Estimation of ligand binding free energy using multi-eGO.

Bruno Stegani[1,2], Emanuele Scalone[1,3], Fran Bačić Toplek[1], Thomas Löhr[4], Stefano Gianni[2], Michele Vendruscolo[4], Riccardo Capelli[1,*], and Carlo Camilloni[1,*]

[1]Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milan, Italy

[2]Dipartimento di Scienze Biochimiche "A. Rossi Fanelli", Istituto Pasteur-Fondazione Cenci Bolognetti and Istituto di Biologia e Patologia Molecolari del CNR, Sapienza Università di Roma, Rome, Italy

[3]Department of Chemistry, Dartmouth College, Hanover, New Hampshire 03755, United States

[4]Centre for Misfolding Diseases, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK

Corresponding authors: Riccardo Capelli riccardo.capelli@unimi.it, Carlo Camilloni: carlo.camilloni@unimi.it

KEYWORDS. Ligand binding, binding free energy, thermodynamic integration, molecular dynamics, lysozyme, c-Src kinase, aβ42, multi-eGO

ABSTRACT: The computational study of ligand binding to a target protein provides mechanistic insight into the molecular determinants of this process and can improve the success rate of *in silico* drug design. All-atom molecular dynamics (MD) simulations can be used to evaluate the binding free energy, typically by thermodynamic integration, and to probe binding mechanisms, including the description of protein conformational dynamics. The advantages of MD come at a high computational cost, which limits its use. Such cost could be reduced by using coarse-grained models, but their use is generally associated with an undesirable loss of resolution and accuracy. To address the trade-off between speed and accuracy of MD simulations, we describe the use of the recently introduced multi-eGO atomic model for the estimation of binding free energies. We illustrate this approach in the case of the binding of benzene to lysozyme by both thermodynamic integration and metadynamics, showing multiple binding/unbinding pathways of benzene. We then provide equally accurate results for the binding free energy of dasatinib and PP1 to Src kinase by thermodynamic integration. Finally, we show how we can describe the binding of the small molecule 10074-G5 to Aβ42 by single molecule simulations and by explicit titration of the ligand as a function of concentration. These results demonstrate that multi-eGO has the potential to significantly reduce the cost of accurate binding free energy calculations and can be used to develop and benchmark *in silico* ligand binding techniques.

INTRODUCTION

The study of ligand binding plays a pivotal role in the understanding of biochemical pathways and in drug design. Ligands, which can be small molecules, peptides, or other macromolecules, interact with specific targets, modulating their conformations, interactions and biological activities[1–3].

Characterizing a ligand-receptor interaction is a three-level problem (structure, thermodynamics and kinetics) that involves the characterization of the bound configuration (the pose), of the stability of the interaction (the binding free energy), and of the kinetics of the process (the on and off rate constants) including the detailed binding mechanism. Although a wide range of experimental and computational methods have been developed to address such aspects[4–6], they must be scaled up to reduce time and associated costs, in particular when used for drug design.

A common computational approach for drug design is based on the use of molecular docking methods[7–9], including recent machine learning approaches[10,11], to search and score the binding pose for large chemical libraries. The binding mechanism of top-scoring ligands can then be studied by a variety of methods[12–19], including endpoint approaches, such as the well-established thermodynamic integration (TI)[20,21], free energy perturbation (FEP)[22], and more recent single-simulation based[23] or non-equilibrium approaches[24]; pathway-dependent calculations, such as long time-scale MD simulations[25], Markov state models[26]; and reaction coordinate based techniques like umbrella sampling[27] and metadynamics (MetaD)[28]. The intensive computational cost of these methods can be reduced by lowering the resolution of the structural representation of the ligand-receptor system (i.e. by coarse graining), but this usually leads to reduced accuracy [29,30]. There is, therefore, a need to develop methods that achieve a more advantageous trade-off between speed and accuracy.

Recently, we developed multi-eGO to simulate protein self-assembly processes[31]. Multi-eGO simplifies the description of protein interactions by deriving effective Lennard-Jones

(LJ) parameters (C_6, C_{12} or ε, σ equivalently), by combining training conventional MD simulations of a system with prior models representing baseline interactions. We showed that multi-eGO provides accurate, atomic-level data on protein folding, intrinsically disordered proteins (IDPs) conformational dynamics and peptide aggregation, with highly reduced computational costs [32,33].

Here we extended the applicability of multi-eGO to ligand binding simulations (Figure 1). The training data consists of simulating a protein in its apo and holo states. The holo training is performed with additional free ligands in the box that allows simultaneous sampling of the bound pose and estimation of non-specific interactions (both protein-ligand and ligand-ligand). The apo training intramolecular contact probabilities are reweighted, following Bayes, by those obtained from a prior simulation, which represents a protein as a self-avoiding polymer. To reweight the protein-ligand and ligand-ligand contact probabilities, we introduce, as prior, a simulation performed at the same ligand concentration of the training data but with ligands and protein interacting only by hard-core repulsions (i.e. LJ C_{12} interactions). This prior simulation captures the rotational and translation entropy associated to the free ligand at the training set concentration. We expect this approach to have a twofold advantage: (i) the cost of training simulations, while not negligible, is often in part offset before free energy calculations are performed, to generate initial configurations and to test the stability of a binding pose; (ii) once the training cost is paid, the resulting multi-eGO model can be used to perform multiple simulations, using multiple free energy techniques as well as unbiased simulations, in a fraction of the time with respect to conventional force-fields (cf. Table 1).

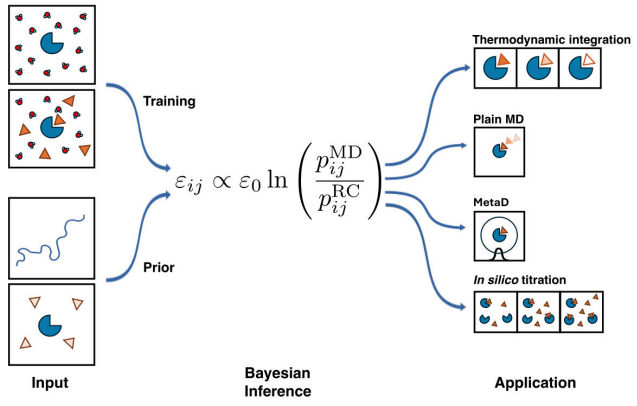

Figure 1. Schematic representation of the multi-eGO approach for ligand binding. Interatomic contact probabilities (P_{ij}^{MD}) from training simulations for the apo and holo states, including free ligands in the box, are reweighted, following Bayes, for prior models that represent a self-avoiding polymer for the protein and random rotations, translations and clashes of the ligands with the folded proteins (P_{ij}^{RC}). The resulting model can be fine-tuned (ε_0) and then employed to efficiently perform multiple simulation techniques, including TI, MetaD, plain MD and concentration dependent simulations.

To test the applicability of the multi-eGO approach, we applied it to four systems of increasing complexity. The first test case is the binding of benzene (BNZ) to the L99A mutant of the bacteriophage T4 lysozyme (LYZ)[34,35], which is widely used to benchmark computational methods of protein-ligand binding[36–39]. LYZ is a small enzyme (162 amino acids in the case of T4) that catalyzes the hydrolysis of 1,4-β-linkages of cell-wall peptidoglycans and consists of two domains (N-terminal and C-terminal). The L99A mutation produces a small apolar cavity in the C-terminal domain, which allows the binding of small nonpolar ligands such as BNZ with an unbinding free energy $\Delta G = 21.7 \pm 0.8$ kJ/mol[34]. The second and third systems regard the binding of dasatinib (DAS) and PP1 to the c-Src kinase (SRC). Dasatinib is a chronic myeloid leukemia drug that has nM affinity for several tyrosine kinases, including Src, whereas PP1 is a Src kinase inhibitor with a lower IC_{50} value[40–44]. These two molecules allow us to explore the effect of complex (in terms of internal degrees of freedom) ligands and strong binding free energies[25]. The fourth system follows the binding of the small molecule 10074-G5[45] to monomeric Aβ42, which is an intrinsically disordered protein (IDP) whose aggregation is associated with Alzheimer's diseases, allowing a binding process, dominated by conformational dynamics, to be tested [46,47].

Our results show that multi-eGO can quantitatively estimate the binding free energies of benzene, dasatinib and PP1, reproduce their hypothesized binding mechanisms and, importantly, allow a quantitative and systematic comparison of different simulation techniques. For Aβ42:10074-G5, we demonstrate that multi-eGO can describe extremely flexible systems, and single-molecule simulations and concentration-dependent simulations may be compared, i.e. models that closely mimic an *in vitro* experimental setup.

Based on these results, we propose multi-eGO as an *in-silico* platform for ligand binding studies for applications in method development, benchmarking, and integration into drug screening pipelines.

Table 1. Training simulations and performances

System	# replicates (Training)	Tot sampling time (Training)	Performances (conventional force-fields, ns/day) *	Performances (multi-eGO, ns/day) *
LYZ	1	1 µs	15.9	2,136
LYZ:BNZ	10	10 µs	15.0	2,128
SRC	1	1 µs	12.4	621
SRC:DAS	4	8 µs	11.1	580
SRC:PP1	4	8 µs	12.3	614

* Performances were calculated with GROMACS 2022 using 4 OpenMP threads on an i7-9700K CPU @ 3.60GHz.

