Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations

Luca Mollica[a] and Sergio Decherchi[b,c], Syeda Rehana Zia[b], Roberto Gaspari[b], Andrea Cavalli[a,d], Walter Rocchia[b]

[a]*CompuNet, Istituto Italiano di Tecnologia, via Morego, 30, I-16163 Genova, Italy*

[b]*CONCEPT Lab, CompuNet, Istituto Italiano di Tecnologia, via Morego, 30, I-16163 Genova, Italy*

[c]*BiKi Technologies s.r.l., Via XX Settembre, 33/10, I-16121 Genova, Italy*

[d]*Department of Pharmacy and Biotechnology, Alma Mater Studiorum, University of Bologna, via Belmeloro 6, I-40126 Bologna, Italy*

Supplementary Information

Supplementary Figures:

1. Distribution of residence times for HSP90 (1a), GRP78 (1b) and A2A (1c)

2. RMSDs example for HSP90

Supplementary table on the computing performance and execution time of the reported simulations

Supplementary note on the observed variance and the statistical analysis method of the residence times

Supplementary Methods

Supplementary References

Supplementary Movie 1
This movie shows the results of the bootstrap analysis for a variable number of replicas of the HSP90 series from 1 to 27

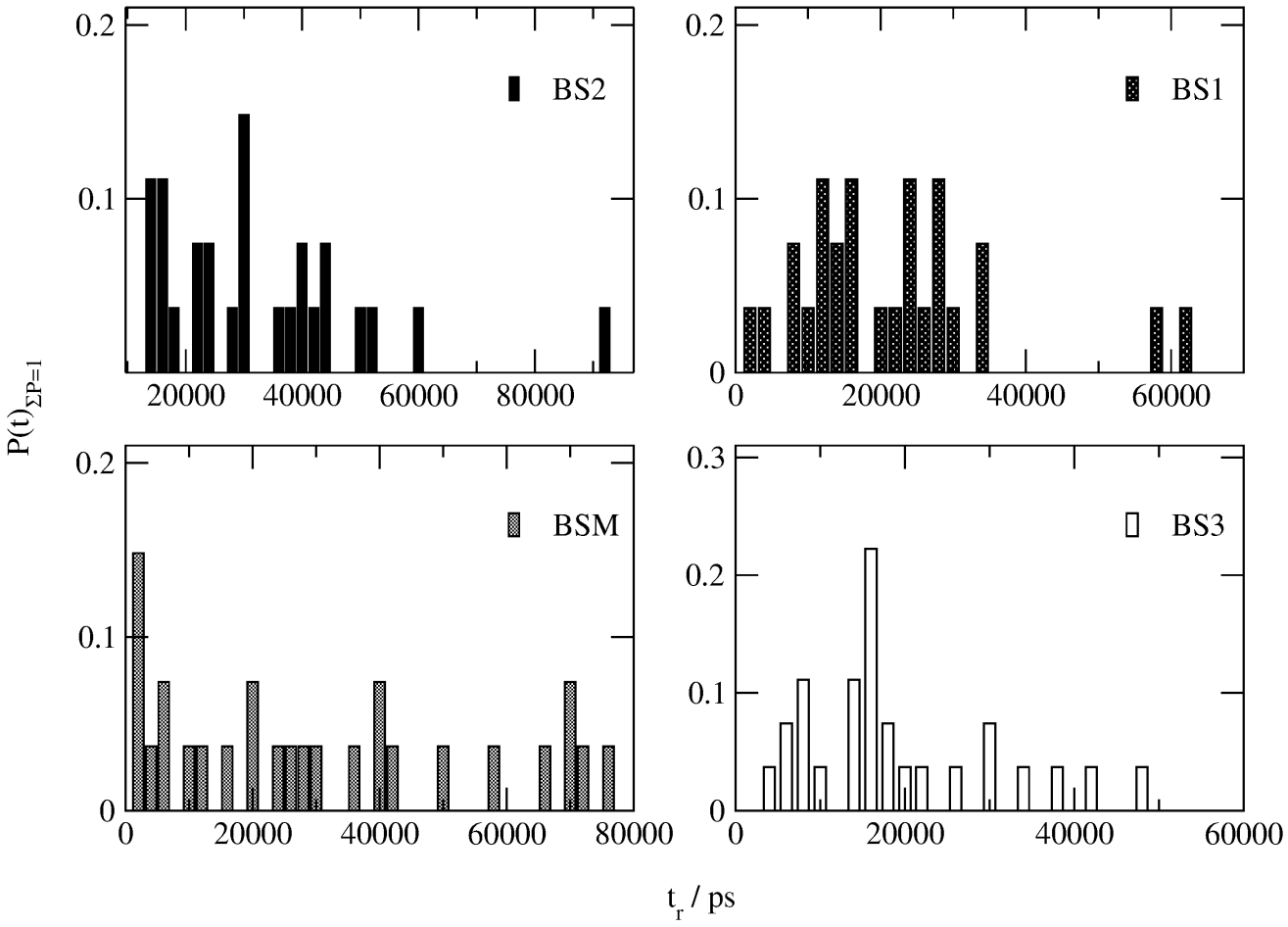
Supplementary Movie 2
This movie shows the results of the bootstrap analysis for a variable number of replicas of the Grp78 series from 1 to 20

