Contact map fingerprints of protein-ligand unbinding trajectories reveal mechanisms determining residence times computed from Scaled Molecular Dynamics

Marc Bianciotto,[*,†] Paraskevi Gkeka,[‡] Daria B. Kokh,[¶] Rebecca C. Wade,[¶,§,‖] and Hervé Minoux[⊥]

†*Molecular Design Sciences, Sanofi R&D, Vitry-sur-Seine, France*
‡*Molecular Design Sciences, Sanofi R&D, Chilly-Mazarin, France*
¶*Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*
§*Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance, Heidelberg University, Heidelberg,Germany*
‖*Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany*
⊥*Data and Data Science, Sanofi R&D, Chilly-Mazarin, France*

E-mail: marc.bianciotto@sanofi.com
Phone: +33 (0)1 58933050

Abstract

The binding kinetic properties of potential drugs may significantly influence their subsequent clinical efficacy. Predictions of these properties based on computer simulations provide a useful alternative to their expensive and time-demanding experimental counterparts, even at an early drug discovery stage. Herein, we perform Scaled Molecular Dynamics (ScaledMD) simulations on a set of 27 ligands of HSP90 belonging to more than 7 chemical series in order to estimate their relative residence time. We introduce two new techniques for the analysis and the classification of the simulated unbinding trajectories. The first technique, which helps in estimating the limits of the free energy well around the bound state and the second one, based on a new contact map fingerprint, allows the description and the comparison of the paths that lead to unbinding. Using these analyses, we find that ScaledMD's relative residence time generally enables the identification of the slowest unbinders. We propose an explanation for the underestimation of the residence times of a subset of compounds and we investigate how the biasing in ScaledMD can affect the mechanistic insights that can be gained from the simulations.

1 Introduction

During the past decade, different binding kinetic properties of potential drugs have emerged as one of the key factors that characterize their subsequent clinical efficacy.[1–3] It is thus advisable, if not necessary, to obtain good estimates of these properties even at the early drug discovery stage, for example when selecting chemical series found through virtual screening approaches. In this context, experiments to measure these properties cannot always be performed and *in silico* predictions of these properties provide a useful alternative. The interest of these kinetic properties in drug discovery, as well as the various methods that have been proposed to evaluate them, have been reviewed by several groups.[4–10] The molecular factors that are believed to influence binding kinetics include, among others, binding site accessibility, ligand and protein conformational fluctuations, as well as electrostatic and hydrophobic effects.[11] More recently, ligand desolvation has been pointed out as a key factor influencing residence time[12–14] (abbreviated as τ). While several of these factors can be relatively straightforward to calculate, the actual estimation and evaluation of drug binding kinetics using computational methods remains challenging, indeed more so than drug binding thermodynamics, as it requires a sufficient sampling of the *a priori* unknown unbinding paths.

As members of the Kinetics for Drug Discovery (K4DD) consortium,[15] we were interested in the development and evaluation of molecular simulation approaches to compute drug-binding kinetic properties in the context of hit-to-lead or lead optimization projects. It led some of us to develop the τ-Random Acceleration Molecular Dynamics (τ-RAMD)[16] method for the evaluation of τ of a drug binding to its pharmacological target. It has been coupled with ML analysis[17] and MD-IFP (Molecular Dynamics Interaction Fingerprints)[18] for providing robust estimates of the residence time and insights into the features important for residence time using Machine Learning on the ligand's exit trajectories and applied to systems of varying levels of complexity.[19,20] Related less computationally intensive approaches that rely on the availability of a training set, such as COMBINE analysis,[21] demonstrate remarkable performance by extracting protein-specific multilinear relationships between dissociation rate k_{off} and Lennard-Jones and Coulombic per-residue descriptors. Other approaches have the potential to rank and eventually prioritize new leads by estimating τ. Some of them, based on Metadynamics,[22] are able to accurately order[23] or even quantitatively estimate[24–26] τ, albeit only on small sets of compounds. These approaches rely on Collective Variables (CVs) whose number and definition are often system-specific and generally require trial and error iterations before being considered appropriate.[23] Nevertheless, several more general protocols have been proposed recently such as using generic electrostatics-like CVs,[27] Path CVs,[28] Machine Learning techniques methods[29] for finding optimal CVs and efficiently computing kinetics constants using them,[30] or Ratchet&Pawl simulations for finding CVs and reconstructing the full free energy landscape.[26] Recently, a method based on Steered Molecular Dynamics and on the evaluation of the desolvation energy during unbinding[31] has shown remarkable performance on a set of adenosine A_{2A} antagonists. One useful characteristic of this method is that it leads to simulation times that are approximately independent of the computed residence time of the ligand considered.

In another approach, the use of scaled potentials during Molecular Dynamics (MD) simulations, coined ScaledMD or smoothed potential MD, on multiple replicas of the system, combined with a statistical treatment for calculating the confidence in the estimations, has been shown to perform well for ranking τ in congeneric series on several targets in both a retrospective and a prospective manner.[32–35] The above-mentioned procedure has been integrated within the Biki Life Sciences suite[36] in the Biki-Netics package.

Our initial goal here was to evaluate the performance of Biki-Netics in quantitatively predicting τ and qualitatively ranking compounds within series and across series in the context

of a given industrially-relevant drug discovery project. For this purpose, we performed a series of ScaledMD simulations on a set of 27 ligands belonging to several chemical series and binding to human $HSP90_\alpha$ N-terminal domain (N-HSP90) (Figure 1) in two different conformations. We report our results in light of those obtained with τ-Random Acceleration Molecular Dynamics (τ-RAMD)[16] using the same input structures, topologies and forcefield parameters, and of those obtained with non-equilibrium Targeted MD simulations[37] on a common subset of our dataset. The qualitative comparison between the computed relative residence time and the experiment shows that, within a chemical series, compounds are generally correctly ranked by ScaledMD. Quantitatively, two distinct trends emerge in the ScaledMD results: the τ estimation for the fastest unbinders (defined as $\tau \lesssim \tau_{lig3}$, see below) is correct, but τ is almost systematically underestimated for the slowest unbinders ($\tau > 10$ min). We analyze the unbinding events with two focuses: *the early unbinding event*, which is the crossing of the transition state (TS) before diffusing out of the protein, and *the late unbinding event*, where the ligand separates from the protein. Since no energy profile of the dissociation pathway was generated, we locate approximately the TS by correlating the computed τ and the time needed to reach a certain distance cutoff from the bound state. For describing the late unbinding event, we developed an *ad hoc* protein-ligand contact fingerprint (coined as cFP) in which are listed the last residues close to the ligand during its exit. The clustering of the cFP highlights the two main exit pathways already reported in several other studies,[16,35,37] the front route (1) and the back route (2 in Figure 1), and we discuss their mechanistic relevance. We then dissect the early unbinding event and find that what is key for the exit time is the distance to the actual binding shell, defined as the residues in contact with the ligand in the bound complex. Through an in-depth examination of the unbinding trajectories in a molecular matched pair of compounds, **1** and **2**, we relate the underestimation of τ to the interactions of the ligand at the TS with parts of the protein that are left unstructured by the scaling potential. We also compare several aspects of our work with results obtained by other teams[32,35] on close analogs of the compounds considered here. Overall, our results highlight the strengths and limitations of ScaledMD in the evaluation of protein-ligand binding kinetics and their mechanistic analysis in the perspective of prioritization of small molecules at the early stage of drug discovery process. Moreover, the protein-ligand contact fingerprint defined and used herein (cFP) provides an efficient way for the mechanistic description of ligand unbinding process and a potential measure for direct comparison between different approaches.

2 Methods

2.1 Dataset overview

A set of 27 inhibitors is considered in the present study (Figure 2). It is diverse in terms of chemical structure: the Tanimoto similarity of the ligands to their nearest neighbor in the set computed using Morgan fingerprint[38] is 0.46 ± 0.25. They display different bound protein-ligand conformations (both 'helix' and 'loop' conformations are represented, Figure S1) and a wide range of kinetic and thermodynamic properties, as summarized in the kinetic plot in Figure 3 (See the Supplementary Information - SI - for the structures, experimental data, and respective names in ref. 16 and 37 of the compounds considered in this work). The ligands can be clustered into 8 chemical series, seven based on their central heterocyclic moiety and one containing three singletons. Two of the ligands, compounds **7** and **21**, can be clustered either in Cluster 1 or Cluster 3.

The two N-HSP90 conformations considered differ by the secondary structure of residues 105 to 113: these residues from the sequence subset C form the L1 loop that covers the binding site in the conformation represented in Figure 1 that will be called 'loop' in the text[39] (the sequence subsets A to G are defined in Figure 1). These residues are involved in the α_3 he-

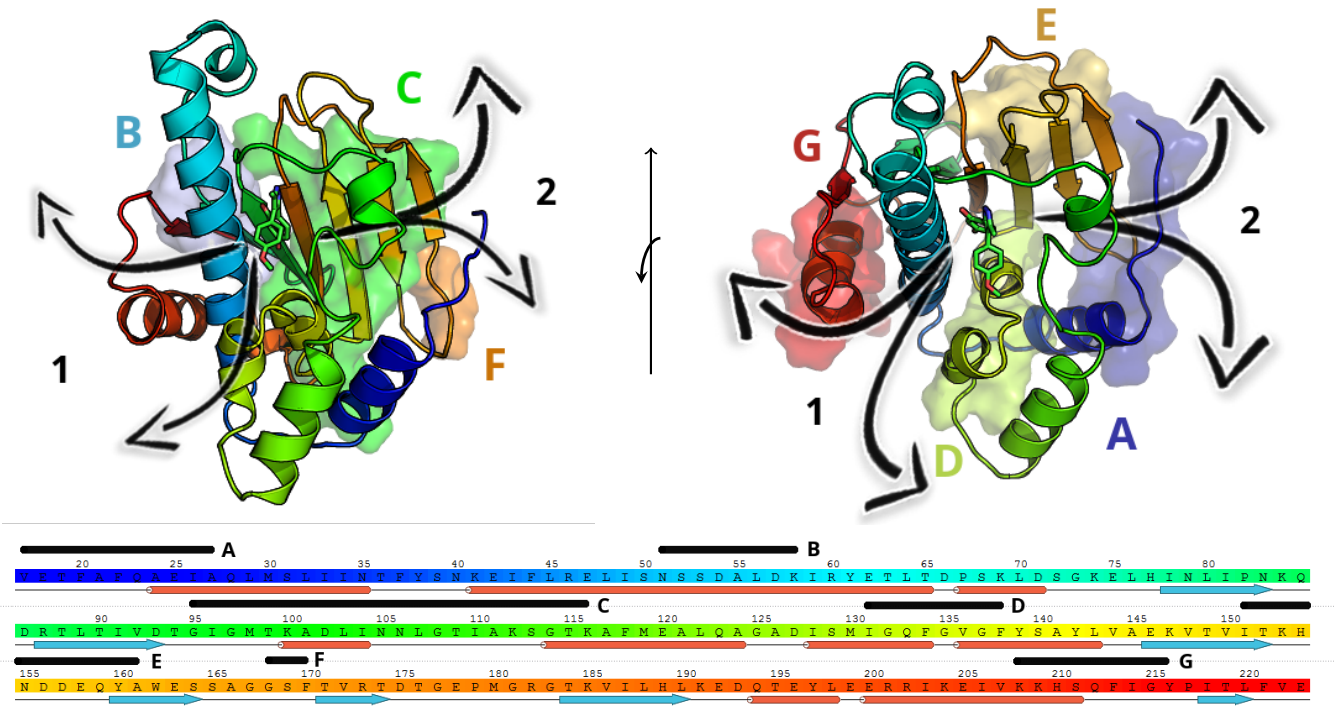


Figure 1: Top: Two views of the N-terminal domain of HSP90 (N-HSP90) in 'loop' conformation in complex with compound **8** (pdb 6ELN[16]). The right view is rotated by c.a. 90° towards the reader with respect to the left view. Major ligand exit routes are illustrated by arrows. 1: Front route 2: Back route. Sequence subsets A to G are highlighted in the 3D structures. Subset A (res. 17-27, in deep blue) is from N-Ter up to the beginning of α-helix 1. Subset B (res. 51-58, in light blue) is located at the center of α-helix 2. Subset C (res.95-116 in green) encompasses the loop at the top of the binding site, α-helix 3, the L1 stretch known to be involved in the conformational flexibility of N-HSP90, up the beginning of α-helix 4. Subset D (res.131-138, in lime green) is α-helices 5 and 6. Subset E (res. 151-161, in light orange) is the loop between β-sheet 1 and 2. Subset F (res.168-170, in deep orange) is second half of the loop between β-sheet 2 and 3. Subset G (res. 208-216, in red) is α-helix 7. Bottom: Subsets definition and residue coloring convention projected on to the N-HSP90 sequence.

lix spanning residues 101 to 124 in the 'helix' conformation of L1 (see Figure S1 for a visual comparison of the two conformations). In four of the chemical clusters, all compounds in the series bind to the same N-HSP90 conformation; for the other three clusters, the N-HSP90 conformation is different depending on the compound. For all but two inhibitors (**7**, **22**), X-ray crystal structures were used as the starting point for system preparation. The starting structures of the complexes for the remaining two compounds were prepared as described in ref. 16.