METHODS

Conventional molecular dynamics simulations. All-atom, explicit solvent, training simulations of LYZ:BNZ, SRC:DAS and SRC:PP1, were performed using the DES-Amber force field[48] in TIP4P-D water[49], with ligands parameterized using GAFF2. To match the scaled electrostatic interactions of

DES-Amber force field, we rescaled the AM1-BCC-derived charges for the ligands by a factor of 0.9.

All systems were prepared at pH 7.4 and in the presence of Na^+ and Cl^- ions to maintain physiological salinity (150mM) and to neutralize the total charge of the system. All systems were subjected to energy minimization using the steepest descent algorithm until the maximum force converges to a value < 1000 kJ/(mol nm), followed by a conjugate-gradient minimization until the maximum force converges to a value < 100 kJ/(mol nm) and a LBFGS algorithm until the maximum force converges to a value < 10 kJ/(mol nm). Subsequently, the minimized configuration was relaxed for 2 ns at a constant pressure of 1 bar and constant temperature, keeping the protein atoms restrained to their minimum energy configurations. The simulations used the leap-frog algorithm with a time step of 2 fs and LINCS restraints[50] for hydrogen atoms. The cut off for non-bonded interactions was 1 nm, using PME for long-range electrostatics. Temperature and pressure were controlled by stochastic velocity rescaling[51] and cell rescaling[52] algorithms, respectively. The number of replicates and simulation times are reported in Table 1. All simulations were run using GROMACS 2022[53].

LYZ was modelled starting from the PDB code 1L84 structure[54] and simulated at 300 K. Training the LYZ:BNZ complex was performed by inserting one BNZ in the binding pocket, and 4 BNZ outside, in order to explore its interaction with the protein surface.

SRC was modelled using the PDB code 1Y57 structure[55] and simulated at 310 K. To train the holo complex of SRC:DAS/PP1, 1 molecule (either DAS or PP1) was inserted in the binding pose and 4 ligands were added to the solvent to explore the interaction with the protein surface. To train our multi-eGO model, the holo training trajectories were removed from the non-equilibrium portions, consisting of irreversible aggregation of three or more ligands with each other, as previously observed[25].

MD simulations of apo and holo Aβ42:10074-G5 were carried out as previously described[46,56]. Briefly, the apo simulation was carried out at 278 K using the CHARMM22* forcefield[57] and the TIP3P water model[58]. The holo simulation was performed under the same conditions, parameterizing the ligand 10074-G5 using the Force Field Toolkit[59].

Multi-eGO Simulations. Multi-eGO is a multi-ensemble hybrid transferable/non-transferable force field[31]. The functional form is shared with conventional molecular mechanics force-fields except for using only LJ for non-bonded interactions. Bonded interactions are transferable, derived from the GROMOS force-field, apart for dihedrals that have been newly derived. Default C_{12} LJ parameters were also modified starting from the GROMOS one[31,32]. The non-transferable (structure-based) part of the force field is obtained from state-of-the-art simulation(s) of the system of interest, sampling the fluctuations of specific states (i.e. monomer, apo, fibril, etc.), and is then weighted with a prior (or reference) simulation, obtained with a simplified description of the system. From the training and the reference simulations, we extract pair-wise distances and contact probabilities P_{ij}^{MD} and P_{ij}^{RC}, respectively. Then, by Bayesian reweighting, an estimation of the atom-pairs non-bonded interaction energy as:

$$\varepsilon_{ij} = -\frac{\varepsilon_0}{\ln P_{threshold}^{RC}} \cdot \ln \frac{P_{ij}^{MD}}{\max\left(P_{ij}^{RC}, P_{threshold}^{RC}\right)},$$

is estimated, where *i,j* are atom indices, ε_0 is the energy scale parameter, and $P_{threshold}^{RC}$ is the minimum probability used for regularization. A detailed description of the model can be found in ref.[32] and the associated code and parameters are available on GitHub.

For each apo system, we performed a prior simulation consisting of a self-avoiding chain, obtained with bonded interactions and C_{12} repulsion from the multi-eGO transferable potential. For multi-domain proteins like LYZ, a second prior simulation was introduced. This consists of a multi-eGO simulation in which the inter-domain attractions, namely between residues 1-71 and 72-162 are turned off, while preserving intra-domain and local interactions, in the alpha-helix connecting the two domains, to maintain its integrity. The two sets of prior probabilities allowed inter- and intra-domain interactions to be correctly decoupled and to assign them specific energy scale (ε_0). For intermolecular interactions, the reference prior simulation corresponds to protein:ligand complex at the same concentration of the training simulation (considering only unbound ligands) with only hard-core repulsions between protein and ligands. This allows the concentration-dependent, roto-translational entropy of the ligands to be accounted for by removing the effect of box size on the contact probability estimation. Multi-eGO ligands were parametrized starting from the training topology, removing hydrogens and optimizing the bonded parameters to reproduce the local geometries of the training simulations.

All multi-eGO MD simulations were performed using stochastic dynamics integration at a 5 fs timestep and 25 ps relaxation time. The cutoff for the LJ interactions was set to $2.5\sigma_{max}$, corresponding to 1.44 nm. A 10% larger radius was used for the neighbor lists, which were updated every 20 steps. Different values of ε_0 were tested to maximize the overall agreement between training and multi-eGO simulations. All simulations were run using GROMACS 2022[53].

Thermodynamic Integration. For the multi-eGO unbinding free energy estimation, we used TI and the Bennett's accepting ratio (BAR) algorithm[20,21]. For each system a set three-cross linked restraints, namely CL1, CL2, and CL3, to keep the ligand in the binding pose during the decoupling. The effect of this restraining potential can be removed analytically a posteriori[39]. The thermodynamic cycle was computed through a set of λ values scaling the restraint and non-bonded interactions (Table S1 to S3). For LYZ:BNZ, a steepest descent energy minimization was performed at each λ, followed by a 500 ps relaxation and a 1 ns run at 300K in the NVT ensemble. For SRC:DAS and SRC:PP1, a steepest descent energy minimization was performed at each λ, followed by a 2 ns relaxation and a 1 ns run at 310K in the NVT ensemble. All simulations were run using GROMACS 2022[53].

Volume-based Metadynamics. VMetaD[38] simulations LYZ:BNZ were run using well-tempered metadynamics[60] considering 3 CVs, namely the relative position in spherical coordinates (ρ, θ, φ) of the benzene center of mass, with respect to the center of mass of the lysozyme binding domain (residues 72-162). We set an initial height of the Gaussians of 1.2 kJ/mol, the widths at 1 Å, π/16 rad, and π/8 rad for ρ, θ, and φ, respectively and a bias factor of 20, depositing a

Gaussian every 10 ps. 4 different 1.5 μs-long replicas of the simulations were run, employing the same MD parameters used in the unbiased multi-eGO simulations.

To avoid an extensive sampling time of the unbound state, a repulsive spherical potential was set at a distance of ρ_{sph} (here set at 30 Å) from the center of mass of LYZ:

$$U_{sph}(\rho) = \begin{cases} 0 \text{ if } \rho \leq \rho_{sph} \\ \frac{1}{2}k(\rho - \rho_{sph})^2 \text{ if } \rho > \rho_{sph} \end{cases}$$

with k set to 10000 kJ/mol/Å^2. To consider the loss of entropy caused by the imposition of such restraints and to calculate the binding free energy difference in standard conditions, we apply the following entropic correction

$$\Delta G^0 = \Delta G_{MetaD} - RT \log\left(\frac{V^0}{\frac{4}{3}\pi\rho_{sph}^3 - V_{prot}}\right),$$

where ΔG_{MetaD} is the binding free energy of VMetaD, R is the gas constant, T is the temperature of the system, V^0 is the standard volume (1660 Å^3), and V_{prot} is the volume of the protein included in the sphere restraint. Protein volume inside the sphere was calculated using the double cubic lattice method[61] available in GROMACS. To compute the ΔG_{MetaD}, we reweighted the computed free energy surface using the Tiwary-Parrinello algorithm[62], removing the initial 200 ns of trajectory on CVs that allow the precise definition of the bound and the unbound states. Following ref.[38] the distance ρ and the coordination number c, between the set of the ligand atoms A and the set of the protein atoms B, were selected as:

$$c = \sum_{i \in A}\sum_{j \in B} \frac{1 - \left(\frac{r_{ij}}{r_0}\right)^6}{1 - \left(\frac{r_{ij}}{r_0}\right)^{12}}.$$

The error estimation was performed using the standard deviation of the mean of the four replicas. All simulations were run using GROMACS 2022[53] and PLUMED2[63,64].

***In Silico* Titration Experiments.** Six 5 μs long Aβ42: 10074-G5 MD multi-eGO simulations were performed, placing 10 monomers of Aβ42 in a 255 nm cubic box, corresponding to a concentration of 1 μM Aβ42, and adding 10, 60, 100, 200, 400 or 800 10074-G5 at random, respectively. Simulations were run using GROMACS 2022[53].

RESULTS

Preparation of a multi-eGO model for lysozyme:benzene: all-atom training. To train a multi-eGO model for LYZ:BNZ binding, we performed multiple apo and holo simulations using the DES-Amber[48] and the GAFF2 force field (see Methods and Table 1). With respect to the apo state, the simulations reproduced a breathing motion of the two domains over the microsecond time scale that is well characterized experimentally by residual dipolar couplings by nuclear magnetic resonance (NMR) spectroscopy[65] (Fig. 2A), as shown by the distribution of the root mean square deviation of the atomic positions (RMSD) calculated with respect to the backbone of a reference structure (PDB 1L84). The simulations of the holo state were performed with one BNZ molecule in the binding cavity and four additional molecules in solution. We found that BNZ is more likely to be found in regions surrounding the binding cavity, suggesting a role of surface interactions in directing BNZ to its binding site (Fig. 2B).