Supplementary Movie 3
This movie shows the results of the bootstrap analysis for a variable number of replicas of the A_{2A} series from 1 to 20

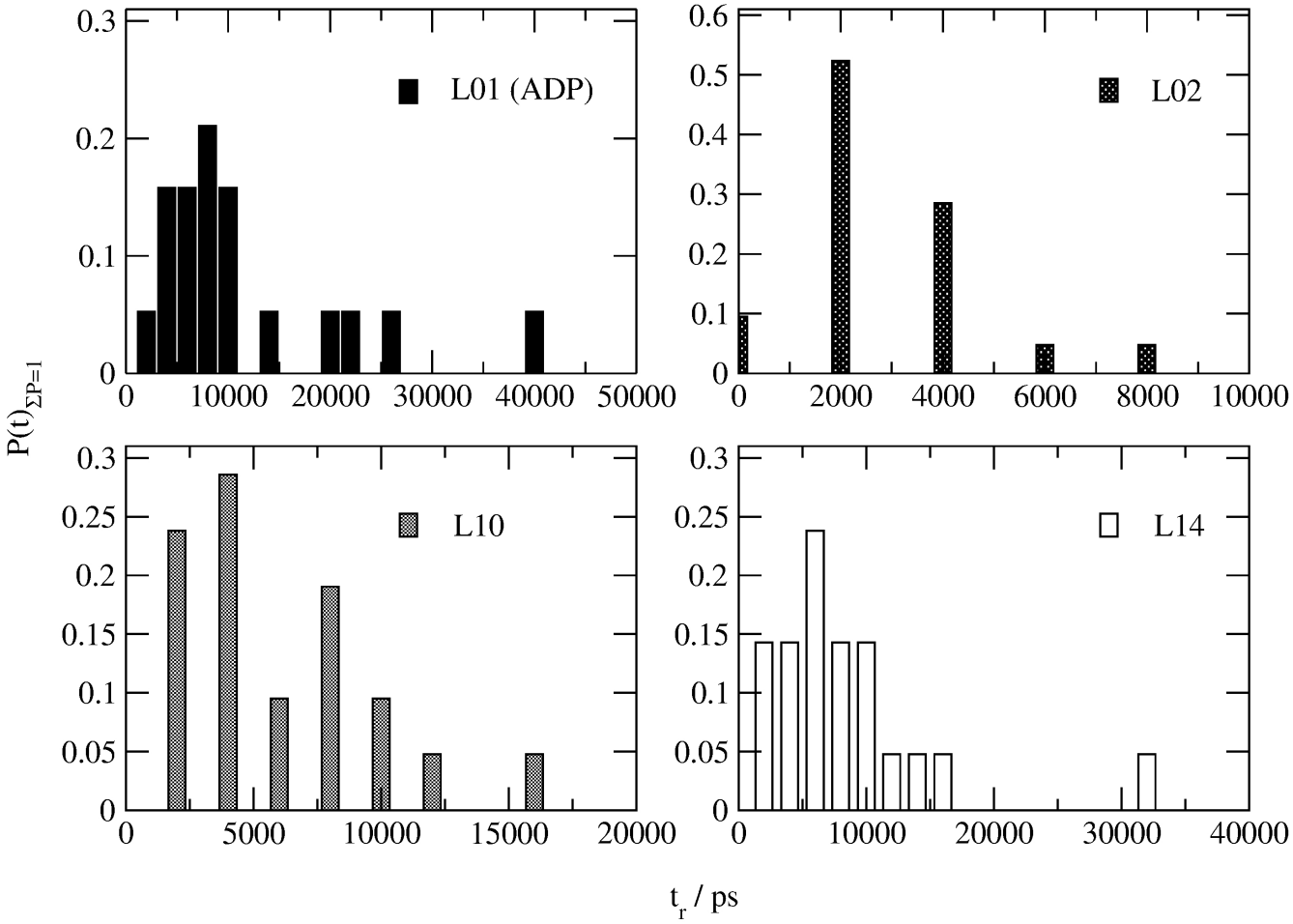
Supplementary Movie 4
This movie shows an example of the unbinding of BS1 from HSP90 (total simulated time: 34ns)