2.2 System preparation

The input structures and topologies were obtained from ref. 16 in Amber format and were then converted to Gromacs format using Acpype[40] and in-house scripts. A detailed step-by-step description of the conversion process and all scripts can be found on the KBbox website.[41,42] All the ligands were protonated using Epik[43,44] at pH 7.5 and parametrized for the GAFF force-field using Antechamber.[45] RESP partial charges[46,47] were fitted from electrostatic potentials generated at the HF/6-31G*(1d) level using GAMESS.[48] The Amber14 force field[49] was used for the protein, a 10Å buffer of TIP3P[50] water molecules was

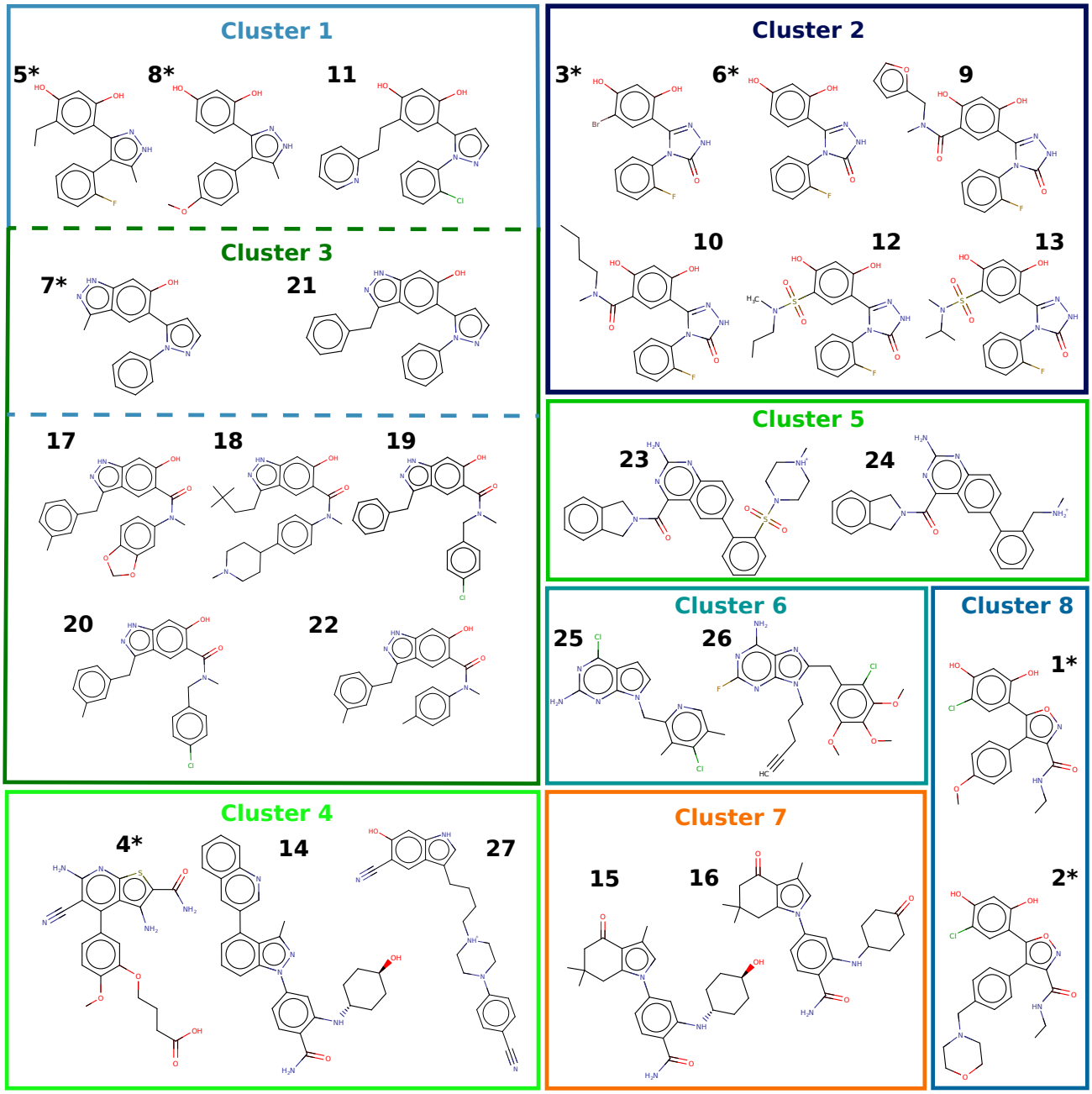


Figure 2: Chemical structures of the ligands used in this study. The compounds are grouped in clusters based on their chemical similarity. The same coloring per cluster is used throughout the publication. Starred ligands are bound in loop conformation, the others are bound in helix conformation.

built around the complex with tLeap,[51] and Na^+ or Cl^- ions were added to ensure system neutrality and 150mM salt concentration to mimic physiological conditions. All protein-ligand systems were then minimized, heated, and relaxed for 10ns before the ScaledMD simulations.

2.3 Scaled molecular dynamics simulations

In ScaledMD, the conformational space sampled in MD is increased by smoothing the system's potential energy. This is particularly useful when we are trying to observe protein-ligand dissociation using MD simulations. In

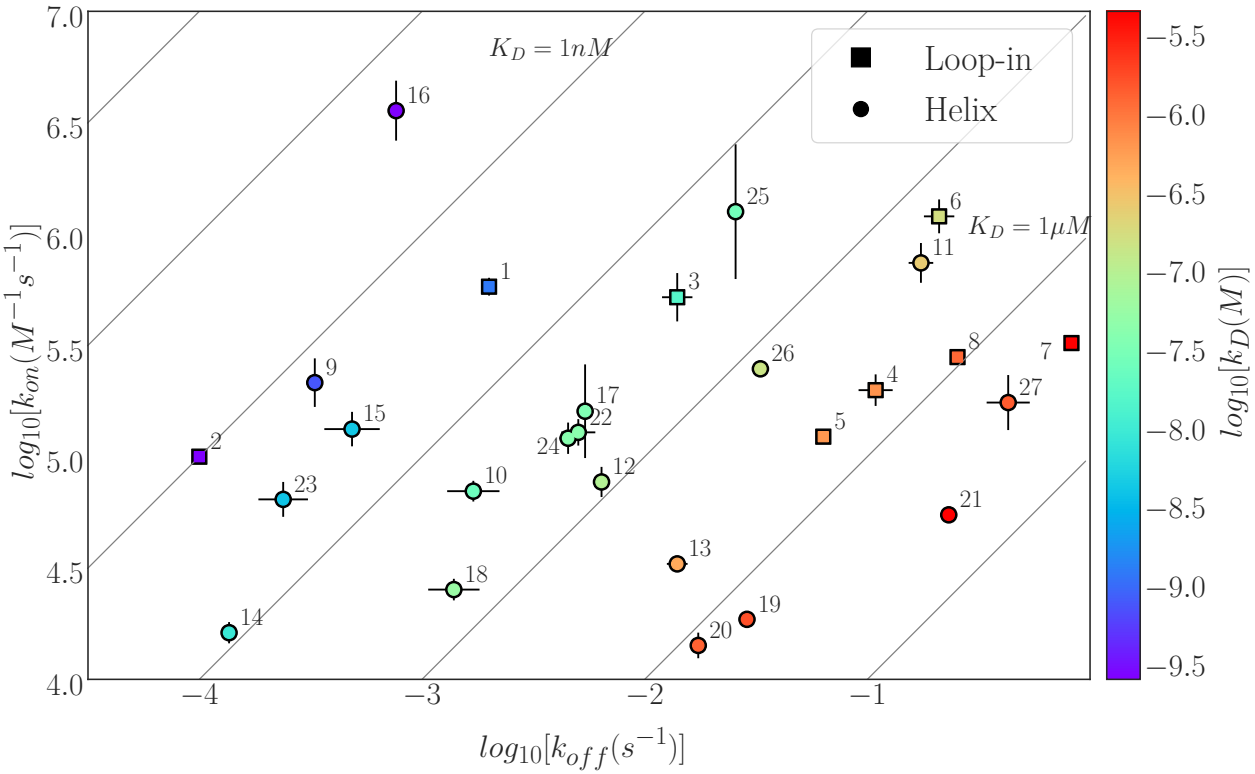


Figure 3: Kinetic map relating the measured association rate constant k_{on} with the dissociation rate constant k_{off} for the 27 compounds under study. The corresponding error bars are also shown. Squares indicate the loop N-HSP90 bound conformation, circles the helical N-HSP90 bound conformation. The color of the symbols is related to each ligand's k_D ; iso-affinity (k_D) lines are shown in gray.

our study, each of the prepared protein-ligand systems was used as the starting point of circa 20 independent MD simulations (with the exception of compound **16**, for which only 7 simulations led to an unbinding event) using GROMACS 4.6.1[52] modified for ScaledMD[53] as implemented in the Biki suite (version 1.0.7).[36] The importance of using multiple replicas in MD simulations has been previously discussed in ref. 54 and thus a sufficiently large number of replicas was used. It is however important to note that some simulations did not led to a complete dissociation and ligand exit from the binding site and were discarded from the analysis.

To prevent protein unfolding during protein-ligand dissociation, it is common practice in ScaledMD to restrain all protein atoms outside the binding site. In our systems, in order for the ScaledMD simulations to be comparable between different ligands, the same set of unrestrained protein residues was used throughout the simulations of all compounds and defined as follows. The protein residues closer than 6Å to all ligand heavy atoms in their bound structure were considered first. The selection was then expanded if needed in order to include single residues located between two selections. It led to the following residues being unrestrained : 47-56 (similar to subset B), 93-114 (similar to subset C), 132-138 (similar to subset D) and 183-186. Despite the fact that the initial 6Å selection criterion is the same as in ref. 32 (it is 5Å in ref. 35), in the present study, the final selection of unrestrained residues is larger, reflecting the greater diversity of the bound conformations in our dataset. Nevertheless, the whole subset C is left free of restraints during the simulations, which translates into the sampling of several unfolded conformations of α-helix 3 and L1 during the ScaledMD simulations. Positional harmonic restraints with a force con-

stant of 50 kJ/mol/nm^2 were imposed on the backbone of all restrained protein residues. As in ref. 32, the unbinding time t_{off} was defined as the first time when the ligand is surrounded by two full solvation shells, defined as a union of spheres of 6Å radius, centered on the ligand atoms and free of protein atoms. For each ligand, the unbinding time was computed from the average unbinding time over all replicas and its uncertainty was estimated by a 1000-fold bootstrapping (see Table S3 in SI).

One of the crucial parameters in ScaledMD simulations is the choice of the scaling factor λ. Similarly to ref. 32, different λ values were tested in order to identify the optimal one as the best balance between a reasonable computing time and unbinding times ranging from a few nanoseconds to tens of nanoseconds. For this evaluation, we simulated eight ligands (**2**, **3**, **8**-**13**) covering a large range of measured residence times (from four seconds to more than 2 hours), bound to two different conformations of N-HSP90 and belonging to different chemical series.

