The role of water in host-guest interaction

Valerio Rizzi
Department of Chemistry and Applied Biosciences, ETH Zurich, 8092 Zurich, Switzerland
https://orcid.org/0000-0001-5126-8996
Luigi Bonati
Department of Physics, ETH Zurich, 8092 Zurich, Switzerland https://orcid.org/0000-0002-9118-6239
Narjes Ansari
Department of Chemistry and Applied Biosciences, ETH Zurich, 8092 Zurich, Switzerland
Michele Parrinello (michele.parrinello@phys.chem.ethz.ch)
Department of Chemistry and Applied Biosciences, ETH Zurich, 8092 Zurich, Switzerland

Article

Keywords: atomistic computer simulation, ligand binding energies, host-guest interaction, water

Posted Date: June 29th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-37647/v1

Version of Record: A version of this preprint was published on January 4th, 2021. See the published version at https://doi.org/10.1038/s41467-020-20310-0.

The role of water in host-guest interaction

Valerio Rizzi[1,3], Luigi Bonati[2,3], Narjes Ansari[1,3], and Michele Parrinello[1,3,4,*]

[1]Department of Chemistry and Applied Biosciences, ETH Zurich, 8092 Zurich, Switzerland
[2]Department of Physics, ETH Zurich, 8092 Zurich, Switzerland
[3]Facoltà di Informatica, Istituto di Scienze Computazionali, Università della Svizzera Italiana, Via G. Buffi 13, 6900 Lugano, Switzerland
[4]Italian Institute of Technology, Via Morego 30, 16163 Genova, Italy
[*]michele.parrinello@phys.chem.ethz.ch

June 23, 2020

Abstract

One of the main applications of atomistic computer simulations is the calculation of ligand binding energies. The accuracy of these calculations depends on the force field quality and on the thoroughness of configuration sampling. Sampling is an obstacle in modern simulations due to the frequent appearance of kinetic bottlenecks in the free energy landscape. Very often this difficulty is circumvented by enhanced sampling techniques. Typically, these techniques depend on the introduction of appropriate collective variables that are meant to capture the system's degrees of freedom. In ligand binding, water has long been known to play a key role, but its complex behaviour has proven difficult to fully capture. In this paper we combine machine learning with physical intuition to build a non-local and highly efficient water-describing collective variable. We use it to study a set of of host-guest systems from the SAMPL5 challenge. We obtain highly accurate binding energies and good agreement with experiments. The role of water during the binding process is then analysed in some detail.

Host-guest interactions regulate the workings of proteins and have been intensively studied [1, 2]. Atomistic simulations have been widely used [3, 4, 5] to calculate key parameters like ligand affinity and residence time, and to gain a microscopic understanding of how protein-ligand binding works. The accuracy of these simulations depends crucially on the quality of the model used to describe the interatomic interactions and on the thoroughness of the statistical sampling [6, 7]. We will show that sampling can be much improved if the role of water in the binding-unbinding processes is duly taken into account.

Binding processes take place on a timescale that is unreachable with current computer resources, thus the use of enhanced sampling methods is mandatory. We will frame our discussion in the context of Metadynamics (MetaD) [8, 9, 10] or, more precisely, of its most recent evolution, the on-the-fly probability-enhanced sampling method (OPES) [11]. OPES, like MetaD and many other methods [12, 13, 14], relies on the identification of suitable order parameters or collective variables (CVs). For such methods to be accurate and accelerate sampling, the CVs must be able to describe the slow degrees of freedom of the system. Here we will identify one such powerful CV of general applicability aimed at describing the role of water in the ligand binding process.

Water is expected to play an important role since, upon entering the binding site, the ligand has to shed its solvation shell *in toto* or in part, while the water that originally was in the binding site has to rearrange and negotiate its way out of the binding cavity. Not surprisingly much effort has been devoted on the role of water in ligand-host binding [15, 16, 17, 18, 19]. In the context of enhanced sampling many attempts have been made at capturing the role of water in a CV, leading to an improvement in binding energy estimation [20, 4, 21, 22, 23]. We show here that there is room for a further decisive step as none

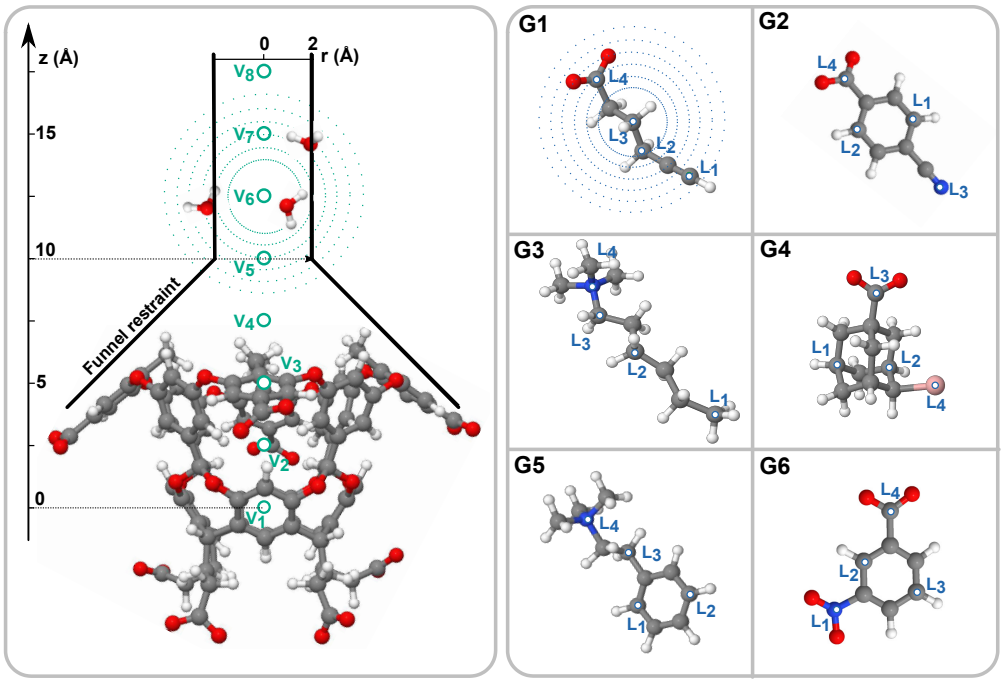

Figure 1: **Sketch of the octa-acid host OAMe with the funnel restraint geometry and the guest molecules from the SAMPL5 challenge.** We indicate the position of the points where the descriptors are centred and hint at their spatial outreach by drawing surfaces at a constant radius around some of them.

of these water-related CVs has been able to accurately describe the highly non-local changes in water structure that take place during binding, both in the vicinity of the ligand and in and around the binding pocket.