Supplementary Figure 1a: Distribution of residence times for HSP90.

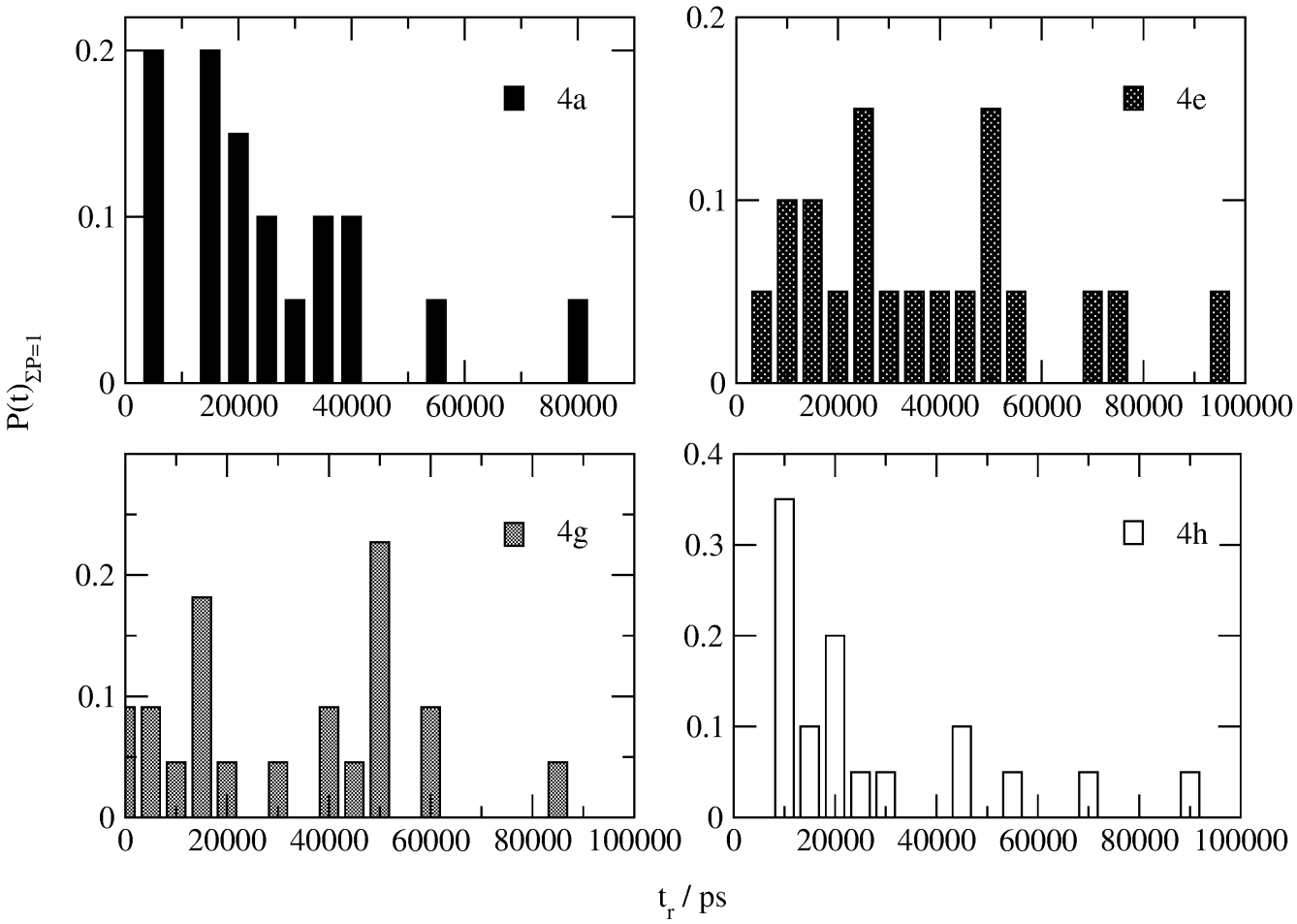
Residence time distributions of ligands in their binding pocket have been mapped after binning them with a width of 2000 ps: a total number of 108 simulations have been taken into account for HSP90[1] (1a.), 84 for Grp78[2] (1b.) and 80 for A_{2A}[3] (1c.). Residence times have been identified with occurrence times of unbinding events, i.e. configurations where no longer interactions between the ligand and the binding site are present. The distributions are reported in terms of normalized populations. More details are provided in the Supplementary Methods section.



