Supporting Information

Mapping Functional Group Free Energy Patterns at Protein Occluded Sites: Nuclear Receptors and G-Protein Coupled Receptors

Sirish Kaushik Lakkaraju[1], Wenbo Yu[1], E Prabhu Raman[1], Alena V Hershfeld[2], Lei Fang[1], Deepak D Deshpande[3], Alexander D MacKerell, Jr.[1*]

[1]Department of Pharmaceutical Sciences, School of Pharmacy; [2]Division of Pulmonary and Critical Care Medicine, School of Medicine, University of Maryland, 20 Penn St, Baltimore, MD 21201.

[3]Center for Translational Medicine, Department of Medicine, Thomas Jefferson University, 1020 Locust Street, Philadelphia, PA 19107.

Corresponding Author email: alex@outerbanks.umaryland.edu

S1. Protein Preparation

AR & PPARγ: PDB coordinates of 2AM9[1] (resolution 1.64 Å) and 3U9Q[2] (resolution 1.52 Å) with the ligands testosterone (TES) and decanoic acid (DA), respectively, were used following the deletion of the ligands. Missing residues 262-275 in the PDB 3U9Q were build using MODELLER[3]. A total of 100 models were generated and ranked using the Discrete Optimized Protein Energy (DOPE) method[4] and the highest ranking model was used as a starting structure. Crystal water molecules were retained, as were any structurally important ions. An in-house preparation script utilized GROMACS utilities to generate the simulation system involving protein, water, and small molecules, with the size of the system so as to have the protein extrema separated from the edge by 12 Å on all sides. The net system charge was made neutral by replacing random water molecules with the appropriate number of sodium or chloride

ions. The proteins were minimized for 500 steps with the steepest descent (SD)[5] algorithm in the presence of periodic boundary conditions (PBC). This was followed by a 100 ps equilibration during which temperature was adjusted by velocity rescaling[6]. During the minimization and equilibration, harmonic positional restraints with a force constant of 2.4 kcal mol^{-1} $Å^{-2}$ were applied to protein non-hydrogen atoms. The final coordinates at the end of equilibration are used as the starting conformations for the GCMC/MD simulation.

mGluR: PDB coordinates of 4OR2[7] (resolution 2.80 Å) with the negative allosteric modulator (NAM), 4-fluoro-N-(4-(6-(isopropylamino)pyrimidin-4-yl)thiazol-2-yl)-Nmethylbenzamide (FITM) was used following the deletion of the ligand. The apocytochrome b562RIL that was fused to the N-terminus of the GPCR was removed. Only one of the monomers in the parallel dimeric 7 transmembrane (TM) regions was used. Intracellular loop 2 (ICL2, 688-695) that was missing in the crystal structure was built with MODELLER following the same procedure as with the PPARγ. The crystal water molecules and the cholesterol molecules were maintained from the crystal structure. The disulphide bridge between residues 657-746 was also maintained. The TM monomer was then inserted into a rectangular palmitoyl oleoyl phosphatidyl choline (POPC) membrane containing ~90 lipids with 10% cholesterol using the CHARMM-GUI[8]; the remainder of the system was filled with ~0.15 M NaCl aqueous solution based on the TIP3P water model[9] to neutralize the system. 3000 steps of minimization were run to remove bad contacts: of these 1500 steps were with SD[5] and the remainder with the adopted basis Newton-Raphson (ABNR) algorithm[10]. During minimization, positions of the protein backbone were harmonically restrained using a force constant of 10 kcal/mol/$Å^2$ on the non-hydrogen atoms. The side chains were restrained with a force constant of 5 kcal/mol/$Å^2$. Influx of water molecules into the hydrophobic core of the TM region was prevented using a harmonic restraining potential of 2.5 kcal/mol/Å along the x,y planes at a +/-11 Å of the z axis from the center of the receptor. The same force was also used to keep the heads and tails of the lipids in place, and configurations of the lipids

were maintained with harmonic dihedral restraints with a force constant of 250 kcal/mol/rad^2. The same restraints were used during the 375 ps of equilibration with a 1 fs time-step, but the restraint forces were gradually reduced as shown in Table S1.

The first 50ps of equilibration was performed using Langevin dynamics with a friction coefficient of 3 ps^{-1}. The remaining 325 ps of the equilibration used constant pressure-temperature (NPT) dynamics using the Langevin Piston integrator[11]. Covalent bonds involving hydrogens were fixed at the equilibrium length by the SHAKE algorithm[12]. Lennard-Jones (LJ) interactions were smoothed from 10 to 12 Å by a switching function and the non-bonded pair list was generated out to 14 Å and updated heuristically. Electrostatic interactions were calculated by the particle-mesh Ewald [13] summation method with a real space cutoff of 12 Å.

Table S1. Harmonic force constants used in the equilibration cycle of the mGluR GPCR preparation MD.

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Time (ps)	25	25	25	100	100	100
Timestep (fs)	1	1	1	2	2	2
Backbone non-hydrogen atoms[a]	10.0	5.0	2.5	1.0	0.5	0.1
Side-chain non-hydrogen atoms[b]	5.0	2.5	1.0	0.5	0.1	0.0
Water oxygens[c]	2.5	2.5	1.0	0.5	0.1	0.0
Lipid tail non-hydrogen atoms[d]	2.5	2.5	1.0	0.5	0.1	0.0
Lipid head non-hydrogen atoms[e]	2.5	2.5	1.0	0.5	0.1	0.0
Ion[f]	10.0	0.0	0.0	0.0	0.0	0.0
Lipid cis double bond[g]	250.0	100.0	50.0	50.0	25.0	0.0
Lipid chirality[h]	250.0	100.0	50.0	50.0	25.0	0.0
Dynamics Integrator	Langevin (leapfrog)			Constant P/T (leapfrog)		
a,b and f are positional force restraints (kcal/mol/Å^2) restraining atoms to their original coordinates, c,d and e (kcal/mol/Å^2) restrain atoms outside a distance along the normal of the lipid bilayer or within the planes and g and i are applied to the selected dihedrals (kcal/mol/rad^2).						