Three different values of λ were tested, namely 0.40, 0.42 and 0.45, leading to 480 simulations in total with computed τ values ranging from 8ns to 130ns depending on the ligand and the scaling factor. A value of $\lambda = 0.50$ was also considered, but simulating the ligands with the smallest τ with this scaling factor led to mean t_{off} greater than 150ns, suggesting that for the rest of the set, the simulation times needed would be too long for practical use. $\lambda = 0.40$ was found to be optimal for this set of molecules as the correlation coefficient between the computed and measured residence times was the best and the error was the smallest (See Figure S3 and text in SI). This value of λ was used for ScaledMD simulations for the 19 remaining ligands. All the related analyses can be found in SI. All simulations were performed on Intel processors under Linux on the Marconi facility at Cineca (Italy).

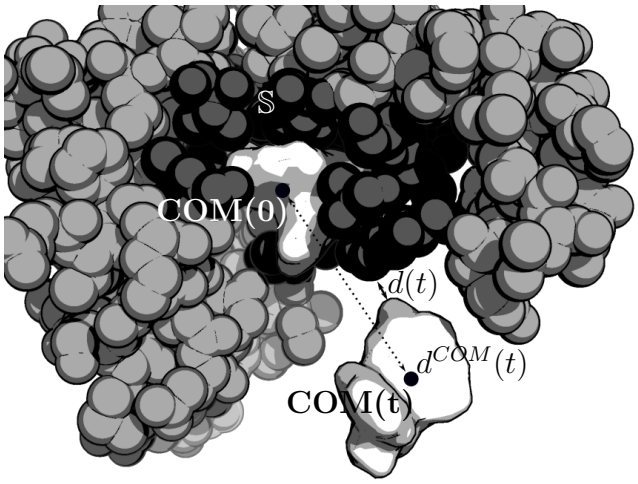


Figure 4: Definition of the metrics used for the analysis of the early exit events. The residues of the initial contact shell 𝕊 are indicated in black. The distance between the nearest heavy atoms of ligand and residues of 𝕊 is indicated as $d(t)$. The COM metric, $d^{COM}(t)$, measures the distance between the COM of the ligand at a given t and at $t = 0$.

2.4 Analysis of the unbinding process

In order to characterize the features of the ligand unbinding trajectories, three complementary approaches were followed and implemented using MD Traj[55] and contact_map[56] python libraries. The first two approaches focus on characterizing the early stage of the unbinding process and the third describes the latest stages of the unbinding trajectories. The notations used in the text are in Table 1.

2.4.1 Early unbinding

Before we describe the process followed for the characterization of the early unbinding, we need to define the metrics used. First, we define the initial contact shell 𝕊 as the set of the protein residues that are close to the ligand l in the bound state (colored black in Figure 4). These are residues whose heavy atoms are found closer than 3.5Å to the heavy atoms of the ligand at the starting point of the simulation. Then, $d(t)$ is the shortest distance between any heavy atom of the ligand and the nearest of the residues in 𝕊

Table 1: Notations used in the text

τ	Measured residence time
t_{off}	Computed unbinding time: the first time when the ligand is further than d_{far} from any heavy atom of the protein
$\tau_{comp} = (t_{off}/t_{off,lig3})^{\lambda}$	Computed relative residence time (wrt **3**)
$\tau_{rel} = (\tau/\tau_{lig3})^{\lambda}$	Measured relative residence time (wrt **3**)
$t_{fpt}(d_c)$	First passage time: the first time when $d(t, i \in \mathbb{S}) \geq d_c$
$t_{fpt}^{COM}(d_c)$	First passage time in COM metric: the first time when $d^{COM}(t) \geq d_c$
$d(t)$	Distance between the ligand and the initial contact shell $\mathbb{S}$ at time t
$d^{COM}(t)$	Distance between the COM of ligand at time t and at $t = 0$
d_c	Cutoff distance used to assess the early unbinding
$d^{\pm}$	Distance where $t_{fpt}(d^{\pm})$ correlates significantly with t_{off} : $R^2(d^{\pm}) > 0.8$
$d_{far} = 6$Å	Distance parameter which defines the full unbinding, see t_{off}
$d_{FP} = 4.5$Å	Distance parameter for cFP generation, only residues closer than d_{FP} to the ligand are taken into account during cFP generation
d_r	Distance between the ligand and a given residue during cFP generation. For d_{FP} and d_r, the closest heavy atoms of the ligand and the residue are considered
$\delta_{CM}(cFP_A, cFP_B)$	Distance in contact map space between the cFP, FP_A and FP_B, of unbinding trajectories A and B
$\mathbb{S}$	Initial contact shell: residues that are closer than 3.5Å to the bound ligand
N_A	Set of the residues that are closer than d_{FP} to the ligand during the late unbinding
$n(N_A)$	Number of residues in N_A
$R^2(d_c)$	Correlation coefficient between the distribution of unbinding times t_{off} and first passage times $t_{fpt}(d_c)$ at the cut-off ditance d_c for a set of simulations
$R^2_{COM}(d_c)$	Correlation coefficient between the distribution of unbinding times t_{off} and first passage times $t_{fpt}^{COM}(d_c)$ at the cut-off ditance d_c for a set of simulations

at time t. Finally, we define the center of mass (COM) metric $d^{COM}(t)$ as the distance between the COM of the ligand at time t and at $t = 0$, which corresponds to the ligand bound state. The above-mentioned metrics are illustrated in Figure 4.

For the characterization of the early unbinding using the first approach, for each ligand and for each exit trajectory, we perform the following steps: first, for each timeframe t of the whole trajectory, the distance $d(i \in \mathbb{S}, t)$ between the heavy atoms of the ligand and those of all residues of $\mathbb{S}$ is computed. Then, given a specific cutoff distance d_c, the first passage time $t_{fpt}(d_c)$ is determined as the first time where $d(t, i \in \mathbb{S}) \geq d_c$.

The second approach for characterizing the early unbinding relies on $d^{COM}(t)$, the displacement of the COM of each ligand l with respect to its initial position at time t along the trajectory, as described in Figure 4. Given a cutoff distance d_c, one defines the first passage time $t_{fpt}^{COM}(d_c)$ as the time where $d^{COM}(t) \geq d_c$.

2.4.2 Late unbinding

As discussed in ref. 35, despite the challenge of characterizing the ligand exit pathways in ScaledMD simulations, these simulations might give indications on relevant structural aspects of their unbinding mechanisms. Herein, we propose a simple procedure for describing these exit pathways using as a metric a protein-ligand

contact fingerprint (cFP) which is built as described in Algorithm 1.

As illustrated in Figure 5, this way of describing the late ligand unbinding process has two features: first, it ensures that the nearest distance between each of the nearest residues and the ligand during the last frames of the late unbinding is kept in the cFP, and second, by limiting the total number of residues selected in the process, it ensures that the subsequent comparison between cFP will not be skewed by the longest exit trajectories. Qualitatively, this fingerprint highlights the nearest residues encountered by the ligand during its unbinding from the protein. Importantly, this is done without the need for explicit consideration of time.

A distance function $\delta_{CM}(cFP_A, cFP_B)$ was devised for comparing the similarity of two unbinding trajectories A and B in contact map space. We define cFP_A and cFP_B as the contact map fingerprints of the two trajectories, N_A and N_B as the sets of the nearest residues, $n(N_A)$ and $n(N_B)$ are the cardinalities of the sets considered, so that the non-zero values of e.g. cFP_A can be written as $(res_{i,A}, d_{i,A})_{1\leq i\leq n(N_A)}$: for example, in Figure 5, $res_{1,A} = 31$ and $d_{1,A} = 4.44$. The distance function $\delta_{CM}(cFP_A, cFP_B)$ is expressed by Eq. 1, using the union, intersection and symmetric difference (Δ) between the two sets. This function simplifies to a normalized Euclidean distance when both trajectories have all their nearest residues in common, and d_{far} is the contribution to δ_{CM} for each of the nearest residues which are not in common in the two trajectories. A rare case happens when, for both trajectories, the ligand exits immediately after having reached d_c and stays beyond d_{far} for all protein residues. In this case, both $n(N_A) = 0$ and $n(N_B) = 0$ and by convention $\delta_{CM}(cFP_A, cFP_B) = d_{far}$.

$$\delta_{CM}(cFP_A,cFP_B)=\sqrt{\frac{\sum_{1\leq i\leq n(N_A\cap N_B)}(d_{A,i}-d_{B,i})^2+n(N_A\Delta N_B)\times d_{far}^2}{n(N_A\cup N_B)}} \quad \text{(Eq. 1)}$$

The function δ_{CM} defined in Eq. 1 is used in order to perform cluster analysis of ensembles of late unbinding trajectories. First, δ_{CM} is used to build a distance matrix between the cFP and a hierarchical clustering is performed using the complete method, with a threshold of 95% of the maximum distance between all clustered trajectories. Finally, the distance matrix and the cFP are sorted according to the similarity between the cFP determined at the clustering step.

3 Results and discussion

3.1 An example of the unbinding events analysis

As described in the Methods section, a set of analyses has been devised in order to best describe the early and late unbinding events and to enable comparison between the different ligands. Below, we describe the outcome of the trajectory analysis for compound **8** using this set of tools. All the related results are depicted in Figure 6.

3.1.1 Early unbinding: unbinding time and distance cutoff

In Figure 6b, we have plotted the unbinding time t_{off} versus the time $t_{fpt}(d_c)$. $t_{fpt}(d_c)$ is the first passage time : it is the time for the ligand to reach the distance cutoff $d_c = 3$Å (in black) or $d_c = 4.5$Å (in colors) from its bound pose. A significant correlation between t_{off} and $t_{fpt}(d_c)$ indicates that d_c is beyond the limits of the kinetic trap which keeps the ligand in the vicinity of the initial bound state, and that the ligand is free to diffuse quickly up to the unbound state. It is thus expected that the corresponding correlation coefficient increases sharply starting from a certain value of d_c. Indeed, in this example, R^2 rises from 0.27 for $d_c = 3$Å to 0.89 for $d_c = 4.5$Å. In order to better characterize the limits of this basin, the correlation coefficient $R(d, l)$ between t_{off} and $t_{fpt}(d_c)$ over all replicas was computed for different values of d_c. It was indeed found that for high enough values of distance d_c, $t_{fpt}(d_c)$ and t_{off} were correlated and that the slope of this

Algorithm 1 cFP building procedure for a particular ligand trajectory.