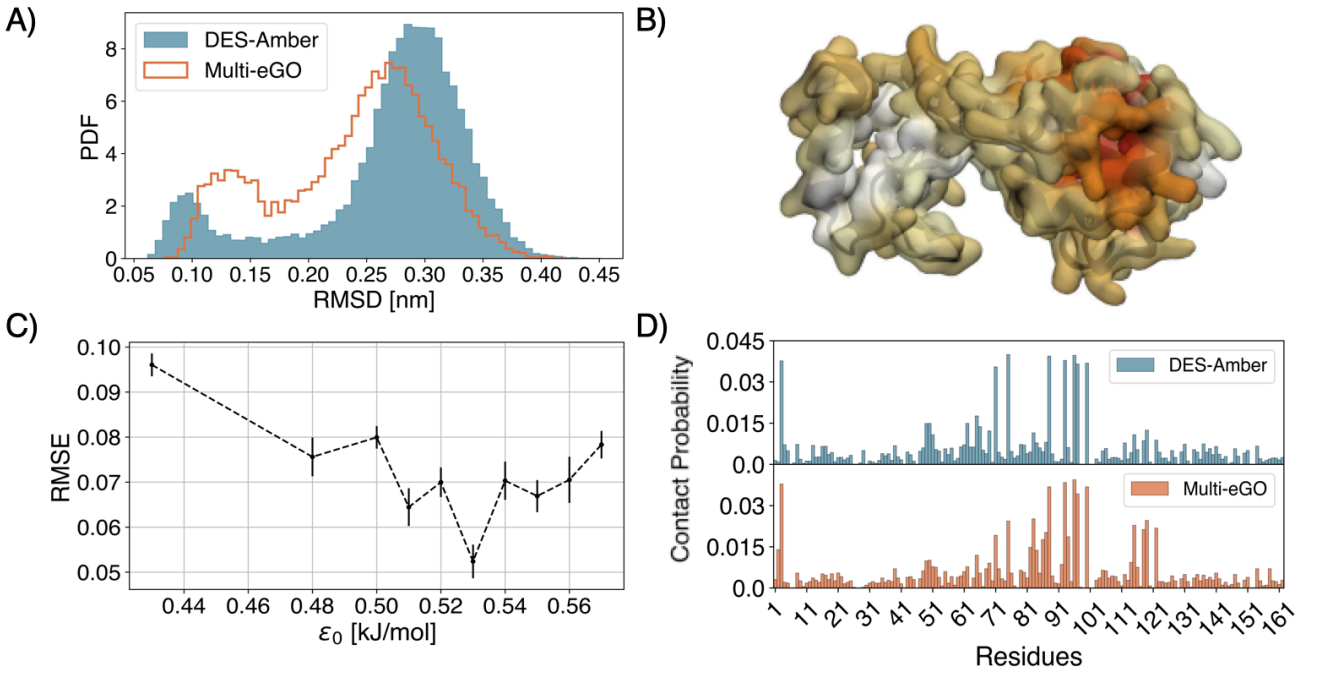

Figure 2. (A) Probability density function (PDF) of the RMSD from the reference crystal structure (PDB 1L84) for the ensemble of structures obtained from the LYZ DES-Amber (training) and multi-eGO simulations. These RMSD distributions illustrate that multi-eGO effectively reproduces the two-state distribution observed in LYZ. (B) Representation of the LYZ:BNZ interaction obtained from training simulations.

Increased red coloring indicates stronger attractions, highlighting the binding pocket in the C-terminal domain and various surface interactions of the protein, some of which correspond to binding entry pathways. (C) Scoring function used to calibrate the energy scale ε_0 of the multi-eGO model for intermolecular interactions. The RMSE was computed by comparing LYZ:BNZ contact probabilities between training and multi-eGO simulations. Error bars were derived by averaging the RMSE over different segments of multi-eGO trajectories. (D) Residue-wise intermolecular contact probabilities from training simulations versus multi-eGO simulations, optimized with an ε_0 of 0.53 kJ/mol.

The apo training simulation described above was analyzed to obtain intramolecular interatomic contact probabilities, which were reweighted using two priors.

Preparation of a multi-eGO model for lysozyme:benzene: prior calculation. First, a reference prior simulation was run, representing the contact probability of a self-avoiding polymer with the same sequence of LYZ. Then, an inter-domain reference was used to decouple the intra-domain and inter-domain multi-eGO energy scales and correctly re-weight the inter-domain training probabilities. The inter-domain prior consisted of a multi-eGO simulation in which all the inter-domain interactions between the N-terminal domain (residues 1-71) and the C-terminal domain (residues 72-162) were described as hard-core repulsions. This procedure allowed an informative prior of the random interdomain collisions to be obtained (see Methods).

Preparation of a multi-eGO model for lysozyme:benzene: final model. The resulting multi-eGO model for LYZ, with an $\varepsilon_0^{intradomain}$ of 0.34 kJ/mol and an $\varepsilon_0^{interdomain}$ of 3 kJ/mol, reproduced the equilibrium between a closed and an open conformation, (see Fig. 2A and Fig. S1). Focusing on the LYZ:BNZ interactions, the intermolecular LYZ:BNZ and BNZ:BNZ contact probabilities were reweighted using an intermolecular reference simulation, consisting of intermolecular (hard-core) clashes between LYZ and BNZ molecules at the same BNZ concentration, used in the training simulation to account for the roto-translational entropy of ligands. The $\varepsilon_0^{intermolecular}$ was obtained by minimizing the root mean square error (RMSE) between inter-molecular residue contact maps as shown in Fig 2C, resulting in an $\varepsilon_0^{intermolecular}$ of 0.53 kJ/mol. The resulting multi-eGO model reproduced the intermolecular contact probabilities between LYZ and BNZ (Fig. 2D), demonstrating that multi-eGO can effectively learn heterogeneous interactions.

As a first test, we performed 40 unbiased binding simulations, starting from random configurations of 4 unbound BNZ molecules at the same concentration as the training simulation (see SI Movie 1 for a representative trajectory). Simulations were stopped when the minimum distance between BNZ atoms and the Cα atom of A99 in the binding pocket fell below 0.4 nm. The extracted binding times were then fitted to a Poisson distribution, yielding a mean binding time τ of 963±20 ns with a p-value of 0.988 obtained from a Kolmogorov-Smirnov (KS) test[66], corresponding to a k_{on} of 7.6×10^7 M^{-1} s^{-1} (see Fig. S2). This rate should be considered as nominal due to the simplified nature of the model, but when compared with the experimental value[67] of 8×10^5 – 10^6 $M^{-1}s^{-1}$ it gives an indication of the intrinsic acceleration of multi-eGO due to removal of degrees of freedom.

Thermodynamic integration and volume-based metadynamics on lysozyme:benzene multi-eGO model. Given the efficiency of multi-eGO, its accuracy in estimating the LYZ:BNZ binding free energy was tested, in comparison with alternative methods. We focused on Thermodynamic Integration (TI) and Volume Based Metadynamics (VMetaD), the former being the industry standard for such calculations[20,21] and the latter, a potentially more informative but also more computationally expensive alternative[38].

TI of LYZ:BNZ was performed in 6 replicates using three alternative restraints (CL1, CL2, and CL3) to keep the ligand in the binding pose and to speed up convergence[39] (see Methods and Fig. S3). The unbinding free energies, calculated as the mean and the standard deviation of the mean, are shown in Fig. 3A. The three estimates of 21.41±0.04 kJ/mol, 21.45±0.05 kJ/mol, and 21.46±0.07 kJ/mol from the three restraints, respectively, show a high statistical precision and a remarkable accuracy when compared to the experimental unbinding free energy of 21.7±0.8 kJ/mol (Fig. 3A).

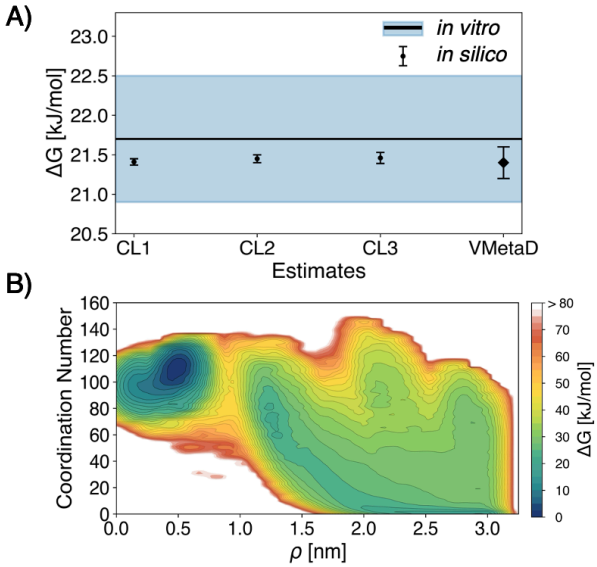

Figure 3. (A) Estimation of the unbinding ΔG obtained with the multi-eGO force field using different techniques: CL1, CL2, and CL3 refer to TI with three different restraints. The error was calculated as the standard error of the mean of six replicates for each restraint. VMetaD denotes the volume-based metadynamics ΔG estimate, with the error calculated as the standard error of the mean of four replicates. (B) Free energy surface (FES) obtained by reweighting of VMetaD in an appropriate collective variable (CV) space. On the x-axis, ρ represents the distance of the BNZ center of mass from the center of the sphere used to define the sampling volume. On the y-axis, the coordination number of BNZ with the residues of LYZ.

An additional unbinding free energy estimate was obtained using VMetaD (see Methods) by running four independent replicates of 1.5 μs each, reaching convergence after approximately 200 ns (Fig. S4). A free energy surface (FES) representing the unbinding process was obtained by reweighting VMetaD simulations as a function of the distance of the BNZ center of mass from the center of the sphere centered on the binding site, which was used to define the sample volume, and the coordination number of BNZ with the

residues of LYZ (see Fig. 3B and Methods section). From the FES, we estimated a ΔG of 21.4±0.2 kJ/mol, obtained as the mean and standard deviation of the mean over the 4 replicates (Fig. 3A), which is within the statistical precision of the TI estimates, demonstrating the reliability of the two different approaches.

The use of VMetaD also allowed us to compare our sampled binding/unbinding pathways with those already reported in the literature (Fig. 4 and Fig. S5). Using a dynamic time-warping clustering algorithm[68], we were able to observe most, if not all (depending on the classification approach), of the previously observed pathways[38,68], further supporting the ability of multi-eGO to describe correctly the binding and unbinding mechanism of a ligand.

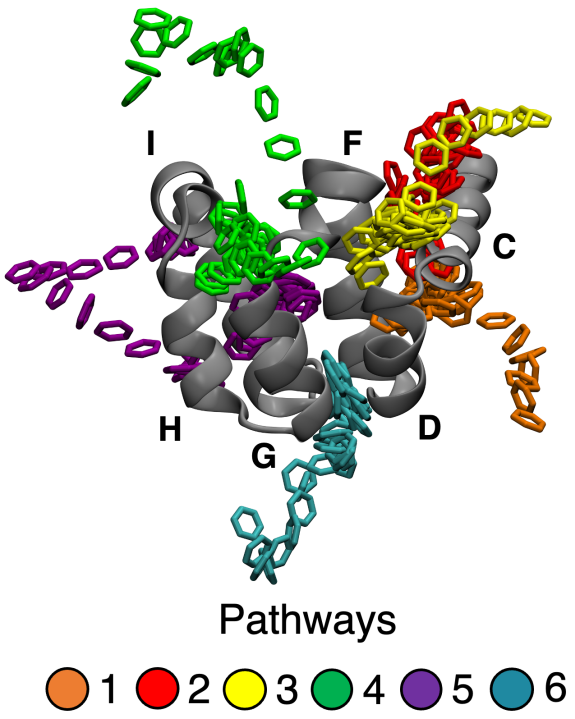

Figure 4. Representation of the main unbinding pathways identified by the dynamic time-warping analysis on the LYZ:BNZ VMetaD simulations. Naming the visible helices with the letters C, D, F, G, H, and I we can define all the six main unbinding pathways. 1) BNZ passes behind D and under C; 2) BNZ passes over C and behind F; 3) BNZ passes over C and under F; 4) BNZ passes in between the top part of H and G; 5) BNZ passes behind H; 6) BNZ passes under the C-terminal domain, between helices G and D. These pathways qualitatively match the ones identified with explicit water and atomistic potentials in ref. [38].

Multi-eGO model of c-Src kinase:dasatininb and c-Src kinase:PP1: all-atom training simulations. To test the transferability of the protocol developed for LYZ:BNZ, we reapplied it to the binding of DAS and PP1 to SRC. First, we performed several training MD simulations for the apo system as well as for the SRC:DAS and SRC:PP1 holo systems, in both cases 4 in-solution ligands were added (see Methods and Table 1). The training simulations showed different interactions between ligands and the protein in correspondence of known binding sites, as previous shown in ref.[25]. Besides the ATP binding site, we found important interactions with the N-lobe (β-sheet), the αG helix, the P-loop, binding in the MYR site for both ligands and binding in the PIF site for PP1 (see Fig. S6).

Multi-eGO model of c-Src kinase:dasatininb and c-Src kinase:PP1: prior model simulations. To parameterize the multi-eGO models, after performing a prior simulation of SRC as a self-avoiding polymer, we set the ε_0 parameter for the protein to match the radius of gyration (Rg), the residue-wise root-mean square fluctuations of the atomic positions (RMSF), and RMSD distributions, obtaining an optimal value of 0.33 kJ/mol (Fig. S7). To describe intermolecular interactions, we first performed an intermolecular reference prior simulation at the same ligand concentration as the training simulation to account for random collisions between ligand molecules. We then optimized the corresponding ε_0 and obtained a value of 0.43 kJ/mol and 0.6 kJ/mol for DAS:DAS and PP1:PP1, respectively, which could reproduce their training contact probabilities. A second intermolecular prior simulation was then performed in which the ligands were allowed to interact with each other, while the protein and ligands only interacted with hard-core repulsion. Finally, by minimizing the RMSE between the training and multi-eGO residue-wise intermolecular contact probabilities, we obtained an optimal ε_0 of 0.32 kJ/mol and 0.35 kJ/mol for SRC:DAS and SRC:PP1, respectively (Fig. S8).

Binding simulation of c-Src kinase:dasatininb and c-Src kinase:PP1: multi-eGO models. The resulting multi-eGO models were used to perform 40 binding simulations for both PP1 and DAS, starting from a random conformation of four unbound ligands (at the same concentration as in training). To establish the binding, we chose three reference atoms of the protein, namely CD:GLU310, N:MET341, CA:THR338 for SRC:DAS and N:ILE294, N:MET341, N:THR338 for SRC:PP1, and we stopped each simulation when the three minimum distances between one of the ligands and the three reference atoms were below 0.5 nm, 0.4 nm, 0.5 nm for SRC:DAS and 0.5 nm, 0.35 nm, 0.5 nm for SRC:PP1.

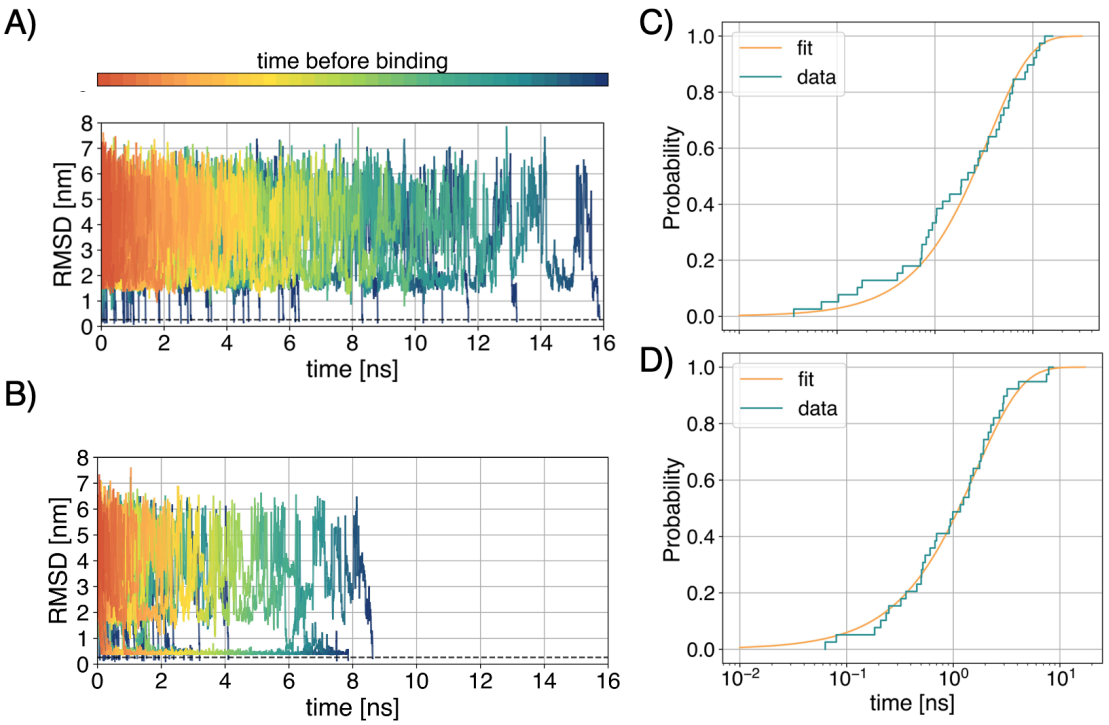

Figure 5. (A) RMSD of DAS with respect to its experimental binding pose over time. The ligand is considered bound when the RMSD is less than 0.26 nm. (B) RMSD of PP1 with respect to its experimental binding pose over time. The ligand is considered as bound when the RMSD is less than 0.26 nm. (C) Cumulative DAS binding times and Poisson fit with mean binding times of 3.5±0.1 ns, p=0.92 from a Kolmogorov-Smirnov test. (D) Cumulative PP1 binding times and Poisson fit with mean binding times of 1.64±0.02 ns, p=0.99992 from a Kolmogorov-Smirnov test.