In order to succeed in our endeavour, we rely on a combination of physical considerations and modern machine learning (ML) techniques. In particular, we use a method that we have recently developed that goes under the name of Deep Linear Discriminant Analysis (Deep-LDA) [24]. Deep-LDA builds efficient CVs from the equilibrium fluctuations of a large set of descriptors, expressing them as a neural network (NN). In this context, the choice of descriptors is essential and we appeal to our physical understanding to introduce one such set that is capable of characterising not only the ligand solvation shell but also the water structure inside and outside the binding cavity. After building such a CV, we use it in OPES for accelerating the sampling of binding-unbinding events.

We measure the performance of our approach on a set of test systems taken from the SAMPL5 competition [25, 26] and study the interaction of six ligands with an octa-acid calixarene host (OAMe) (see Fig. 1). We choose this system because, despite its relative simplicity, it retains most of the key features of a biologically relevant protein-ligand system. Very recently, a closely related system has been used to investigate how water flows in and out of the system in the absence of a ligand [27]. Furthermore, its symmetry simplifies the analysis and comparison can be made to existing experiments [28] and theoretical calculations [29, 30, 31].

Results

Collective variables from equilibrium fluctuations with Deep-LDA

In this work, we are mainly interested in computing the free energy difference ΔG between the bound state (B) in which the ligand sits in the lowest free energy binding pose and the unbound state (U) where

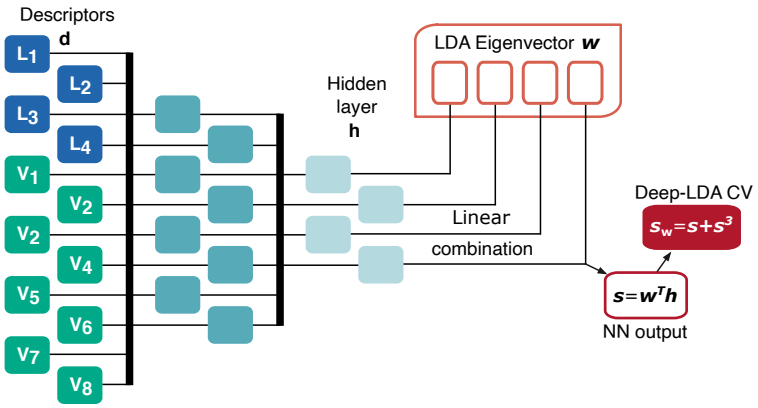

Figure 2: **Schematics of the Deep-LDA architecture used in this work.** The descriptors **d** are fed to a NN that generates s as a linear combination of the last NN hidden layer **h** and the LDA eigenvector $\boldsymbol{w}$. The Deep-LDA CV is then $s_\mathrm{w} = s + s^3$.

the ligand is solvated in water and free to diffuse. In order to obtain a CV able to capture water behaviour we use the recently developed machine learning Deep-LDA method [24].

Deep-LDA is a non-linear evolution of the time-honoured Linear Discriminant Analysis (LDA) classification method [32]. In LDA, one takes two sets of data, in our case the configurations visited in short unbiased simulations in B and U, and defines a set of N_d descriptors **d** that are able to distinguish between the two. The aim of LDA is to find the linear combination of descriptors $s = \boldsymbol{w}^T\mathbf{d}$ that best separates the two sets of data, $\boldsymbol{w}$ being a N_d-dimensional vector.

To this effect, one calculates for each set of data the vectors of the average descriptors values $\boldsymbol{\mu}_\mathrm{B}, \boldsymbol{\mu}_\mathrm{U}$ and their variance matrices $\boldsymbol{S}_\mathrm{B}$, $\boldsymbol{S}_\mathrm{U}$. With these quantities, one then computes the so-called Fisher's ratio:

$$\mathcal{J}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_\mathrm{b} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_\mathrm{w} \boldsymbol{w}}. \tag{1}$$

where one has defined the within scatter matrix $\boldsymbol{S}_\mathrm{w} = \boldsymbol{S}_\mathrm{B} + \boldsymbol{S}_\mathrm{U}$ and the between one $\boldsymbol{S}_\mathrm{b} = (\boldsymbol{\mu}_\mathrm{B} - \boldsymbol{\mu}_\mathrm{U})\,(\boldsymbol{\mu}_\mathrm{B} - \boldsymbol{\mu}_\mathrm{U})^T$. The $\boldsymbol{w}$ that maximises this ratio is the direction that optimally discriminates the two states and gives the best separated projection of the data in the one-dimensional s space. The variable thus obtained has been shown to perform well as CV in many cases, especially if one uses its Harmonic LDA variant [33, 34].

In Deep-LDA, a similar paradigm applies with the key difference that LDA is performed on a non-linear transformation of the descriptors. The non-linearity is introduced by a neural network (NN) (see Fig. 2) whose input is the set of N_d descriptors **d** and the outputs are the N_h components of the last hidden layer **h**. LDA is performed on the components of **h**, so that, after determining the corresponding $\boldsymbol{S}_\mathrm{w}$ and $\boldsymbol{S}_\mathrm{b}$, the NN is optimised using $\mathcal{J}(\boldsymbol{w})$ as loss function. At convergence, one determines the weights of the NN and the N_h-dimensional optimal vector $\boldsymbol{w}$ that produces the Deep-LDA projection:

$$s = \boldsymbol{w}^T \boldsymbol{h}. \tag{2}$$

Deep-LDA is a powerful classifier that tends to compress the data into very sharp distributions which are unsuitable for enhanced sampling applications. To address this issue, we smooth the distributions by applying the following cubic transformation $s_\mathrm{w} = s + s^3$, in the spirit of what done in Ref. [35]. The CV thus obtained will be used to describe water behaviour in our simulations.

Including water in the model

The choice of the descriptors **d** is of paramount importance since it implies the physics that we want to describe. In our case, we are interested in capturing the role of water in the binding process. To this

effect, we choose two sets of points around which we compute the water coordination number. One set is located on the ligand, while the second one is fixed along the host's axis z at regular intervals (see Fig. 1 and the Supporting Information (SI)).

The first set of coordination numbers $\{L_i\}$ describes water solvation around the ligand and is similar in spirit to the ligand solvation variables that have been used in the past [4, 23]. The second one $\{V_i\}$ is aimed instead at capturing the water arrangement inside and outside the binding pocket without any explicit reference to the ligand. The whole set of descriptors $\{L_i, V_i\}$ gives information on the structure of water and its non-local changes on a small to medium length scale during the binding-unbinding process. The use of these descriptors is one of the elements of novelty in our approach and one of the keys to its success.