Require: ligand exit trajectory $A(t = 0, ..., t_{off})$
  $t \leftarrow t_{off}$
  $nearest_A \leftarrow \{\}$ ▷ The set of (nearest residue, nearest distance) along the trajectory A
  $cFP \leftarrow \{\}$
  while $d(t) > d_c$ or $n(nearest_A) < 20$ **do** ▷ No more than 20 residues considered
    $t \leftarrow t - 1$
    Update $d(t)$ ▷ Only the late unbinding trajectory is considered
    Find in $A(t)$ the *list* of protein residues closer than d_{FP} (4.5Å) to the ligand
    for residue r in *list* **do**
      Compute the distance d_r to the ligand
      if r in $nearest_A$ and $d_r < (d_r$ in $nearest_A)$ **then**
        Update d_r in $nearest_A$
      else if r in $nearest_A$ **then**
        Add (r, d_r) to $nearest_A$
      end if
    end for
  end while
  for r in protein residues **do** ▷ c_{FP} building
    if r in $nearest_A$ **then** $cFP(r) \leftarrow d_r$
    else $cFP(r) \leftarrow 0$
    end if
  end for

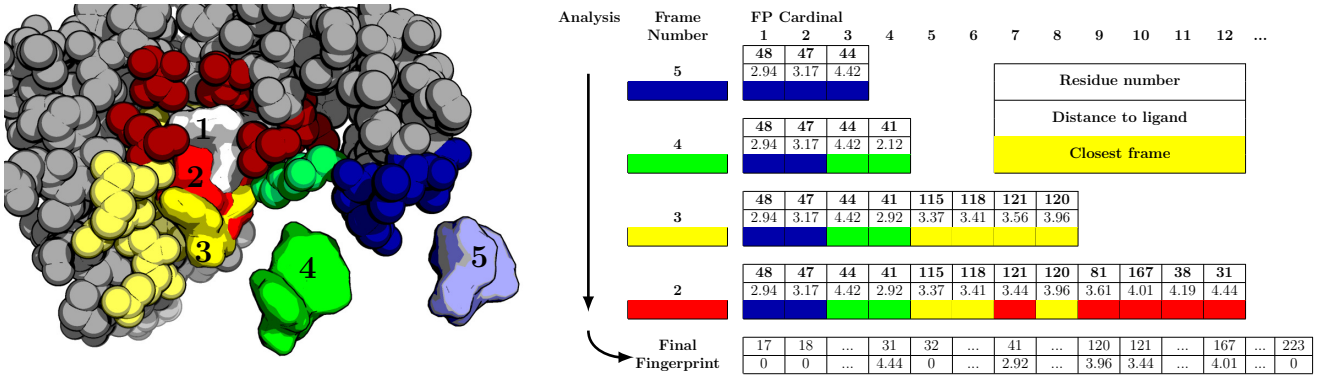


Figure 5: Illustration of the contact map fingerprint generation process. The fingerprint is built by scanning the exit event in reverse starting from the frame corresponding to ligand exit from the protein (defined here as frame 5, in blue) until the first frame after the first passage time t_{fpt} (frame 2 in this example). Residues closer than $d_{FP} = 4.5$Å relative to the ligand in each frame are added to the contact map fingerprint along with the nearest distance to the ligand among all frames.

correlation $R^2(d_c)$ was close to 1. The same procedure was performed for relating the unbinding time t_{off} and the time $t_{fpt}^{COM}(d_c)$ for the ligand COM to reach the distance d_c from its initial position, which allows the computation of the squared correlation coefficient $R^2_{COM}(d_c)$. The trends for all compounds of the set are displayed on Figure 7 and will be discussed later.

3.1.2 Late unbinding: contact maps and exit paths

The cFP of the 20 unbinding trajectories with $d_c = 4.5$Å is represented in Figure 6a. The cFP bits are related with the sequence subsets of HSP90 that are highlighted in Figure 1, and the cFP bits themselves are darker when the

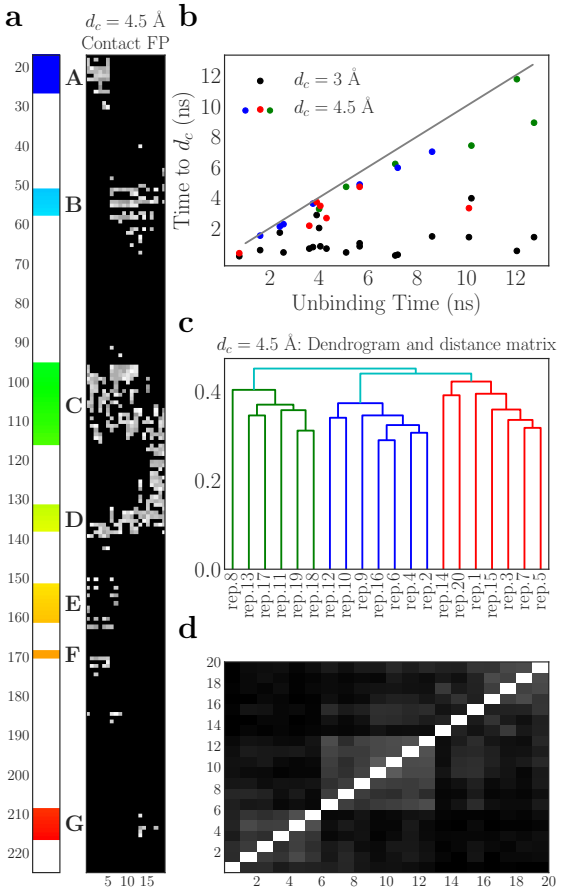


Figure 6: Compound **8** trajectory analysis. **a** Heatmap of the cFP. On the x axis, the trajectory cFP are sorted according to the dendrogram (see below). Pixels are colored by cFP value, white for the nearest residues and black beyond the distance cutoff $d_c = 4.5$Å. The protein sequence numbering is on the y axis, as well as the sequence subsets referred to in the text, using the color code illustrated in Figure 1. **b** First passage time $t_{fpt}(d_c)$ for $d_c = 3$Å (in black) and $d_c = 4.5$Å (in colors) vs unbinding time t_{off}. Each dot represents one trajectory, its color relates to one of the three exit clusters in the dendrogram **c**. **c** Dendrogram of the unbinding trajectories (replicas 1 to 20) according to the distance matrix. The generated ordering is used in **a** and **d**. At the cutoff of 95% of the maximum inter-replica distance, three clusters are identified. **d** Matrix of δ_{CM} distances between the unbinding trajectories sorted as in the dendrogram. Short distances are light grey and long distances black.

associated distance is higher. The distance matrix between the trajectories is displayed in Figure 6d. It indicates the existence of three main exit clusters, as confirmed by the dendrogram in Figure 6c. Its coloring relates to the exit cluster with the unbinding and exit times displayed in Figure 6b. It shows for example that the earliest exit event belongs to the red trajectory cluster, which corresponds to direct exits of the ligand with only a few interactions outside subsets C and D residues. The cFP associated with the green cluster displays interactions with subsets A, C, D, E and F but not with subset B; it corresponds to the back route (route 2) in Figure 1, where the ligand goes below α-helix 3 (subset C) during its exit. In the case of the blue cluster, interactions with subsets B, C and D indicate the front route (route 1).

3.2 Clearly identifiable Transition States are crossed early

The evolution of R^2, R^2_{COM} and of the mean RMSD of the protein (calculated for heavy atoms) with $d(t)$ (resp. $d^{COM}(t)$) is shown for all ligands in Figure 7. First, all mean RMSD plots have a very similar shape: they reach a plateau around $d = 3.5$Å to 4Å regardless of the bound conformation of the ligand. Let us then consider the correlation coefficient R^2, in green. It is close to zero when d is small : small displacements of the ligand out of the binding site do not end up in an exit event. However, in 18 of the 27 cases, R^2 then rises sharply to a plateau whose value is generally between 0.8 and 1.0. It means that beyond a certain distance threshold $d^{\pm}$, the ligand becomes free to diffuse rapidly out of the protein: the time to reach this threshold is highly correlated with t_{off}. In this case, a kinetically relevant TS is located before, but close to $d^{\pm}$. For 9 ligands, **7**, **11**, **14**, **19**, **20**, **23**, **24**, **25**, and **27**, R^2 rises slowly and, in some cases, it stays well below 0.8, so that no specific TS event can be observed. Finally, the trend concerning R^2_{COM}, in blue, is less systematic, the plateau is lower than in the case of R^2 for the same ligand, and, in most cases, the plateau is reached for higher values of d. This last finding demonstrates that residue-specific information describes more accurately the kinetically rele-

vant unbinding event than the displacement of the ligand from its starting point, which is described by R^2_{COM}. Moreover, for several large compounds with rotatable bonds, such as **14**, **18**, **25**, **26**, and **27**, the correlation coefficient R^2_{COM} is close to zero for the whole range of distances considered, indicating the limitation of the COM metric for the most complex compounds.

The comparison of these plots for two pairs of similar compounds highlights some consequences of the scaling on the description of the unbinding process. A first example is the pair of compounds **14** and **15**. They share a common substructure and differ mainly by the quinoline moiety of **14** which interacts with the hydrophobic pocket formed under α-helix 3 (both are helix-binders). The RMSD plots indicate that this hydrophobic packing discussed in Kokh et al.[16] for these two compounds does not prevent destabilization of the protein during the unbinding process, as the RMSD plateau of compound **14** is similarly high and is reached sooner than for compound **15**. The variation of R^2 with d for the two compounds is markedly different and this difference can be related to the differences in structure and illustrates a limit of the definition of the metric. Indeed, in the case of the long and buried compound **14**, the ligand stays close to some residues of the initial binding shell until late in the unbinding process, which is much less the case for compound **15** and leads to an earlier increase of R^2. A second matched pair is composed of compounds **7** and **21**, with the methyl in **7** (loop binder) changed to a phenyl group in compound **21** (helix binder). They have a very similar K_D, but compound **21** has a only a slightly longer residence time.

Overall, the analysis shown in Figure 7 allows for most of the ligands to pinpoint the threshold where the kinetically relevant TS has been crossed and after which the ligand is free to diffuse. This information can be of great interest for identifying the interactions that most influence the kinetics of the unbinding process, i.e. the ones disappearing around the plateau of R^2, especially in cases where the trends in R^2 and R^2_{COM} are different. Finally, the influence of the potential scaling on $d^{\pm}$ is illustrated in Figure S4 (top), where the squared correlation coefficients R^2 between the unbinding time t_{off} and the first passage time $t_{fpt}(d)$ is plotted against the distance d for the three values of λ used for the ligands of the training set. Two situations are observed: for ligands **3** and **8**, $d^{\pm}(\lambda = 0.45) > d^{\pm}(\lambda = 0.42)$, which indicate that the TS is displaced depending on the scaling. For the six other cases, $d^{\pm}$ is similar for the different values of λ, suggesting a smaller effect of the scaling on the description of the TS.

3.3 Most ligands follow nearly equally the two main exit pathways in ScaledMD simulations

The cFP of all unbinding trajectories ordered by cluster is shown in Figure 8. (Similar cFP and dendrograms representations together with the correponding distance matrices were produced for all ligands, see SI). 27 clusters were identified by the clustering analysis, the end of all exit trajectories were visually reviewed together cluster by cluster and annotated as Front- or Back-like trajectories at the cluster level. These visualizations confirm that unbinding trajectories belonging to the same cluster display similar features. In a few cases, the limitation of the cFP in describing topology makes it unable to assign the front or the back route e.g. where a ligand with a front-like cFP rotates and goes behind the unfolded α-helix 3 just when exiting.

In most trajectories, the ligand interacts with the beginning of subset C (bright green), which is the loop before α-helix 3 which surrounds the binding pocket. As discussed previously for ligand **8**, other frequent combinations of interaction hotspots during ligand exit are subsets A, E and F on one hand (back route), and subsets B and D on the other (front route). It is interesting to note that for all ligands in the set, between c.a. 25% and 70% of all exit trajectories involve the back exit route (see individual heatmaps in SI). It is a greater proportion than in ref. 35 where the back route is