The calculated RMSD of the ligand to its bound conformation after alignment of the protein is shown in Fig. 5A,B (see also SI Movie S2 and S3). Binding times (RMSD less than 0.26 nm) were fitted to a Poisson distribution, yielding mean binding times of 3.5±0.1 ns and 1.64±0.02 ns for SRC:DAS and SRC:PP1, respectively, with corresponding p-values of 0.92 and 0.99992 obtained with KS-test (Fig. 5C,D). The corresponding k_{on} for SRC:DAS and SRC:PP1 are 2.51×10^{10} $M^{-1}s^{-1}$ and 5.39×10^{10} $M^{-1}s^{-1}$, respectively. Despite the fact that, as for LYZ:BNZ, the time scales are nominal and cannot be compared directly with experiments (i.e., the *in vitro* k_{on} for DAS is $\sim 5 \times 10^{6}$ $M^{-1}s^{-1}$), their ratio is consistent with the ratio of those obtained by previously published MD simulations[25] (1.9×10^{6} $M^{-1}s^{-1}$ and 4.3×10^{6} $M^{-1}s^{-1}$ for DAS and PP1, respectively), indicating for this case a speed up of around four order of magnitudes.

Estimation of c-Src kinase:dasatininb and c-Src kinase:PP1 unbinding free energy via Thermodynamic Integration: multi-eGO models. Having shown that the model can simulate the correct binding of the two ligands, we estimated their unbinding free energy by TI, using three alternative restraints as before (see Methods and Fig. S9 and S10). For each restraint we performed 4 TI replicates. For SRC:DAS we estimated an unbinding free energy of 55.0±0.5 kJ/mol, 55.5±0.3 kJ/mol , 55.6±0.4 kJ/mol for the three restraints, respectively (Figure 6A), while for SRC:PP1 we estimated an unbinding free energy of 38.34±0.06 kJ/mol, 38.4±0.1 kJ/mol, 38.3±0.1 kJ/mol (Figure 6B). As for LYZ:BNZ all estimates are compatible and show high statistical precision.

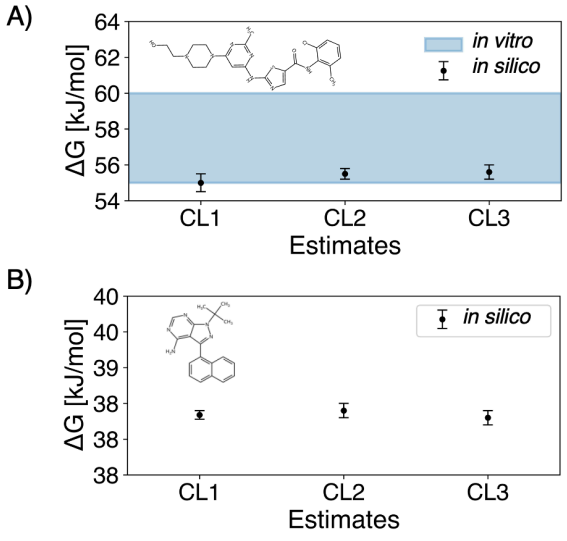

Figure 6. (A) TI estimates of the unbinding ΔG of c-Src kinase and dasatinib using multi-eGO. The three estimates (CL1, CL2, CL3) correspond to different restraints used to keep the ligand in the binding pose during decoupling. Errors were calculated as the standard error of the mean of four replicates for each restraint. The blue band represents the interval of *in vitro* measurements value[40–44]. (B) TI estimates of the unbinding ΔG of c-Src kinase and PP1 using multi-eGO. CL1, CL2, CL3 corresponds to different

restraints. Errors were calculated as the standard error of the mean of four replicas for each constraint.

Importantly, the ΔG obtained for SRC:DAS is compatible with values reported in the literature for *in vitro* measurements value[40–44], ranging from 55.17 kJ/mol to 60.23 kJ/mol (Figure 5A). For SRC:PP1, we compared the experimental ratio of the IC_{50} of SRC:DAS (0.5 nM)[69] and SRC:PP1 (170 nM)[44] with the ratio of the $K_D = \exp\left(-\frac{\Delta G}{k_B T}\right)$ estimates obtained with multi-eGO. The ratio of the multi-eGO K_D values was found to be in qualitative agreement with the IC_{50} ratio, namely $\frac{K_D^{PP1}}{K_D^{DAS}} = 740 \pm 114$ and $\frac{IC_{50}^{PP1}}{IC_{50}^{DAS}} = 340$.

Preparation of the Aβ42:10074-G5 multi-eGO model. The above binding processes represent the most common scenario where a pose is known, or hypothesized, under the lock-and-key mechanism, where the ligand minimally perturbs the structure of the receptor. At the other end of the spectrum is the case of ligand binding to IDPs, which do not entail a well-defined binding site due to the absence of a stable tertiary structure. To test multi-eGO in this scenario, we used the large publicly available dataset for Aβ42 simulated in its apo and 10074-G5 holo states[46,56]. After parametrizing the Aβ42 monomer, as previously published, with an intramolecular ε_0 of 0.33 kJ/mol, we parametrized the multi-eGO 10074-G5 starting from its GROMOS ATB-derived model[70] and optimized its bonded parameters to match those of the training simulation (see Methods). We then performed a reference prior simulation in which mutual interactions between Aβ42 and 10074-G5 are represented only by steric clashes in the same training condition. After this simulation, we set the intermolecular interaction between Aβ42 and 10074-G5 by minimizing the RMSE of the intermolecular residue-wise contact maps (see Fig. S11), obtaining a value of ε_0 of 0.385 kJ/mol. The multi-eGO holo-simulation showed the same Rg behavior observed in the training simulation (Fig. S12), qualitatively reproducing the same peak shift and narrowing of the distribution. Note that this effect is only due to ligand binding, since the intramolecular interactions are only learned from the apo simulation.

Single-molecule simulation of the Aβ42:10074-G5 multi-eGO model. With the parameters found above, we ran a 2 μs multi-eGO MD simulation and calculated the probability of binding vs. unbinding to estimate the K_D. In the case of single molecule simulations, direct estimation of K_D as $\frac{[L][P]}{[LP]}$ is affected by the finite size of the simulation box[71]. To correctly calculate the dissociation constant, considering the box size, we followed Lopez et al. and calculated K_D as

$$K_D = \frac{1}{V\, N_A\, p_B(V)} \frac{1 - p_v(V)}{1 - \frac{v}{V}},$$

where V is the box volume, N_A is Avogadro's number, $p_B(V)$ is the bound probability, v is the interacting volume (the volume where protein and ligand interact but the ligand is not in the bound pose), and p_v is the probability of finding the ligand within the volume V. In the case of ligand binding with an IDP, the difference between the interacting and bound volumes is subtle (with no precise binding pose).

By defining the bound state considering the configurations where the minimum distance between Aβ42 and 10074-G5 is less than 0.33 nm and as interacting the state where the minimum distance is less than 0.4 nm, a dissociation constant of 683±7 μM was calculated for multi-eGO (with the error calculated as standard deviation of the mean of 4 replicates of 2 μs each) (see Figure 7A), in comparison with 381±5 μM obtained by analyzing the training simulation (where the error is calculated using a bootstrapping method with 95th percentile). To note, the difference between the training and multi-eGO dissociation constants corresponds to 1.5 kJ/mol in free energy, less than the thermal fluctuation energy (~2.5 kJ/mol at 300 K). However, these values are not in quantitative agreement with the experimental value of 7-40 μM, most likely because of limitations in force-field accuracy. In this case, a substantial advantage of using multi-eGO is represented by the possibility of updating the model to reproduce the experimental affinity by changing a single parameter, ε_0, which represents the energy scale for the Aβ42:10074-G5 interaction, without the requirement to repeat the computationally intensive all-atom training simulation. A ε_0 of 0.46 kJ/mol corresponded to a dissociation constant of 18±1 μM (Figure 7A), before and after volume correction, where the data represent the mean, and the standard deviation of the mean obtained from 5 replicates of 2 μs each.