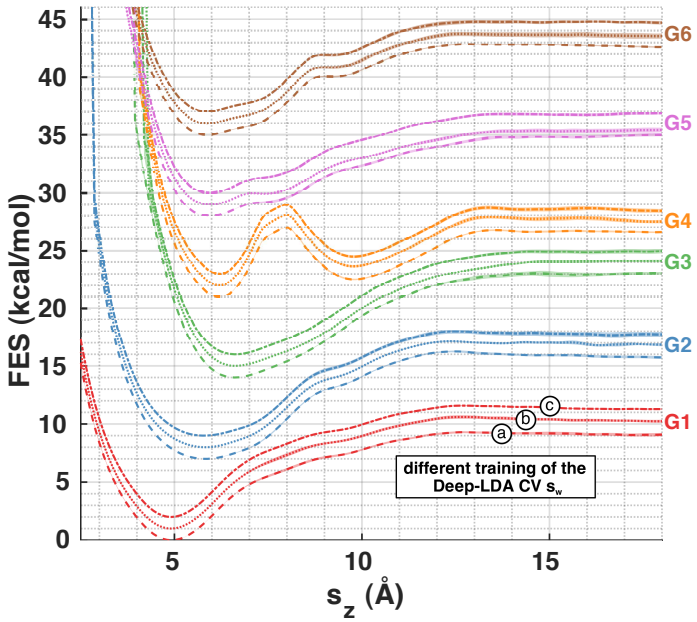

Figure 3: **Free energy surfaces projected along the host-guest distance.** For each of the six ligands, we compute the free energy along the s_z variable using a standard umbrella-sampling-like reweighting formula to recover the unbiased distribution [11]. The shaded areas indicate the errors, whose calculation is detailed in the SI. To ensure that the results do not depend on a specific realisation of the Deep-LDA CV, we repeat the training three times by using different initial weights of the NN. The resulting CVs are denoted as s_w^a, s_w^b and s_w^c and the corresponding FES are indicated respectively by dashed, dotted and dash-dotted lines. For clarity, curves related to the same ligand but with different CVs are shifted by 1 kcal/mol, while the shift between different ligand curves is 5 kcal/mol.

Binding free energies from enhanced sampling simulations

We perform OPES simulations to estimate the binding free energies of all the six ligands of Fig. 1. We use the Deep-LDA CV s_w together with a second CV s_z, that is the projection of the ligand centre of mass on the binding axis z. In the ligand binding context, using the latter is a natural choice [4, 31] as it has a clear physical interpretation and helps in clearly distinguishing B from U. Furthermore, we employ a funnel-like restraint potential [3] to encourage the ligand to find its way back to the binding site once it is out in the solution. The entropic correction to the free energy due to the imposition of the funnel can be calculated analytically (see Eq. S-4) and is taken into account when computing the binding free energies ΔG. We refer the interested reader to the SI for further details.

The combined use of these two CVs leads to a very efficient sampling, which is reflected in a high number of binding-unbinding events per unit time (see for example Fig. S-16). We notice a clear improvement over a more standard set of CVs [31], namely s_z itself and the cosine of the angle θ between the binding axis z and the ligand orientation (see Fig. S-17). The introduction of a water-based CV in enhanced sampling simulations allows the system to reach a regime where it diffuses effortlessly from one metastable state to another, yielding a high accuracy in estimating ensemble averages of physical quantities.

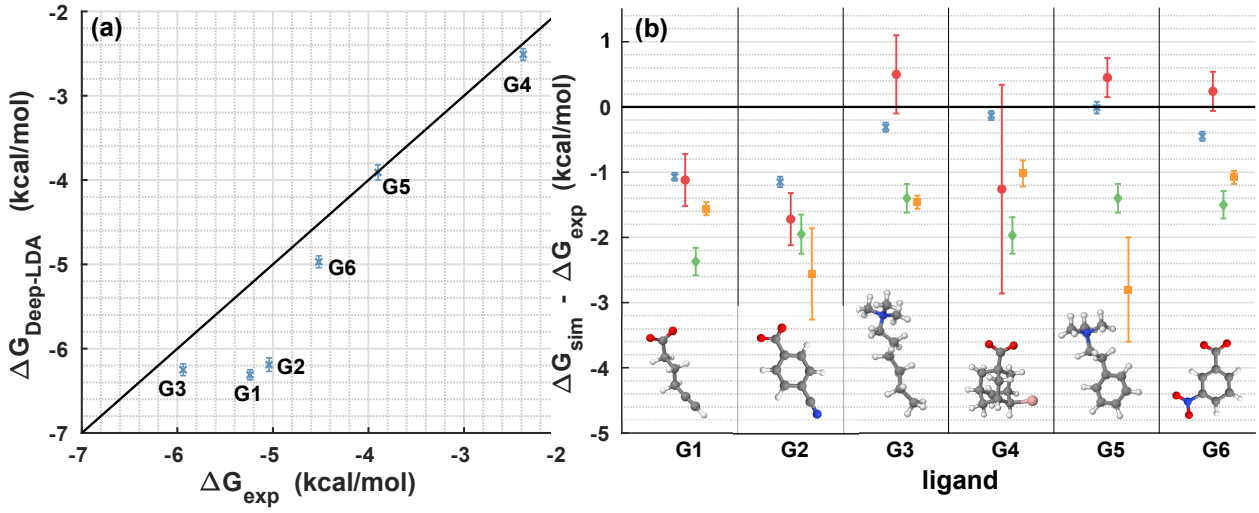

Figure 4: **Comparison of the binding free energies with experiments and other calculations.** In (a), we plot the value of ΔG obtained from the Deep-LDA simulations (in blue crosses) for every ligand versus the experimental values. In (b), we report their difference with the experimental values and compare them with other computational results performed using the same simulation setup. Results from [31] are indicated with red circles, from [30] in green diamonds and from [29] in yellow squares.

Table 1: **Binding free energies.**We show the mean binding energy ΔG (kcal/mol) for every ligand and the corresponding experimental value. We calculate ΔG as a weighted block average over the simulations with all Deep-LDA CVs (see SI for further details).