Supplementary Figure 1b: Distribution of residence times for Grp78.

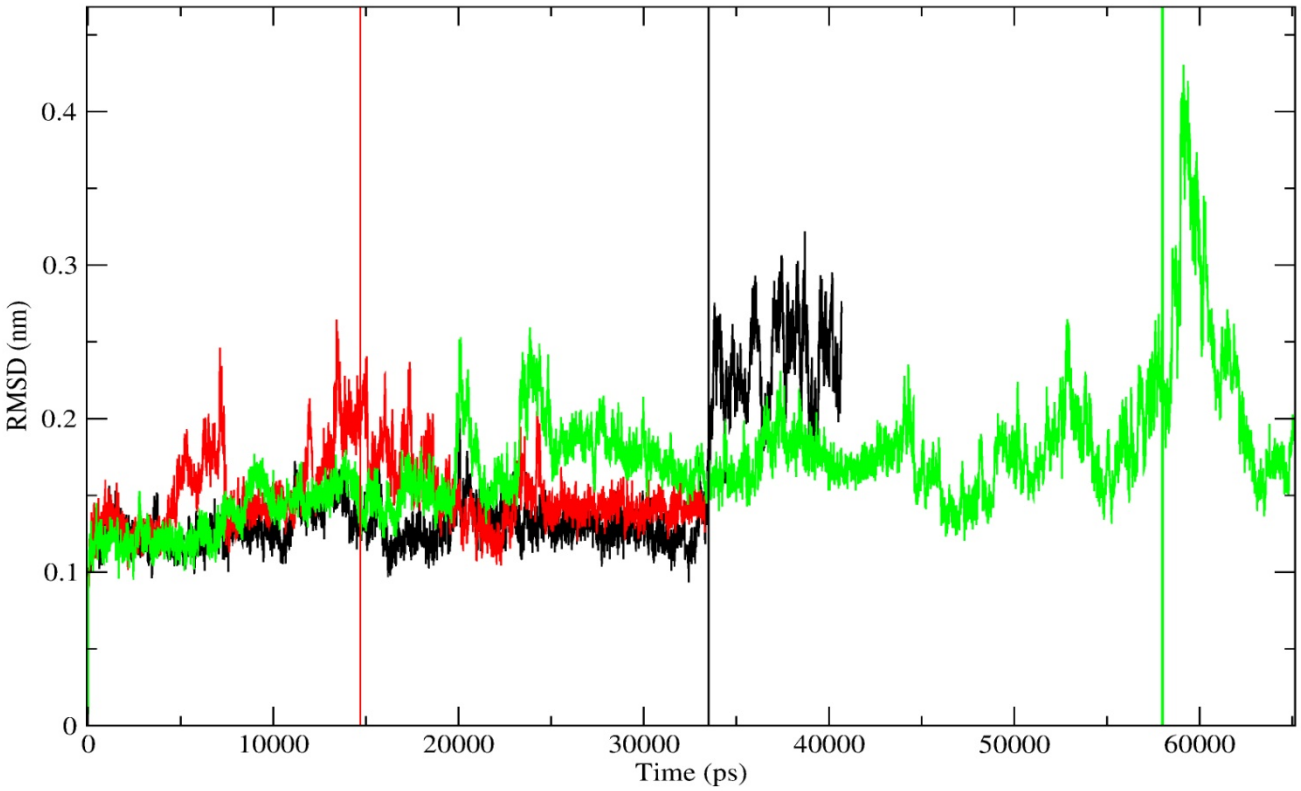


Supplementary Figure 1c: Distribution of residence times for A2A.