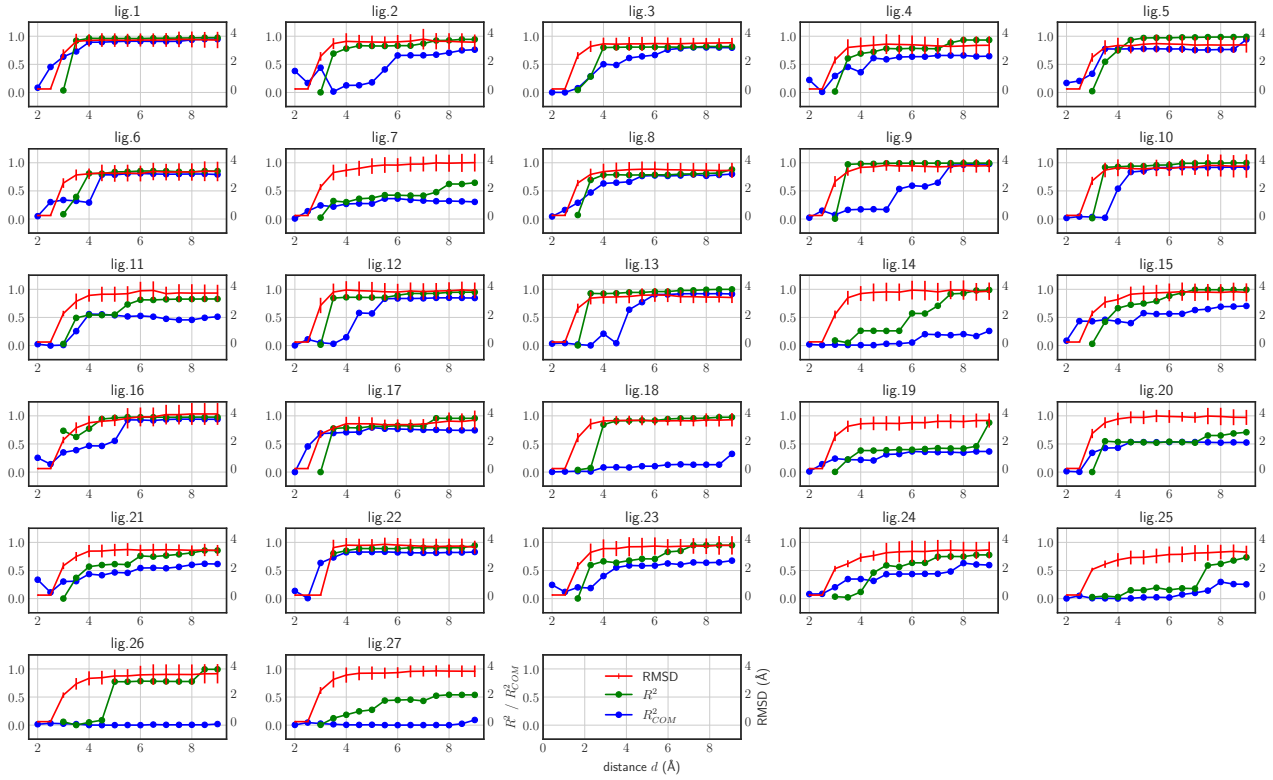


Figure 7: Squared correlation coefficients R^2 (green) and R^2_{COM} (blue) between the unbinding time t_{off} and the first passage time $t_{fpt}(d)$ (resp. $t^{COM}_{fpt}(d)$)as a function of the distance d for each ligand considered. In red, mean RMSD of the protein heavy atoms as a function of d (error bars are SD).

found, but more anecdotally. This might be due to the fact that in our work the whole subset C is left free of restraints and adopts unfolded conformations during the ScaledMD simulations that open space for the ligands to exit by the back route. This further emphasizes the importance of choosing wisely the restraints used in ScaledMD simulations in order to balance between the ability to compare a large set of ligands using a larger set of unrestrained residues and the ability to obtain structurally relevant insights from the simulations by restricting the set of unrestrained residues. Indeed, in τRAMD simulations,[16] the ligand exit channel that goes under the α-helix 3 is only open when α-helix 3 has a pure helical conformation and when the transient hydrophobic sub-pocket under this helix is open, which means that loop-binding compounds were not using the back route. Wolf et al.[37] have also compared those two routes using ligand **2** and found that the front route was favored in terms of work required to pull the ligand into the solvent. Nevertheless, the exit trajectories corresponding to the two groups display roughly the same mean exit time within statistical error in ScaledMD. This might be surprising at first, but it is less so considering the fact that the fingerprinting considers only residues that are beyond the $d_c = 4.5$Å distance cutoff to the initial contact shell $\mathbb{S}$. At this distance from $\mathbb{S}$, as suggested by Figure 7, in most of the cases the ligand has already crossed the limits of the kinetic basin and the exact interactions with its surrounding do not have a notable effect on its exit time. From this analysis, it is tempting to question the relevance of studying the latest stages of the exit pathways when using ScaledMD for understanding the specifics of the Structure-Kinetics Relationships, as the interactions which have the most effect on exit time during simulations are those that are broken around the threshold in $R^2_{COM}(d)$, which is often closer than d_c to the initial contact shell residues.

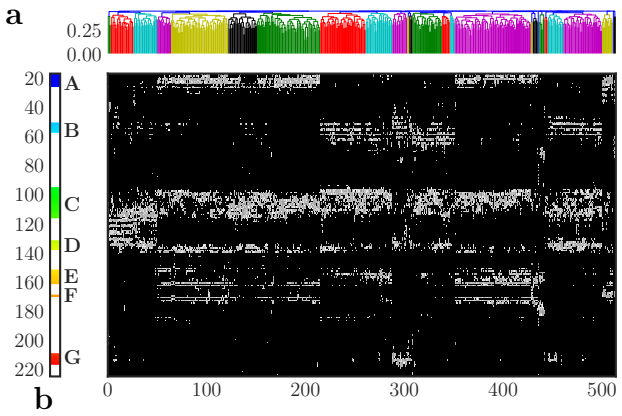


Figure 8: Clustering of all 520 unbinding trajectories cFP for all ligands, using $d_c = 4.5$Å. **(a)** Dendrogram of the hierarchical clustering for all trajectories, leading to 27 different clusters. **(b)** Heatmap of the cFPs, using the same conventions as in Fig. 6. The full figure with the matrix of δ_{CM} distances between the unbinding trajectories and the coloring of the dendrogram leaves depending on the chemical cluster of the ligand is in SI.

3.4 Estimation of Relative Residence Time

Relative residence times, τ_{comp}, were computed as the ratio of the mean unbinding time during the ScaledMD simulations to the mean unbinding time of ligand **3**. τ_{comp} were related to the measured residence times relative to ligand **3** $\tau_{rel}(\lambda) = (t_{off}/t_{off,lig3})^{\lambda}$ and plotted in Figure 9.

The correlation obtained is modest when considering the whole set, but can be analyzed as the superposition of two trends. The first trend is related to the subset of compounds for which $\tau_{comp} \sim \tau_{rel}$ in Figure 9. The second trend is represented by a set of compounds, generally with larger relative residence time ($\tau_{rel}(\lambda) > 2$, i.e. $t_{off} > 7.9.10^{-2}s$), where the slope between the relative residence time and τ_{comp} is lower than 1. This trend is followed by larger and more rigid compounds such as **2**, **9**, **10**, **14**, the pair **15** and **16** from Cluster 7, **18**, from Cluster 3 and the pair **23** and **24** from Cluster 5. In Schuetz et al.,[35] compound 6, which is the largest of the dataset and has the longest measured residence time, was also underestimated in ScaledMD unbinding simulations performed with very similar settings. Similarly to their studies, this compound was the only compound whose τ is significantly underestimated in this dataset, it was rightfully considered as an outlier and discarded from analysis. As described in equation 3 in SI, under simple approximations τ_{comp} can be rewritten as $e^{(\zeta(l_2,\lambda)-\zeta(l_1,\lambda))/R} \times \left(\frac{\tau(l_2)}{\tau(l_1)}\right)^{\lambda}$ where $\zeta(l,\lambda)$ is the deviation of the activation entropy from linearity with respect to λ. The contributions $\zeta(l,\lambda)$ of the two ligands might be different in cases where the smoothing of the Potential Energy Surface by the scaling favors a different ensemble of TSs for the two ligands. A similar effect with respect to the scaling factor is illustrated in Figure S4 in SI, where the squared correlation coefficients $R^2(d)$ between the unbinding time t_{off} and the first passage time $t_{fpt}(d)$ is plotted against the distance d for the three values of λ used for the ligands of the training set. For compounds **3** and **8**, the profiles are markedly different between $\lambda = 0.45$ and the two lower values, which suggests that the TS is located much further from the binding site with this scaling factor than with the lower ones.

In the next sections, we will describe and discuss the two trends in τ_{comp} found from the analysis of several chemical series in this dataset: Clusters 1, 2, 3, 7 and 8 (see SI for description of Clusters 4-6). Based on the analysis of the pair from Cluster 8, we will propose an interpretation of the systematic underestimation of τ in one of the two subsets.

3.5 ScaledMD distinguishes short versus medium τ ligands within series and underestimates τ for ligands with long τ

Figure 10 summarises the structures as well as the τ_{rel} and τ_{comp} values for the compounds of Cluster 1 and 3. Cluster 1 (in light blue) contains five compounds which have short residence times. They are close (structurally and in terms of kinetic constants) to the set of four

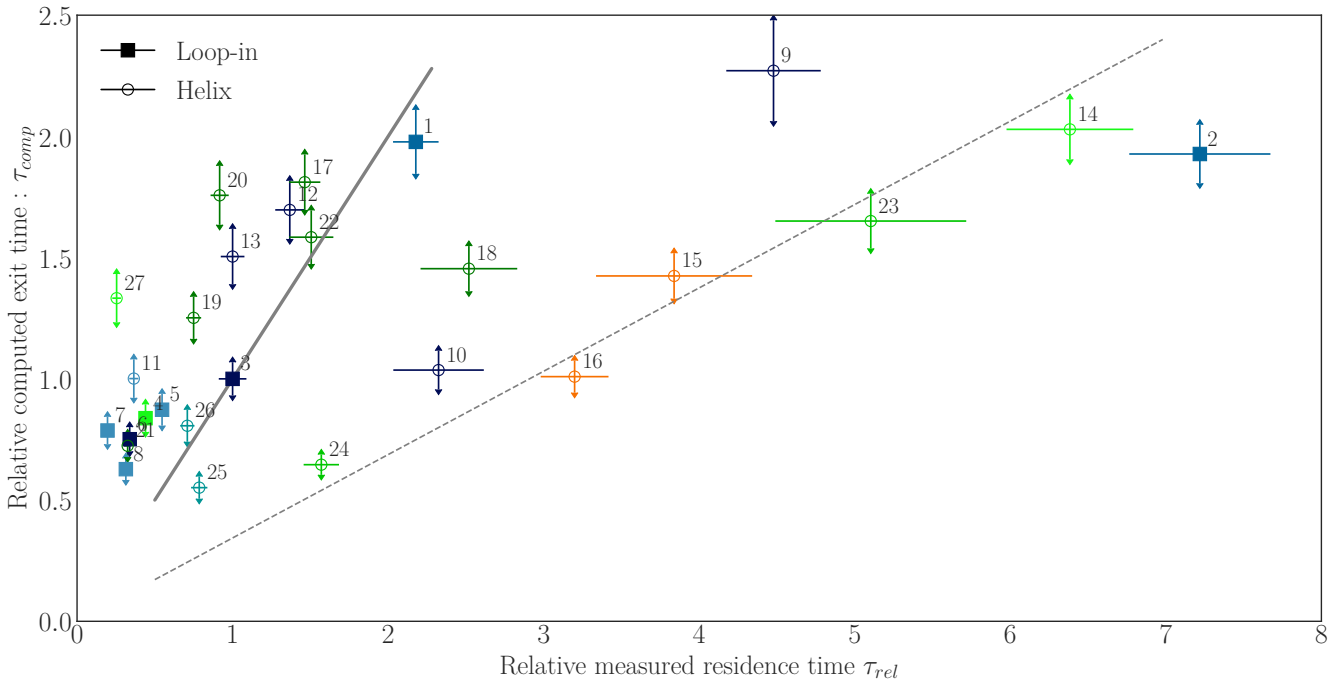


Figure 9: τ_{comp} vs τ_{rel} for $\lambda = 0.4$. Vertical error bars show standard deviation, and horizontal error bars show standard error. The color of the symbols corresponds to the chemical series (see Fig. 2) and their shape to the bound HSP90 conformation. The bold line indicates the line where τ_{comp} and τ_{rel} are equal and the dashed line shows the trend followed by a subset of the ligands with higher residence time.

HSP90 resorcinol ligands studied with the same methodology in ref. 32. In our case, however, the relative computed residence times are not linearly related to the relative measured unbinding constants. Their relative computed residence times are all very close and they do not allow ranking the compounds when the error estimated by bootstrapping is taken into account. This difference might be due to the fact that the compounds in this set are more diverse structurally than in reference 32. This structural diversity in the ligand structures also leads to interactions with a diverse set of residues in the bound state, leading to a larger set of unconstrained residues used in our work. Three of the Cluster 1 compounds (**5**, **8** and **11**, see also Table S2) have also been simulated by Wolf et al.[37] with a very similar outcome: the three ligands are among the fastest unbinders but they are not differentiated quantitatively.