Ligand	Deep-LDA	Exp
G1	-6.31 ± 0.06	-5.04
G2	-6.19 ± 0.08	-5.24
G3	-6.27 ± 0.07	-5.94
G4	-2.51 ± 0.07	-2.38
G5	-3.91 ± 0.09	-3.90
G6	-4.97 ± 0.07	-4.52

Performing enhanced sampling simulations allows retrieving the equilibrium distribution $P(s)$ of any collective variable s [12]. Here we focus on the free energy surface (FES), defined as $\mathrm{FES}(s) = -k_B T \log P(s)$ where k_B is the Boltzmann constant and T is the temperature of the system. In the context of ligand binding, it is customary to look at the FES as a function of the host-guest distance s_z. For each of the six ligands we compute the FES and estimate the errors with a block average analysis. We report these results in Fig. 3 in which we also assess the robustness of the Deep-LDA CV by showing the results corresponding to three different Deep-LDA training.

We then report the binding energies ΔG corrected for the presence of the funnel in Tab. 1. In Fig. 4 we compare them with experimental values and theoretical calculations performed on the same model but with different sampling techniques [29, 30, 31]. Our results are by and large in agreement with a previous MetaD calculation [31]. However in our case there is a dramatic reduction in the errors.

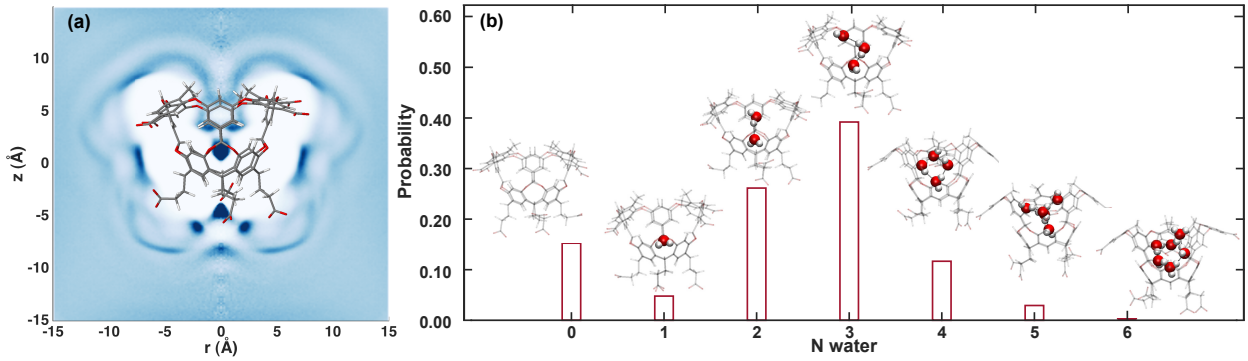

Figure 5: **Water distribution analysis in the presence of the host without a guest.** In (a), histogram in cylindrical coordinate z, r representing the presence of the water oxygen atoms around the host without any guest molecule being present. Darker colours correspond to a higher water density. In (b), probability distribution of the number of water molecules inside the pocket in the absence of a guest molecule. We show the snapshots of typical configurations for each case.

Water behaviour in the ligand-free state

To gain a better understanding of the role that the water plays in host-guest interaction, we first investigate how water interacts with the host in the absence of a ligand. For such analysis, we run a plain molecular dynamics (MD) simulation and find that the number of water molecules inside the cavity fluctuates with a bi-modal distribution between wet and dry states (see Fig. 5 (b)), as observed in a similar system [27]. A typical wet configuration is the one in which three water molecules form a linear cluster inside the cavity, in agreement with results on an analogous system [16].

Another way of representing cavity solvation is to calculate the water oxygen density averaged over the angles around the binding axis (see Fig. 5 (a)), taking advantage of the host's symmetry (see Fig. 1). We observe that there is a high probability of finding a water molecule at the centre of the cavity in proximity of the 8 equatorial oxygen atoms. An analysis of the charge distribution shows that this position is a minimum of the electrostatic potential (see Fig S-2 in the SI). Starting from this position a short wire of hydrogen-bonded water molecules can form inside the cavity. This wire can possibly link up with water outside the pocket as indicated by the density bands in Fig. 5 (a).

The role of water in ligand binding: the case of G4

The use of the Deep-LDA CV s_w not only allows us to obtain accurate binding energies but also a detailed insight into water behaviour during the binding process. We now describe the results obtained for the case of the host in the presence of the ligands. We illustrate here the case of G4, the guest that shows the most complex behaviour, and refer the interested reader to the SI for a detailed analysis of all the other ligands.

In Fig. 6 we show the FES of G4 and the cylindrically averaged water density in the different metastable states. We find that the system presents two binding poses B and B1. The lowest free energy binding pose B is the same as the one found in the experiments and contains no water. Our simulation discovered a second binding pose B1 that differs from B for the presence of a water molecule at the centre of the cavity. This second pose is $\approx 2\,k_B T$ higher in free energy and thus it is occupied with a much lower probability.

When the ligand exits the pocket, before being fully solvated, it can pass through two intermediate short lived states I and I1. In I, the cavity is dry and the ligand is free to rotate in front of the cavity entrance. In I1, the ligand sits again in front of the host entrance but its rotations favour configurations in which the ligand bromine atom points towards the cavity forming a linear arrangement where a water at the centre of the cavity is bridged by another water to the Br^- anion (see Fig. S-20 in SI). We underline that neither B1 nor I and I1 were part of the Deep-LDA training.

The ability of the Deep-LDA CV s_w to capture the non-local water structural changes that appear in our system is the main reason behind our capability to study the system's FES and its metastable states at this level of detail. Non-locality manifests itself in a collective action at a distance on the water in the enhanced sampling simulations, allowing the water to be moved in and out of the pocket even while the ligand is fully solvated and far from the host. Local CVs that only describe the average ligand solvation can only partially take into account these non-local effects. Moreover, the use of CVs that concentrate solely on the position of the ligand with respect to the binding site such as s_z would clearly lead to an incomplete picture. In fact, B and B1 (and similarly I and I1) cannot be distinguished properly by s_z alone and, without the presence of a bias pushing the bound water out of the cavity, B1 would erroneously be over represented in the sampling.

Conclusions

We have shown that, even in the relatively simple systems studied here, a complex and subtle re-organisation of water structure takes place and our strategy is able to capture it. Our calculations not only lead to binding free energies of remarkable accuracy but also offer a powerful analysis tool, proving once more that choosing the right CV is not a mere technical issue but is, in a sense, *the* solution of the problem. Having been able to reduce this much the sampling error, we might even be tempted to claim that the discrepancies with respect to experiments can be blamed mainly on the inaccuracy of the force field. The method is very robust and defines a protocol that can be naturally applied to larger and more complex systems. In fact, the sampling proficiency of our method will prove even more crucial in complex scenarios where a large number of water molecules can be trapped in multiple pocket locations.