Supplementary Figure 2: Backbone RMSD examples.

Example of backbone RMSDs measured along the simulation (HSP90-BS2). The reported curves represent the three types of behavior observed in all the simulations, whereas the vertical lines correspond to the unbinding events. Behavior a) the protein keeps its RMSD stable and low, then suddenly the RMSD increases when the ligand leaves the site (black); b) the RMSD has significant fluctuations, which reduce down to a range between 1 and 2 Å after the ligand detaches (red); c) the RMSD fluctuates up to 2 Å but the ligand is not released for a long time, then the release happens with the RMSD increasing suddenly and eventually coming back rapidly to low values (green).



Supplementary table on the computing performance and execution time

For each system considered in the present work the (average) number of atoms has been reported, the calculation performance (expressed in ns/day) on a in house machine equipped with 2GPUs and 2 Intel esacore CPUs, the minimum and maximum time required for running a single simulation on each system.

System	Number of atoms	CPU performance	Min. time / simulation	Max. time / simulation
HSP90	33500	52 ns/day	4h	42h
Grp78	72200	27 ns/day	2h	35h
A2A	90000	22 ns/day	2h	70h

Supplementary note on the observed variance and the statistical analysis method of the residence times

Microscopic unbinding events with different underlying mechanisms, such as reactive paths, cannot be experimentally distinguished due to the average nature of macroscopic observables. Conversely, *in silico* MD runs inherently simulate individual events. Statistically speaking, the variance of the physical experiment is associated to a huge number of unbinding events that unavoidably mix the different mechanisms that might occur, whereas the variance of the computationally obtained residence time relates to a number of single unbinding events. The outcome of one physical experiment hence corresponds to the ensemble average of the observed quantity over many replicas of the same system and the experimental variance is the variance of a mean value. In contrast, the variance of the residence time coming from an MD simulation campaign is the variance of a microscopic quantity before any ensemble averaging process occurs. Moreover, irrespective of the individual mechanism followed, one is however considering a complex multifactorial process occurring at finite temperature; hence possibly significant fluctuations in microscopic observables have to be expected. Moreover, having lowered the energetic barriers between nearby local minima also induces an increase of the fluctuations amplitude. This explains why a direct comparison between computed (microscopic) and experimental (macroscopic) residence time variances is not straightforward and why microscopic standard deviation is not a direct measure of error and it is not expected to tend to zero as the number of MD runs increases.

In order to partially account for these aspects, we did a statistical treatment of the data (all reported in the Table 1), namely we computed the standard error of the mean and applied the bootstrap method to the estimated means.

The standard error of the mean quantifies the uncertainty characterizing the estimate of the mean residence time, which is the observable of our interest here. It takes the form: $\sigma_e = \sigma / \sqrt{N}$, where N is the number of collected samples and σ is the standard deviation of the simulated residence time.

The bootstrap method has been developed in the 80s[4] for inferentially assigning measures of accuracy to sample estimates: it is based upon random iterative drawings with replacement of the elements of a dataset with elements from the same dataset. In this case the samples are the unbinding times; practically one generates several virtual samples set (of the size equal to the original one) and from of each of them estimates the mean; upon those means estimation one can estimate the standard deviation of the mean itself.

As one can note in Table 1, the standard error of the mean and the deviation provided by the bootstrap method provide very similar figures, as expected.