Cluster 3 comprises seven compounds (including compounds **7** and **21**, which are shared with Cluster 1) coming from a rescaffolding of the compounds from Cluster 1 and 2. More specifically, the resorcinol ring has been fused with the central heterocyclic ring to form an hydroxybenzopyrazole, which is modified on two different positions. In this series, the group in R2 position binds to the same hydrophobic pocket as the diverse groups in Cluster 2. On the other hand, the distal part of the group in R1 points towards the outside of the binding site, so that it does not interact much with N-HSP90 in the bound state. One large subseries of five compounds (**17-20** and **22**) is of N-methyl-amides bound to a diverse set of substituted phenyl or benzyl groups in R1. For **17, 19** and **20**, the kinetic estimation is roughly correct, while it is underestimated for compounds **18** and **22** which are N-phenyl substituted compounds very close structurally to **17**. These series are also close structural analogs of the four compounds in the main series studied in Schuetz et al.,[35] which displays a similar trend, where the ligand with a long measured residence time is underestimated.

The two last compounds (**7** and **21**) are common between Cluster 1 and Cluster 3. They

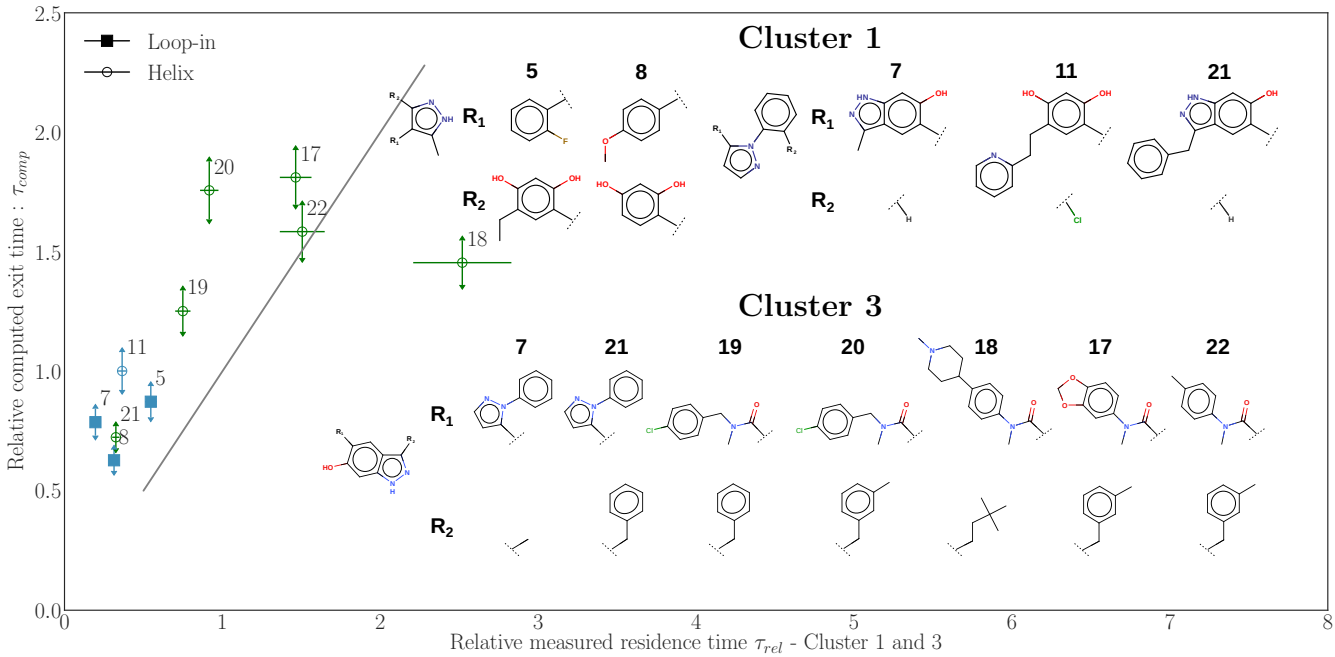


Figure 10: τ_{comp} vs τ_{rel} and their respective standard deviations for Clusters 1 (light blue) and 3 (green). Compound **3** is used as a reference. The color of the symbols corresponds to the chemical series (see Fig. 2) and their shape to the bound HSP90 conformation. Compounds **7** and **21** are depicted twice in 2D as they belong to both series.

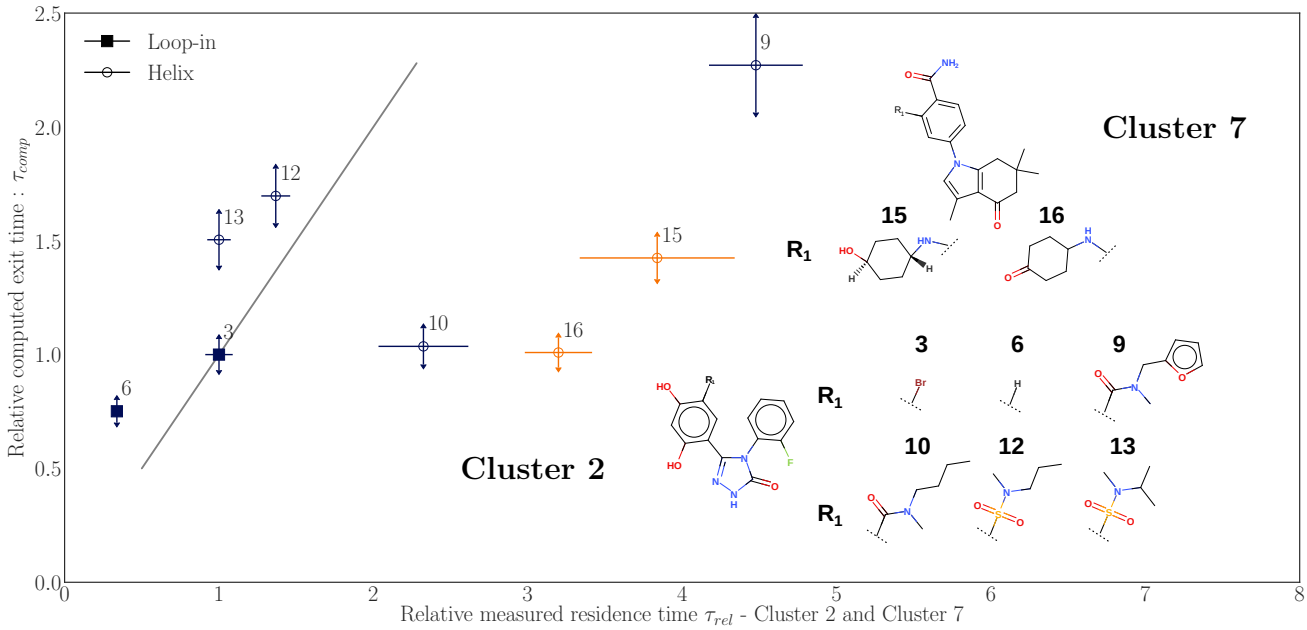


Figure 11: τ_{comp} vs τ_{rel} and their respective standard deviation for Clusters 2 (blue) and 7 (orange). Compound **3** is used as a reference.

have the lowest τ and K_D within Cluster 3, probably because they neither accept the hydrogen bond formed with the Thr184 side-chain that is in the large Cluster 3 subseries, nor donate an hydrogen bond to the Gly97 backbone C=O bond, as for **5** and **8** from Cluster 1, which are better binders than **7** and **21** (see Figure 12). They differ structurally only by a phenyl group in the R2 position. This structural modification leads to a modest difference in τ which

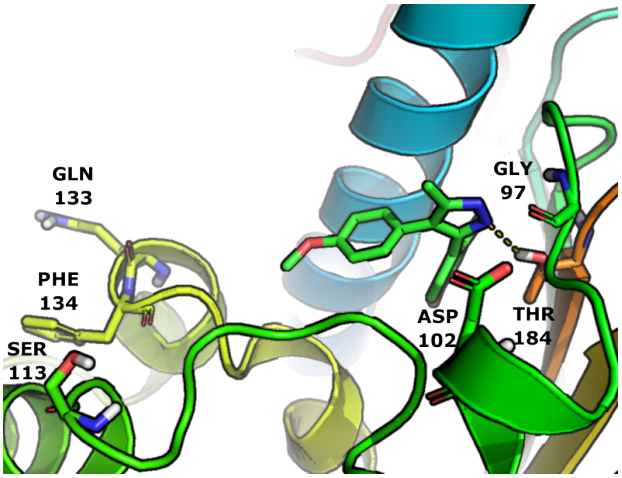


Figure 12: Detail of the N-terminal domain of HSP90 (N-HSP90) in complex with compound **8** : residues cited in the text are labelled, H-bond shown by dashed line.

is not captured by the simulations. Taken together, Clusters 1 and 3, which are close structurally, illustrate a general observation: neither the shortest τ compounds (with an measured relative residence time of less than 0.7) nor the longest τ ones (greater than than 0.7) are ranked correctly within each group. However, the longest τ compounds are clearly separated from the shortest τ compounds in Figure 10.

The relative residence times for compounds from Cluster 2 and 7 are displayed in Figure 11. In this case, the computed residence time for the shortest τ compounds (those with measured relative τ less than two) and for the longest τ ones (the rest of the set) is similar, but compounds can be ranked quite well within each set. Cluster 2 contains six compounds that are structurally very close as they represent single point variations, from R1=H (compound **3**) and bromine (compound **6**) to bulky amides and sulfonamides. For the latter compounds, the different R1 groups are located in the hydrophobic pocket beneath helix 3. **3** and **6** bind to N-HSP90 in its loop conformation, they are differentiated by the ScaledMD simulations and follow the expected quantitative trend, but τ_{comp} for **6** is overestimated with respect to **3**, which is used as a reference for the plot. It might be that the force-field parametrization of the bromine atom of **3** in GAFF, which does not include its polarization, leads to an underestimation of τ_{comp} with respect to **6**. Compounds **12** and **13**, which bind to the helix conformation of N-HSP90, differ only by the position of a methyl group on a sulfonamide and are nearly indistinguishable by their relative τ_{comp}. Compounds **9** and **10** are correctly differentiated, but their τ_{comp} are underestimated with respect to τ_{rel}. Interestingly, the τ of **9** is also underestimated by τRAMD.[16]

The two compounds in Cluster 7 only differ by the reduction of the ketone **16** to the alcohol **15** and their bound structures are superimposable. The difference in kinetic behaviour associated with this very subtle structural modification is correctly reproduced by the simulation but, in this case too, τ_{comp} underestimates τ_{rel} (Figure 11). Interestingly, this pair simulated using TMD[37] had equal unbinding mean work within the error bars.

3.6 Non-specific interactions at TS with unstructured α-helix 3 leads to τ underestimation

The structures of **1** and **2**, the two molecules that belong to Cluster 8, as well as their τ_{rel} and τ_{comp} are given in Figure 13. Their central heterocyclic core is unique in this dataset, but it is decorated similarly to compounds from Clusters 1 and 2. Their only structural difference is in the R1 position, which changes from a methoxy group in compound **1** to a methylmorpholine in compound **2**. These structural differences are associated with a large difference in τ_{rel}, which is not reflected by the simulations. Experimentally, the presence of the morpholine substituent lowers k_{off} by a factor of more than 20 compared to the methoxy: compound **2** is the ligand in our dataset with the lowest k_{off}, to the point that only an upper limit has been measured. On the other hand, the relative computed residence time of the slowest compound **2** is underestimated as it is found equal to **1** within the estimated error. In the bound state, the crystal structure indicates that the morpholino group of **2** is located outside the binding pocket, with no protein residue closer than 3Å.

From Figure 7, the kinetically relevant unbinding event happens similarly early for the two compounds. Hence, snapshots of all unbinding trajectories between $d = 3$ and 3.5Å were extracted for further analysis. Clearly, for both ligands, the main event in these short extracts is the breaking of the hydrogen bond between the central NH of the ring and Thr184, while auxilliary polar interactions, e.g. with the two phenolic OH groups, are often maintained at this point of the unbinding process. However, it is notable that the morpholino group of ligand **2** interacts with residues from α-helix 3, much more than the smaller methoxy of ligand **1**. As the conformational space spanned by subset C in the snapshots is large for both compounds, there is no indication of a specific interaction between alpha-helix 3 and **2**. In order to better quantify the protein-ligand interactions taking place around the TS, we extracted from circa 20 unbinding trajectories of compounds **1** and **2** the part where d is between 3 and 3.5Å.