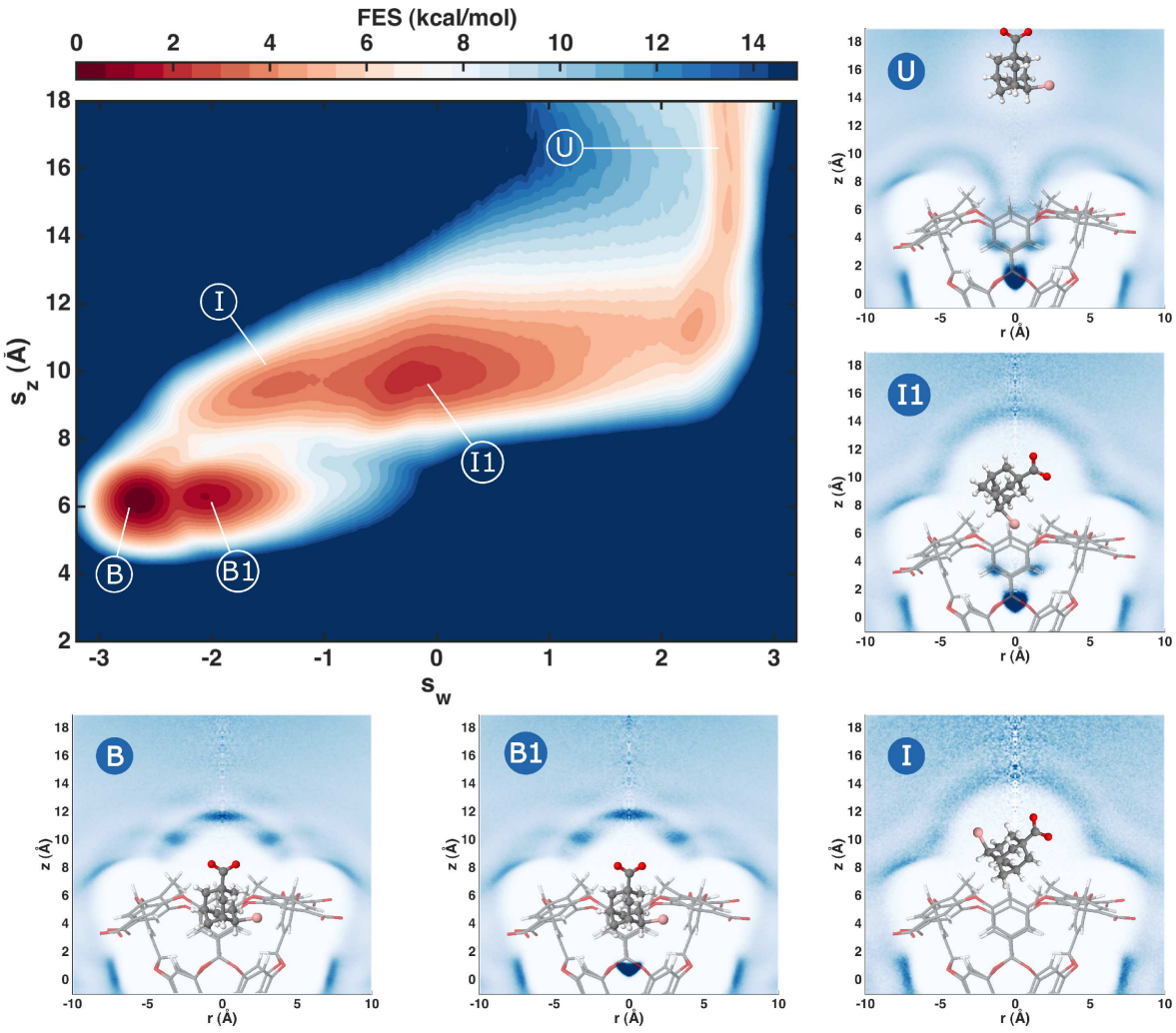

Figure 6: **Binding FES of ligand G4 with a study of the water presence in the visited states.** We show the two-dimensional FES of the ligand G4 with respect to s_z and Deep-LDA CV s_w. Different adjacent colours corresponds to a free energy difference of 1 $k_B T \approx 0.6$ kcal/mol. We highlight some relevant states over which we perform plain MD simulations to measure the presence of water. We show histograms of the water oxygen atoms density in cylindrical coordinates z, r. Each histogram is normalised by the density value in its top right corner and darker colours correspond to higher water density regions. The position of the ligand in these plots is illustrative.

Methods

The simulations inputs were taken from https://github.com/michellab/Sire-SAMPL5. We perform the simulations with GROMACS 2019.4 [36] using the GAFF force field [37] with RESP charges [38] and the TIP3P water model [39]. For enhanced sampling we use a custom version of the PLUMED plugin 2.5.4 [40] where we include OPES [11] and the Pytorch library 1.4 [41]. For each ligand, we first perform plain MD simulations of about 20 ns of state B and U. These trajectories are used for training 3 different Deep-LDA CVs that are then employed in subsequent enhanced sampling simulations. These simulations utilise the multiple walkers feature and include 4 walkers in every calculation, with each walker lasting 140 ns. Average properties are calculated with blocks of 100 ns. More details can be found in the SI.

Data Availability

Simulations data is available from the authors upon request.

Code Availability

The inputs and instructions to reproduce the results presented in this manuscript are deposited in the PLUMED-NEST repository as plumID:XXX. A tutorial about the Deep-LDA training can be found at this link.

References

[1] Michel, J. & Essex, J. W. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *Journal of Computer-Aided Molecular Design* **24**, 639–658 (2010). URL http://link.springer.com/10.1007/s10822-010-9363-3.

[2] Mobley, D. L. & Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annual Review of Biophysics* **46**, 531–558 (2017). URL http://www.annualreviews.org/doi/10.1146/annurev-biophys-070816-033654.

[3] Limongelli, V., Bonomi, M. & Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences* **110**, 6358–6363 (2013). URL http://www.pnas.org/cgi/doi/10.1073/pnas.1303186110.

[4] 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.

[5] Evans, R. *et al.* Combining Machine Learning and Enhanced Sampling Techniques for Efficient and Accurate Calculation of Absolute Binding Free Energies. *Journal of Chemical Theory and Computation* acs.jctc.0c00075 (2020). URL https://pubs.acs.org/doi/10.1021/acs.jctc.0c00075.

[6] Mobley, D. L. Let's get honest about sampling. *Journal of Computer-Aided Molecular Design* **26**, 93–95 (2012). URL http://link.springer.com/10.1007/s10822-011-9497-y.