To this equilibrated system, the solutes, each at 0.25 M were added. The system was again minimized for an additional 1000 steps with the SD[5] algorithm. A short equilibration was

performed for 250 ps. The leapfrog version of the Verlet integrator[14] was used. This phase of minimization and equilibration in the presence of the SILCS solutes was performed using GROMACS[15], with the same CHARMM force field[16, 17] as described above; and applying harmonic potential restraints with a force constant of 2.4 kcal/mol/Å^2 on the protein non-hydrogen atoms. As detailed previously[18, 19], to prevent aggregation/ion pairing of hydrophobic and charged solutes, thereby promoting faster convergence, a repulsive energy term was introduced only between benzene:benzene, benzene:propane, propane:propane, acetate:acetate, acetate:methylammonium, and methylammonium:methylammonium molecular pairs. This was achieved by adding a massless particle to the center of mass of benzene and the central carbon of propane, and methyl carbon of acetate, and methylammonium. Each such particle does not interact with any other atoms in the system but with other particles on the hydrophobic or charged molecules through the Lennard-Jones (LJ) force field term with parameters (ε = 0.01 kcal/mol; R_{min} =12.0 Å).

β_2AR: PDB coordinates of 2RH1[20] (resolution 2.40 Å) and 3P0G[21] (3.50 Å) were used to represent the inactivated (B2I) and the activated (B2A) states of the receptor, following the deletion of their respective ligands, carazolol and BI-167107. The region of the structures corresponding to T4 lysozyme, which was inserted between TM5 and TM6 to facilitate crystallization, was deleted and substituted with the third intracellular loop (ICL3). ICL3 for B2A was built in the presence of the G-protein heterotrimer in the following way: the β_2-adrenergic+G-protein heterotrimer complex was first generated by superposing the helices H1-H3 of the TM region of the PDB 3SN6 onto B2A. MODELLER[22] was then used to generated 100 models of the ICL3 for B2A in the presence of the G-protein heterotrimer. The models were ranked using the DOPE method and the highest ranking model was selected. The G-protein heterotrimer was then removed from B2A. For B2I, the procedure was repeated but without the G-protein heterotrimer. The crystal water molecules and the cholesterol molecules found in the

crystal structure of the B2I were maintained. Disulphide bridges between residues 106-191 and 184-190 were also maintained. Both B2I and B2A were then inserted into rectangular POPC membranes containing ~198 and ~252 lipids, respectively, with 10% cholesterol using the CHARMM-GUI[8]; the remainder of the systems were filled with ~0.15 M NaCl aqueous solution based on the TIP3P water model to neutralize the system. These structures were then prepared for the GCMC/MD run through the two-stage minimization and equilibration process following the same procedure as described for the mgluR GPCR.

S2. GCMC

During GCMC, fragments and water are exchanged between their gas-phase reservoirs and a cubic region of 20 Å side (25 Å for the PPARγ and AR) encompassing the ligand binding pocket of the protein. The excess chemical potential (μ_{ex}) supplied to drive fragment exchanges is periodically fluctuated over every 3 cycles, such that the average μ_{ex} over the 100 cycles is close to the values in SI Table S2. These values are the magnitude of the μ_{ex} required to maintain 0.25 M of a solute in a bulk aqueous mixture devoid of any protein, and are approximately equal to the hydration free energy.

The 100,000 steps of GCMC are divided evenly between each of the SILCS fragments and water. There are 4 possible GCMC moves: insertion, deletion, translation and rotation. The probabilities of these moves are governed by the Metropolis criteria[23].

$$
\begin{aligned}
\mathrm{P}_{\mathrm{insert}} &= \min\left\{1, \frac{f_n}{n+1} e^{\mathrm{B}-\beta\Delta\mathrm{E}}\right\} \\
\mathrm{P}_{\mathrm{delete}} &= \min\left\{1, \frac{n}{f_{n-1}} e^{-\mathrm{B}-\beta\Delta\mathrm{E}}\right\} \\
\mathrm{P}_{\mathrm{trans/rot}} &= \min\left\{1, e^{-\beta\Delta\mathrm{E}}\right\} \\
\text{where } \mathrm{B} &= \beta\mu_{ex} + \ln\bar{\mathrm{n}} \text{ and } \bar{\mathrm{n}} = \overline{\rho v}
\end{aligned}
\qquad (1)
$$

where μ_{ex} is the excess chemical potential, $\bar{n}$ is the expected number of molecules, $\bar{\rho}$ is the density, $\bar{v}$ is the volume of system A, f_n is the fractional volume of the subspace where the insertion attempts are made, ΔE is the change in energy due to a move, β is $1/k_BT$, k_B is the Boltzmann constant and T is temperature (300 K in the present study). Through the GCMC simulation, the volume of the simulation system A, the total energy and the total number of particles between the system A and its reservoir are fixed. Over the 100,000 steps, the order in which the four possible GCMC moves are attempted, and the molecule (solutes or water) on which the move is performed is randomized. GCMC was run using an in-house C++ code that implemented the grid-based GCMC scheme[23] with the cavity-bias algorithm[24, 25] to drive solute and water exchanges between their reservoirs and the systems.

Since GCMC simulations in the presence of explicit solvent suffer from convergence problems due to low acceptance rates encountered for the fragment insertions[24, 26], the excess chemical potential (μ_{ex}) supplied to the solutes and water is varied periodically, such that $d\mu_{ex}$ was alternately added and subtracted to μ_{ex} following every three cycles of GCMC/MD to maintain the target concentration, N^{tgt}. The magnitude of the fluctuation, $d\mu_{ex} = \frac{N}{N^{tgt}}$, where N is the number of molecules of a given solute type within the simulation system. Thus, the frequency of oscillation was six cycles, and by maintaining this oscillation through the length of the simulation, the average μ_{ex} remains at the values in Table S3.

S3. MD

The configuration at the end of the GCMC is used as the starting configuration in the MD. Before the production, a 500 step SD minimization and a 100 ps equilibration is run as described during the protein preparation simulations. Production runs were performed with the leapfrog integrator (GROMACS integrator “md”) with a time step of 2 fs. The Nose−Hoover

method[27, 28] was used to maintain the temperature at 298 K with the protein and the remainder of the system separately coupled to the heat bath. Pressure was maintained at 1 bar using the Parrinello−Rahman[29] barostat. The time constant used for temperature and pressure coupling was uniformly 1 ps. The LINCS[30] algorithm was used to constrain water geometries and all covalent bonds involving a hydrogen atom. vdW interactions were switched off smoothly in the range of 10−12 Å, and the particle mesh Ewald (PME) method[13] was used to treat long-range electrostatics with a real space cut off of 12 Å, with the order of B-spline interpolation set to 4 and the maximum grid spacing set to be 1.2 Å. Long-range dispersion correction to the energy and pressure was applied. Weak restraints were applied only on the backbone Cα carbon atoms with a force constant (k in 1/2 $k\delta x^2$) of 0.12 kcal mol Å. This was done to prevent the rotation of the protein in the simulation box and potential denaturation due to the presence of small molecules in the aqueous solution. The last conformation from the production MD is used as the starting conformation of the next GCMC cycle.

