* **118** (2021). URL http://arxiv.org/abs/2107.03943http://www.pnas.org/lookup/doi/10.1073/pnas.2113533118https://pnas.org/doi/full/10.1073/pnas.2113533118.

[15] Rizzi, V., Bonati, L., Ansari, N. & Parrinello, M. The role of water in host-guest interaction. *Nature Communications* **12**, 93 (2021). URL http://dx.doi.org/10.1038/s41467-020-20310-0http://www.nature.com/articles/s41467-020-20310-0.

[16] Grubmüller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E* **52**, 2893–2906 (1995). URL https://link.aps.org/doi/10.1103/PhysRevE.52.2893.

[17] Voter, A. F. A method for accelerating the molecular dynamics simulation of infrequent events. *The Journal of Chemical Physics* **106**, 4665–4677 (1997). URL http://aip.scitation.org/doi/10.1063/1.473503.

[18] Tiwary, P. & Parrinello, M. From Metadynamics to Dynamics. *Physical Review Letters* **111**, 230602 (2013). URL https://link.aps.org/doi/10.1103/PhysRevLett.111.230602.

[19] McCarty, J., Valsson, O., Tiwary, P. & Parrinello, M. Variationally Optimized Free-Energy Flooding for Rate Calculation. *Physical Review Letters* **115**, 070601 (2015). URL https://link.aps.org/doi/10.1103/PhysRevLett.115.070601.

[20] Debnath, J. & Parrinello, M. Gaussian Mixture-Based Enhanced Sampling for Statics and Dynamics. *The Journal of Physical Chemistry Letters* **11**, 5076–5080 (2020). URL https://pubs.acs.org/doi/10.1021/acs.jpclett.0c01125.

[21] Welling, M. Fisher Linear Discriminant Analysis. Tech. Rep., Dep. Comput. Sci. Univ. Toronto (2005).

[22] Mendels, D., Piccini, G. & Parrinello, M. Collective Variables from Local Fluctuations. *The Journal of Physical Chemistry Letters* **9**, 2776–2781 (2018). URL http://pubs.acs.org/doi/10.1021/acs.jpclett.8b00733.

[23] Karmakar, T., Invernizzi, M., Rizzi, V. & Parrinello, M. Collective variables for the study of crystallisation. *Molecular Physics* **119**, e1893848 (2021). URL http://arxiv.org/abs/2101.03150https://www.tandfonline.com/doi/full/10.1080/00268976.2021.1893848.

[24] Ansari, N., Rizzi, V., Carloni, P. & Parrinello, M. Water-Triggered, Irreversible Conformational Change of SARS-CoV-2 Main Protease on Passing from the Solid State to Aqueous Solution. *Journal of the American Chemical Society* **143**, 12930–12934 (2021). URL http://biorxiv.org/content/early/2021/05/21/2021.05.21.445090.abstracthttps://pubs.acs.org/doi/10.1021/jacs.1c05301.

[25] Bjelobrk, Z. *et al.* Naphthalene crystal shape prediction from molecular dynamics simulations. *CrystEngComm* **21**, 3280–3288 (2019).

[26] Naritomi, Y. & Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *The Journal of Chemical Physics* **134**, 065101 (2011). URL http://aip.scitation.org/doi/10.1063/1.3554380.

[27] Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics* **139**, 015102 (2013). URL http://aip.scitation.org/doi/10.1063/1.4811489.

[28] Wu, H. *et al.* Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *The Journal of Chemical Physics* **146**, 154104 (2017). URL http://aip.scitation.org/doi/10.1063/1.4979344.

[29] McCarty, J. & Parrinello, M. A variational conformational dynamics approach to the selection of collective variables in metadynamics. *The Journal of Chemical Physics* **147**, 204109 (2017). URL http://arxiv.org/abs/1703.08777http://aip.scitation.org/doi/10.1063/1.4998598.

[30] Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology* **3**, 973–980 (1996). URL https://linkinghub.elsevier.com/retrieve/pii/S1074552196901647.

[31] Homans, S. Water, water everywhere — except where it matters? *Drug Discovery Today* **12**, 534–539 (2007). URL https://linkinghub.elsevier.com/retrieve/pii/S135964460700222X.

[32] Ewell, J., Gibb, B. C. & Rick, S. W. Water inside a hydrophobic cavitand molecule. *Journal of Physical Chemistry B* **112**, 10272–10279 (2008).

[33] Mahmoud, A. H., Masters, M. R., Yang, Y. & Lill, M. A. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Communications Chemistry* **3**, 19 (2020). URL http://dx.doi.org/10.1038/s42004-020-0261-xhttp://www.nature.com/articles/s42004-020-0261-x.

[34] Hummer, G. Under water's influence. *Nature Chemistry* **2**, 906–907 (2010).

[35] Mitusińska, K., Raczyńska, A., Bzówka, M., Bagrowska, W. & Góra, A. Applications of water molecules for analysis of macromolecule properties. *Computational and Structural Biotechnology Journal* **18**, 355–365 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S2001037019304556.

[36] Maurer, M. & Oostenbrink, C. Water in protein hydration and ligand recognition. *Journal of Molecular Recognition* **32** (2019).

[37] Bodnarchuk, M. S. Water, water, everywhere... it's time to stop and think. *Drug Discovery Today* **21**, 1139–1146 (2016).