Not only we performed the bootstrap analysis of the full dataset but also on artificially reduced samples sets to assess the change of the rankings with varying samples set size. We did this kind of analysis trying to estimate the probability of each outcomes (i.e. ranking) that could have been obtained having a different number of replicas for each system, in the 1 to 20 range. The results are summarized in the Supplementary Movies 1, 2 and 3, for HSP90, GRP78 and A2A, respectively. In the case of 20 replicas, the probability to get the experimental ranking for HSP90 was 39% and for GRP78 83%. For A_{2A} there are two possible rankings which prevail over the others, namely the experimental one, with a 29%, and the one obtained by swapping the first two ligands, with 25%, indicating that probably a larger number of replicas is needed if a better discrimination is desired.

Overall, those results confirm that this method is able to rapidly (small samples size) show which are the feasible rankings and to exclude most of them because their probability is less than the flat a priori probability (in this case its value is 1/24 and in the movies is represented by a red line).

In conclusion, the method seems to be effective on quite diverse bio-molecular systems, with a robustness that can be system dependent and require a different number of replicated simulations to achieve similar reproducibility values. The ranking probabilities are expected to converge with a rate proportional to the root square of the number of replicas as usual in this kind of estimation problems.

Supplementary Methods

Details of the scaled MD and the restraining procedure

Scaled MD has been described in several publications[5,6,7] and consists of the following modification with respect to a plain MD simulation:

- replacing the original potential energy function U(**x**), where **x** is the set of atomic positions, with a scaled version, λU(**x**), where λ is a positive constant between 0 and 1. The lower the λ value, the higher the acceleration but also the bigger is the loss of detail.

As described in the main text, a restraining is performed over the degrees of freedom of the system that do not play a major role to the unbinding process but preserve, for instance, the correct fold of the protein. The overall potential energy adopted V(**x**) is thus:

$$V(\mathbf{x}) = \lambda U(\mathbf{x}) + B(\mathbf{y})$$

where B(**y**) represents a set of positional harmonic restraints acting over the set { **y** } of the restrained atoms.

Mathematical derivation of the k_{off} and residence time scaling

Potential-scaled molecular dynamics (or, simply, scaled molecular dynamics, SMD) is a simulation technique introduced in 1993 by Hirono and coworkers[8] and recently reconsidered and implemented by McCammon and coworkers[9] that facilitates a more efficient sampling of molecular conformations.
According to Statistical Mechanics, the probability of any configuration is given by:

$$P(\bar{r}) = \frac{e^{-\beta V(\bar{r})}}{\sum_{r=1}^{n} e^{-\beta V(\bar{r})}} \qquad \text{(Eq. 1)}$$

where $\beta = (k_B T)^{-1}$, with k_B the Boltzmann constant and T the system absolute temperature. The evaluation of the partition function Z, i.e. the denominator of Eq.1, can hardly be performed analytically and also computationally it is rather costly due to the difficulty in estimating each microstate probability. In Molecular Dynamics, it is possible to estimate the non-normalized probability $p(\bar{r}) = e^{-\beta V(\bar{r})}$ of the microstates that can be easily accessed from the initial configuration. However, to accurately estimate the partition function Z, the probability of all the microstates should be estimated. This is practically impossible, since it would require infinitely long simulations, in order to fulfill the ergodic hypothesis. Among the many modifications that can be introduced in the MD protocols in order to facilitate the exploration of the configurational space there are those that attempt to flatten or smoothen the potential energy surface. For instance, SMD simply introduces a linear scaling of the system potential energy $V(\bar{r})$ that hence affects the potential energy surface, altering the probabilities of each microstate as follows:

$$p^*(\bar{r}) = e^{-\lambda\beta V(\bar{r})} \qquad \text{(Eq. 2)}$$

and resulting in an easy way of getting the canonical distribution by a suitable reweighting procedure:

$$p(\bar{r}) = p^*(\bar{r})^{1/\lambda} \ . \qquad \text{(Eq. 3)}$$

The applied scaling includes the potential that governs the interaction between the protein and the ligand. Assuming that the dissociation process obeys an Arrhenius-like relationship and that the entropic contributions of the binding processes is similar for all the ligands of the same series, a very simple preprocessing model can be used to infer actual residence times given one or more experimental reference figures in the dataset.
Let us consider the standard form of an unbinding rate following Arrhenius law:

$$k_{off} = Ae^{-E/RT} \qquad \text{(Eq. 4)}$$

with R the universal gas constant, T the system temperature and A an empirically derived pre-exponential factor that slightly depends on T and it is basically influenced by the collision frequency of the considered event. Considering the unbinding rates k_1 and k_2 of two ligands, namely 1 and 2, their relationship can be expressed as follows:

$$\frac{k_{off,1}}{k_{off,2}} = \frac{A_1 e^{-E1/RT}}{A_2 e^{-E2/RT}} \qquad \text{(Eq. 5)}$$