[7] Rizzi, A. *et al.* The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations. *Journal of Computer-Aided Molecular Design* **34**, 601–633 (2020). URL https://doi.org/10.1007/s10822-020-00290-5http://link.springer.com/10.1007/s10822-020-00290-5.

[8] 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.

[9] 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.

[10] Bussi, G. & Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics* **2**, 200–212 (2020). URL http://dx.doi.org/10.1038/s42254-020-0153-0http://www.nature.com/articles/s42254-020-0153-0.

[11] Invernizzi, M. & Parrinello, M. Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *The Journal of Physical Chemistry Letters* **11**, 2731–2736 (2020). URL http://arxiv.org/abs/1909.07250https://pubs.acs.org/doi/10.1021/acs.jpclett.0c00497.

[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] Tiwary, P. & van de Walle, A. A Review of Enhanced Sampling Approaches for Accelerated Molecular Dynamics. In *Multiscale Materials Modeling for Nanomechanics*, chap. 6, 195–221 (Springer, 2016). URL http://link.springer.com/10.1007/978-3-319-33480-6_6.

[14] 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.

[15] 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.

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

[17] Abel, R., Young, T., Farid, R., Berne, B. J. & Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *Journal of the American Chemical Society* **130**, 2817–2831 (2008). URL https://pubs.acs.org/doi/10.1021/ja0771033.

[18] Wang, L., Berne, B. J. & Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proceedings of the National Academy of Sciences* **108**, 1326–1330 (2011). URL http://www.pnas.org/cgi/doi/10.1073/pnas.1016793108.

[19] 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.

[20] 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).

[21] Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *Journal of the American Chemical Society* **139**, 4780–4788 (2017).

[22] 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.

[23] 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.

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

[25] Bannan, C. C. *et al.* Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* **30**, 927–944 (2016). URL http://link.springer.com/10.1007/s10822-016-9954-8.

[26] Yin, J. *et al.* Overview of the SAMPL5 host-guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design* **31**, 1–19 (2017).

[27] Barnett, J. W. *et al.* Spontaneous drying of non-polar deep-cavity cavitand pockets in aqueous solution. *Nature Chemistry* (2020). URL http://dx.doi.org/10.1038/s41557-020-0458-8http://www.nature.com/articles/s41557-020-0458-8.

[28] Sullivan, M. R., Sokkalingam, P., Nguyen, T., Donahue, J. P. & Gibb, B. C. Binding of carboxylate and trimethylammonium salts to octa-acid and TEMOA deep-cavity cavitands. *Journal of Computer-Aided Molecular Design* **31**, 21–28 (2017).

[29] Bosisio, S., Mey, A. S. & Michel, J. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* **31**, 61–70 (2017).

[30] Yin, J., Henriksen, N. M., Slochower, D. R. & Gilson, M. K. The SAMPL5 host–guest challenge: computing binding free energies and enthalpies from explicit solvent simulations by the attach-pull-release (APR) method. *Journal of Computer-Aided Molecular Design* **31**, 133–145 (2017). URL http://link.springer.com/10.1007/s10822-016-9970-8.

[31] Bhakat, S. & Söderhjelm, P. Resolving the problem of trapped water in binding cavities: prediction of host–guest binding free energies in the SAMPL5 challenge by funnel metadynamics. *Journal of Computer-Aided Molecular Design* **31**, 119–132 (2017).

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

[33] 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.

[34] Capelli, R. *et al.* Chasing the Full Free Energy Landscape of Neuroreceptor/Ligand Unbinding by Metadynamics Simulations. *Journal of Chemical Theory and Computation* **15**, 3354–3361 (2019). URL http://pubs.acs.org/doi/10.1021/acs.jctc.9b00118https://pubs.acs.org/doi/10.1021/acs.jctc.9b00118.

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

[36] 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.

[37] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157–1174 (2004). URL http://doi.wiley.com/10.1002/jcc.20035.

[38] Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993). URL https://pubs.acs.org/doi/abs/10.1021/j100142a004.

[39] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983). URL http://aip.scitation.org/doi/10.1063/1.445869.

[40] 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.

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

Acknowledgements

We acknowledge the Swiss National Science Foundation Grant Nr. 200021_169429/1 and the European Union Grant Nr. ERC-2014-AdG-670227/VARMET for funding. This research was also supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. The simulations were performed on the ETH Euler cluster. Many people helped us during the process of developing and writing this article. We give our sincere thanks to Sergio Pérez, Pablo Piaggi, Riccardo Capelli, Michele Invernizzi, Zoran Bjelobrk, Sandro Bottaro, Yue-Yu Zhang and Tarak Karmakar.

Author contributions statement

V.R. performed the simulations. All authors discussed the results and reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figures

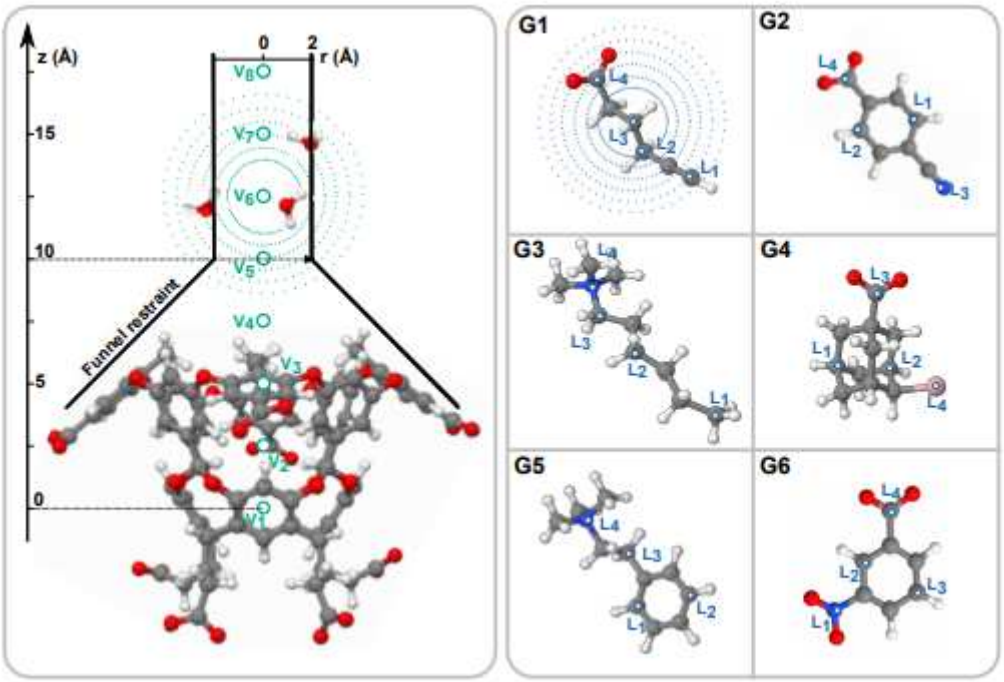

Figure 1

Sketch of the octa-acid host OAMe with the funnel restraint geometry and the guest molecules from the SAMPL5 challenge. We indicate the position of the points where the descriptors are centred and hint at their spatial outreach by drawing surfaces at a constant radius around some of them.