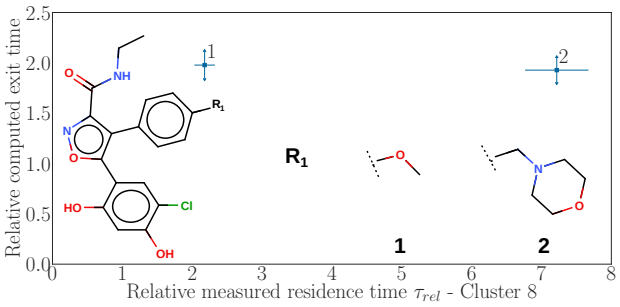


Figure 13: Cluster 8: τ_{comp} vs τ_{rel} for $\lambda = 0.40$ (Error bars show standard deviation).

For each of these trajectories we extracted 20 regularly spaced snapshots of the system and computed the mean of the cFP of those snapshots. This procedure was performed on ScaledMD and τRAMD trajectories from ref. 16 and the results are depicted in Figure 14. The mean cFP of compounds **1** and **2** simulated by ScaledMD, on one hand, and by RAMD, on the other hand, are very similar. However the cFP obtained for ScaledMD simulations and the one obtained for τRAMD simulations for the same compound are markedly different: whereas the ligand simulated with τRAMD interacts specifically with its surroundings, the ScaledMD fingerprint is blurred, reflecting the variety of the interactions occuring during the simulations. Moreover, a distinct signature differenciated the unbinding of **2** with respect to **1** using τRAMD, which was not the case with ScaledMD: **2** in its neutral or protonated form interacts specifically with Ser113 and Gln133 during unbinding, and also with Asp102 when in protonated form. This difference in unbinding patterns translates into the difference in computed residence times obtained with τRAMD, which correctly predicts a longer τ for **2** than for **1**. These observations lead to a plausible interpretation of the underestimation of the residence time by ScaledMD for our set of outliers based on a qualitative evaluation of $\zeta(\mathbf{2}, \lambda) - \zeta(\mathbf{1}, \lambda)$ in these simulations. The contributions to the activation entropy can be separated between those arising strictly from the protein and from the ligand degrees of freedom, and from contributions which are specific to the protein-ligand interactions, such as the restrictions of the conformational space due to a set of specific H-bonds or steric clashes. The degrees of freedom (DoF) specific to **1** and **2** are very similar, even if the morpholino group, as a non-aromatic ring, can have a DoF of its own. What differs more are the number and the relative population of the possible states in the complete space of the protein-ligand degrees of freedom. Indeed, while the interactions of **1** and **2** with the core of the binding site are very similar in the snapshots, the size of the conformational space spanned by residues of subset C in interactions with the morpholino group of **2** is large for $\lambda = 0.40$. It is expected to be much smaller when λ is increased, at least to the point where α-helix 3 keeps its secondary structure. The α-helix 3 behaviour would then be similar to what is observed in conventional or τRAMD simulations. More generally, we could speculate that in cases where, around the unbinding TS, the ligand interacts non-specifically with parts of the protein that are rendered unstructured by the scaling potential, there is a risk of underestimation of the residence time due to non-linearities in the entropy contribution. This hypothesis is testable, but it is beyond the scope of the present study.

Overall, several trends summarize our find-

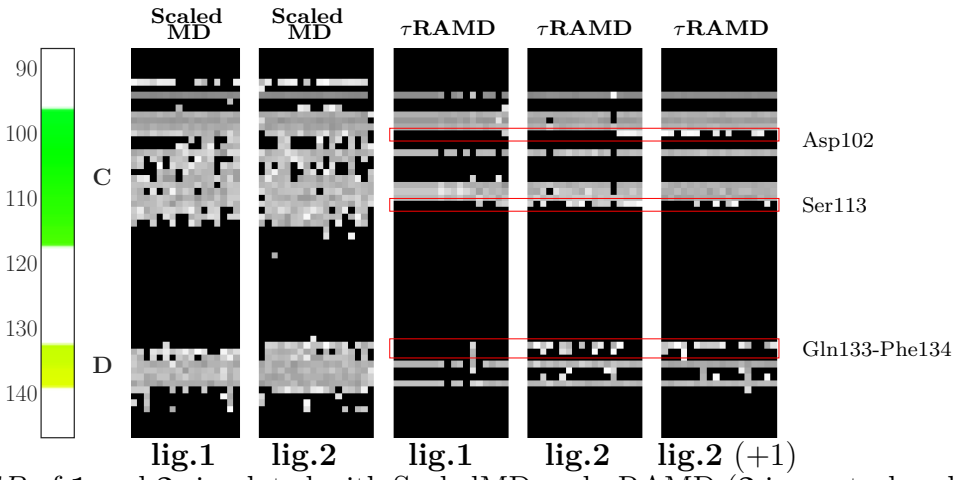


Figure 14: Mean cFP of **1** and **2** simulated with ScaledMD and τRAMD (**2** in neutral and charged states.) Residues which differentiate **1** and **2** unbinding in τRAMD simulations are highlighted.

ings in terms of residence time estimation. First, the compounds with short residence times were not ranked correctly even when in congeneric series, as seen in Clusters 1, 4 and 6. Second, in the largest chemical series displaying a wide range of τ (resorcinols from Cluster 2, indazoles from Cluster 3), the trends in predicted residence times are globally correct despite the presence of outliers. Third, in two cases of pairs of compounds with point modifications, the relative residence times are well reproduced while this is not the case for a third pair (Cluster 8) and compound **2**, whose residence time, the longest of the whole dataset, is underestimated in the ScaledMD simulations. The very long measured τ of ligand **2** and the underestimation of its residence time might be due to a strong interaction between the morpholino group and the α-helix 3 at the TS that is correctly described with τRAMD, but not with ScaledMD, where these interactions are blurred, at least with the simulation settings used here.

4 Conclusion

In this work, we have evaluated the ability of ScaledMD to qualitatively and quantitatively reproduce the residence times of a diverse and sizeable set of ligands binding to two different conformations of their protein target in a context reasonably close to a realistic drug discovery project. In order to describe the structural aspects of the unbinding events, we have developed simple procedures for (1) estimating the limits of the basin around the bound state using statistics on the first passage time $t_{fpt}(d)$, and (2) describing and comparing the paths which lead to unbinding using contact-map fingerprints. These procedures are not specific to ScaledMD simulations, they can be used to analyze unbinding trajectories generated with other protocols. For many ligands, the evolution of the squared correlation coefficients R^2 between the unbinding time t_{off} and the first passage time $t_{fpt}(d)$ can be used to locate the kinetically relevant TSs at 3.5 to 4.5Å from the ligand's binding shell residues. The clustering of the late unbinding parts of the trajectories from our simulations revealed that the two exit routes from the N-HSP90 binding site that were reported previously by several other groups[16,35,37] were indeed followed by most ligands. What is more, each ligand followed both routes in sizeable proportions, and both led to comparable mean exit times.

In order to estimate the quantitative character of the τ prediction by ScaledMD, we have related τ_{comp} to the measured normalized residence times relative to ligand **3** $\tau_{rel,lig3}(\lambda) = (t_{off,ligand}/t_{off,lig3})^{\lambda}$. We found that τ_{comp} allows the identification of the slowest unbinders of the set, and that it gives generally good trends within chemical series even when the ligands bind to different protein conformations. For a subset of the ligands which includes many of the slowest unbinders, the residence time is consistently underestimated in the simulations.

For some of them, this underestimation might be due to specific structural or force-field issues, as **9** and the quinazolines **23** and **24** are also underestimated by τRAMD. On the other hand, the apparent consistency of the underestimation and the fact that it was observed in other works using ScaledMD[35] required a proper investigation.

We have compared the structure of two homologous compounds, **1** whose τ is correctly predicted, and **2** whose τ is underestimated, around the unbinding TSs, and found a plausible explanation of this underestimation. The bias induced by the scaling of the interactions, together with the choice of residues to keep restrained during the simulations, increase the mobility of α-helix 3 up to a point where its secondary structure elements are no longer conserved, leading to an incorrect description of the interaction taking place and an artefactually low τ_{comp} for ligands that interact with unstructured residues (here α-helix 3) around the unbinding TS.

Our work is complementary to other recent studies on the unbinding kinetics of N-HSP90 ligands.[16,35,37] It confirms the main results found by Schuetz et al.[35] using ScaledMD on a smaller dataset, such as the identification of two main exit routes or the underestimation of the τ for the slowest binders. With respect to TMD,[37] ScaledMD residence time estimations seem to be less sensitive to the initial protein conformation. On the other hand, the scaling of all interactions in ScaledMD leads to a greater perturbation of the whole system compared to TMD or in τRAMD[16] where external forces are applied on the ligand. The perturbations due to the scaling makes relevant a detailed analysis of the unbinding trajectories generated by these techniques, for example when comparing possible exit routes. The quantitative performance we obtained in predicting τ using ScaledMD is not on par with τRAMD's, which also has the advantage of requiring less parameters, but it is still decent. Still, several directions might help in improving this performance: First, a careful and series-specific selection of the restrained residues is of great importance and actually recommended by Biki when setting up ScaledMD simulations. Nevertheless, using a broader set of unrestrained residues allowed the comparison of ligands across the series, which was one of the objectives of this work. Second, a higher scaling factor necessarily improves the description of the physics in ScaledMD, at the cost of much longer simulations. Reanalysis of the training dataset used for λ determination actually suggests that this selection might have been biased by the slowest unbinders whose τ_{comp} were underestimated for the reasons explained above. Third, the selective application of the scaling to only a subset of the system is a recent and appealing variation of all-atoms ScaledMD.[57] Selectively scaledMD is a simple restraint-free approach which might alleviate some of the difficulties raised above.

In a drug design perspective, the use of unbinding simulations might be of use at the lead optimization stage, when high-quality structural information is available to perform high-quality low-throughput comparisons within chemical series. It is also tempting to insert unbinding simulations at some point of a virtual screening with the objective of weeding out false positives. Our results suggest that ScaledMD might be useful in this respect as it allowed us to classify correctly most of the compounds with a long τ in our dataset using τ_{comp}. Among the approaches discussed above for the evaluation of protein-ligand binding kinetics, the ones that perform best and are amenable to medium to high throughput calculations will be a welcome addition to the Computer-Aided Drug Design practitioner's toolbox.

Acknowledgement The authors thank Giovanni Bottegoni, Sergio Decherchi and Andrea Spitaleri from Biki for their help in accessing and using Biki-Netics and for insightful discussions. This work was supported by EU/EFPIA Innovative Medicines Initiative (IMI) Joint Undertaking, K4DD (grant no. 115366). The authors acknowledge PRACE for awarding access to Marconi computational facilities based in Italy at Cineca (Project Pra12_3089). DK and RCW acknowledge the support of the Klaus Tschira Foundation and the funding from the European Union's Horizon 2020 Framework

Programme for Research and Innovation under Specific Grant Agreement No.945539 (Human Brain Project SGA3)

Supporting Information Available

- Views of the loop and helix conformations of N-HSP90
- Derivation of τ_{comp} as a function of $\zeta(l_{1/2}, \lambda)$
- Ligands considered: naming, structures, experimental data
- Determination of the optimal λ on the training set
- Effect of λ on $R^2(d, l)$ and $R^2_{COM}(d, l)$ on the training set
- Description of SKR in the structural clusters 4 to 6
- Ligands exit statistics
- Ligands exit trajectories fingerprints and detail of the clusters

References

(1) Tummino, P. J.; Copeland, R. A. Residence Time of Receptor−Ligand Complexes and Its Effect on Biological Function. *Biochemistry* **2008**, *47*, 5481–5492.

(2) Swinney, D. C. The role of binding kinetics in therapeutically useful drug action. *Current Opinion Drug Discov Devel* **2009**, *22 (1)*, 23–34.

(3) Tonge, P. J. Drug–Target Kinetics in Drug Discovery. *ACS Chemical Neuroscience* **2018**, *9*, 29–39.