Considering protein-ligand systems in similar conditions (e.g., same collision rate) like the cases treated in the presented work, the terms presented in Eq. 4 can be approximated and simplified as follows:

$$\frac{k_{off,1}}{k_{off,2}} = \frac{A_1 e^{-E1/RT}}{A_2 e^{-E2/RT}} \approx \frac{Ae^{-E1/RT}}{Ae^{-E2/RT}} = \frac{e^{-E1/RT}}{e^{-E2/RT}} = e^{-\Delta E/RT} \qquad \text{(Eq. 6)}$$

If we consider the effect of scaling:

$$\frac{k_{off,1\lambda}}{k_{off,2\lambda}} = \frac{A_1 e^{-\lambda E1/RT}}{A_2 e^{-\lambda E2/RT}} \approx \frac{Ae^{-\lambda E1/RT}}{Ae^{-\lambda E2/RT}} = \frac{e^{-\lambda E1/RT}}{e^{-\lambda E2/RT}} = e^{-\lambda \Delta E/RT} \qquad \text{(Eq. 7)}$$

Hence, under our assumptions, the relationship between unscaled and scaled k_{off} can be written as follows:

$$\frac{k_{off,1\lambda}}{k_{off,2\lambda}} = e^{-\lambda \Delta E/RT} = (e^{-\Delta E/RT})^{\lambda} = \left(\frac{k_{off,1}}{k_{off,2}}\right)^{\lambda} \qquad \text{(Eq. 8)}$$

and, converting the rate constants to residence times:

$$\frac{t_{r,2\lambda}}{t_{r,1\lambda}} = \left(\frac{t_{r,2}}{t_{r,1}}\right)^{\lambda} \qquad \text{(Eq. 9)}$$

If experimental binding data are available for more than one sample in a series, a linear regression can be performed and this simplified model can be improved by incorporating the slope of the regression line.

Supplementary References

[1] Schmidtke, P., Luque, F.J., Murray, J.B., Barril, X., Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design, J. Am. Chem. Soc., 133(46), 18903-10 (2011).

[2] Macias, A.T., et al., Adenosine-derived inhibitors of 78 kDa glucose regulated protein (Grp78) ATPase: insights into isoform selectivity, J Med Chem, 54(12), 4034-41 (2011).

[3] Congreve, M., et al., Discovery of 1,2,4-triazine derivatives as adenosine A(2A) antagonists using structure based drug design, J Med Chem, 55(5), 1898-903 (2012).

[4] Efron, B., Bootstrap methods: Another look at the jackknife, Ann. Statist., 7, 1–26 (1979).

[5] Mark, A.E., Van Gunsteren, W.F. & Berendsen, H.J. Calculation of Relative Free-Energy via Indirect Pathways, *J. Chem. Phys.* **94**, 3808−3816 (1991).

[6] Tsujishita, H, Moriguchi, I. & Hirono, S. Potential-Scaled Molecular Dynamics and Potential Annealing: Effective Conformational Search Techniques for Biomolecules, *J. Phys. Chem.* **97**, 4416−4420 (1993).

[7] Sinko, W., Miao, Y., de Oliveira, C.A.F. & McCammon, J.A. Population Based Reweighting of Scaled Molecular Dynamics, *J. Phys. Chem. B*, **117**, 12759-12768 (2013).

[8] Tsujishita, H, Moriguchi, I. & Hirono, S. Potential-Scaled Molecular Dynamics and Potential Annealing: Effective Conformational Search Techniques for Biomolecules, *J. Phys. Chem.* **97**, 4416−4420 (1993).

[9] Sinko, W., Miao, Y., de Oliveira, C.A.F. & McCammon, J.A. Population Based Reweighting of Scaled Molecular Dynamics, *J. Phys. Chem. B*, **117**, 12759-12768 (2013).