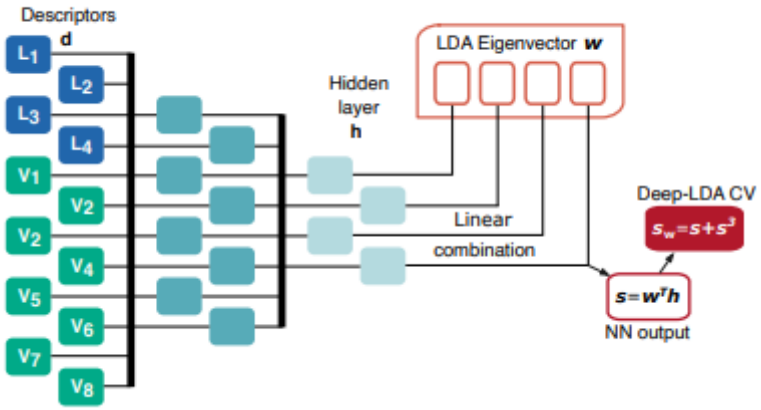

Figure 2

Schematics of the Deep-LDA architecture used in this work. The descriptors d are fed to a NN that generates s as a linear combination of the last NN hidden layer h and the LDA eigenvector w. The Deep-LDA CV is then sw = s + s3.

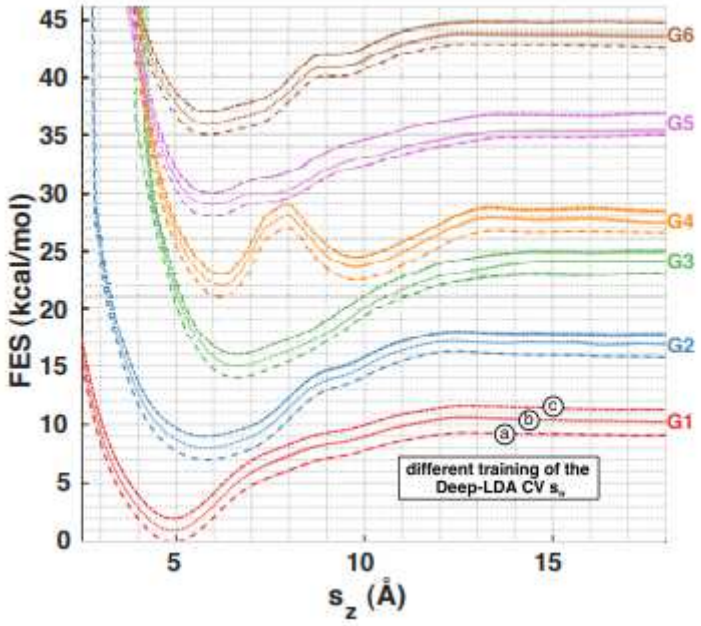

Figure 3

Free energy surfaces projected along the host-guest distance. For each of the six ligands, we compute the free energy along the sz variable using a standard umbrella-sampling-like reweighting formula to recover the unbiased distribution [11]. The shaded areas indicate the errors, whose calculation is detailed in the SI. To ensure that the results do not depend on a specific realisation of the Deep-LDA CV, we repeat the training three times by using different initial weights of the NN. The resulting CVs are denoted as saw, sbw and scw and the corresponding FES are indicated respectively by dashed, dotted and dash-dotted lines. For clarity, curves related to the same ligand but with different CVs are shifted by 1 kcal/mol, while the shift between different ligand curves is 5 kcal/mol.

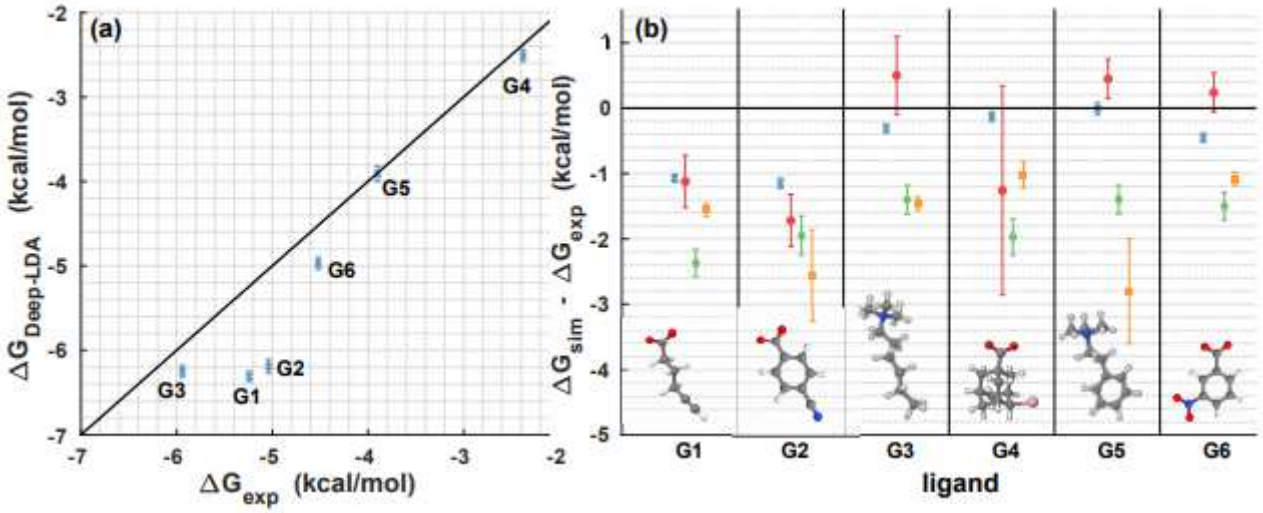

Figure 4

Comparison of the binding free energies with experiments and other calculations. In (a), we plot the value of ΔG obtained from the Deep-LDA simulations (in blue crosses) for every ligand versus the experimental values. In (b), we report their difference with the experimental values and compare them with other computational results performed using the same simulation setup. Results from [31] are indicated with red circles, from [30] in green diamonds and from [29] in yellow squares.