[38] Bellissent-Funel, M.-C. *et al.* Water determines the structure and dynamics of proteins. *Chemical Reviews* **116**, 7673–7697 (2016).

[39] Pietrucci, F., Marinelli, F., Carloni, P. & Laio, A. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *Journal of the American Chemical Society* **131**, 11811–11818 (2009). URL https://pubs.acs.org/doi/10.1021/ja903045y.

[40] Tiwary, P., Mondal, J. & Berne, B. J. How and when does an anticancer drug leave its binding site? *Science Advances* **3**, e1700014 (2017). URL https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1700014.

[41] Pérez-Conesa, S., Piaggi, P. M. & Parrinello, M. A local fingerprint for hydrophobicity and hydrophilicity: From methane to peptides. *The Journal of Chemical Physics* **150**, 204103 (2019). URL http://dx.doi.org/10.1063/1.5088418http://aip.scitation.org/doi/10.1063/1.5088418.

[42] Brotzakis, Z. F., Limongelli, V. & Parrinello, M. Accelerating the Calculation of Protein–Ligand Binding Free Energy and Residence Times Using Dynamically Optimized Collective Variables. *Journal of Chemical Theory and Computation* **15**, 743–750 (2019). URL https://pubs.acs.org/doi/10.1021/acs.jctc.8b00934.

[43] Hüfner-Wulsdorf, T. & Klebe, G. Role of Water Molecules in Protein–Ligand Dissociation and Selectivity Discrimination: Analysis of the Mechanisms and Kinetics of Biomolecular Solvation Using Molecular Dynamics. *Journal of Chemical Information and Modeling* **60**, 1818–1832 (2020). URL https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00156.

[44] Ansari, N. & Raucci, U. Hydration spot. https://github.com/narjesansari/hydration_spot (2022).

[45] Talhout, R., Villa, A., Mark, A. E. & Engberts, J. B. F. N. Understanding binding affinity: A combined isothermal titration calorimetry/molecular dynamics study of the binding of a series of hydrophobically modified benzamidinium chloride inhibitors to trypsin. *Journal of the American Chemical Society* **125**, 10570–10579 (2003).

[46] Talhout, R. & Engberts, J. B. F. N. Probing the effect of the amidinium group and the phenyl ring on the thermodynamics of binding of benzamidinium chloride to trypsin. *Organic and Biomolecular Chemistry* **2**, 3071 (2004).

[47] Yamane, J. *et al.* In-crystal affinity ranking of fragment hit compounds reveals a relationship with their inhibitory activities. *Journal of Applied Crystallography* **44**, 798–804 (2011).

[48] Invernizzi, M. & Parrinello, M. Exploration vs Convergence Speed in Adaptive-bias Enhanced Sampling. *arXiv* (2022). URL http://arxiv.org/abs/2201.09950.

[49] Salvalaglio, M., Tiwary, P. & Parrinello, M. Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *Journal of Chemical Theory and Computation* **10**, 1420–1425 (2014). URL http://pubs.acs.org/doi/10.1021/ct500040r.

[50] Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **334**, 517–520 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/22034434https://www.sciencemag.org/lookup/doi/10.1126/science.1208351.

[51] Peña Ccoa, W. J. & Hocky, G. M. Assessing models of force-dependent unbinding rates via infrequent metadynamics. *The Journal of Chemical Physics* **156**, 125102 (2022). URL http://arxiv.org/abs/2112.03249https://aip.scitation.org/doi/10.1063/5.0081078.

[52] Guillain, F. & Thusius, D. Use of proflavine as an indicator in temperature-jump studies of the binding of a competitive inhibitor to trypsin. *Journal of the American Chemical Society* **92**, 5534–5536 (1970).

[53] Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S2352711015000059.

[54] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **185**, 604–613 (2014). URL http://linkinghub.elsevier.com/retrieve/pii/S0010465513003196.

[55] Paszke, A. *et al.* Automatic differentiation in PyTorch. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).

[56] Maier, J. A. *et al.* ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation* **11**, 3696–3713 (2015).

[57] Essmann, U. *et al.* A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593 (1995). URL http://aip.scitation.org/doi/10.1063/1.470117.

[58] Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182–7190 (1981).

[59] Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).

[60] Brotzakis, Z. F. Vac metad applications. https://gitlab.e-cam2020.eu/brotzakis/vac_metad_applications/-/tree/master/TRYPSIN/GROMACS_INPUTe (2019).

[61] Limongelli, V. *et al.* Sampling protein motion and solvent effect during ligand binding. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1467–1472 (2012).

Acknowledgements

Part of this work was performed while the authors were associated with ETH Zürich and Università della Svizzera italiana, Lugano.

The authors thank Faidon Brotzakis for providing the simulations initial input and they are grateful to Michele Invernizzi, Luigi Bonati, Jayashrita Debnath and Umberto Raucci for many useful conversations. The authors thank Michele Invernizzi for the implementation of OPES flooding. N.A. thanks Umberto Raucci for the support in developing the convex hull script. M.P. thanks Maria Ramos and Paolo Carloni for enlightening discussions on protein regulation.

Author contributions statement

N.A. performed the simulations. N.A., V.R. and M.P. discussed the results and reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf