Computing Ligand Binding Free Energy in a Large Flexible Pocket of a Large Protein

Evgeny Gutkin, Chamali Narangoda, and Maria G. Kurnikova[*]

Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, 15213

[*] Corresponding author, kurnikova@cmu.edu

Abstract

Protein-ligand standard binding free energy (SBFE) calculations based on molecular dynamics simulations have recently emerged as an efficient tool for computer-aided drug design. However, the majority of studies have focused on well-defined ligand binding pockets in small proteins, while computing SBFE of a small molecule ligand to a large, flexible binding site remains problematic. Numerous proteins with large flexible binding pockets play an important role in biological processes and are of pharmaceutical importance. Therefore, the calculation of the ligand SBFE with such proteins is an important challenge. In this work, we developed a hierarchical approach to compute SBFE of a flexible multi-conformational system as an ensemble average of individual local SBFEs to specific conformational states of the protein-ligand complex. This approach allows us to simulate a truncated portion of a large protein, which brings an intractable system within the reach of modern computational tools. Our approach also accounts for the differences in conformational preferences between a ligand-bound and an apo states of the protein. The approach is validated using the T4 lysozyme mutant in complex with a small molecule inhibitor. Binding energies of a non-competitive antagonist (GYKI) with the GluA2 glutamate receptor of AMPA type (AMPAR) are computed for several poses of GYKI in the binding pocket to help facilitate structure validation at relatively low resolution.

Introduction

Prediction of a ligand binding affinity to a specific protein is a major challenge in computer-aided drug design[1, 2]. One of the best existing approaches for such prediction is to compute a protein-ligand standard binding free energy (SBFE), also termed absolute binding free energy (ABFE),

1

using molecular dynamics (MD) simulations[3, 4]. Computational methods based on sampling states of a system along a non-physical (alchemical) thermodynamic pathway were demonstrated to provide accurate SBFE predictions[5, 6]. SBFE calculations have recently emerged as a promising strategy for the early stages of drug discovery, as reported in a number of studies using a variety of approaches: thermodynamic integration (TI)[7-10], free energy perturbation (FEP)[7-9, 11-13], non-equilibrium approaches[14, 15], alchemical transfer method (ATM)[16], and with different force fields including OPLS[17, 18], GAFF[19, 20], CGenFF[21], and OpenFF[22, 23].

Most prior studies dealt with well-formed ligand binding pockets in relatively small globular proteins which ensures a single conformation of the protein-ligand complex. However, computing SBFE for proteins with large flexible binding pockets and multiple conformations of the complex remains a challenge. Examples of such proteins are cytochrome P450 enzymes[24], G protein-coupled receptors (GPCRs)[25], glutamate receptors[26], multidrug resistance proteins[27], and others. These proteins are of pharmaceutical importance due to their crucial role in various biological processes. This work aims to develop an approach for computing SBFE for such protein-ligand systems.

Due to conformational rearrangements of flexible protein regions, the binding pocket can exist in multiple conformational states, some of which may exhibit considerable differences in structure and energy. At the same time, current computational approaches to simulated SBFEs rely on the system to remain in the vicinity of a single main conformation of the protein. However, due to SBFE sensitivity to a local environment of a ligand, such computed energies may vary significantly between distinct conformations of the binding pocket. Hence, computing SBFE only for a single conformation may provide an unreliable estimate of the total protein-ligand binding affinity. For

computed SBFE to accurately reflect the affinity of the ligand to the protein, it needs to account for the whole ensemble of conformations of the protein-ligand complex.

Moreover, conformational preferences of the same binding pocket in the ligand-bound state and the apo state of the protein may differ as well. Sufficient sampling of all relevant degrees of freedom for both states is required for computing statistically converged SBFEs. Due to the slow rate of the protein conformational transitions, achieving convergence of the SBFE calculations for a large system would require prohibitively long MD simulations. Recent studies reported the successful use of replica exchange[28] or metadynamics[29] to sample through conformations of small protein systems concurrently with an SBFE calculation. However, such an approach significantly increases the computational cost of SBFE calculations and can even make these calculations impractical for larger systems.

In this work we developed a hierarchical approach to compute the SBFE of a ligand to a large flexible binding pocket in a large protein complex. We consider the SBFE as an ensemble average of SBFEs for individual conformational states of the protein-ligand complex. Our approach also accounts for the difference between conformational ensembles for complex and apo states. Furthermore, we developed an automated procedure for the SBFE calculation for a single conformation via GPU-accelerated MD TI simulations of a truncated protein-ligand complex system. To test the performance of the developed approach, we computed the SBFE of a small molecule inhibitor to T4 lysozyme mutant. We also applied our approach to compute the SBFE of a non-competitive antagonist to the allosteric pocket of the AMPA-subtype GluA2 receptor (AMPAR) that was modeled in our previous work[26]. We have also developed an automatic procedure for the proposed hierarchical protocol, which includes preparation of the truncated systems, based on a procedure developed in our previous work[26]. The total computational time of

3

the MD TI simulations for a single truncated system was approximately 4 hours on a single GPU (NVIDIA GeForce RTX 2080).

AMPARs play a central role in the excitatory synaptic transmission in the central nervous system. Dysregulation of the AMPAR activity has been implicated in a range of neurological and neuropsychiatric disorders, including epilepsy[30], depression[31], and Alzheimer's disease[32]. Consequently, AMPAR inhibitors attracted a significant interest as potential drugs for treatment of neurological disorders[33]. Noncompetitive allosteric antagonists of AMPAR have been proven to be the most potent and selective inhibitors with one of them, pyridone perampanel (PMP), approved as a therapy for epilepsy[34, 35]. The X-ray crystal structures revealed that PMP binds to four equivalent allosteric binding sites (one for each AMPAR subunit)[36]. Another AMPAR non-competitive antagonist, GYKI 53655[37], was demonstrated to bind to the same allosteric pocket[36]. However, due to the limited resolution of these structures, details of the binding mode of ligands and their specific interactions with protein residues remained unclear. Our previous molecular modeling study demonstrated that SBFEs of representative structures of GYKI varied significantly for different AMPAR subunits and ligand binding modes[26]. In this work, we performed extensive MD simulations of the AMPAR-GYKI complex, and computed ensemble-averaged SBFEs using the developed approach, which allowed us to estimate the difference in SBFE between distinct binding modes of GYKI.

Approach

Two major challenges can be identified in computing the SBFE of a ligand to a flexible binding pocket of a large protein: 1) the high computational cost of simulating the full protein system and 2) the flexibility of the binding pocket. The first challenge is due to the big size of the full protein-

4

ligand complex. For example, the crystal structure of AMPAR in complex with GYKI includes four subunits, each of which contains TMD, LBD and TMD-LBD linkers (1628 protein residues for all subunits in total). Computing the SBFE with MD TI simulations for a system of this size will take a significant time. Therefore, the calculation of the SBFE for the full protein-ligand complex system can be infeasible, especially for multiple protein conformations. To address this challenge, we used truncated protein-ligand complex systems for all SBFE calculations. A truncated system included the ligand, the protein and lipids within a local proximity of the allosteric binding pocket of a single subunit (see Methods section for details).

The second challenge is the flexibility of the protein binding pocket which is mainly due to conformational rearrangements of the protein. There are multiple protein backbone conformations of the binding pocket of both the protein-ligand complex and the apo protein. For brevity, we will refer to a protein backbone conformation of the binding pocket as just the "conformation" in the following text. Computing the SBFE only for a single conformation may provide an unreliable estimate of the total SBFE. To address this challenge, we computed SBFEs for multiple conformations with harmonic restraints imposed on protein C_α atoms at all MD TI simulations for the protein-ligand complex. For these ends, we extended and automated the protocol for MD TI calculations developed in our previous work[26] (see the Methods section for details). These restraints serve two goals: firstly, they are needed to prevent the destabilization of the truncated system during MD TI simulations. Since the truncated systems contain several protein fragments that may not be bound covalently, imposing restraints stabilizes the system during simulations. Secondly, these restraints impose a fixed specific conformation, thus facilitating simulation convergence.

Therefore, MD TI simulations for truncated systems with imposed restraints yield SBFE of a ligand to a restrained conformation of the binding pocket, $\Delta G_{b,r}^{\circ}$. The ensemble averaged SBFE is then computed by taking a mean of SBFEs over multiple conformations from an equilibrium ensemble

$$\langle \Delta G_{b,r}^{\circ} \rangle = \sum_i^N \Delta G_{b,r,i}^{\circ} , \tag{1}$$

where i is the index of a conformation and N is the total number of conformations. In this work, conformations were extracted equidistantly from an equilibrium portion of a long MD trajectory.

The ensemble averaged SBFE $\langle \Delta G_{b,r}^{\circ} \rangle$ is a biased estimation due to the presence of the restraints in MD TI simulations. The effect of restraints in an ensemble averaged SBFE $\langle \Delta G_{b,r}^{\circ} \rangle$ should be unbiased to obtain the final SBFE ΔG_{b}°. The approach that we used for unbiasing requires the estimation of two additional free energy terms: the free energy of the addition of the protein restraints for the protein-ligand complex ΔG_r^{com} and the analogous quantity in the apo state ΔG_r^{apo}. The total unbiased SBFE ΔG_{b}° is expressed as follows:

$$\Delta G_{b}^{\circ} = \Delta G_r^{apo} + \langle \Delta G_{b,r}^{\circ} \rangle - \Delta G_r^{com} . \tag{2}$$

The eq 2 can be schematically illustrated by the thermodynamic cycle presented in Figure 1. The cycle includes four states (numbered 1-4 in Figure 1) each of which is characterized by a set of protein backbone conformations. Each set of conformations is shown by 2D distributions where the axes are arbitrary conformational coordinates of the protein, and each protein conformation is represented by a single point. The total unbiased SBFE ΔG_{b}° is the free energy difference between the protein-ligand complex (Figure 1, state 1) and the apo protein (Figure 1, state 4) and free ligand is solvent (not shown). For these states, the corresponding sets of conformations are obtained from unrestrained equilibrium MD trajectories (see Figure 1, state 1, black points for protein-ligand complex, and state 4, grey points for apo state). The ensembled-averaged SBFE $\langle \Delta G_{b,r}^{\circ} \rangle$ can be

6

considered as the free energy difference between a subset of the protein-ligand complex ensemble (state 2) and the corresponding subset of the apo protein ensemble (state 3) and free ligand. The protein conformations of state 2 are the conformations of the protein-ligand complex structures selected for free energy calculations and highlighted by red in Figure 1. State 3 is obtained from state 2 by annihilating ligand in the binding pocket by MD TI simulations. Since the protein backbone is restrained during these simulations, the conformations of state 3 are the same as the conformations of state 2. State 2 is a subset of state 1, and the free energy difference between these states is referred to as the free energy of the addition of the protein restraints for the protein-ligand complex ΔG_r^{com}. Analogously, the free energy difference between state 3 and state 4 is referred to as the free energy of the addition of the protein restraints for the apo state ΔG_r^{apo}. The conformations of the state 3 are obtained from the conformations of the protein-ligand complex and, in general, do not exactly match conformations of apo protein (state 4) but the distributions of these states are overlapping (see grey and red points in Figure 1, state 4).
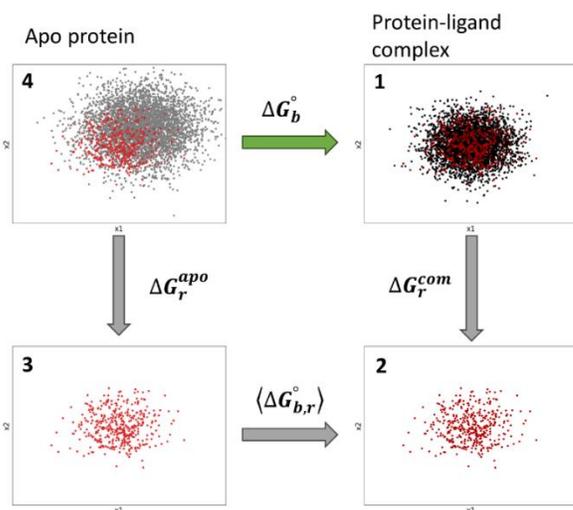
Figure 1. Thermodynamic cycle for unbiasing of the contribution of restraints in SBFE. Each state of the system (1-4) is illustrated by 2D plot where each point represents a single conformation. Axes x1 and x2 correspond to arbitrary coordinates of the conformational space. States 1 and 4 are the protein-ligand complex and the apo protein, respectively. Conformations of the apo protein and the protein-ligand complex are represented by grey and black points, respectively.

7

Conformations selected for MD TI calculations are represented in red. States 2 and 3 correspond to the set of conformations selected for MD TI calculations in the protein-ligand complex and the apo protein, respectively.

Calculation of both ΔG_r^{apo} and ΔG_r^{com} is based on the relationship between probabilities of the thermodynamic states of the system and the free energy difference between them. For two arbitrary states A and B, this relationship can be expressed as follows:

$$\Delta G_{A \to B} = G_B - G_A = -kT \ln \frac{P_B}{P_A}, \tag{3}$$

where P_A and P_B are the probabilities of the system to be in State A and B, respectively, and $\Delta G_{A \to B}$ is the free energy difference between these states.

Therefore, for any ensemble of conformations, we can define the free energy difference between the subset of conformations selected for MD TI calculations (S) and the rest of the conformations (rest):

$$\Delta G_{rest \to S} = G_S - G_{rest} = -kT \ln \frac{P_S}{P_{rest}} = = -kT \ln \frac{P_S}{1 - P_S}. \tag{4}$$

This equation is true for the ensemble of both the apo protein and the protein-ligand complex. Therefore, ΔG_r^{apo} can be calculated as the free energy difference between the subset of conformations selected for TI (S) and the full ensemble (F) :

$$\Delta G_r^{apo} = G_S^{apo} - G_F^{apo} = G_S^{apo} - \left(P_S^{apo} G_S^{apo} + \left(1 - P_S^{apo}\right) G_{rest}^{apo}\right) =$$

$$= \left(1 - P_S^{apo}\right)\left(G_{rest}^{apo} + \Delta G_{rest \to S}\right) - \left(1 - P_S^{apo}\right) G_{rest}^{apo} = \left(1 - P_S^{apo}\right) \Delta G_{rest \to S} =$$

$$= -kT \left(1 - P_S^{apo}\right) \ln \frac{P_S^{apo}}{1 - P_S^{apo}}, \tag{5}$$

where the subscripts S and F correspond to the subset of selected conformations and the full ensemble respectively, and P_S^{apo} is the probability of a conformation from the full ensemble (F) to be in the subset of the conformations selected for TI (S).

Analogously, ΔG_r^{com} can be calculated as follows:

8

$$\Delta G_{\text{r}}^{\text{com}} = -kT(1 - P_S^{com}) \ln \frac{P_S^{com}}{1 - P_S^{com}}. \tag{6}$$

It is important to note that conformations included in the subset S are selected only from the full ensemble for protein-ligand complex. Therefore, the conformations from the subset S themselves are absent in the full ensemble for the apo protein. However, both full ensembles can contain conformations with high similarity (see below) to those from the subset S. These conformations, together with the conformations from S, form a new subset $\tilde{S}$. Each of probabilities P_S^{apo} and P_S^{com} can be calculated as a share of the subset $\tilde{S}$ in the full ensemble F:

$$P_S^{apo} = \frac{N_{\tilde{S}}^{apo}}{N_F^{apo}}, \tag{7}$$

$$P_S^{com} = \frac{N_{\tilde{S}}^{com}}{N_F^{com}}, \tag{8}$$

where $N_{\tilde{S}}^{com/apo}$ and $N_F^{com/apo}$ are numbers of conformations in the subset $\tilde{S}$ and in the full ensemble correspondingly.

The similarity between a pair of conformations can be quantitatively characterized by the root mean square deviation (RMSD) of the specified atoms. Since the restraints are imposed on protein C_α atoms, it is reasonable to use these atoms for RMSD calculations. Also, since protein residues located relatively far from the ligand will not have a significant effect on the SBFE, the selection of atoms used for RMSD can be limited to only protein C_α atoms within a local proximity of the ligand. To select conformations with high similarity to a certain reference conformation, it is sufficient to select all conformations with RMSD not greater than a specified threshold. The subset $\tilde{S}$ consists of conformations which have RMSD within the threshold to at least one of the conformations from the subset S.

Therefore, our approach for SBFE calculation can be summarized in the following way:

1) perform unrestrained MD simulations for protein-ligand complex and apo protein;

9

2) select conformations from protein-ligand complex trajectory (subset S);

3) prepare truncated systems and compute SBFE $\Delta G_{b,r}^{\circ}$ for these conformations with restraints on protein C_α atoms;

4) average computed SBFEs over all snapshots to obtain ensemble averaged SBFE $\langle \Delta G_{b,r}^{\circ} \rangle$ (see eq 1);

5) calculate shares P_S^{apo} and P_S^{com} from the unrestrained MD trajectories using eq 7 and 8;

6) calculate ΔG_r^{apo} and ΔG_r^{com} using eqs 5 and 6;

7) calculate the final unbiased SBFE ΔG_b° using eq 2.

Steps 5-7 are independent from Steps 3-4 and can be performed directly after Step 2. The most time-consuming steps are Steps 1 and 3 because they include MD simulations. The total computational cost of this approach depends mainly on the simulation time of unrestrained MD simulations (Step 1), the number of conformations selected at Step 2, and computational cost of the MD TI simulation protocol (Step 3).

Methods

Standard Binding Free Energy Calculation

A thermodynamic cycle for the standard binding free energy (SBFE) calculation is shown in Figure 2. The free energy difference between a protein-ligand complex and a free protein and ligand in solution ($1 \rightarrow 2$: ΔG_b°) is calculated by completing an alchemical pathway between these two endpoint states. A key component of these simulations is the virtual bond, which is a set of harmonic restraints between three protein backbone heavy atoms and three ligand heavy atoms (see Figure S1). The application of the virtual bond is described below. SBFE simulations typically consist of three alchemical transformations: 1) within the protein-ligand complex, the addition of

10

a virtual bond between three protein residue backbone atoms and three ligand heavy atoms ($2 \rightarrow$ 4: $\Delta G_{\text{int}}^{\text{solv}}$); 2) within the protein-ligand complex with the virtual bond, the removal of electrostatic and van der Waals interactions either stepwise or simultaneously, often termed "annihilation", of the ligand ($4 \rightarrow 5$: $\Delta G_{\text{int}}^{\text{prot}}$); and 3) the annihilation of the free ligand in solution ($1 \rightarrow 3$: $\Delta G_{\text{int}}^{\text{solv}}$). As Boresch *et al.*[38] demonstrated, once the ligand with the virtual bond has been annihilated, the free energy of the virtual bond can be calculated analytically as the ligand does not interact with the environment ($5 \rightarrow 6$: $\Delta G_{-\text{VB}}^{\text{prot}^\circ}$); thus, after its removal, there is no free energy cost to remove the ligand from the binding pocket ($6 \rightarrow 3$). Thus, the standard binding free energy $\Delta G_{\text{b}}^\circ$ can be calculated as a sum of the terms described above by the following equation:

$$\Delta G_{\text{b}}^\circ = \Delta G_{\text{int}}^{\text{solv}} - \Delta G_{+\text{VB}}^{\text{prot}} - \Delta G_{\text{int}}^{\text{prot}} - \Delta G_{-\text{VB}}^{\text{prot}^\circ}, \tag{9}$$
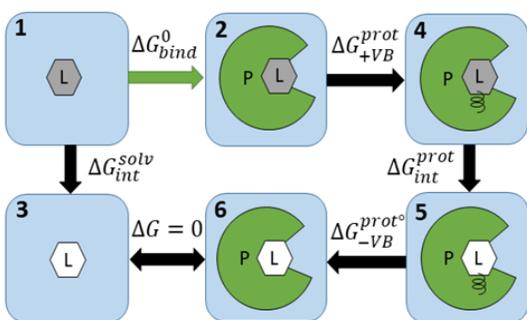
Figure 2. Alchemical thermodynamic cycle for standard binding free energy calculations. See the description in text.

The free energy of the removal of virtual bond between the annihilated ligand and the protein can be computed analytically using the following equation derived by Boresch et al.[38]:

$$\Delta G_{-\text{VB}}^{\text{prot}^\circ} = -kT \ln \left[\frac{8\pi^2 V^\circ}{r^2 \sin\theta_1 \sin\theta_2} \frac{\sqrt{K_r K_{\theta_1} K_{\theta_2} K_{\phi_1} K_{\phi_2} K_{\phi_3}}}{(2\pi kT)^3} \right], \tag{10}$$

where V° is the standard volume (1660 Å³), r, θ_1 and θ_2 are the distance and angles needed to restrain the ligand orientation with respect to protein (see Figure S1) and $K_r, \ldots, K_{\phi_3}$ are force constants of the corresponding harmonic potentials.

11

Other free energy terms ($\Delta G_{\text{int}}^{\text{solv}}$, $\Delta G_{\text{int}}^{\text{prot}}$, $\Delta G_{\text{+VB}}^{\text{prot}}$) can be computed by alchemical free energy calculation methods. One of the most common alchemical methods is thermodynamic integration (TI) which computes the free energy of transforming one state of molecular system into another via performing stratified MD simulations along a pre-defined reaction pathway. Typically, the pathway is defined as a linear interpolation between the potential energy functions of two states A and B:

$$V(\lambda) = \lambda V_A + (1 - \lambda)V_B, \tag{11}$$

where the coupled potential function $V(\lambda)$ defines the reaction coordinate, $V_{A/B}$ is the potential energy of state A/B correspondingly, and the parameter $\lambda \in [0, 1]$ defines the weights of the superposition. In TI, the free energy difference between endpoints A and B are derived by integrating the derivative of the coupled potential function with respect to λ. In practice, this is performed numerically:

$$\Delta G_{A \rightarrow B} = \int_0^1 \langle \frac{dV}{d\lambda} \rangle_\lambda \, d\lambda \cong \sum_i^N w_i \langle \frac{dV}{d\lambda} \rangle_i, \tag{12}$$

where $\langle \frac{dV}{d\lambda} \rangle_i$ is the average of the derivative of the coupled potential function over the MD simulation at window $\lambda = i$, w_i is the statistical weight of the window determined by the selected integration scheme, and N is the number of windows employed.

Molecular dynamics simulations

System preparation. MD simulations were performed for AMPAR TMD-LBD in complex with GYKI in crystal and docked (flipped) orientations. The X-ray crystal structures of the AMPAR in apo state (PDB: 5L1B) and in complex with GYKI (PDB: 5L1H)[36] were used as the starting structures for MD simulations. Docked poses of GYKI in each AMPAR subunit were taken from our previous work[26]. Apo protein and protein−ligand complexes were inserted into pre-

equilibrated membranes of POPC lipids and solvated with water and neutralizing ions. Parameters for MD were obtained using the FF99SB-ILDN force field[39] for the protein, the general AMBER force field (GAFF)[40] for the ligand, the AMBER Lipid14 force field[41] for lipids and the TIP3P model[42] for water using the tleap and antechamber programs from AmberTools[43].

MD simulations. GPU-accelerated MD simulations were carried out with AMBER 18 using the pmemd.cuda program. A simulation time step of 2 fs was used and all hydrogen bonds were constrained via the SHAKE algorithm[44]. Periodic boundary conditions were applied in all directions with a cutoff radius of 10 Å. Electrostatic interaction calculations were performed using the Particle Mesh Ewald (PME) method[45]. A Langevin thermostat and a Berendsen barostat[46], as implemented in AMBER, were used to maintain the temperature and pressure, respectively. All simulations were carried out in the NPT ensemble using anisotropic scaling. Each system was equilibrated at 1 atm and 300 K with gradually decreasing harmonic restraints on the protein and ligands over 60 ns. Production simulations were carried out for approximately 450 ns for the AMPAR in apo state, 600 ns for AMPAR in complex with the crystal orientation of GYKI and 1100 ns for AMPAR with the docked orientation of GYKI.

Standard binding free energy calculations using thermodynamic integration (TI)

System preparation. For TI simulations of the solvated ligand, the ligand was solvated in a TIP3P water box with a 15 Å distance between the ligand and the edge of the box using tleap[43]. For TI simulations of the protein-ligand complex, an automated procedure including several steps was developed (see Figure 3). First, 100 frames from a 500 ns equilibrium portion of AMPAR-GYKI trajectories were extracted equidistantly (the offset between frames was 5 ns). The extracted frames were then used to prepare truncated systems for TI simulations for each AMPAR subunit. Each

13

truncated system included the ligand, the protein and water within 15 Å of the ligand, and the seven closest lipid fragments to the ligand. All other atoms of the extracted frames were removed. The truncated systems were then solvated with 12000 TIP3P water molecules with a minimum distance of 8 Å between the ligand and the edge of the box and neutralized with Na^+/Cl^- ions. 400 truncated systems were prepared for each AMPAR-GYKI-C and AMPAR-GYKI-D systems.

MD TI simulations. Standard binding free energies were computed for each truncated system using TI based on the alchemical thermodynamic cycle (Figure 1). GPU-accelerated MD TI simulations were performed using the AMBER 20 pmemd.cuda program[47]. To compute ΔG_{int}^{solv} and ΔG_{int}^{prot}, a λ-schedule of 11 equally distributed windows were used (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). The second-order smoothstep softcore potential (SSC(2)) was utilized as implemented in AMBER 20[48]. The orientation of the ligand with respect to the protein was restrained using the virtual bond approach[38]. Force constants of 10 kcal/(mol*Å^2) and 200 kcal/(mol*rad^2) were used for distance and both angle and dihedral angle restraints, respectively. For the addition of the virtual bond restraints (ΔG_{+VB}^{prot}), 7 unequally distributed λ-windows were used (0.0, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0). For each λ-window, the system was minimized with 2000 steps of steepest descent, then heated for 50 ps in NVT ensemble and then equilibrated for 300 ps in NPT ensemble using 2 fs timestep. Production simulations were done for 3 ns with 4 fs timestep to reduce the computational cost. Topologies for production simulations were modified by hydrogen mass repartitioning using the parmed program[43]. To retain the protein conformation in the binding pocket and prevent the system from falling apart, protein C_α atoms and all carbon atoms of the lipids were restrained at their initial positions with a force constant of 40 kcal/mol in all TI simulations. Free energies were obtained by numerical integration of average dV/dλ using the trapezoid rule (eq 12). Standard errors were calculated using the alchemlyb python library[50].

14

The free energy of adding virtual bond restraints for the non-interacting ligand was calculated using the Boresch formula[38] (eq 10). The total SBFEs were computed for each AMPAR subunit using Equation 9. The total standard errors were computed as a square root of a sum of squares of standard errors of individual free energy terms included in Equation 9.
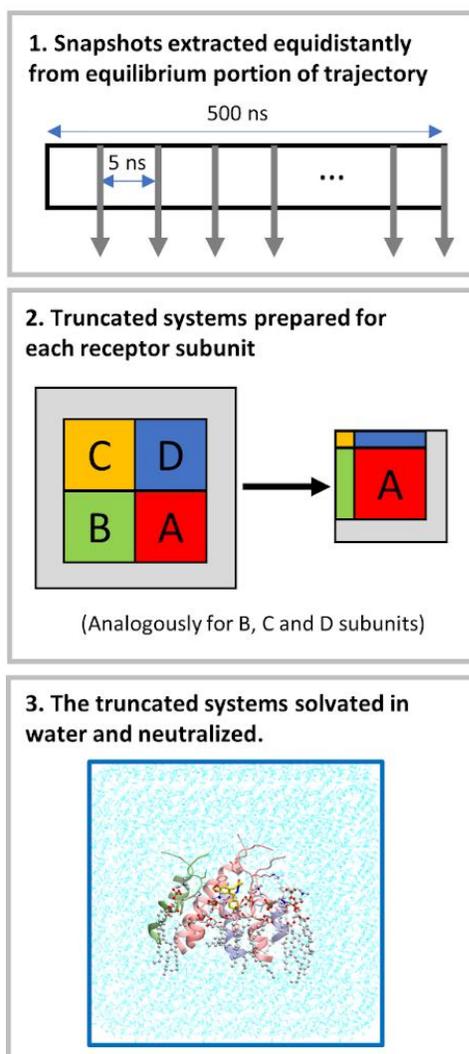
Figure 3. The automated procedure for preparation of truncated systems.

Clustering of AMPAR binding pocket conformations. Average linkage hierarchical clustering[51] was performed for the conformations extracted from the 500-ns equilibrium portions of the MD trajectories using the cpptraj program[52] from AmberTools[43]. The RMSD of C_α atoms of protein

15

residues within 8 Å from the ligand was used as a distance metric. The Davies-Bouldin Index (DBI){Davies, 1979 #584} and the pseudo-F statistic (pFS){Milligan, 1985 #586} were used as clustering performance metrics. Clustering was performed for the distance threshold varying from 0.5 to 2.0 Å with the increment of 0.1 Å. The DBI and pFS were computed for each distance threshold (see Figure S2). The threshold of 1.6 Å was selected as optimal since it provided the minimum of DBI, and the maximum of pFS. The clustering results obtained for this threshold were used for further analysis.

Results and Discussion

Approach performance for a benchmark system

To evaluate the performance of our approach for computing SBFEs (see Approach section), we performed the corresponding calculations for the T4 lysozyme L99A/M102Q mutant in complex with its small-molecule inhibitor, N-phenylglycinonitrile (PDB ID: 2RBN{Graves, 2008 #587}; Figure 4A). Lysozyme is a 14 kDa enzyme with antibacterial activity which occurs in high concentrations in egg whites as well as secretions including tears, saliva, and milk. It catalyzes the hydrolysis of 1,4-beta-linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine residues in peptidoglycan, which is the major component of gram-positive bacterial cell walls. The lysozyme from bacteriophage T4 was previously investigated extensively as a model system to study protein-ligand binding[12]. For most ligands, no substantial protein motions occur upon ligand binding, allowing for SBFE calculations to reach convergence. Furthermore, the size of the protein is relatively small compared to other proteins used in computational studies of protein-ligand binding. For these reasons, lysozyme is used commonly as a benchmark system for protein-ligand binding free energy calculations[53]. Mutation of methionine to glutamine in T4 lysozyme

16

L99A/M102Q mutant introduces polarity and a potential for hydrogen bonding in an originally nonpolar binding site. This enables small polar ligands to bind to the binding pocket while the protein can still bind a variety of nonpolar ligands such as toluene. The SBFE of N-phenylglycinonitrile to this protein was previously determined both experimentally and computationally[13] which allows for a direct estimation of the accuracy of our calculations.
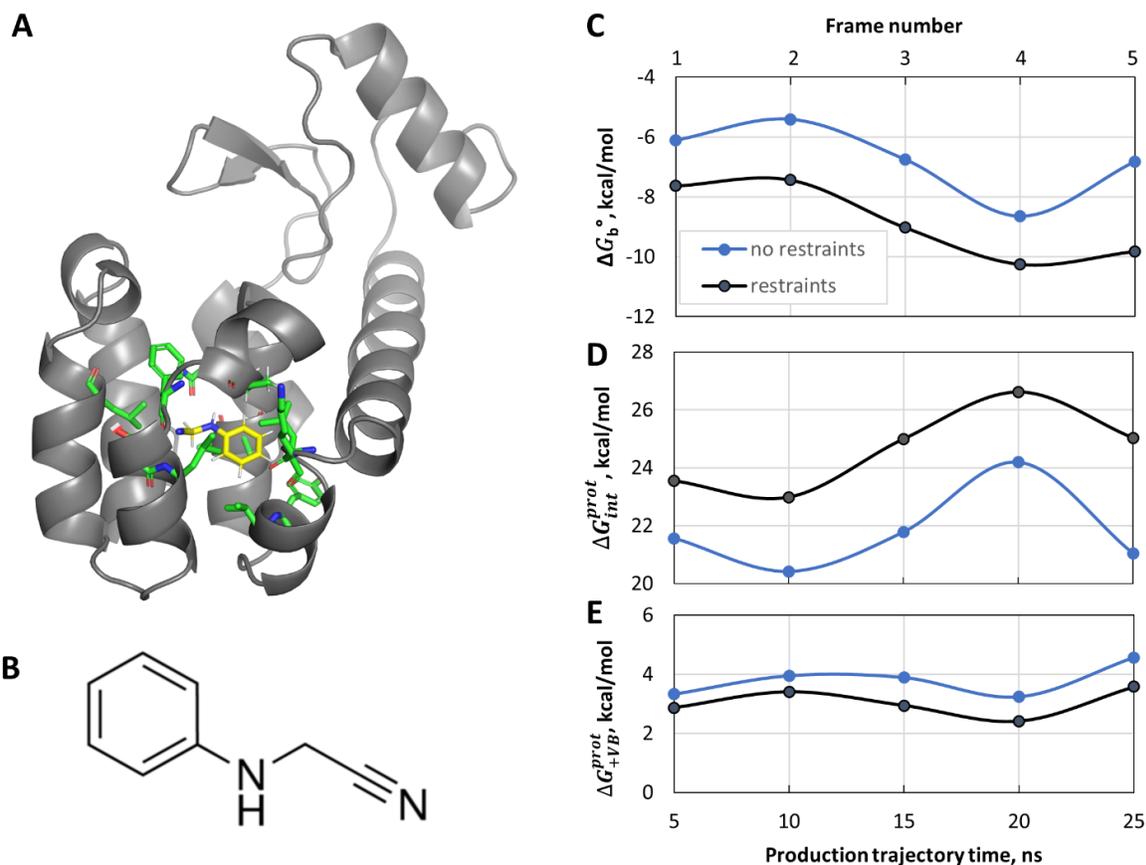
Figure 4. Structure and SBFE calculations for T4 lysozyme L99A/M102Q mutant in complex with N-phenylglycinonitrile. **A.** Equilibrated structure of the protein-ligand complex (PDB ID: 2RBN). The protein is shown by gray cartoon, the ligand is shown by yellow sticks, and the residues within 4 Å from the ligand are shown by green sticks. **B.** Structure of N-phenylglycinonitrile. **C-E.** SBFE (**C**) and its components ΔG_{int}^{prot} (**D**) and ΔG_{+VB}^{prot} (**E**) for the five selected snapshots computed with and without restraints on protein C_α atoms (shown by black and blue, respectively). Each snapshot is labeled by the time it occurred in the equilibrium portion of MD production trajectory of the protein-ligand complex (X axis).

17

Following our approach for SBFE calculation (see Approach section) we performed MD simulations of lysozyme in ligand-bound and apo states and used 25 ns equilibrium portion of each MD trajectory for the following calculations and analysis (see Methods section for details). After that, we selected five snapshots equidistantly with the offset of 5 ns from the MD trajectory for protein-ligand complex. Each snapshot was then computed with two MD TI protocols: with and without restraints on protein C_α atoms. For consistency, all other parameters of both MD TI protocols were the same as for the protocol used for AMPAR – GYKI complex. Comparison of the computed SBFEs for the selected snapshots are presented in Figure 4C. Our results demonstrate that SBFEs computed with restraints on protein C_α atoms are systematically lower than those computed without these restraints. The trajectory averaged SBFEs are -8.83 and -6.75 kcal/mol for calculations with and without the restraints, respectively. These results suggest that the presence of the restraints on protein C_α atoms during MD TI simulations decreases the averaged SBFE by ~2.1 kcal/mol. Given that the experimental SBFE is -5.5 kcal/mol for this system[54], the unrestrained MD TI simulations provide more accurate estimation of the SBFE.

SBFEs computed with and without the restraints as well as their difference vary for distinct snapshots (Figure 4C). This suggests that the effect of restraints on the SBFE depends on the structure of the system. Decomposition of computed SBFEs into free energy terms according to the TI thermodynamic cycle(see Figure 6 and eq 9) allows for a more detailed inspection of these results (see Table 1 and Figure 4D, E). The free energy of annihilating the ligand in the protein-ligand complex, ΔG_{int}^{prot}, is a major contributor to the difference between SBFEs computed with and without the restraints. The presence of restraints increases this term by ~2.8 kcal/mol on average compared to unrestrained MD TI simulations. In contrast, the free energy of the addition of the virtual bond, ΔG_{+VB}^{prot}, is lower by ~0.7 kcal/mol on average for restrained MD TI simulations

18

compared to those without restraints. Tree energy of the removal of the virtual bond ΔG_{+VB}^{prot} and of annihilation of the ligand in water ΔG_{int}^{solv} does not depend on the presence of the protein restraints: the former is computed analytically (see eq 10) and depends only on the reference orientation of the ligand which is the same for both restrained and unrestrained MD TI simulations; the latter is computed in the absence of the protein.

Table 1. SBFE and its components in kcal/mol for the five selected snapshots computed with and without restraints on protein C_α atoms. The presence and absence of restraints is denoted as "r." and "no r." respectively. N is the conformation number. The SBFE (ΔG_b°), the free energy of annihilating the ligand in the protein-ligand complex and solvent (ΔG_{int}^{prot} and ΔG_{int}^{solv}), and the free energy of addition and removal of the virtual bond (ΔG_{+VB}^{prot} and $\Delta G_{-VB}^{prot\circ}$) are related as follows: $\Delta G_b^\circ = \Delta G_{int}^{solv} - \Delta G_{+VB}^{prot} - \Delta G_{int}^{prot} - \Delta G_{-VB}^{prot\circ}$

N	time, ns	ΔG_b°		ΔG_{int}^{prot}		ΔG_{+VB}^{prot}		$\Delta G_{-VB}^{prot\circ}$	ΔG_{int}^{solv}
		no r.	r.	no r.	r.	no r.	r.		
1	5	-6.11	-7.63	21.56	23.55	3.32	2.86	-10.82	7.96
2	10	-5.41	-7.44	20.42	22.98	3.94	3.41	-10.99	7.96
3	15	-6.76	-9.03	21.78	25.00	3.88	2.94	-10.95	7.96
4	20	-8.65	-10.25	24.19	26.62	3.23	2.41	-10.81	7.96
5	25	-6.82	-9.82	21.05	25.03	4.56	3.57	-10.83	7.96
Average		6.75(54)	8.83(57)	21.8(64)	24.64(64)	3.79(24)	3.04(21)	-10.88(4)	7.96
Difference		2.09(78)		2.84(91)		0.75(32)		0	0
ΔG_b°		6.75(54)	7.56(57)						
Difference		0.81(78)							

According to our approach, the free energy terms ΔG_r^{apo} and ΔG_r^{com} should be calculated to obtain the total unbiased SBFE (see eq 2). To calculate these terms, we performed unrestrained MD

simulations for the lysozyme mutant in the apo form and used its 25-ns equilibrium portion for further analysis. The RMSD of C_α atoms of the protein residues of the binding pocket during equilibrium MD trajectories with respect to the equilibrated structure of protein-ligand complex was comparable for both the apo protein and the protein-ligand complex (Figure 5A). The binding site residues of the complex deviated slightly less from the reference structure compared to the apo protein in the first 17 ns of the equilibrium MD trajectory. However, they shifted during the 17th ns of the trajectory and the RMSD remained at the same level for both systems from 20th to 25th ns of the trajectory. Pairwise RMSDs indicated that the binding site conformations of the complex are overall more similar to those from the same system than to the conformations of the apo protein (Figure 5B). Conformations from 18th-21th ns of the complex trajectory differ significantly from both the rest of complex conformations and apo conformations. While PCA did not distinguish these conformations from the rest of the conformations (Figure 5D), t-SNE projection separated these conformations from the rest of the structures (Figure 5C). The computed SBFE for a conformation from this subset (N = 4 in the Table 1) is the lowest compared to the rest of conformations. Therefore, a significant difference in SBFE for this conformation is apparently due to a shift in the backbone of the binding pocket.

The calculation of ΔG_r^{apo} and ΔG_r^{com} requires the probabilities P_S^{apo} and P_S^{com} (see eqs 5 and 6). To calculate these probabilities, we estimated shares of apo and complex conformations with RMSD not greater than 0.5 Å with respect to the conformations with computed SBFE (see eq 7 and 8). This resulted in P_S^{apo} and P_S^{com} of 0.0076 and 0.055, respectively. Calculated ΔG_r^{apo} and ΔG_r^{com} values were 2.85 and 1.58 kcal/mol, respectively, and hence the total unbiased SBFE was -7.56 kcal/mol according to eq 2. Therefore, the total unbiased SBFE is significantly closer to the ensemble-average unrestrained SBFE (-6.75 kcal/mol; see Table 1) compared to the ensemble-

averaged restrained SBFE. Specifically, the unbiasing allowed for accounting 1.27 kcal/mol out of 2.09 kcal/mol difference between restrained and unrestrained SBFEs. At the same time, while the unbiased SBFE is still 0.81 ± 0.78 kcal/mol lower compared to the unrestrained SBFE, the absolute value of this difference is comparable with its error.
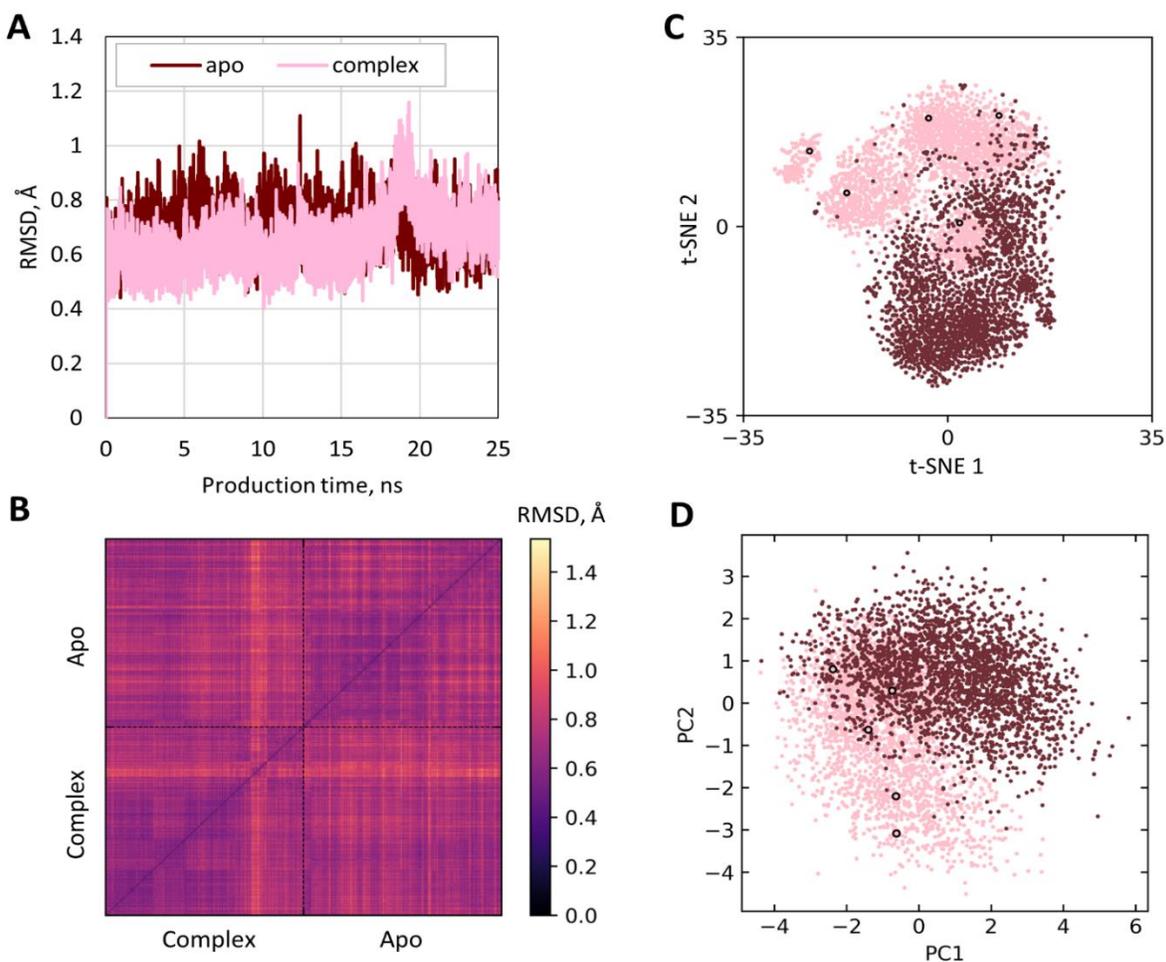
Figure 5. MD simulations for ligand-bound (shown in pink) and apo (shown by dark red) states of the lysozyme mutant. **A**. RMSD of protein C_α atoms of the binding pocket versus MD production simulation time with respect to the equilibrated structure of the protein-ligand complex. **B**. Matrix of pairwise RMSD of protein C_α atoms of the binding pocket for the frames of production MD trajectories of protein-ligand complex and apo protein. **C-D**. Frames from MD production trajectories projected onto 2D space using t-SNE (**C**) and PCA (**D**). Frames for which SBFE was computed by MD TI simulations are encircled in black.

AMPAR-GYKI complex

The AMPAR homotetramer consists of four identical subunits referred to as A, B, C and D. Each subunit is comprised of multiple domains including the ligand-binding domains (LBD) and the transmembrane domain (TMD). The ion channel pore is formed by TMD α-helixes from four subunits of the tetramer. The allosteric binding pocket of each AMPAR subunit (see Figure 6) is located at the interface between the TMD and linkers connecting it to the LBD. The inhibitors bound to this allosteric site are assumed to act as wedges stabilizing the closed state of the receptor and preventing conformational rearrangements required for the ion channel opening[36].
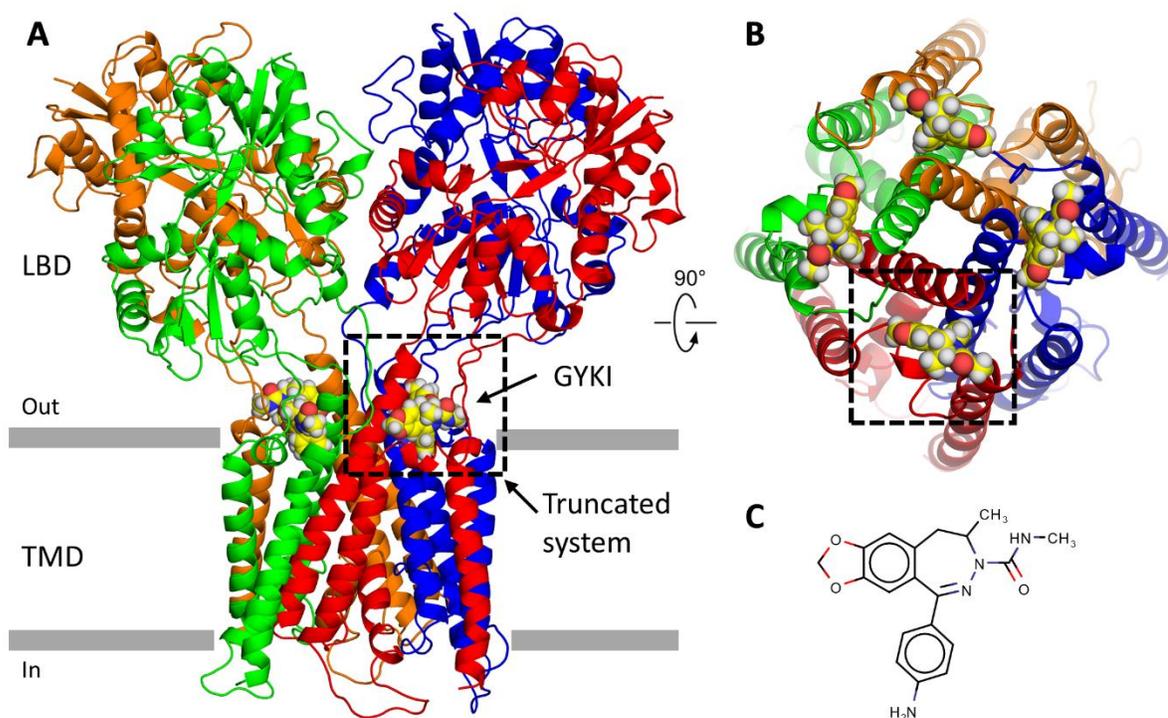
Figure 6. Structure of AMPA receptor in complex with GYKI. **A**. Crystal structure of AMPAR TMD-LBD system in complex with GYKI (PDB ID: 5L1H) presented parallel to membrane. The inner and outer sides of membrane are indicated by parallel gray bars. The protein is presented by a cartoon with each subunit shown in a different color. GYKI molecules are shown in space-filling representation. A truncated system is highlighted by a black dashed rectangular outline. **B**. Top view of AMPAR-GYKI binding sites. The style and colors of representations are consistent with Figure 1 A. **C**. Chemical structure of GYKI.

22

## MD simulations of AMPAR in complex with GYKI and in the apo state

The AMPAR allosteric binding pocket of each subunit is formed by the residues of TMD α-helixes (M1, M3 and M4), short α-helical fragment preM1, and TMD-LBD linkers S1M1 and S2M4 (Figure 7). Importantly, the residues of S2M4 linker of the neighboring subunit can also interact with the ligand in both binding poses.
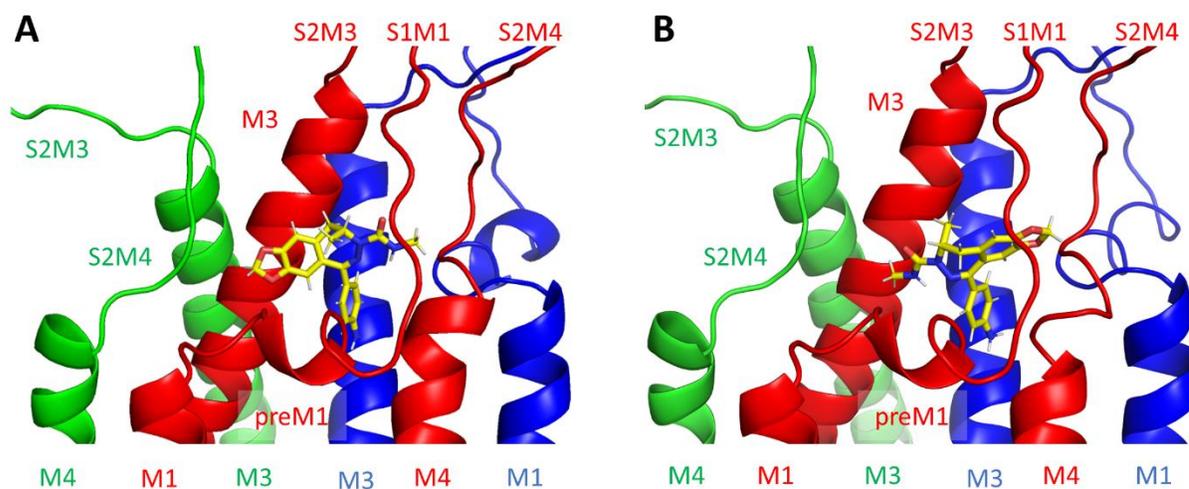
Figure 7. Binding modes of GYKI in AMPAR allosteric pocket. Initial minimized structures of crystal (**A**) and docked (**B**) poses of GYKI in the binding pocket of AMPAR subunit A. Subunits A, B, and D are shown by red, green, and blue cartoon, respectively. Protein fragments forming the allosteric binding pocket are labeled. The ligand bound to subunit A is shown by yellow sticks. Ligands bound to the neighboring subunits are not shown.

MD simulations of the AMPAR-GYKI truncated complexes performed previously[26] included the transmembrane domains and the TMD-LBD linker peptides loosely restrained at their ends. An absence of the LBD domains and constraints at the points of truncation may have influenced protein dynamics as well as the ligand binding pose energy. Here we carried out long equilibrium unrestrained MD simulations of the full length TMD-LBD AMPAR in complex with the crystal and docked (flipped) GYKI poses. Each simulated system included four molecules of GYKI bound to distinct AMPAR subunits, as well as the lipid bilayer, water and counter ions to represent a nearly physiologically native environment. An equilibrium MD trajectory for the apo form

23

AMPAR TMD-LBD system (450 ns) was also used for further analysis. To assess overall stability of the AMPAR-GYKI complex in the equilibrium MD simulations, we computed an RMSD of the protein C_α atoms and heavy atoms of the ligand with respect to the initial minimized structure (Figure 8). The GYKI crystal pose was relatively stable for all subunits, with RMSD fluctuating around ~2-3 Å depending on the subunit. In contrast, the GYKI docked pose was relatively stable only for subunit B, while for subunits A, C, and D the ligand shifted significantly after the first ~50-200 ns of the MD simulation, retained the new pose for ~50-400 ns, then shifted again, and stabilized at this pose for the rest of the MD trajectory (Figs. 8, A and D). The time needed for the docked pose to stabilize varied significantly between subunits. The protein backbone also shifted more significantly for the AMPAR-GYKI docked complex compared to the crystal-like complex (Figs. 8, B and E). To characterize deviations of distinct residues we computed the RMSD per residue C_α atom over all frames of the MD trajectory (time-averaged RMSD; Figs. 6, C and F). For the crystal system, the residues of the TMD α-helices M1, M3 and M4 (see Figure 9) exhibited less deviation from the initial structure compared to the residues of the TMD-LBD linkers and LBD. For the docked system, the deviation of the residues of S1M1, preM1, M1 and M4 are larger compared to those in the crystal system; the deviations of the residues of M3 are similar or slightly larger. Thus, the residues of the AMRAR allosteric pocket shifted from their initial positions more significantly for the GYKI docked pose compared to the GYKI crystal pose.
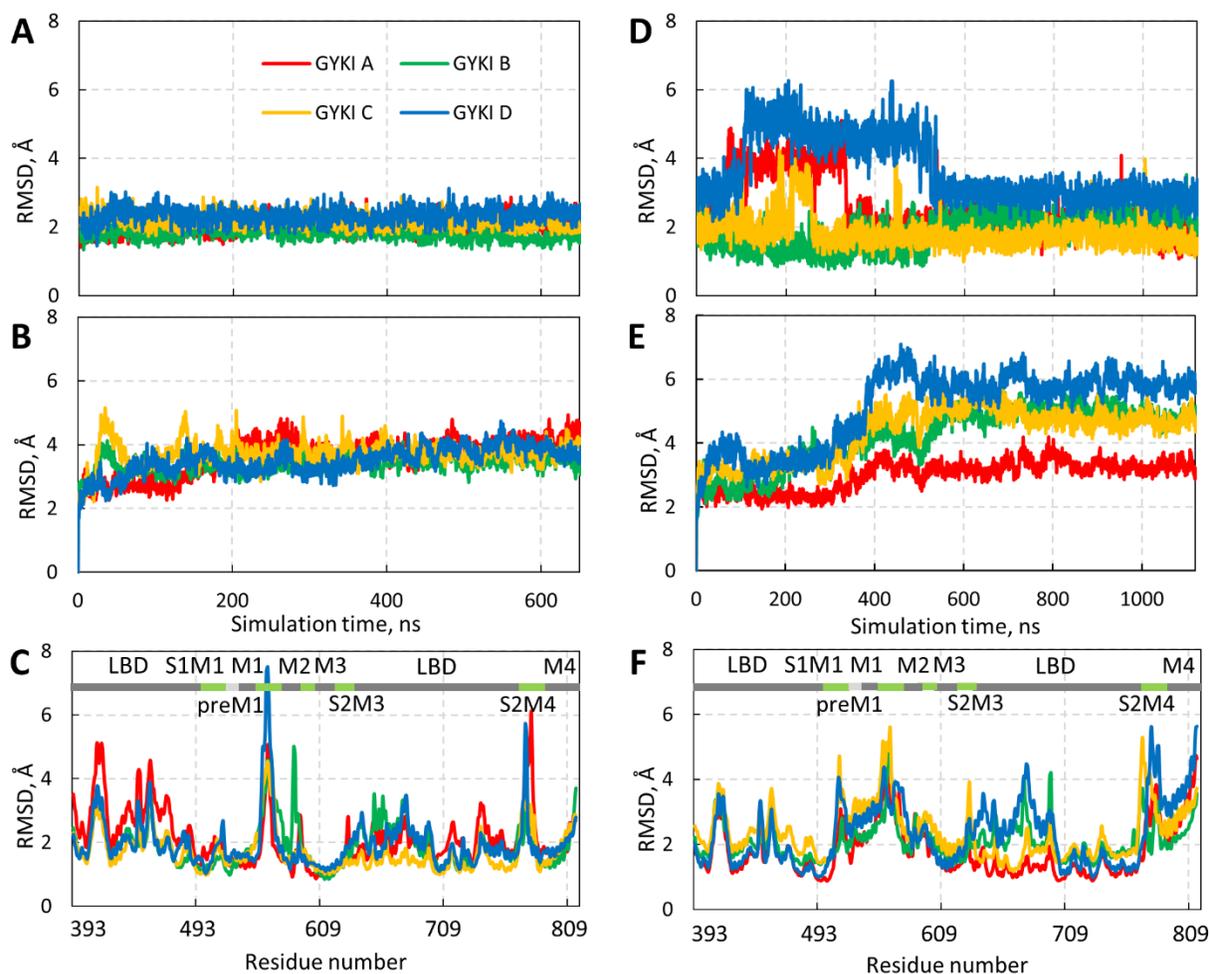
Figure 8 Ligand RMSD and protein RMSD during equilibrium MD simulations for the crystal (left column, **A-C**) and docked (right column, **D-F**) poses of GYKI. AMPAR subunits A, B, C, and D are shown by red, green, yellow, and blue, respectively. **A**, **D**: RMSD of ligand heavy atoms with respect to the initial minimized structure versus simulation time. **B**, **E**: RMSD of protein C_α atoms with respect to the initial minimized structure versus simulation time. **C**, **F**: Time-averaged RMSD of protein C_α atoms with respect to the initial minimized structure. Protein fragments are labeled and highlighted by color: LBD and TMD shown in dark gray, preM1 shown in light gray and TMD-LBD linkers shown in light green.

For further analysis and SBFE calculations we used 500 ns equilibrium portions of the MD trajectories: 100 - 600 ns for the crystal system and 600 - 1100 ns for the docked system. The corresponding representative structures were also extracted from these equilibrium trajectories (Figure 9A, B). Figure 9C shows root mean square fluctuation (RMSF) of the binding pocket residue C_α atoms during the equilibrium MD trajectories. The residues of the preM1 are more

stable compared to those of the TMD-LBD linkers for all systems (RMSFs are ~0.5-1 Å). The RMSF of the residues of S2M3 and S2M4 increases with distance from the M3 and M4, respectively (M4 follows S2M4). However, for S1M1 this is true only for the crystal system, while in the docked and apo systems the largest fluctuations are observed for S510 and K511 that are only 2-3 residues from the preM1. The C-terminal residues of S1M1 (residues 510-513), the full preM1 and most residues of S2M4 (residues 784-790) are more stable in the crystal system compared to the apo system. This stabilization is apparently due to interactions of these residues with ligand. The same effect is partly observed in the docked system, but the corresponding residues are less stable compared to the crystal system. This trend is not observed for S2M3 but the residues of this region are located further from the ligand compared to those of S1M1, preM1, and S2M4.

Calculation of the ensemble-averaged standard binding free energy of GYKI to AMPAR

To quantitatively characterize binding of GYKI to AMPAR, we computed the SBFE for both crystal and docked poses of GYKI. Following our approach (see Approach section) for each pose we selected 100 snapshots from the equilibrium portion of AMPAR-GYKI trajectory (500 ns) equidistantly with the offset of 5 ns, prepared truncated systems for individual subunits (see Methods section for the details) and performed MD TI simulations for all the systems. Therefore, SBFEs were computed for 800 systems in total (100 snapshots * 4 subunits * 2 ligand binding poses). Results of SBFE calculations are presented in Figure 10 and Table 2.

26
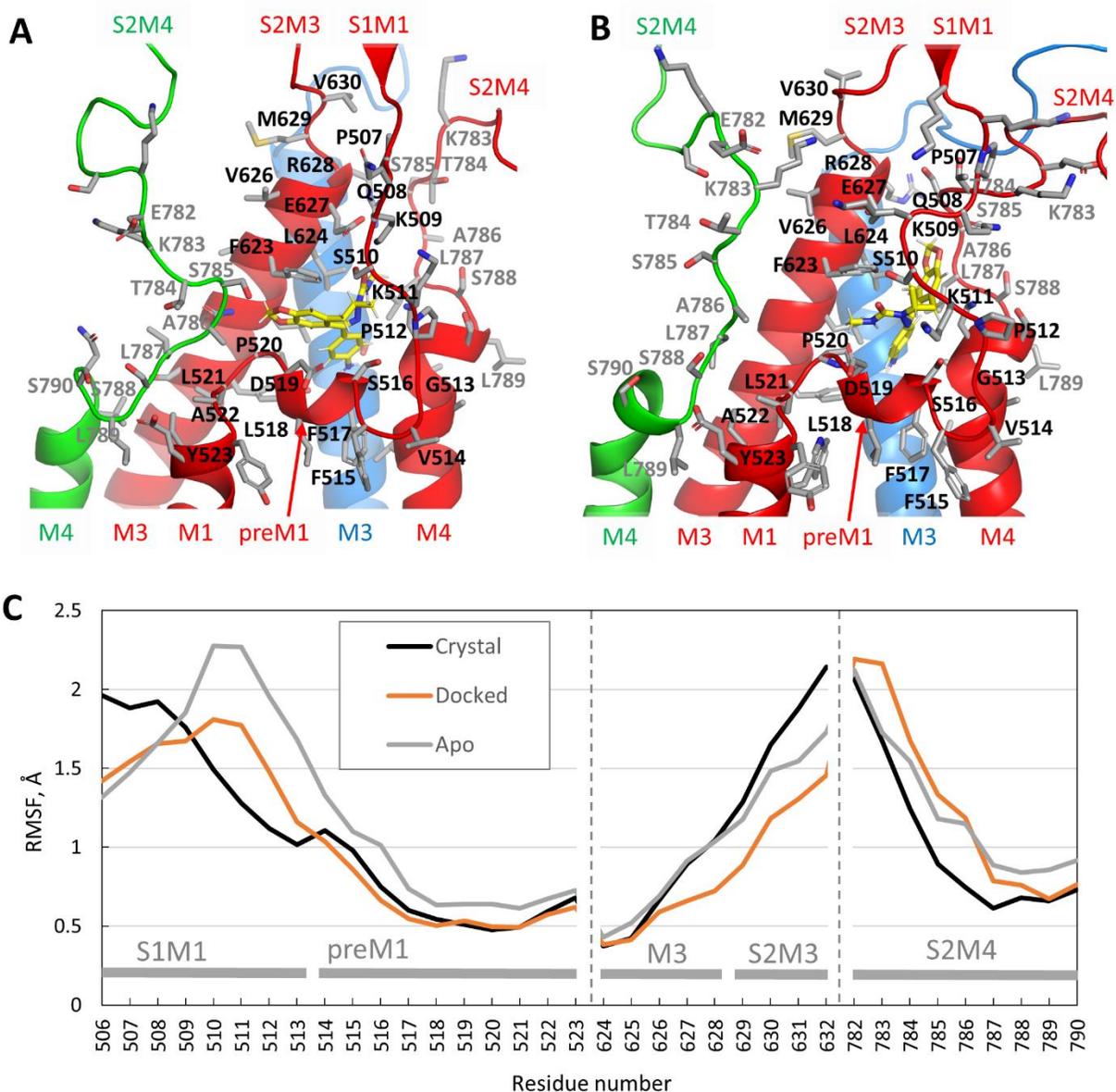
Figure 9. Representative structures of AMPAR-GYKI complex and RMSF of protein residues of the AMPAR allosteric binding pocket. **A-B**. Representative structure of GYKI in the allosteric pocket of subunit A in the crystal (**A**) and the docked (**B**) pose. Colors and representations are consistent with Figure 7. The residues of the S2M4 are labeled in grey and other residues are labeled in black. **C**. RMSF of protein residues of the AMPAR the allosteric binding pocket averaged over subunits for AMPAR in complex with the GYKI crystal pose, the docked pose and in the apo form. The RMSF is calculated on protein C$_\alpha$ atoms using equilibrium portions of the MD trajectories.
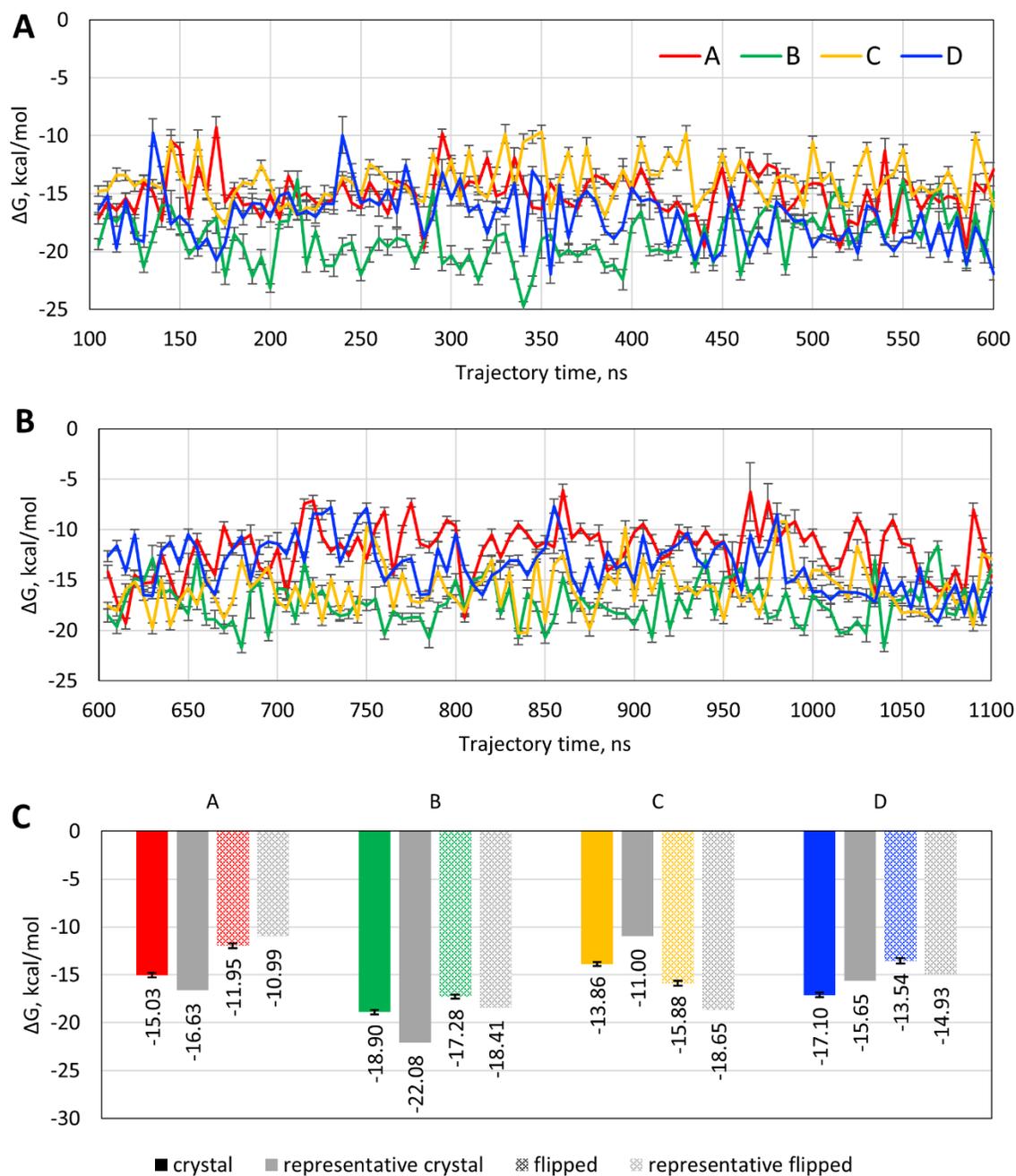
Figure 10. Standard binding free energies computed for truncated systems of AMPAR-GYKI complex for individual subunits. **A-B**. SBFEs vs production trajectory time for the GYKI crystal pose (**A**) and the GYKI docked pose (**B**). **C**. Ensemble-averaged SBFEs and SBFEs of representative structures of AMPAR-GYKI complex.

Table 2. Standard binding free energies of GYKI to AMPAR (kcal/mol). $\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$ and $\Delta G^\circ_{\mathrm{b,r}}(\mathrm{rep})$ are the ensemble-averaged SBFE and the SBFE for representative poses, respectively, and $\Delta\Delta G^\circ_{\mathrm{b,r}} = \Delta G^\circ_{\mathrm{b,r}}(\mathrm{rep}) - \langle\Delta G^\circ_{\mathrm{b,r}}\rangle$. "Difference" corresponds to the difference between SBFEs for the docked and crystal GYKI pose. "A"-"D" corresponds to ensemble-averaged SBFEs for individual subunits. "Average" corresponds to SBFEs averaged over all subunits. $\Delta G^\circ_{\mathrm{b}}$ corresponds to the total unbiased SBFEs obtained according to eq 2.

AMPAR subunit	Crystal			Docked			Difference	
	$\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$	$\Delta G^\circ_{\mathrm{b,r}}(\mathrm{rep})$	$\Delta\Delta G^\circ_{\mathrm{b,r}}$	$\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$	$\Delta G^\circ_{\mathrm{b,r}}(\mathrm{rep})$	$\Delta\Delta G^\circ_{\mathrm{b,r}}$	$\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$	$\Delta G^\circ_{\mathrm{b,r}}(\mathrm{rep})$
A	-15.03(2)	-16.63(58)	-1.60(61)	-11.95(27)	-10.99(46)	0.96(53)	3.08(33)	5.64(74)
B	-18.90(22)	-22.08(58)	-3.18(62)	-17.28(22)	-18.41(41)	-1.13(47)	1.62(32)	3.67(71)
C	-13.86(19)	-11.00(53)	2.86(56)	-15.88(24)	-18.65(34)	-2.77(42)	-2.02(31)	-7.65(63)
D	-17.10(24)	-15.65(39)	1.45(55)	-13.54(26)	-14.93(43)	-1.39(50)	3.56(35)	0.72(58)
Average	**-16.22(14)**	**-16.34(27)**	**-0.12(27)**	**-14.66(16)**	**-15.75(79)**	**-1.08(79)**	**1.56(21)**	**0.60(24)**
$\Delta G^{\mathrm{apo}}_{\mathrm{r}}$	0.57			2.10			1.53	
$\Delta G^{\mathrm{com}}_{\mathrm{r}}$	-0.02			-0.01			-0.01	
$\Delta G^\circ_{\mathrm{b}}$	**-15.63(14)**			**-12.55(16)**			**3.08(21)**	

Computed SBFEs vary significantly between different poses and different subunits (Table 2). The difference between the lowest and the highest ensemble-averaged SBFE ($\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$) for a subunit is around 5 kcal/mol for both crystal and docked poses of GYKI. The lowest $\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$ is for subunit B for both GYKI poses. Subunits A, B and D are ranked the same for both GYKI poses: $\langle\Delta G^\circ_{\mathrm{b,r}}(\mathrm{B})\rangle < \langle\Delta G^\circ_{\mathrm{b,r}}(\mathrm{D})\rangle < \langle\Delta G^\circ_{\mathrm{b,r}}(\mathrm{A})\rangle$. However, the rank of subunit C is not consistent between GYKI poses: $\langle\Delta G^\circ_{\mathrm{b,r}}(\mathrm{C})\rangle$ is the highest SBFE for the GYKI crystal pose and the second lowest SBFE for the GYKI docked pose. For all subunits, the difference in $\langle\Delta G^\circ_{\mathrm{b,r}}\rangle$ between crystal and docked pose of

29

GYKI is greater than 1.6 kcal/mol in absolute value. Binding of GYKI in the crystal pose is more favorable compared to the docked pose for all subunits except subunit C. Overall, our results indicate that binding of GYKI in crystal pose is more favorable by ~1.6 kcal/mol on average. The standard error of this quantity is relatively small (~0.2 kcal/mol) which allows us to discriminate between the two binding modes of GYKI.

Furthermore, SBFEs computed for representative poses ($\Delta G_{b,r}^{\circ}(\text{rep})$) differ significantly (more than 0.9 kcal/mol in absolute value) from ensemble-averaged SBFEs ($\langle\Delta G_{b,r}^{\circ}\rangle$) for all subunits. There is no clear trend in these differences: for GYKI crystal pose, $\Delta G_{b,r}^{\circ}(\text{rep})$ are lower then $\langle\Delta G_{b,r}^{\circ}\rangle$ for subunits A and B, and higher for subunits C and D; for the GYKI docked pose, $\Delta G_{b,r}^{\circ}(\text{rep})$ are lower than $\langle\Delta G_{b,r}^{\circ}\rangle$ for subunits B, C, and D, and higher for the subunit A. Despite these differences for individual subunits, average $\Delta G_{b,r}^{\circ}(\text{rep})$ and $\langle\Delta G_{b,r}^{\circ}\rangle$ match well for the GYKI crystal pose with a difference of only -0.12 kcal/mol. However, this is not observed for the GYKI docked pose where average $\Delta G_{b,r}^{\circ}(\text{rep})$ is lower than average $\langle\Delta G_{b,r}^{\circ}\rangle$ by -1.08 kcal/mol, although this discrepancy lies within the standard error of $\Delta G_{b,r}^{\circ}(\text{rep})$ of ~1.8 kcal/mol. The standard errors are relatively large for average $\Delta G_{b,r}^{\circ}(\text{rep})$ because of small sample sizes.

Differences in $\Delta G_{b,r}^{\circ}(\text{rep})$ between poses for individual subunits are consistent with those in $\langle\Delta G_{b,r}^{\circ}\rangle$ by sign but differ significantly by their absolute values. Overall SBFE calculations for representative structures underestimated the difference in average SBFE between ligand poses: for representative structures, this quantity is 0.6 kcal/mol which is ~1 kcal/mol less than the difference obtained from ensemble-averaged SBFEs (see Table 2, "Average" row, last two columns). Given that the standard error in this quantity is relatively large (~3 kcal/mol), SBFE calculations for

representative poses do not allow us to distinguish between crystal and docked poses of GYKI in contrast to ensemble-averaged SBFE calculations.

Calculation of the free energy of restraints and total unbiased standard binding free energy of GYKI to AMPAR

Ensemble-averaged SBFE $\langle \Delta G_{b,r}^{\circ} \rangle$ were computed using MD TI simulations with restraints on the protein C_α atoms. Hence, to obtain final SBFEs we need to account for the free energy cost of these restraints in apo (ΔG_r^{apo}) and complex (ΔG_r^{com}) states. To compute these terms, we performed clustering of all conformations from equilibrium portions of the MD trajectories for the crystal, docked and apo systems using average linkage hierarchical clustering (see Methods for details) which resulted in identifying 12 clusters (Figure 11D). The RMSD of C_α atoms of residues within 8 Å from the ligand was used as a distance metric for clustering. The linkage distance threshold was selected based on comparison of the clustering metrics for different thresholds (see Methods and Figure S2) and set to 1.6 Å.

In general, most of conformations from the same subunit were included in the same cluster. Subunit C of the apo system was the only subunit which formed two roughly equal clusters. The RMSD between conformations of the same subunit is significantly lower compared to the RMSD between conformations from the different subunits which can be seen from the matrix of pairwise RMSDs between conformations (Figure 11A). Accordingly, most conformations with computed SBFE were included in the cluster formed by the rest of conformations of their subunit. To further analyze these data, we projected conformations to two-dimensional space using the t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA). The t-SNE projections showed a clear distinction between conformations of the individual subunits (Figure 11B). However, t-SNE projections do not allow inference regarding the relative distances between

31

groups of conformations since this method does not preserve the global structure of the data. PCA of the same set of conformations did not provide clear distinction between them but indicated that there are groups of conformations from different subunits located relatively close to each other (Figure 11C). Indeed, the two most populated clusters included conformations from two subunits simultaneously: the first included conformations from subunits B and D of the crystal system, and the second included conformations from subunit A of the crystal system and D of the apo system (Figure 11D). The third and the fourth most populated clusters were mainly formed by conformations from subunits A and D of the docked system, respectively, but also included a noticeable share of conformations from subunit A of the crystal system and B of the apo system.

The clustering results allowed us to estimate probabilities P_S^{apo} and P_S^{com} required for calculation of ΔG_r^{apo} and ΔG_r^{com}, respectively (see eqs 5 and 6). First, we identified clusters formed by conformations with computed SBFEs for the crystal and docked systems (SBFE clusters). These are clusters 0, 1 and 5 for the crystal system, and clusters 2, 3, 4, and 6 for the docked system (see Figure 11D). Next, the probability P_S^{apo} and P_S^{com} were calculated separately for the crystal and docked systems according to Equations 7 and 8. For example, P_S^{apo} for the crystal system was calculated as a share of all conformations of the apo system included in the SBFE clusters of the crystal system. The probability P_S^{com} for the crystal system was calculated as a share of all conformations of the crystal system included in the SBFE clusters of the crystal system. For the docked system, P_S^{apo} and P_S^{com} were calculated analogously. The calculated probabilities P_S^{apo} and P_S^{com} were ~0.224 and ~0.994 for the crystal system, and ~0.0255 and ~0.997 for the docked system. After that, we calculated free energies ΔG_r^{apo} and ΔG_r^{com} (see eqs 5 and 6) which were 0.57 and -0.02 kcal/mol for the crystal system and 2.10 and -0.01 kcal/mol for the docked system.

Finally, we calculated the total unbiased SBFE ΔG_b° (see eq 2) for both systems which were around

-15.6 and -12.5 kcal/mol for the crystal and docked systems, respectively (Table 2).
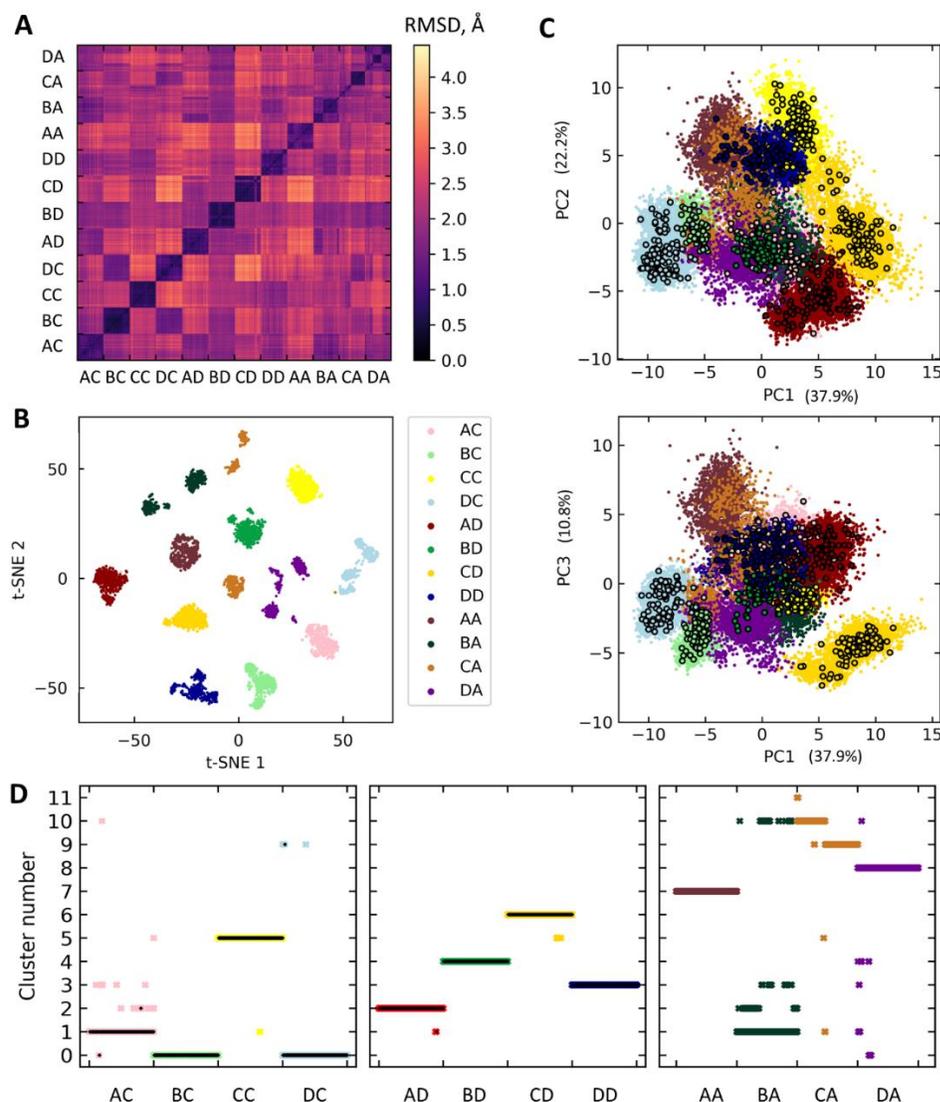
Figure 11. Clustering of conformations of the allosteric binding pocket of individual subunits from equilibrium portions of MD simulations of AMPAR-GYKI complex and apo protein. The systems are labeled as follows: the first letter of the label corresponds to a subunit and the last letter corresponds to the system: C is used for the GYKI crystal pose, D – for the GYKI docked pose, and A – for the apo protein. **A**. Matrix of pairwise RMSDs for all pairs of conformations. The RMSD is calculated on the protein C_α atoms located within 8 Å of GYKI. Conformations are grouped by systems; for each system, conformations are ordered as they occur in the MD trajectories. **B-C.** Conformations projected onto 2D space using t-SNE (**B**) and PCA (**C**). **D**. Clusters obtained for conformations for the GYKI crystal pose (left panel), the GYKI docked pose (middle panel), and the apo protein (right panel). The clusters are sorted in descending order by their populations. Conformations for which SBFE were computed are shown in black and the rest of the conformations are shown in colors according to panels **B** and **C**.

33

Thus, our free energy calculations suggest that binding of GYKI to AMPAR in the crystal pose is ~3 kcal/mol more favorable compared to the docked pose. Interestingly, the contributions of ensemble-averaged SBFE $\langle\Delta G_{b,r}^{\circ}\rangle$ and free energy of restraints in the apo protein ΔG_r^{apo} to this difference are approximately the same (~1.5 kcal/mol). Therefore, our calculations suggest that binding of GYKI in the crystal pose is preferable both because of the stronger protein-ligand interactions and because conformational behavior of the AMPAR allosteric pocket for the crystal pose is more similar to that in its apo form compared to the docked pose.

Conclusions

In this work, we presented a hierarchical approach for computing the SBFE of a ligand to a large, flexible binding pocket of a large protein. Our approach is general and can be used for computing the standard binding free energy for any protein-ligand system. Since this approach has been developed for the system where a significant number of dissimilar conformations of the protein binding pocket exist in both apo and ligand-bound states, the maximum efficiency of our approach can be expected for these cases. In this study, we validated our approach on the T4 lysozyme mutant in complex with a small molecule inhibitor and then applied it to investigate the energetics of binding of a small molecule antagonist GYKI to the allosteric site of the AMPAR.

In our approach, we consider the SBFE as an average SBFE over the ensemble of distinct conformations of the binding pocket. The ensemble of conformations is represented by a set of conformations selected equidistantly from an equilibrium MD trajectory for the full protein-ligand complex. The SBFEs are computed with MD TI simulations for these conformations using truncated systems with restraints on the protein C_α atoms. Our approach allows for both computing

34

the SBFE for a large protein-ligand system within a reasonable computational cost and accounting for conformational flexibility of the binding pocket.

To assess the performance of the developed approach, we computed SBFE for a benchmark system (lysozyme mutant in complex with a small-molecule inhibitor) with and without protein restraints. Our results demonstrated that our unbiasing procedure significantly increases the accuracy of computed SBFEs.

Using the developed approach, we computed the SBFE of GYKI to AMPAR for two ligand poses: a crystal pose, which is observed in the AMPAR-GYKI crystal structure and a flipped pose, which was obtained by docking in our previous work. Our long MD simulations demonstrated that the GYKI crystal pose is stable in all subunits while the GYKI docked pose shifted significantly with respect to the initial pose and stabilized only after a significant time (up to 600 ns depending on a subunit). We performed massive MD TI simulations for 800 conformations of AMPAR binding pocket using an automated MD TI preparation and simulation protocol which we developed.

The SBFEs of the restrained conformations of AMPAR were in the range of -11 to -17 kcal/mol. The SBFEs also varied significantly for individual conformations of the same subunit. The computed binding affinity of GYKI to AMPAR, namely the ensemble-averaged SBFE, was 3 kcal/mol lower for the crystal pose than the docked pose. This computed energy difference was statistically significant. At the same time, computed SBFEs for the representative poses did not result in a statistically significant difference, making it impossible to distinguish between the GYKI poses.

We believe that the results of this work will be valuable for research studies aimed at designing noncompetitive inhibitors of the AMPAR. Furthermore, the developed approach is generalizable and can be expanded to perform SBFE calculations for other proteins with flexible binding pocket.

Supplementary Information
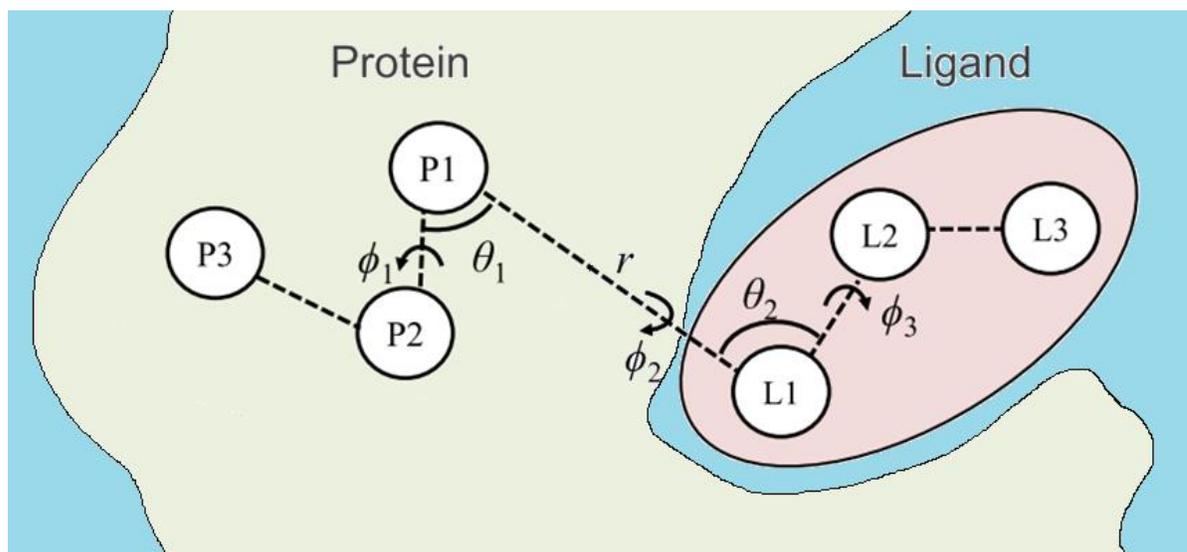
Supplementary figures

Figure S1. Schematic representation of the virtual bond[25] used for restraining the orientation of ligand in protein.
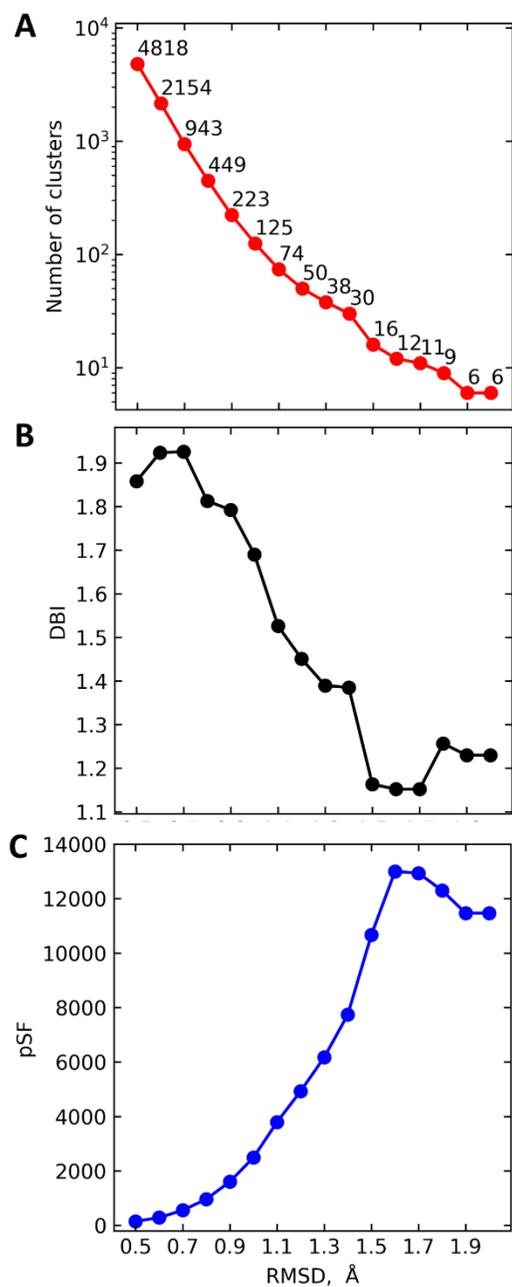
Fig S2. Number of clusters, DBI index and pseudo F statistic for the different linkage distance thresholds (RMSD) for average linkage clustering of AMPAR binding pocket conformations.

References

1. Kairys, V.; Baranauskiene, L.; Kazlauskiene, M.; Matulis, D.; Kazlauskas, E., Binding affinity in drug design: experimental and computational techniques. *Expert Opin Drug Discov* **2019,** *14* (8), 755-768.
2. Sadybekov, A. V.; Katritch, V., Computational approaches streamlining drug discovery. *Nature* **2023,** *616* (7958), 673-685.
3. Aldeghi, M.; Bluck, J. P.; Biggin, P. C., Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide. *Methods Mol Biol* **2018,** *1762*, 199-232.
4. Gumbart, J. C.; Roux, B.; Chipot, C., Standard binding free energies from computer simulations: What is the best strategy? *J Chem Theory Comput* **2013,** *9* (1), 794-802.
5. Klimovich, P. V.; Shirts, M. R.; Mobley, D. L., Guidelines for the analysis of free energy calculations. *J Comput Aided Mol Des* **2015,** *29* (5), 397-411.
6. King, E.; Aitchison, E.; Li, H.; Luo, R., Recent Developments in Free Energy Calculations for Drug Discovery. *Front Mol Biosci* **2021,** *8*, 712085.
7. Feng, M.; Heinzelmann, G.; Gilson, M. K., Absolute binding free energy calculations improve enrichment of actives in virtual compound screening. *Sci Rep* **2022,** *12* (1), 13640.
8. Heinzelmann, G.; Gilson, M. K., Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation. *Sci Rep* **2021,** *11* (1), 1116.
9. Lee, T. S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D., Jr.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H. C.; Xu, H.; Sherman, W.; York, D. M., Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J Chem Inf Model* **2020,** *60* (11), 5595-5623.
10. Gutkin, E.; Gusev, F.; Gentile, F.; Ban, F.; Koby, S. B.; Narangoda, C.; Isayev, O.; Cherkasov, A.; Kurnikova, M. G., In silico screening of LRRK2 WDR domain inhibitors using deep docking and free energy simulations. **2023**.
11. Chen, W.; Cui, D.; Jerome, S. V.; Michino, M.; Lenselink, E. B.; Huggins, D. J.; Beautrait, A.; Vendome, J.; Abel, R.; Friesner, R. A.; Wang, L. L., Enhancing Hit Discovery in Virtual Screening through Absolute Protein-Ligand Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling* **2023,** *63* (10), 3171-3185.
12. Kim, S.; Oshima, H.; Zhang, H.; Kern, N. R.; Re, S.; Lee, J.; Roux, B.; Sugita, Y.; Jiang, W.; Im, W., CHARMM-GUI Free Energy Calculator for Absolute and Relative Ligand Solvation and Binding Free Energy Simulations. *Journal of Chemical Theory and Computation* **2020,** *16* (11), 7207-7218.
13. Jespers, W.; Esguerra, M.; Aqvist, J.; Gutierrez-de-Teran, H., QligFEP: an automated workflow for small molecule free energy calculations in Q. *J Cheminform* **2019,** *11* (1), 26.
14. Gapsys, V.; Yildirim, A.; Aldeghi, M.; Khalak, Y.; van der Spoel, D.; de Groot, B. L., Accurate absolute free energies for ligand-protein binding based on non-equilibrium approaches. *Commun Chem* **2021,** *4* (1), 61.
15. Khalak, Y.; Tresadern, G.; Aldeghi, M.; Baumann, H. M.; Mobley, D. L.; de Groot, B. L.; Gapsys, V., Alchemical absolute protein-ligand binding free energies for drug design. *Chem Sci* **2021,** *12* (41), 13958-13971.
16. Wu, J. Z.; Azimi, S.; Khuttan, S.; Deng, N. J.; Gallicchio, E., Alchemical Transfer Approach to Absolute Binding Free Energy Estimation. *Journal of Chemical Theory and Computation* **2021,** *17* (6), 3309-3319.

17. Harder, E.; Damm, W.; Maple, J.; Wu, C. J.; Reboul, M.; Xiang, J. Y.; Wang, L. L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016,** *12* (1), 281-296.

18. Roos, K.; Wu, C. J.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L. L.; Abel, R.; Friesner, R. A.; Harder, E. D., OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *Journal of Chemical Theory and Computation* **2019,** *15* (3), 1863-1874.

19. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

20. He, X. B.; Man, V. H.; Yang, W.; Lee, T. S.; Wang, J. M., A fast and high-quality charge model for the next generation general AMBER force field. *Journal of Chemical Physics* **2020,** *153* (11).

21. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *Journal of Computational Chemistry* **2010,** *31* (4), 671-690.

22. Qiu, Y. D.; Smith, D. G. A.; Boothroyd, S.; Jang, H.; Hahn, D. F.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C. D.; Rizzi, A.; Tjanaka, B.; Tresadern, G.; Lucas, X.; Shirts, M. R.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Mobley, D. L.; Wang, L. P., Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field. *Journal of Chemical Theory and Computation* **2021,** *17* (10), 6262-6280.

23. Boothroyd, S.; Behara, P. K.; Madin, O. C.; Hahn, D. F.; Jang, H.; Gapsys, V.; Wagner, J. R.; Horton, J. T.; Dotson, D. L.; Thompson, M. W.; Maat, J.; Gokey, T.; Wang, L. P.; Cole, D. J.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Shirts, M. R.; Mobley, D. L., Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *Journal of Chemical Theory and Computation* **2023,** *19* (11), 3251-3275.

24. Poulos, T. L., Cytochrome P450 flexibility. *Proc Natl Acad Sci U S A* **2003,** *100* (23), 13121-2.

25. Kobilka, B. K.; Deupi, X., Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol Sci* **2007,** *28* (8), 397-406.

26. Narangoda, C.; Sakipov, S. N.; Kurnikova, M. G., AMPA Receptor Noncompetitive Inhibitors Occupy a Promiscuous Binding Site. *ACS Chem Neurosci* **2019,** *10* (11), 4511-4521.

27. Altenberg, G. A., Structure of multidrug-resistance proteins of the ATP-binding cassette (ABC) superfamily. *Curr Med Chem Anticancer Agents* **2004,** *4* (1), 53-62.

28. Procacci, P., Dealing with Induced Fit, Conformational Selection, and Secondary Poses in Molecular Dynamics Simulations for Reliable Free Energy Predictions. *J Chem Theory Comput* **2023,** *19* (23), 8942-8954.

29. Raniolo, S.; Limongelli, V., Ligand binding free-energy calculations with funnel metadynamics. *Nat Protoc* **2020,** *15* (9), 2837-2866.

30. Rogawski, M. A., AMPA receptors as a molecular target in epilepsy therapy. *Acta Neurol Scand Suppl* **2013,** (197), 9-18.

31. Alt, A.; Nisenbaum, E. S.; Bleakman, D.; Witkin, J. M., A role for AMPA receptors in mood disorders. *Biochem Pharmacol* **2006,** *71* (9), 1273-88.

32. Babaei, P., NMDA and AMPA receptors dysregulation in Alzheimer's disease. *Eur J Pharmacol* **2021,** *908*, 174310.

33. Hansen, K. B.; Wollmuth, L. P.; Bowie, D.; Furukawa, H.; Menniti, F. S.; Sobolevsky, A. I.; Swanson, G. T.; Swanger, S. A.; Greger, I. H.; Nakagawa, T.; McBain, C. J.; Jayaraman, V.; Low, C. M.; Dell'Acqua, M. L.; Diamond, J. S.; Camp, C. R.; Perszyk, R. E.; Yuan, H.; Traynelis, S. F., Structure, Function, and Pharmacology of Glutamate Receptor Ion Channels. *Pharmacol Rev* **2021,** *73* (4), 298-487.

34. Hanada, T., The discovery and development of perampanel for the treatment of epilepsy. *Expert Opin Drug Discov* **2014,** *9* (4), 449-58.

35. Patsalos, P. N., The clinical pharmacology profile of the new antiepileptic drug perampanel: A novel noncompetitive AMPA receptor antagonist. *Epilepsia* **2015,** *56* (1), 12-27.

36. Yelshanskaya, M. V.; Singh, A. K.; Sampson, J. M.; Narangoda, C.; Kurnikova, M.; Sobolevsky, A. I., Structural Bases of Noncompetitive Inhibition of AMPA-Subtype Ionotropic Glutamate Receptors by Antiepileptic Drugs. *Neuron* **2016,** *91* (6), 1305-1315.

37. Balannik, V.; Menniti, F. S.; Paternain, A. V.; Lerma, J.; Stern-Bach, Y., Molecular mechanism of AMPA receptor noncompetitive antagonism. *Neuron* **2005,** *48* (2), 279-88.

38. Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M., Absolute binding free energies: A quantitative approach for their calculation. *Journal of Physical Chemistry B* **2003,** *107* (35), 9535-9551.

39. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-8.

40. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-74.

41. Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C., Lipid14: The Amber Lipid Force Field. *J Chem Theory Comput* **2014,** *10* (2), 865-879.

42. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **1983,** *79* (2), 926-935.

43. Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E., III; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M., Jr., AmberTools. *Journal of Chemical Information and Modeling* **2023,** *63* (20), 6183-6191.

44. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** *23* (3), 327-341.

45. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.

46. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.

47. Case, D. A.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, I., T.E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kovalenko, A.; Krasny, R.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Man, V.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wilson, L.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Kollman, P. A., AMBER 20, University of California, San Francisco. **2020**.

48. Lee, T. S.; Lin, Z.; Allen, B. K.; Lin, C.; Radak, B. K.; Tao, Y.; Tsai, H. C.; Sherman, W.; York, D. M., Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore Potentials. *J Chem Theory Comput* **2020,** *16* (9), 5512-5525.

49. Efron, B.; Tibshirani, R. J., *An introduction to the bootstrap*. CRC press: 1994.

50. Chodera, J. D., A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J Chem Theory Comput* **2016,** *12* (4), 1799-805.

51. Xu, R.; Wunsch, D., 2nd, Survey of clustering algorithms. *IEEE Trans Neural Netw* **2005,** *16* (3), 645-78.

52. Roe, D. R.; Cheatham, T. E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-95.

53. Mobley, D. L.; Gilson, M. K., Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu Rev Biophys* **2017,** *46*, 531-558.

54. Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K., Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site. *Journal of Molecular Biology* **2009,** *394* (4), 747-763.