(4) Ferruz, N.; De Fabritiis, G. Binding Kinetics in Drug Discovery. *Molecular Informatics* **2016**, *35*, 216–226.

(5) Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Wade, R. C. New approaches for computing ligand–receptor binding kinetics. *Current Opinion in Structural Biology* **2018**, *49*, 1–10.

(6) Bruce, N. J.; Ganotra, G. K.; Richter, S.; Wade, R. C. KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding. *Journal of Chemical Information and Modeling* **2019**, *59*, 3630–3634.

(7) Bernetti, M.; Masetti, M.; Rocchia, W.; Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annual Review of Physical Chemistry* **2019**, *70*, 143–171.

(8) Ribeiro, J. M. L.; Tsai, S.-T.; Pramanik, D.; Wang, Y.; Tiwary, P. Kinetics of Ligand–Protein Dissociation from All-Atom Simulations: Are We There Yet? *Biochemistry* **2019**, *58*, 156–165.

(9) Limongelli, V. Ligand binding free energy and kinetics calculation in 2020. *WIREs Computational Molecular Science* e1455.

(10) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Recent progress in molecular simulation methods for drug binding kinetics. 2020.

(11) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today* **2013**, *18*, 667 – 673.

(12) Mondal, J.; Friesner, R. A.; Berne, B. J. Role of Desolvation in Thermodynamics and Kinetics of Ligand Binding to a Kinase. *Journal of Chemical Theory and Computation* **2014**, *10*, 5696–5705.

(13) Schuetz, D. A.; Richter, L.; Amaral, M.; Grandits, M.; Grädler, U.; Musil, D.; Buchstaller, H.-P.; Eggenweiler, H.-M.; Frech, M.; Ecker, G. F. Ligand Desolvation Steers On-Rate and Impacts Drug Residence Time of Heat Shock Protein 90 (Hsp90) Inhibitors. *Journal of Medicinal Chemistry* **2018**, *61*, 4397–4411.

(14) Magarkar, A.; Schnapp, G.; Apel, A.-K.; Seeliger, D.; Tautermann, C. S. Enhancing Drug Residence Time by Shielding of Intra-Protein Hydrogen Bonds: A Case Study on CCR2 Antagonists. *ACS Medicinal Chemistry Letters* **2019**, *10*, 324–328.

(15) Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H.; Segala, E.; Amaral, M.; Guo, D.; Andres, D.; Georgi, V.; Stoddart, L. A.; Hill, S.; Cooke, R. M.; Graaf, C. D.; Leurs, R.; Frech, M.; Wade, R. C.; de Lange, E. C. M.; Ijzerman, A. P.; Müller-Fahrnow, A.; Ecker, G. F. Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug Discovery Today* **2017**, *22*, 896 – 911.

(16) Kokh, D. B.; Amaral, M.; Bomke, J.; Grädler, U.; Musil, D.; Buchstaller, H.-P.; Dreyer, M. K.; Frech, M.; Lowinski, M.; Vallée, F.; Bianciotto, M.; Rak, A.; Wade, R. C. Estimation of drug-target residence times by tau-random acceleration molecular dynamics simulations. *Journal of Chemical Theory and Computation* **2018**, *14*, 3859–3869.

(17) Kokh, D. B.; Kaufmann, T.; Kister, B.; Wade, R. C. Machine Learning Analysis of τRAMD Trajectories to Decipher Molecular Determinants of Drug-Target Residence Times. *Frontiers in Molecular Biosciences* **2019**, *6*, 36.

(18) Kokh, D. B.; Doser, B.; Richter, S.; Ormersbach, F.; Cheng, X.; Wade, R. C. A workflow for exploring ligand dissociation from a macromolecule: Efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *The Journal of Chemical Physics* **2020**, *153*, 125102.

(19) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Ligand unbinding mechanisms and kinetics for T4 lysozyme mutants from tauRAMD simulations. 2020.

(20) Berger, B.-T.; Amaral, M.; Kokh, D. B.; Nunes-Alves, A.; Musil, D.; Heinrich, T.; Schröder, M.; Neil, R.; Wang, J.; Navratilova, I.; Bomke, J.; Elkins, J. M.; Müller, S.; Frech, M.; Wade, R. C.; Knapp, S. Structure-kinetic relationship reveals the mechanism of selectivity of FAK inhibitors over PYK2. *Cell Chemical Biology* **2021**, 2451–9456.

(21) Ganotra, G. K.; Wade, R. C. Prediction of Drug–Target Binding Kinetics by Comparative Binding Energy Analysis. *ACS Medicinal Chemistry Letters* **2018**, *9*, 1134–1139.

(22) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(23) Callegari, D.; Lodola, A.; Pala, D.; Rivara, S.; Mor, M.; Rizzi, A.; Capelli, A. M. Metadynamics Simulations Distinguish Short- and Long-Residence-Time Inhibitors of Cyclin-Dependent Kinase 8. *Journal of Chemical Information and Modeling* **2017**, *57*, 159–169.

(24) 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* **2015**, *112*, E386–E391.

(25) Sun, H.; Li, Y.; Shen, M.; Li, D.; Kang, Y.; Hou, T. Characterizing Drug–Target Residence Time with Metadynamics: How To Achieve Dissociation Rate Efficiently without Losing Accuracy against Time-Consuming Approaches. *Journal of Chemical Information and Modeling* **2017**, *57*, 1895–1906.

(26) Capelli, R.; Bochicchio, A.; Piccini, G.; Casasnovas, R.; Carloni, P.; Parrinello, M. Chasing the Full Free Energy Landscape

of Neuroreceptor/Ligand Unbinding by Metadynamics Simulations. *Journal of Chemical Theory and Computation* **2019**, *15*, 3354–3361.

(27) Gobbo, D.; Piretti, V.; Di Martino, R. M. C.; Tripathi, S. K.; Giabbai, B.; Storici, P.; Demitri, N.; Girotto, S.; Decherchi, S.; Cavalli, A. Investigating Drug–Target Residence Time in Kinases through Enhanced Sampling Simulations. *Journal of Chemical Theory and Computation* **2019**, *15*, 4646–4659.

(28) Bernetti, M.; Masetti, M.; Recanatini, M.; Amaro, R. E.; Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *Journal of Chemical Theory and Computation* **2019**, *15*, 5689–5702.

(29) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications* **2019**, *10*, 3573.

(30) 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* **2019**, *15*, 743–750.

(31) Potterton, A.; Husseini, F. S.; Southey, M. W. Y.; Bodkin, M. J.; Heifetz, A.; Coveney, P. V.; Townsend-Nicholson, A. Ensemble-Based Steered Molecular Dynamics Predicts Relative Residence Time of A2A Receptor Binders. *Journal of Chemical Theory and Computation* **2019**, *15*, 3316–3330.

(32) Mollica, L.; Decherchi, S.; Zia, S. R.; Gaspari, R.; Cavalli, A.; Rocchia, W. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Scientific Reports* **2015**, *5*.

(33) Mollica, L.; Theret, I.; Antoine, M.; Perron-Sierra, F.; Charton, Y.; Fourquez, J.-M.; Wierzbicki, M.; Boutin, J. A.; Ferry, G.; Decherchi, S.; Bottegoni, G.; Ducrot, P.; Cavalli, A. Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein–Ligand Residence Times. *Journal of Medicinal Chemistry* **2016**, *59*, 7167–7176.

(34) Bernetti, M.; Rosini, E.; Mollica, L.; Masetti, M.; Pollegioni, L.; Recanatini, M.; Cavalli, A. Binding Residence Time through Scaled Molecular Dynamics: A Prospective Application to hDAAO Inhibitors. *Journal of Chemical Information and Modeling* **2018**, *58*, 2255–2265.

(35) Schuetz, D. A.; Bernetti, M.; Bertazzo, M.; Musil, D.; Eggenweiler, H.-M.; Recanatini, M.; Masetti, M.; Ecker, G. F.; Cavalli, A. Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics. *Journal of Chemical Information and Modeling* **2019**, *59*, 535–549.

(36) Decherchi, S.; Bottegoni, G.; Spitaleri, A.; Rocchia, W.; Cavalli, A. BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Chemical Information and Modeling* **2018**, *58*, 219–224.

(37) Wolf, S.; Amaral, M.; Lowinski, M.; Vallée, F.; Musil, D.; Güldenhaupt, J.; Dreyer, M. K.; Bomke, J.; Frech, M.; Schlitter, J.; Gerwert, K. Estimation of Protein–Ligand Unbinding Kinetics Using Non-Equilibrium Targeted Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2019**, *59*, 5135–5147.

(38) Landrum, G. RDKit: Open-source cheminformatics. 2006; `http://www.rdkit.org`.

(39) Krukenberg, K. A.; Street, T. O.; Lavery, L. A.; Agard, D. A. Conforma-

tional dynamics of the molecular chaperone Hsp90. *Quarterly Reviews of Biophysics* **2011**, *44*, 229–255.

(40) da Silva, A. W. S.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **2012**, *5*, 367.

(41) Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Richter, S.; Wade, R. C. KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding. https://kbbox.h-its.org/toolbox/, 2017.

(42) Bianciotto, M. Estimation of relative residence times of protein-ligand complexes using Scaled Molecular Dynamics. https://kbbox.h-its.org/toolbox/tutorials/estimation-of-relative-residence-times-of-protein-ligand-complexes-using-scaled-molecular-dynamics-scmd-bikinetics/, 2020.

(43) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *Journal of Computer-Aided Molecular Design* **2007**, *21*, 681–691.

(44) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design* **2010**, *24*, 591–604.

(45) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247–260.

(46) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* **1993**, *115*, 9620–9631.

(47) 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* **1993**, *97*, 10269–10280.

(48) Gordon, M. S.; Schmidt, M. W. *Theory and Applications of Computational Chemistry*; Elsevier, 2005; pp 1167–1189.

(49) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015**, *11*, 3696–3713.

(50) 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* **1983**, *79*, 926–935.

(51) Case, D.; Betz, R.; Botello-Smith, W.; Cerutti, D.; T.E. Cheatham, I.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D.; Roitberg, A.; Sagui, C.; Simmerling, C.; Botello-Smith, W.; Swails, J.; Walker, R.; Wang, J.; Wolf, R.; Wu, X.; Xiao, L.; Kollman, P. AMBER 2016. University of California: San Francisco, 2016.

(52) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E.

GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(53) Sinko, W.; Miao, Y.; de Oliveira, C. A. F.; McCammon, J. A. Population based reweighting of scaled molecular dynamics. *The Journal of Physical Chemistry B* **2013**, *117*, 12759–12768.

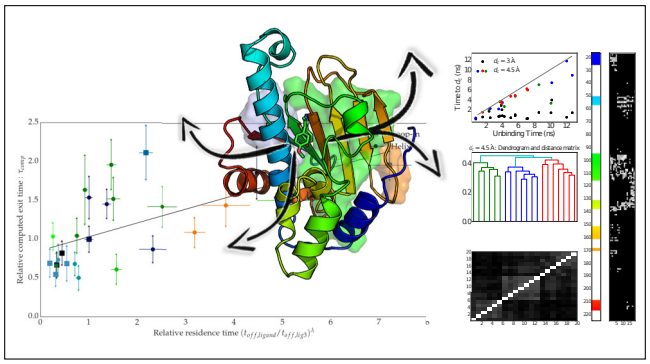
(54) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *Journal of Chemical Theory and Computation* **2018**, *14*, 6127–6138.

(55) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **2015**, *109*, 1528 – 1532.

(56) Swenson, D. W. Contact_map : Contact map analysis for biomolecules; based on MDTraj. https://contact-map.readthedocs.io/en/latest/, 2017.

(57) Deb, I.; Frank, A. T. Accelerating Rare Dissociative Processes in Biomolecules Using Selectively Scaled MD Simulations. *Journal of Chemical Theory and Computation* **2019**, *15*, 5817–5828.

TOC Graphic


Relative computed exit time : τ_{comp}
Relative residence time $(k_{off,ligand}/k_{off,high})^{\lambda}$
$d_c = 3$ Å
$d_c = 4.5$ Å
Time to d_c (ns)
Unbinding Time (ns)
$d_c = 4.5$ Å: Dendrogram and distance matrix