S4. Bulk-phase GCMC/MD simulations

To determine the μ_{ex} needed to maintain 0.25 M of the solutes in an aqueous system devoid of protein and to normalize the probabilities of solutes in the presence of a protein to their bulk occupancies, bulk-phase GCMC/MD simulations were run, with a protocol that was developed previously[31]. Ten 50 Å simulation boxes were built that contained 0.25 M of each of the solutes immersed in water. 200 cycles GCMC/MD simulations were run on each of the boxes. Each cycle of GCMC/MD involves 100,000 steps of GCMC with both the solutes and water, and 250 ps of MD, yielding 10x50 ns of MD and 10x20 million steps of GCMC. GCMC with solutes and water were limited to a smaller 40 Å cubic region within the center of the box. In the first cycle, during GCMC, μ_{ex} supplied to all solutes and water was set to 0. At the end of 100,000 steps of GCMC, for each of the fragments and water, the number of molecules within the GC area was compared to the number required to achieve concentrations of 0.25 M and 55 M, respectively

and used to modulate the μ_{ex} to be supplied in the next cycle. The last configuration at the end of the GCMC was used as the starting configuration of a 250 ps NPT MD simulation following the same protocol as with the protein systems. μ_{ex} converged over about 50 cycles for fragments and water. Average μ_{ex} across the cycles 150-200 over the 10 runs is reported in Table S2. Average voxel occupancies for each of the solutes were also calculated over the GC area from these last 50 cycles over the 10 runs.

Table S2. Average μ_{ex} of the solutes and the water required to maintain 0.25 M and 55 M in the aqueous protein simulation systems during GCMC for the solutes and water, respectively. Hydration free energies using free energy perturbation simulations, HFE^{fep} , were obtained from our previous work [31]. μ_{ex} is periodically fluctuated every 3 cycles around these values.

Fragment	HFE^{fep} [31] (kcal/mol)	μ_{ex} (kcal/mol)
Benzene	-0.71	-0.79
Propane	1.60	1.96
Acetaldehyde	-4.43	-3.23
Methanol	-6.16	-5.62
Formamide	-10.71	-10.92
Imidazole	-12.55	-14.18
Acetate	-96.5	-97.31
Methylammonium	-52.0	-68.49
Water	-	-5.6

S5. FragMap preparation

3D probability distributions of the selected atoms from the solutes, called “FragMaps”, from the GCMC/MD and GCMC-only simulations were constructed for the following atom types: benzene carbons, propane carbons, methanol polar hydrogen, methanol oxygen, formamide polar hydrogens, formamide oxygen, imidazole unprotonated nitrogen, imidazole polar hydrogen, acetaldehyde oxygen, methylammonium polar hydrogens, and acetate oxygens. For the

GCM/MD simulations, atoms from snapshots output every 10 ps of the production run in every cycle were binned into 1 Å × 1 Å × 1 Å cubic volume elements (voxels) of a grid spanning the GCMC box and the voxel occupancy for each FragMap atom type was calculated. For the GCMC-only simulations, voxel occupancies for each FragMap atom type were obtained every 1000 steps of GCMC in every cycle. These were then normalized by the voxel occupancies of the fragments in a bulk-phase system devoid of the protein. Bulk-phase occupancies were equal across the GCMC/MD and GCMC-only simulations. The voxel occupancies of the eleven atom types were merged in the following manner to create the following five generic FragMap types: (1) generic nonpolar, APOLAR (benzene and propane carbons); (2) generic neutral hydrogen bond donor, HBDON (methanol, formamide and imidazole polar hydrogens); (3) generic neutral hydrogen bond acceptor, HBACC (methanol, formamide, and acetaldehyde oxygen and imidazole unprotonated nitrogen) (4) positive donor, POS (methylammonium polar hydrogens); and (5) negative acceptor, NEG (acetate oxygens). Probability distributions were normalized based on the respective distributions of the atom types in the absence of proteins and converted to free energies via a Boltzmann-based transform of the normalized probability to yield a "grid free energy (GFE)" for each fragment type T for the coordinates x,y,z, referred to as GFE FragMaps. The convergence of the FragMaps was monitored by calculating overlap coefficients (Section S7).

S6. LGFE scoring

The use of GFE FragMaps has the advantage that it allows for quantitative evaluation of the relative affinity of the ligands, based on LGFE scores, as previously described[32]. LGFE quantifies the overlap of atoms in ligands in the LBP with the corresponding GFE FragMaps. Ligand atoms were classified into FragMap types, for which an assignment map was created. The rules for assignment are presented in our previous work[32]. Each classified atom of a ligand with coordinates ($x_i.y_i,z_i$) was assigned a score equal to the GFE value of the corresponding

FragMap type (f), $GFE^{f}_{xi,yi,zi}$, of the voxel it occupies in the LBP. LGFE is then the sum of each of these GFE values for all the classified ligand atoms. Instead of a single crystallographic conformation of the ligand in the LBP, LGFE is calculated as Boltzmann weighted average over an ensemble of conformations obtained by MC sampling of the ligand in LBP, in the field of FragMaps.

An in-house suite of programs was used to setup and run the MC simulations. The ligand had rotational, translational and intramolecular degrees of freedom. The ligand had no rotational restraints, but its center of mass (CoM) was restrained to lie within 2.5 Å of the CoM of the ligand crystal conformation using a flat bottom restraint. Intramolecular degrees of freedom consisted of rotatable bonds, which were automatically detected based on the CGenFF molecular topology. All acyclic non-terminal bonds were considered rotatable, with the exception of bonds ending in methyl or NH_3^+ groups. The force-field terms corresponding to the intramolecular degrees of freedom comprised of dihedral, van der Waals (vdW) and electrostatic terms. Due to the absence of protein and solvent during these simulations a distance dependent dielectric (=4|r|) was used to evaluate the intramolecular electrostatic contributions to prevent their overestimation. The energy computed during the Metropolis MC can be written as follows.

$$E = E_{vdw,intra} + E_{elec,intra} + E_{dihe,intra} + LGFE \quad (2)$$

MC sampling of crystallographic ligands were performed using the crystal structure as the starting conformation and position in the LBPs. Simulations for the crystallographic ligands were run for 200,000 steps, with snapshots recorded every 1000 steps. Translation, rotation and dihedral rotations were randomly chosen between (-0.2 Å, 0.2 Å) and (-36°, 36°), respectively. Boltzmann weighted average LGFE was calculated over the resulting 200 conformations. For each of the non-crystallographic ligands, 10 different MC simulations (each for 10,000,000

steps; snapshots recorded every 10,000) were run, and for each of the runs, the ligand position and orientation in the LBP was assigned randomly. Average LGFE was first calculated over each of the 10 MC simulations. $LGFE^{MC}$ is then the Boltzmann weighted average over these LGFE values obtained across the 10 MC simulations.

S7. Experimental Binding affinity (ΔG_{bind})

In the case of PPARγ and β_2AR, ΔG_{bind} is calculated from K_I as $\Delta G_{bind} = -RT*\log(K_I)$, where R is the universal gas constant (=1.987×10^{-3}), T is the temperature (=300K). For mGluR, ΔG_{bind} is estimated from IC_{50} of the ligand, $\Delta G_{bind} = RT*\log(IC_{50})$.

S8. Overlap Coefficient

The convergence of the FragMaps was monitored by calculating overlap coefficients (OC). The ten trajectories were divided into two groups (group 1, trajectories 1−5; group 2, trajectories 6−10), and FragMaps of each group were separately computed. OC relates the overlap between group 1 and group 2 FragMaps to a number between 0 and 1, with 1 reflecting completely identical maps.

$$\mathrm{OC} = \sum_{i=1}^{N} \min\left(\frac{Q_i^i}{\sum_{j=1}^{N} Q_j^i} ; \frac{Q_i^2}{\sum_{j=1}^{N} Q_j^2} \right)$$

(3)

In equation 3 N is the number of voxels in the FragMaps and Q_i^1 and Q_i^2 are occupancies for the *i*th voxel from group 1 and 2 generated FragMaps, respectively. The ratios in the parentheses are computed to normalize the occupancy of each voxel by the sum of occupancies of all voxels in the corresponding FragMap. For each voxel, the smaller values (the conserved part) from group 1 and 2 are summed over all voxels to get the OC. It should be noted that the OC index

does not behave linearly, such that a relatively small difference in the two distributions leads to a decrease from 1 to approximately 0.8, and values of >0.5 indicate a high degree of similarity (Fig. S1). As shown Table S3, reasonable overlap coefficients are recorded for FragMaps of all the tested systems.

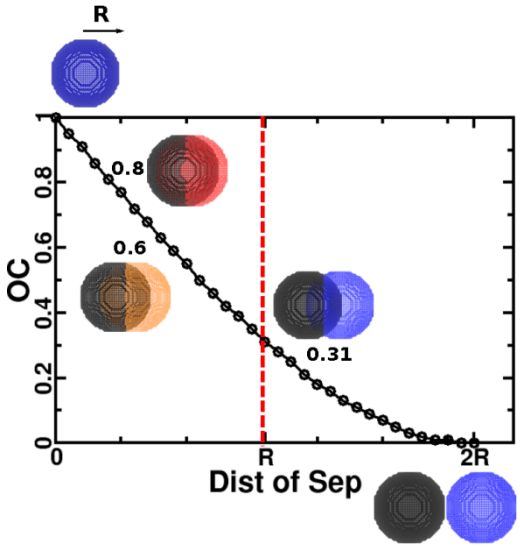

Fig S1. Overlap coefficient (OC) between two spheres of radius R, as a function of distance of separation between them.

Table S3. Overlap coefficients of FragMaps from trajectories 1-5 and 6-10 across the GCMC/MD simulations for the four protein systems.

Receptor	APOLAR	POS	NEG	HBDON	HBACC
AR	0.68	0.65	0.63	0.41	0.47
PPARγ	0.78	0.66	0.75	0.58	0.63
mGluR	0.75	0.52	0.55	0.55	0.49
B2A	0.73	0.79	0.49	0.53	0.51

B2I	0.77	0.73	0.62	0.49	0.62

S9. Supplementary Tables

Table S4. Occludedness of a ligand binding pocket (LBP) measured as a ratio of the solvent accessible surface area (SASA) of a ligand in bulk vs. in the LBP for the NRs and GPCRs in this work (AR, PPARγ, mGluR and β_2AR) compared to proteins studied in our previous works[18, 32] (FKBP12, FactorXA and DHFR).

Receptor	Ligand	Bulk SASA	LBP SASA	Ratio
AR	Testosterone	402.1	0	0
PPARγ	Rosiglitazone	506.8	68.6	0.13
	Decanoic acid	341.1	28.8	0.08
mGluR	FITM	576.8	43.8	0.07
β_2AR	BI-167107	534.8	36.32	0.07
	Carazalol	478.6	62.8	0.13
FKBP12	SB3	621.2	231.2	0.37
FactorXa	ZK-807834	679.5	171.3	0.25
DHFR	NDP	818.9	420.7	0.50

Table S5. Binding affinities of ligands for the PPARγ.

	PDB	Compound	ΔG_{bind}	Reference
1	3U9Q	Decanoic acid	-5.46	[2]
2	2I4Z,2I4P	S-1(LT-127)	-7.88	[33]
3	3FEJ	CTM	-8.33	[34]
4	1NYX	Ragaglitazar	-8.58	[35]
5	3DZU	PLB	-8.69	[36]
6	1I7I	Tesaglitazar	-8.92	[37]
7	3T03	GQ-16	-9.23	[38]
8	2I4J	R-1	-9.32	[33]
9	1FM6, 2PRG, 1ZGY, 3CS8, 3DZY	Rosiglitazone	-9.40	[39]
10	3H0A	D30	-10.16	[40]
11	3LMP	Cerco-23	-10.44	[41]
12	2YFE	Amorfrutin	-11.48	[42]
13	1K74	GW409544	-11.71	[35]
14	1FM9	GI262570 (Farglitazar)	-12.23	[39]
15	2P4Y	SPPARgM2	-12.36	[43]
16	3B1M	Cerco-A	-12.66	[44]

Table S6. Binding affinities of ligands for the mGluR.

Compound	**ΔG_{bind}**	**Reference**
4-fluoro-N-(4-(6-(isopropylamino)pyrimidin-4-yl)thiazol-2-yl)-Nmethylbenzamide **(FITM)**	-11.27	[7]
Compound 14 **(CPD14)**	-9.01	[7]
Compound 17 **(CPD17)**	-10.87	[7]
Compound 22 **(CPD22)**	-7.74	[7]
2-Methyl-6-(phenylethynyl)pyridine **(MPEP)**	-11.56	[45]

Table S7. Binding affinities of ligands for the β_2AR.

	PDB	Compound	Function	ΔG_{bind}	Reference
1		Salbutamol	Agonist/ Partial Agonist	-8.09	[46]
2	4LDO	Adrenaline(epinephrine)		-8.60	[47]
3		Isoproterenol		-9.63	[47]
4	4LDL	Hydroxybenzyl isoproterenol		-9.72	[48]
5		Indacaterol		-10.19	[49]
6		Formoterol		-10.60	[49]
7		THRX-144877(AA1)		-11.04	[49]
8		Picumeterol		-12.23	[49]
9		Salmeterol		-12.00	[50]
10	3P0G	BI-167107		-13.69	[21]
11		Practolol	Antagonists/ inverse agonist	-6.77	[50]
12		Atenolol		-8.14	[50]
13		Acebutalol		-8.26	[50]
14		Bisoprolol		-9.10	[50]
15		Labetolol		-10.91	[50]
16		Propranolol		-12.34	[50]
17		Pindolol		-12.43	[51]
18		ICI-118551		-12.58	[50]
19		Carvedilol		-12.77	[50]
20	3D4S	Timolol		-13.15	[50]
21	2RH1	Carazolol		-13.35	[52]

S9. Supplementary Figures

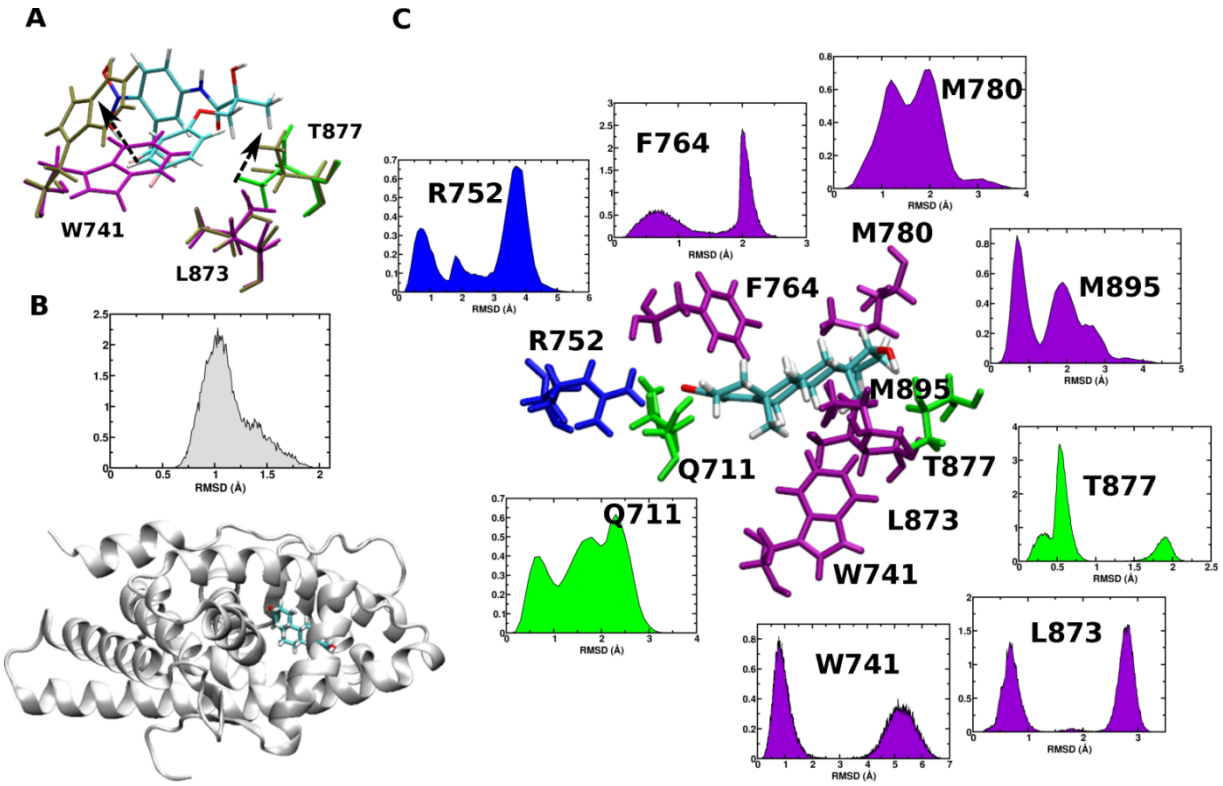

Figure S2. AR structural analysis. A) Change in the side-chains of W741, L873 and T877 needed to accommodate the second phenyl ring of the ligands S-1 and EM-5744. B) The global conformation of AR is maintained through the simulations as evidenced by the distribution of the backbone RMSD. C) Distributions of the RMSD of selected side chains identified to be pertinent for the ligand binding through the 10x100 ns GCMC-MD simulations of the AR. Positively charged side chain is colored blue, polar side chains are colored green, and hydrophobic side chains are colored purple. TES in the crystallographic orientation is shown in cyan.

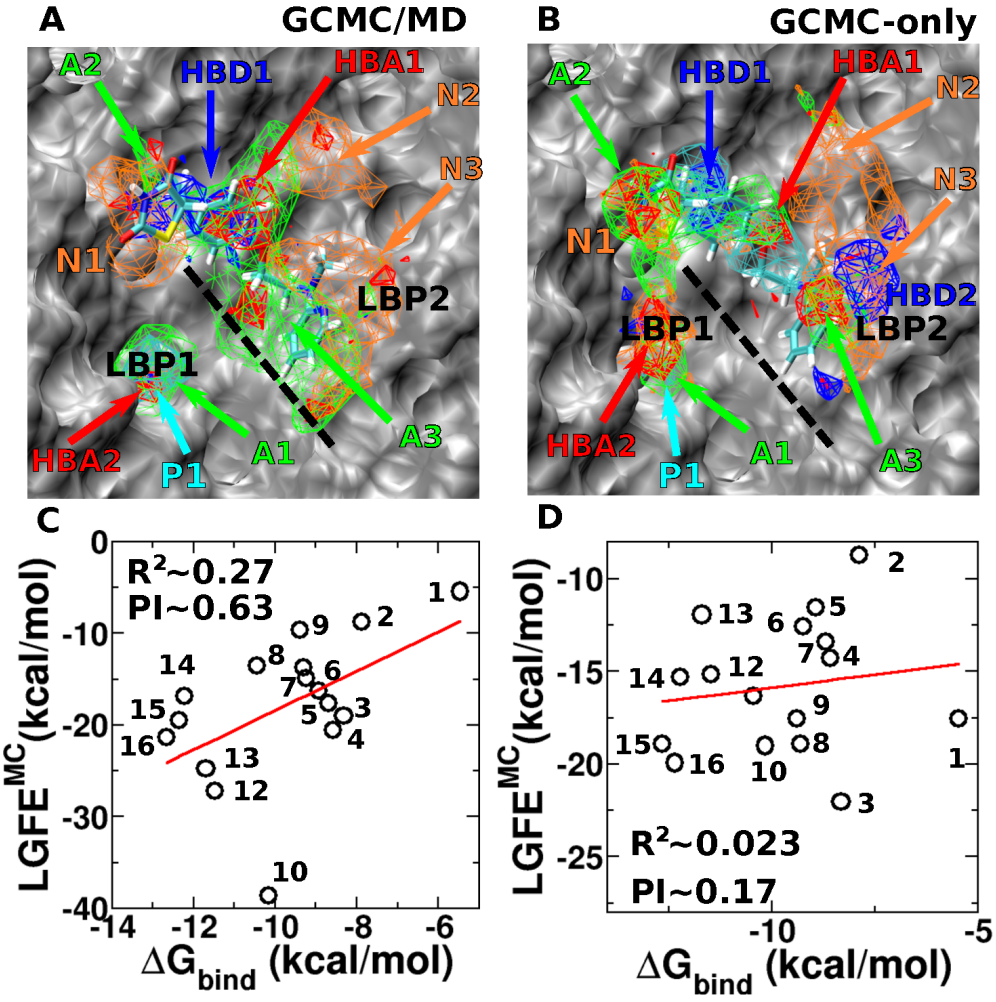

Figure S3. Comparison of FragMaps between A) GCMC/MD vs. B) GCMC-only simulations for the PPARγ. In A) HBACC and HBDON FragMaps are set to a cutoff of -0.5 kcal/mol, while APOLAR, NEG and PDON FragMaps are set to a cutoff of -1.2 kcal/mol. In B) all FragMaps contours are displayed at -1.2 kcal/mol. C) LGFEs calculated using the FragMaps from GCMC/MD correlated well with ΔG_{bind} (Table S5) with the correlation being lost when D) the LGFEs were calculated using the FragMaps from GCMC-only simulations.

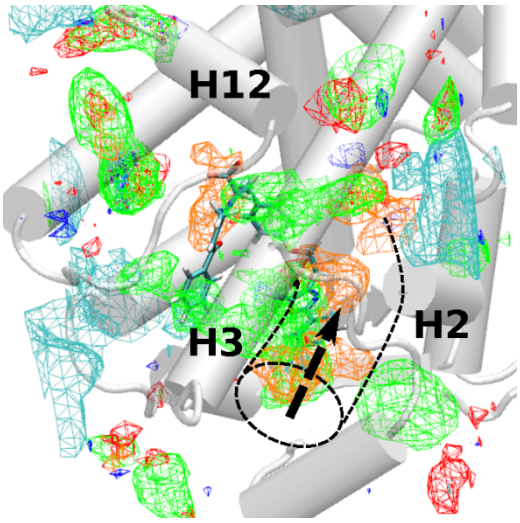

Figure S4. FragMaps trace the possible passage of entry of a ligand in the PPARγ. Ligand GW40944 is shown in the LBP.

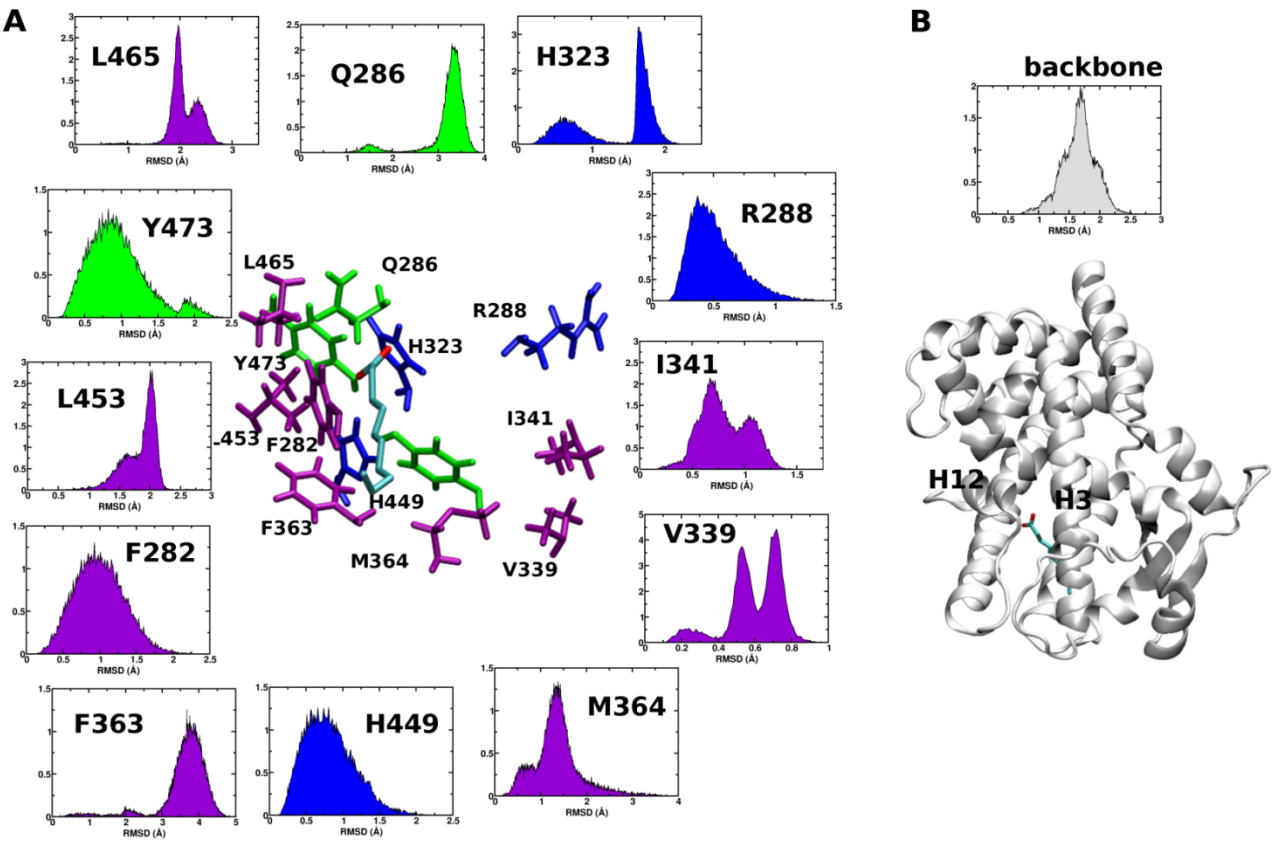

Figure S5. PPARγ structural analysis. A) Distributions of RMSD of the some of the side chains identified to be pertinent for ligand binding through the 10x50ns GCMC/MD simulation of PPAR. Positively charged side chains are colored blue, polar side chains are colored green, and hydrophobic side chains are colored purple. Decanoic acid in the crystallographic orientation is shown in cyan. B) The global conformation of PPAR is maintained through the simulations as evidenced by the distribution of the backbone RMSD. Helices H12 and H3 that flank the LBP1 and LBP2 are marked.

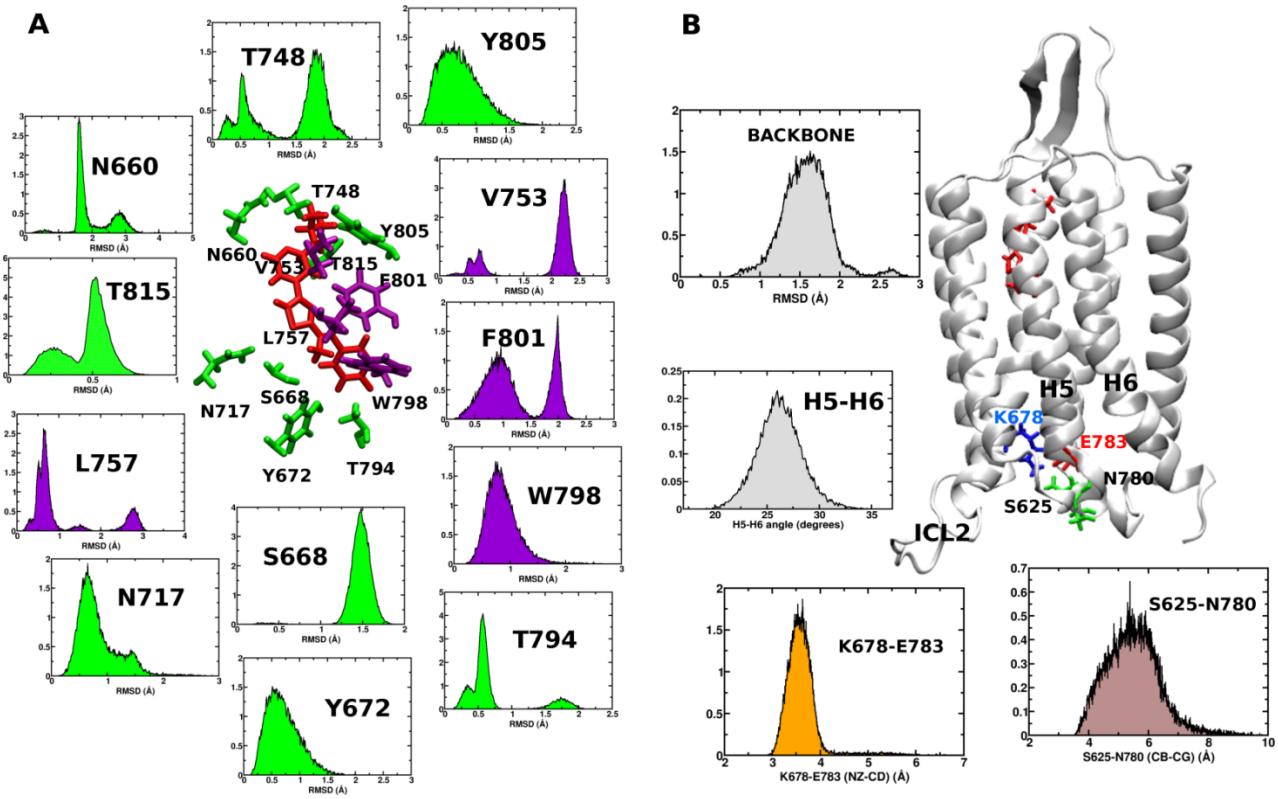

Figure S6. mGluR structural analysis. A) Distributions of RMSD of selected side chains identified to be pertinent for ligand binding through the 10x50ns GCMC/MD simulations of the mGluR. Polar sidechains are colored green and hydrophobic side chains are colored purple. FITM in the crystallographic orientation is shown in red. B) The inactivated conformation of the GPCR is maintained through the simulations as evidenced by the distribution of angle between the TM helices H5 and H6 and the distances between atoms NZ and CD across the salt-bridge forming residues K678-E783 and atoms CB and CG of the hydrogen bond forming residues S625-N780. Tails of the distributions in A & B are not from the end of simulation and occur as random events across the GCMC/MD cycles, possibly an outcome of the molecular fragments influencing the conformational ensembles [53].