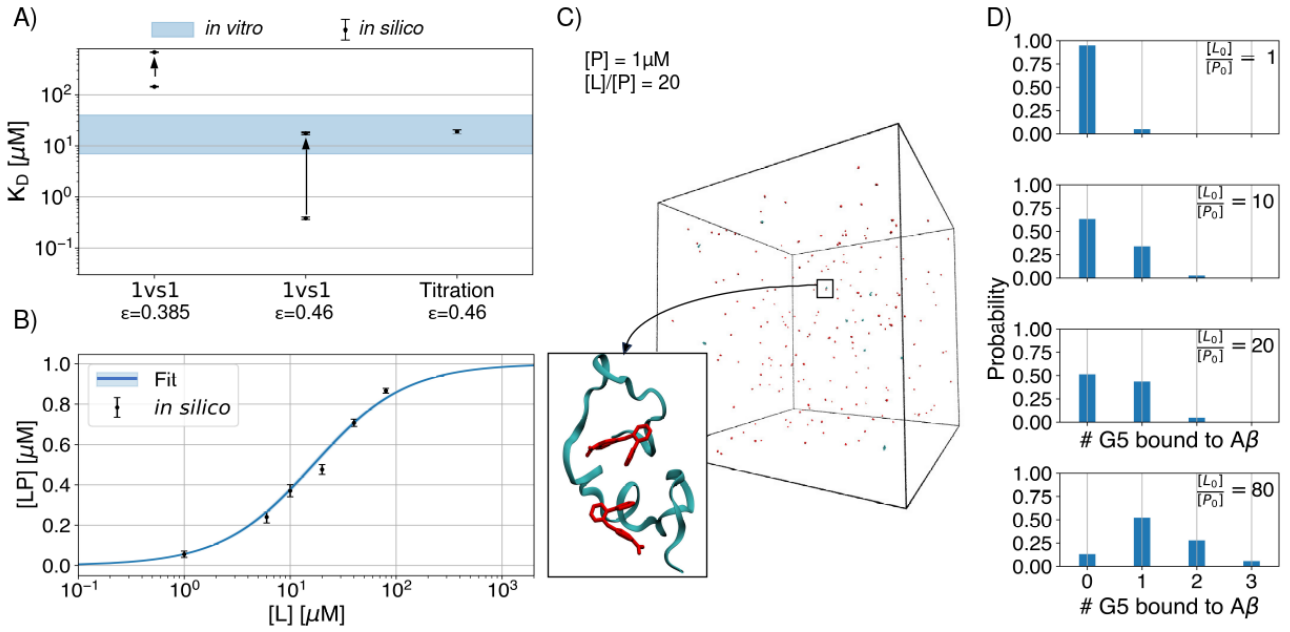

Figure 7. (A) Comparison of dissociation constants obtained from single molecule and titration multi-eGO simulations. The arrows indicate the correction for the finite size effects. (B) Aβ42:10074-G5 bound state population as a function of 10074-G5 concentration as obtained from multi-eGO 10074-G5 titration simulations, the line is the fit using the Hill equation. (C) Display of a simulation box for 10 Aβ42 molecules at 1 μM concentration mixed with 200 10074-G5 molecules. The inset shows an Aβ42 configuration bound to two 10074-G5 molecules. (D) Probability of observing 0 to 3 10074-G5 molecules simultaneously bound to an Aβ42 molecule as a function of 10074-G5 concentration.

In silico titration simulation of the Aβ42:10074-G5 multi-eGO model. The updated model was then used to study Aβ42:10074-G5 binding in the presence of an increasing number of 10074-G5 molecules. Following *in vitro* experiments, we performed an *in-silico* titration experiment mixing 1 μM Aβ42 (consisting of 10 monomers in a 255 nm cubic box, cf. Fig. 7C) with increasing concentrations of 10074-G5, namely 1:1, 1:6, 1:10, 1:20, 1:40 and 1:80 Aβ42:10074-G5 ratios. At each ratio, 5 μs long multi-eGO simulation were run (see Fig. S13 for binding probability as a function of time and block averaging), and the concentration of bound Aβ42:10074-G5 was fitted as a function of ligand concentration using the equation $[LP] = \frac{[L][P]}{[L]+K_D}$, which resulted in a K_D of 16.4±0.7 μM (as shown in Figure 7B with error estimates resulting from 1 μs block averaging), consistent with that obtained with the volume-corrected single molecule approach. This case further emphasizes the usefulness of multi-eGO[71] when estimating dissociation constants from single molecule equilibrium simulations.

In Figure 7D, we report the probability of having multiple concurrent bindings at the different concentrations, showing the presence of 2 or 3 ligands bound to one Aβ42 monomer (see also the inset of Figure 7C for a representative configuration). This analysis suggests that the effect of simultaneous binding is negligible at the concentrations of the experiments. The lack of cooperativity is consistent with the entropic expansion model with minimal structural perturbations[72]. Notably, we anticipate that our simulations could be extended to account for additional Aβ42 interactions, such as those that could be derived from a fibril structure, paving the way to study protein aggregation itself and its inhibition mechanisms.

DISCUSSION

Structure-based models have been introduced to study protein folding, based on the hypothesis that the native protein structure should capture the most relevant interactions across the whole free energy landscape of a protein. They have proven useful to study otherwise, inaccessible folding mechanisms at high resolution[73]. Recently, we have introduced multi-eGO, an ensemble-based model rooted in Bayesian statistics, where one or more conformational ensembles of a protein, representing the fluctuations of relevant free energy minima, are weighted with prior information[31,32]. This approach enables the generation of a model that can simultaneously describe folded and disordered proteins, in addition to self-assembly processes such as peptide amyloidogenic aggregation.

Here, we have extended the multi-eGO approach to include small molecules. Our initial results on four case studies allow us on the one hand to propose multi-eGO as a model for benchmarking and developing *in silico* techniques for ligand binding studies, and on the other hand to suggest that multi-eGO can be used for quantitative estimation of ligand affinities. This is because multi-eGO describes the system at

atomic resolution, with binding free energies compatible with those measured *in vitro*, but with a dramatic acceleration of the kinetics (that we can tentatively quantify between 2 and 4 orders of magnitude). We exploited the LYZ:BNZ system to apply orthogonal free energy calculation techniques (TI with multiple restraining conditions and VMetaD) showing their applicability and robustness on the model. It must be underlined that all simulations are replicated multiple times to obtain statistically meaningful estimates, at minimal computational costs. While Aβ42:10074-G5 represented a case that is inaccessible to current conventional approaches, namely explicit titration as a function of ligand concentration, with direct comparisons with *in vitro* concentrations. The latter case, on the one hand, demonstrated the efficiency of multi-eGO and, on the other hand, underlined the importance of correcting single-molecule simulations for the finite-size effect, following procedures such as those described in Lopez et al.

The use of multi-eGO requires model training, which consists of running training simulations and finding a set of free parameters. Here, we introduced a possible strategy to set our energy scale parameters, excluding the intervention of any possible user bias. In practice, we minimized the RMSE with the protein fluctuation as well as the RMSE for the intermolecular residue-wise contact map, comparing the multi-eGO and training simulations. This approach was first successfully tested on the LYZ:BNZ system and then successfully replicated for the SRC:DAS case. Successfully, because in both cases, knowledge of the experimental binding free energy was not used in the input and instead was accurately obtained by the simulations. In both cases, our training simulations include a ligand in its binding pose as well as several ligands free to sample the surface of the protein. The latter data are used to set the energy scale of the system. For Aβ42:10074-G5, we used a previously generated, large dataset of simulations, from which we inherited an overestimated dissociation constant. This is where the advantage of multi-eGO becomes apparent, as it is sufficient to tune a single parameter to make the model match the experimental value. This allows us to speculate that multi-eGO may be particularly suitable for studying ligand binding processes such as those represented by the LYZ:BNZ, SRC:DAS, and SRC:PP1 cases, where the training data can be generated with very limited computational cost and should represent only the weak (i.e., faster to sample) non-specific interactions of a ligand with the protein surface, given knowledge or a hypothesis about the binding pose. We anticipate that further work will be needed to better understand the strengths and limitations of multi-eGO in ligand binding studies, including cases with more complex binding mechanisms such as induce-fit or conformational selection, as well as to streamline and possibly automatize the simulation setup protocol.

CONCLUSIONS

Ligand binding studies are among the most important applications of MD simulations since this computational technique can capture the conformational dynamics of both the ligand and the receptor. It is thus possible to obtain a relatively accurate description of their physicochemical interactions, and both kinetic and thermodynamic information about the binding process. These results can be obtained if sufficient computational resources are available. While several methods have been developed to reduce the computational costs of obtaining accurate kinetic or thermodynamic data, MD simulations remain hindered by the timescale for sampling the relevant regions of conformational space, to obtain reliable estimates of thermodynamic properties. Our results highlight the potential of multi-eGO both as a model to benchmark and develop *in silico* free energy calculation techniques and as an accurate and efficient framework for ligand binding studies, potentially extending the current capabilities, in terms of time and number of particles, of *in silico* molecular studies.

ASSOCIATED CONTENT

Supporting Information

The supporting information file contains comparison between the training simulations and multi-eGO ones for all the showed systems, extended convergence plots for all the TI and VMetaD calculations performed, the representation of binding and unbinding pathways for LYZ:BNZ, the ε_0 selection details for SRC:DAS, SRC:PP1, and Aβ42:10074-G5; supporting tables with the parameters for all the TI performed on the four systems. Finally, we added 4 movies showing the binding events from multi-eGO simulations in the four complexes.

AUTHOR INFORMATION

Corresponding Author

* Riccardo Capelli, riccardo.capelli@unimi.it; Carlo Camilloni, carlo.camilloni@unimi.it

Author Contributions

BS performed and analyzed all the simulations. ES, with the support of TL, performed proof of principle simulations. BS, ES, FBT, RC, and CC developed the model. RC and CC supervised the work with the support of SG and MV. BS, RC and CC wrote the manuscript with contributions of all authors.

Funding Sources

RC acknowledges financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project P20225ZPYH (RESISTANCE) – CUP G53D23006890001 - Grant Assignment Decree No. 1110 adopted on 20.07.2023 by the Italian Ministry of University and Research (MUR). We acknowledge co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, and the Investment PE8–Project Age-It: "Ageing Well in an Ageing Society". This resource was co-financed by the Next Generation EU [DM 1557 11 October 2022]. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

ACKNOWLEDGMENT

We thank Max Bonomi and Toni Giorgino for stimulating discussions. We thank D. E. Shaw Research for sharing their molecular dynamics trajectories. We thank Louise Gourlay for carefully reading the manuscript and improve its language. The authors

acknowledge CINECA for an award under the ISCRA initiative, for the availability of high-performance computing resources and support.

ABBREVIATIONS

MD, molecular dynamics. TI, Thermodynamic Integration; VMetaD, Volume-based Metadynamics. RMSD, root mean square deviation. RMSF, root mean square fluctuations, RMSE, root mean square error. Rg, radius of gyration. LYZ, lysozyme. BNZ, benzene. KS, Kolmogorov-Smirnov. SRC, c-Src kinase. DAS, dasatinib. G5, 10074-G5. FES, free energy surface. IDP, intrinsically disordered protein.

DATA AND SOFTWARE AVAILABILITY STATEMENT

Simulations data are publicly available via Zenodo with records DOI: 10.5281/zenodo.12800706, 10.5281/zenodo.12800760, 10.5281/zenodo.12800768; the multi-eGO code and parameters are publicly available on GitHub at https://github.com/multi-ego/multi-eGO, use the beta3 tag for a snapshot of the repository associated to this paper.

REFERENCES

(1) Hlavacek, W. S.; Faeder, J. R.; Blinov, M. L.; Perelson, A. S.; Goldstein, B. The Complexity of Complexes in Signal Transduction. *Biotechnol. Bioeng.* **2003**, *84* (7), 783–794. https://doi.org/10.1002/bit.10842.

(2) Hunter, T. Signaling—2000 and Beyond. *Cell* **2000**, *100* (1), 113–127. https://doi.org/10.1016/s0092-8674(00)81688-8.

(3) Wolf, F. A. de; Brett, G. M. Ligand-Binding Proteins: Their Potential for Application in Systems for Controlled Delivery and Uptake of Ligands. *Pharmacol. Rev.* **2000**, *52* (2), 207–236.

(4) Cera, E. D. Mechanisms of Ligand Binding. *Biophys. Rev.* **2020**, *1* (1), 011303. https://doi.org/10.1063/5.0020997.

(5) Pang, X.; Zhou, H.-X. Rate Constants and Mechanisms of Protein–Ligand Binding. *Annu. Rev. Biophys.* **2016**, *46* (1), 1–26. https://doi.org/10.1146/annurev-biophys-070816-033639.

(6) Fersht, A. R. *Structure and Mechanism in Protein Science*; W. H. Freeman and Company; W. H. Freeman and Company, 1998.

(7) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 935–949. https://doi.org/10.1038/nrd1549.

(8) Paggi, J. M.; Pandit, A.; Dror, R. O. The Art and Science of Molecular Docking. *Annu. Rev. Biochem.* **2024**. https://doi.org/10.1146/annurev-biochem-030222-120000.

(9) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813–1818. https://doi.org/10.1126/science.1096361.

(10) Corso, G.; Deng, A.; Fry, B.; Polizzi, N.; Barzilay, R.; Jaakkola, T. DEEP CONFIDENT STEPS TO NEW POCKETS: STRATEGIES FOR DOCKING GENERALIZATION. *biorxiv*. https://doi.org/10.48550/arxiv.2402.18396.

(11) Zhang, X.; Zhang, O.; Shen, C.; Qu, W.; Chen, S.; Cao, H.; Kang, Y.; Wang, Z.; Wang, E.; Zhang, J.; Deng, Y.; Liu, F.; Wang, T.; Du, H.; Wang, L.; Pan, P.; Chen, G.; Hsieh, C.-Y.; Hou, T. Efficient and Accurate Large Library Ligand Docking with KarmaDock. *Nat. Comput. Sci.* **2023**, *3* (9), 789–804. https://doi.org/10.1038/s43588-023-00511-5.

(12) Plattner, N.; Noé, F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* **2015**, *6* (1), 7653. https://doi.org/10.1038/ncomms8653.

(13) Cavalli, A.; Spitaleri, A.; Saladino, G.; Gervasio, F. L. Investigating Drug–Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms. *Acc. Chem. Res.* **2015**, *48* (2), 277–285. https://doi.org/10.1021/ar500356n.

(14) Harvey, M. J.; Fabritiis, G. D. High-Throughput Molecular Dynamics: The Powerful New Tool for Drug Discovery. *Drug Discov. Today* **2012**, *17* (19–20), 1059–1062. https://doi.org/10.1016/j.drudis.2012.03.017.

(15) Jorgensen, W. L.; Thomas, L. L. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **2008**, *4* (6), 869–876. https://doi.org/10.1021/ct800011m.

(16) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72* (3), 1047–1069. https://doi.org/10.1016/s0006-3495(97)78756-3.

(17) Camilloni, C.; Pietrucci, F. Advanced Simulation Techniques for the Thermodynamic and Kinetic Characterization of Biological Systems. *Adv. Phys.: X* **2018**, *3* (1), 1477531. https://doi.org/10.1080/23746149.2018.1477531.

(18) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu Rev Biophys* **2012**, *41* (1), 429–452. https://doi.org/10.1146/annurev-biophys-042910-155245.

(19) Vivo, M. D.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.

(20) Shirts, M. R.; Pande, V. S. Comparison of Efficiency and Bias of Free Energies Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. *J. Chem. Phys.* **2005**, *122* (14), 144107. https://doi.org/10.1063/1.1873592.

(21) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245–268. https://doi.org/10.1016/0021-9991(76)90078-4.

(22) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426. https://doi.org/10.1063/1.1740409.

(23) Perthold, J. W.; Oostenbrink, C. Accelerated Enveloping Distribution Sampling: Enabling Sampling of Multiple End States While Preserving Local Energy Minima. *J. Phys. Chem. B* **2018**, *122* (19), 5030–5037. https://doi.org/10.1021/acs.jpcb.8b02725.

(24) Gapsys, V.; Yildirim, A.; Aldeghi, M.; Khalak, Y.; Spoel, D. van der; Groot, B. L. de. Accurate Absolute Free Energies for Ligand–Protein Binding Based on Non-Equilibrium Approaches. *Commun. Chem.* **2021**, *4* (1), 61. https://doi.org/10.1038/s42004-021-00498-y.

(25) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J Am Chem Soc* **2011**, *133* (24), 9181–9183. https://doi.org/10.1021/ja202726y.

(26) Buch, I.; Giorgino, T.; Fabritiis, G. D. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* **2011**, *108* (25), 10184–10189. https://doi.org/10.1073/pnas.1103547108.

(27) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J Comput Phys* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(28) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc National Acad Sci* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.

(29) Souza, P. C. T.; Thallmair, S.; Conflitti, P.; Ramírez-Palacios, C.; Alessandri, R.; Raniolo, S.; Limongelli, V.; Marrink, S. J. Protein–Ligand Binding with the Coarse-Grained Martini Model. *Nat. Commun.* **2020**, *11* (1), 3714. https://doi.org/10.1038/s41467-020-17437-5.

(30) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163.

(31) Scalone, E.; Broggini, L.; Visentin, C.; Erba, D.; Toplek, F. B.; Peqini, K.; Pellegrino, S.; Ricagno, S.; Paissoni, C.; Camilloni, C. Multi-EGO: An in Silico Lens to Look into Protein Aggregation Kinetics at Atomic Resolution. *Proc. Natl. Acad. Sci.* **2022**, *119* (26), e2203181119. https://doi.org/10.1073/pnas.2203181119.

(32) Toplek, F. B.; Scalone, E.; Stegani, B.; Paissoni, C.; Capelli, R.; Camilloni, C. Multi-eGO: Model Improvements toward the Study of Complex Self-Assembly Processes. *J. Chem. Theory Comput.* **2024**, *20* (1), 459–468. https://doi.org/10.1021/acs.jctc.3c01182.

(33) Pennacchietti, V.; Matteo, S. di; Pagano, L.; Toplek, F. B.; Stegani, B.; Toto, A.; Toso, J.; Puglisi, E.; Capelli, R.; Felice, M. D.; Malagrinò, F.; Camilloni, C.; Gianni, S. A PDZ Tandem Repeat Folds and Unfolds via Different Pathways. *Protein Sci.* **2024**, *33* (12), e5203. https://doi.org/10.1002/pro.5203.

(34) Morton, A.; Baase, W. A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry* **1995**, *34* (27), 8564–8575. https://doi.org/10.1021/bi00027a006.

(35) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. Lessons from the Lysozyme of Phage T4. *Protein Sci.* **2010**, *19* (4), 631–641. https://doi.org/10.1002/pro.344.

(36) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Ligand Unbinding Mechanisms and Kinetics for T4 Lysozyme Mutants from TRAMD Simulations. *Curr. Res. Struct. Biol.* **2021**, *3*, 106–111. https://doi.org/10.1016/j.crstbi.2021.04.001.

(37) Rydzewski, J.; Valsson, O. Finding Multiple Reaction Pathways of Ligand Unbinding. *J. Chem. Phys.* **2019**, *150* (22), 221101. https://doi.org/10.1063/1.5108638.

(38) Capelli, R.; Carloni, P.; Parrinello, M. Exhaustive Search of Ligand Binding Pathways via Volume-Based Metadynamics. *J Phys Chem Lett* **2019**, *10* (12), 3495–3499. https://doi.org/10.1021/acs.jpclett.9b01183.

(39) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003**, *107* (35), 9535–9551. https://doi.org/10.1021/jp0217839.

(40) Sicheri, F.; Moarefi, I.; Kuriyan, J. Crystal Structure of the Src Family Tyrosine Kinase Hck. *Nature* **1997**, *385* (6617), 602–609. https://doi.org/10.1038/385602a0.

(41) Getlik, M.; Grütter, C.; Simard, J. R.; Klüter, S.; Rabiller, M.; Rode, H. B.; Robubi, A.; Rauh, D. Hybrid Compound Design To Overcome the Gatekeeper T338M Mutation in CSrc. *J. Med. Chem.* **2009**, *52* (13), 3915–3926. https://doi.org/10.1021/jm9002928.

(42) Shan, Y.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H.; Jensen, M. Ø.; Dror, R. O.; Kuriyan, J.; Shaw, D. E. A Conserved Protonation-Dependent Switch Controls Drug Binding in the Abl Kinase. *Proc. Natl. Acad. Sci.* **2009**, *106* (1), 139–144. https://doi.org/10.1073/pnas.0811223106.

(43) Tatton, L.; Morley, G. M.; Chopra, R.; Khwaja, A. The Src-Selective Kinase Inhibitor PP1 Also Inhibits Kit and Bcr-Abl Tyrosine Kinases*. *J. Biol. Chem.* **2003**, *278* (7), 4847–4853. https://doi.org/10.1074/jbc.m209321200.

(44) Hanke, J. H.; Gardner, J. P.; Dow, R. L.; Changelian, P. S.; Brissette, W. H.; Weringer, E. J.; Pollok, B. A.; Connelly, P. A. Discovery of a Novel, Potent, and Src Family-Selective Tyrosine Kinase Inhibitor STUDY OF Lck- AND FynT-DEPENDENT T CELL ACTIVATION (∗). *J. Biol. Chem.* **1996**, *271* (2), 695–701. https://doi.org/10.1074/jbc.271.2.695.

(45) Yin, X.; Giap, C.; Lazo, J. S.; Prochownik, E. V. Low Molecular Weight Inhibitors of Myc–Max Interaction and Function. *Oncogene* **2003**, *22* (40), 6151–6159. https://doi.org/10.1038/sj.onc.1206641.

(46) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Small Molecule Stabilizes the Disordered Native State of the Alzheimer's Aβ Peptide. *ACS Chem. Neurosci.* **2022**, *13* (12), 1738–1745. https://doi.org/10.1021/acschemneuro.2c00116.

(47) Heller, G. T.; Aprile, F. A.; Michaels, T. C. T.; Limbocker, R.; Perni, M.; Ruggeri, F. S.; Mannini, B.; Löhr, T.; Bonomi, M.; Camilloni, C.; Simone, A. D.; Felli, I. C.; Pierattelli, R.; Knowles, T. P. J.; Dobson, C. M.; Vendruscolo, M. Small-Molecule Sequestration of Amyloid-β as a Drug Discovery Strategy for Alzheimer's Disease. *Sci. Adv.* **2020**, *6* (45), eabb5924. https://doi.org/10.1126/sciadv.abb5924.

(48) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes. *J Chem Theory Comput* **2020**. https://doi.org/10.1021/acs.jctc.9b00251.

(49) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural

Properties of Disordered Protein States. *J Phys Chem B* **2015**, *119* (16), 5113–5123. https://doi.org/10.1021/jp508971m.

(50) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput* **2008**, *4* (1), 116–122. https://doi.org/10.1021/ct700200b.

(51) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J Chem Phys* **2007**, *126* (1), 014101. https://doi.org/10.1063/1.2408420.

(52) Bernetti, M.; Bussi, G. Pressure Control Using Stochastic Cell Rescaling. *J. Chem. Phys.* **2020**, *153* (11), 114107. https://doi.org/10.1063/5.0020514.

(53) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *Softwarex* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(54) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. A Cavity-Containing Mutant of T4 Lysozyme Is Stabilized by Buried Benzene. *Nature* **1992**, *355* (6358), 371–373. https://doi.org/10.1038/355371a0.

(55) Cowan-Jacob, S. W.; Fendrich, G.; Manley, P. W.; Jahnke, W.; Fabbro, D.; Liebetanz, J.; Meyer, T. The Crystal Structure of a C-Src Complex in an Active Conformation Suggests Possible Steps in c-Src Activation. *Structure* **2005**, *13* (6), 861–871. https://doi.org/10.1016/j.str.2005.03.012.

(56) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Kinetic Ensemble of the Alzheimer's Aβ Peptide. *Nat. Comput. Sci.* **2021**, *1* (1), 71–78. https://doi.org/10.1038/s43588-020-00003-w.

(57) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys J* **2011**, *100* (9), L47-9. https://doi.org/10.1016/j.bpj.2011.03.051.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(59) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, *34* (32), 2757–2770. https://doi.org/10.1002/jcc.23422.

(60) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys Rev Lett* **2008**, *100* (2), 020603. https://doi.org/10.1103/physrevlett.100.020603.

(61) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies. *J. Comput. Chem.* **1995**, *16* (3), 273–284. https://doi.org/10.1002/jcc.540160303.

(62) Tiwary, P.; Parrinello, M. From Metadynamics to Dynamics. *Phys Rev Lett* **2013**, *111* (23), 230602. https://doi.org/10.1103/physrevlett.111.230602.

(63) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613. https://doi.org/10.1016/j.cpc.2013.09.018.

(64) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; Capelli, R.; Carloni, P.; Ceriotti, M.; Cesari, A.; Chen, H.; Chen, W.; Colizzi, F.; De, S.; Pierre, M. D. L.; Donadio, D.; Drobot, V.; Ensing, B.; Ferguson, A. L.; Filizola, M.; Fraser, J. S.; Fu, H.; Gasparotto, P.; Gervasio, F. L.; Giberti, F.; Gil-Ley, A.; Giorgino, T.; Heller, G. T.; Hocky, G. M.; Iannuzzi, M.; Invernizzi, M.; Jelfs, K. E.; Jussupow, A.; Kirilin, E.; Laio, A.; Limongelli, V.; Lindorff-Larsen, K.; Löhr, T.; Marinelli, F.; Martin-Samos, L.; Masetti, M.; Meyer, R.; Michaelides, A.; Molteni, C.; Morishita, T.; Nava, M.; Paissoni, C.; Papaleo, E.; Parrinello, M.; Pfaendtner, J.; Piaggi, P.; Piccini, G.; Pietropaolo, A.; Pietrucci, F.; Pipolo, S.; Provasi, D.; Quigley, D.; Raiteri, P.; Raniolo, S.; Rydzewski, J.; Salvalaglio, M.; Sosso, G. C.; Spiwok, V.; Šponer, J.; Swenson, D. W. H.; Tiwary, P.; Valsson, O.; Vendruscolo, M.; Voth, G. A.; White, A. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16* (8), 670–673. https://doi.org/10.1038/s41592-019-0506-8.

(65) Simone, A. D.; Montalvao, R. W.; Dobson, C. M.; Vendruscolo, M. Characterization of the Interdomain Motions in Hen Lysozyme Using Residual Dipolar Couplings as Replica-Averaged Structural Restraints in Molecular Dynamics Simulations. *Biochemistry* **2013**, *52* (37), 6480–6486. https://doi.org/10.1021/bi4007513.

(66) Salvalaglio, M.; Tiwary, P.; Parrinello, M. Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *J Chem Theory Comput* **2014**, *10* (4), 1420–1425. https://doi.org/10.1021/ct500040r.

(67) Feher, V. A.; Baldwin, E. P.; Dahlquist, F. W. Access of Ligands to Cavities within the Core of a Protein Is Rapid. *Nat. Struct. Biol.* **1996**, *3* (6), 516–521. https://doi.org/10.1038/nsb0696-516.

(68) Ray, D.; Parrinello, M. Data-Driven Classification of Ligand Unbinding Pathways. *Proc. Natl. Acad. Sci.* **2024**, *121* (10), e2313542121. https://doi.org/10.1073/pnas.2313542121.

(69) Lombardo, L. J.; Lee, F. Y.; Chen, P.; Norris, D.; Barrish, J. C.; Behnia, K.; Castaneda, S.; Cornelius, L. A. M.; Das, J.; Doweyko, A. M.; Fairchild, C.; Hunt, J. T.; Inigo, I.; Johnston, K.; Kamath, A.; Kan, D.; Klei, H.; Marathe, P.; Pang, S.; Peterson, R.; Pitt, S.; Schieven, G. L.; Schmidt, R. J.; Tokarski, J.; Wen, M.-L.; Wityak, J.; Borzilleri, R. M. Discovery of N-(2-Chloro-6-Methyl- Phenyl)-2-(6-(4-(2-Hydroxyethyl)- Piperazin-1-Yl)-2-Methylpyrimidin-4- Ylamino)Thiazole-5-Carboxamide (BMS-354825), a Dual Src/Abl Kinase Inhibitor with Potent Antitumor Activity in Preclinical Assays. *J. Med. Chem.* **2004**, *47* (27), 6658–6661. https://doi.org/10.1021/jm049486a.

(70) Koziara, K. B.; Stroet, M.; Malde, A. K.; Mark, A. E. Testing and Validation of the Automated Topology Builder (ATB) Version 2.0: Prediction of Hydration Free Enthalpies. *J. Comput.-Aided Mol. Des.* **2014**, *28* (3), 221–233. https://doi.org/10.1007/s10822-014-9713-7.

(71) Lopez, A. J.; Quoika, P. K.; Linke, M.; Hummer, G.; Köfinger, J. Quantifying Protein–Protein Interactions in Molecular Simulations. *J*

Phys Chem B **2020**, *124* (23), 4673–4685. https://doi.org/10.1021/acs.jpcb.9b11802.

(72) Heller, G. T.; Sormanni, P.; Vendruscolo, M. Targeting Disordered Proteins with Small Molecules Using Entropy. *Trends Biochem Sci* **2015**, *40* (9), 491–496. https://doi.org/10.1016/j.tibs.2015.07.004.

(73) Takada, S. Gō Model Revisited. *Biophysics Physicobiology* **2019**, *16* (0), 248–255. https://doi.org/10.2142/biophysico.16.0_248.