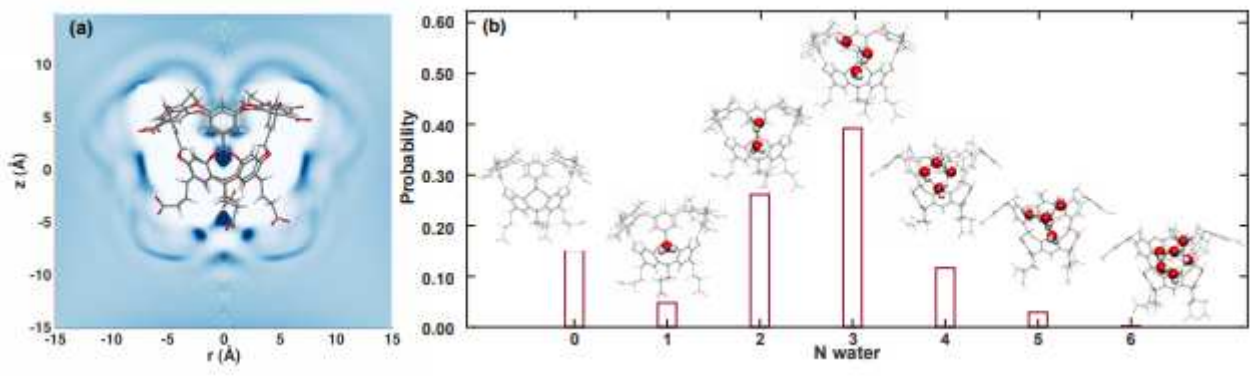

Figure 5

Water distribution analysis in the presence of the host without a guest. In (a), histogram in cylindrical coordinate z,r representing the presence of the water oxygen atoms around the host without any guest molecule being present. Darker colours correspond to a higher water density. In (b), probability distribution of the number of water molecules inside the pocket in the absence of a guest molecule. We show the snapshots of typical configurations for each case.

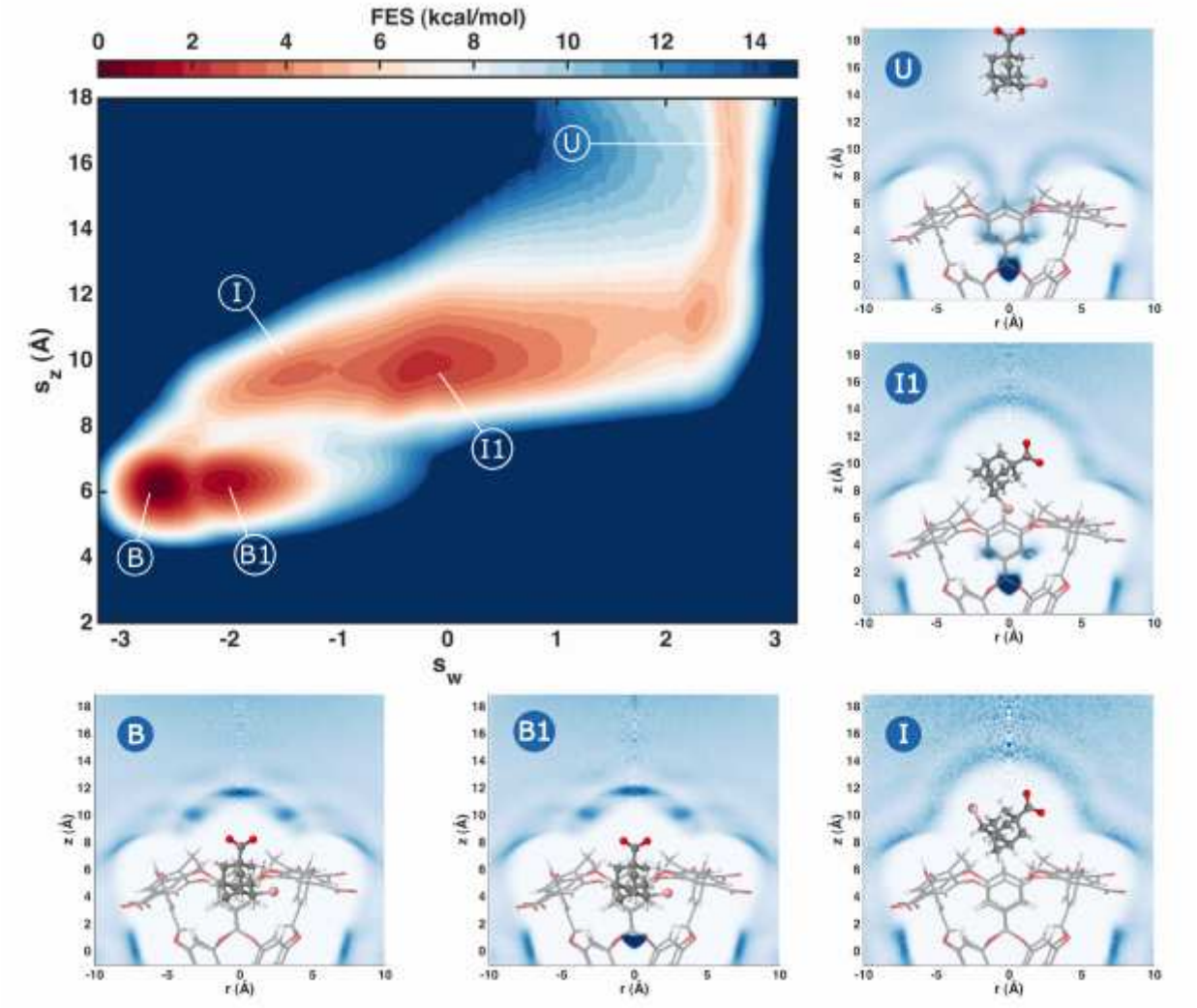

Figure 6

Binding FES of ligand G4 with a study of the water presence in the visited states. We show the two-dimensional FES of the ligand G4 with respect to sz and Deep-LDA CV sw. Different adjacent colours corresponds to a free energy difference of 1 kBT ≈ 0.6 kcal/mol. We highlight some relevant states over which we perform plain MD simulations to measure the presence of water. We show histograms of the water oxygen atoms density in cylindrical coordinates z,r. Each histogram is normalised by the density value in its top right corner and darker colours correspond to higher water density regions. The position of the ligand in these plots is illustrative.

Supplementary Files

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

- DeepLDASI.pdf