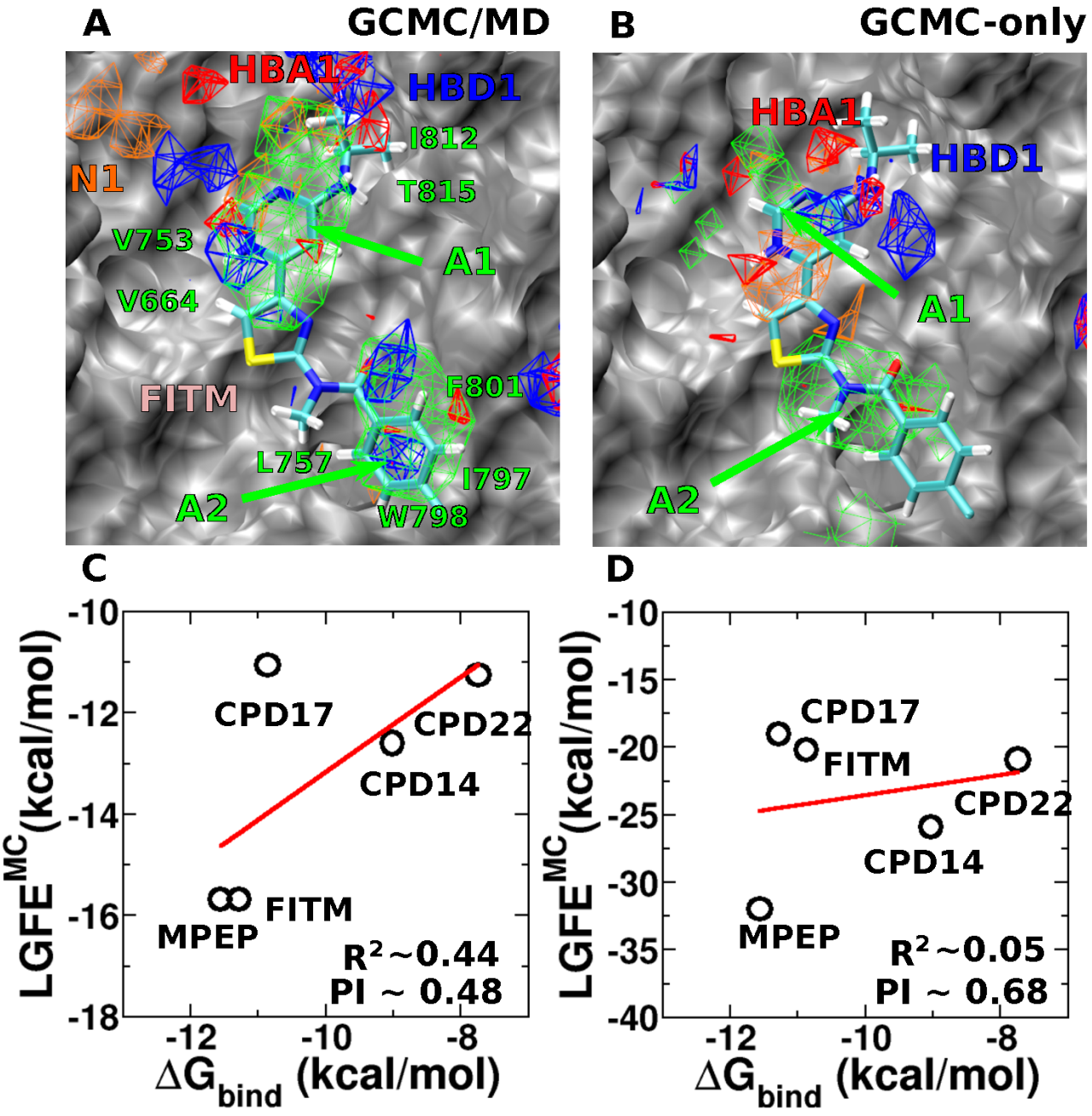

Figure S7. Comparison of FragMaps between A) GCMC/MD vs. B) GCMC-only simulations for the mGluR. All FragMap contours are displayed at -1.2 kcal/mol. C) LGFEs calculated using the FragMaps from GCMC/MD correlated well with ΔG_{bind} (Table S6) with the correlation being lost when D) the LGFEs were calculated using the FragMaps from GCMC-only simulations.

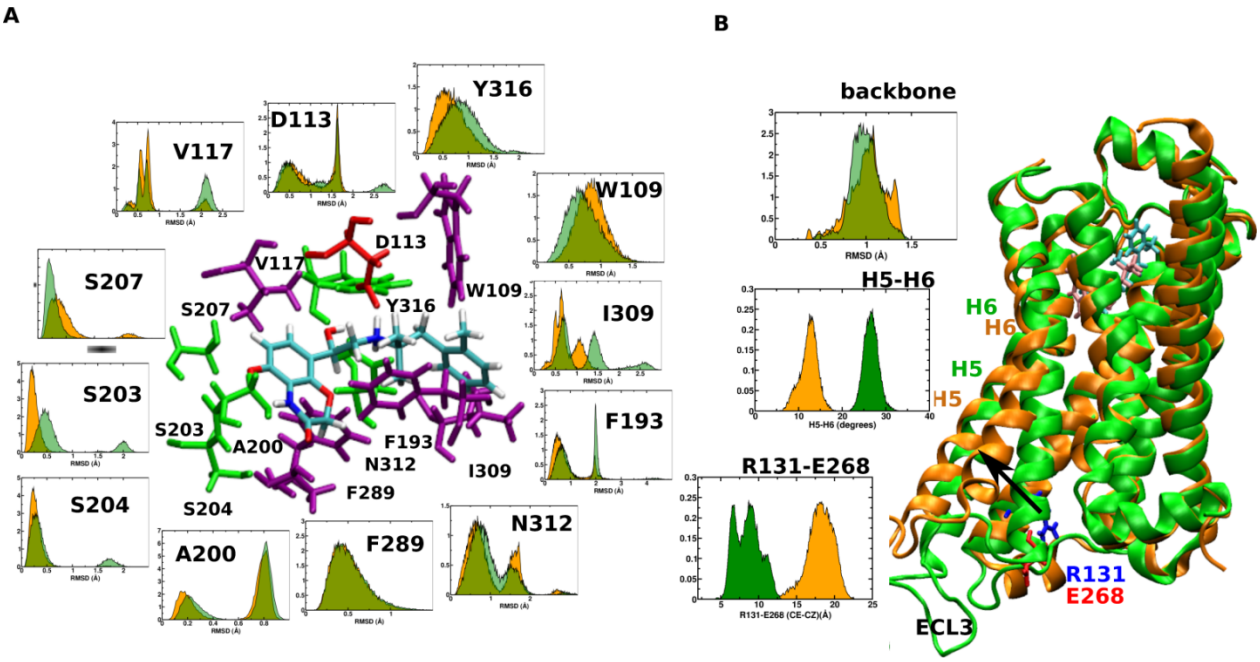

Figure S8. Structural analysis of B2A and B2I. A) Distributions of RMSD of selected side chains identified to be pertinent for the ligand binding through the 10x50 ns GCMC/MD simulation of B2A (orange) and B2I (green). Polar residues are colored green, hydrophobic residues are colored purple and negatively charged Asp is colored red. BI-167107 and the carazolol are colored orange and green respectively. B) The inactive (green) and the active (orange) conformations of the β_2AR are maintained through the simulations as evidenced by the distribution of angle between the TM helices H5 and H6 and the distance between atoms CE and CD across the salt-bridge forming residues R131-E268.

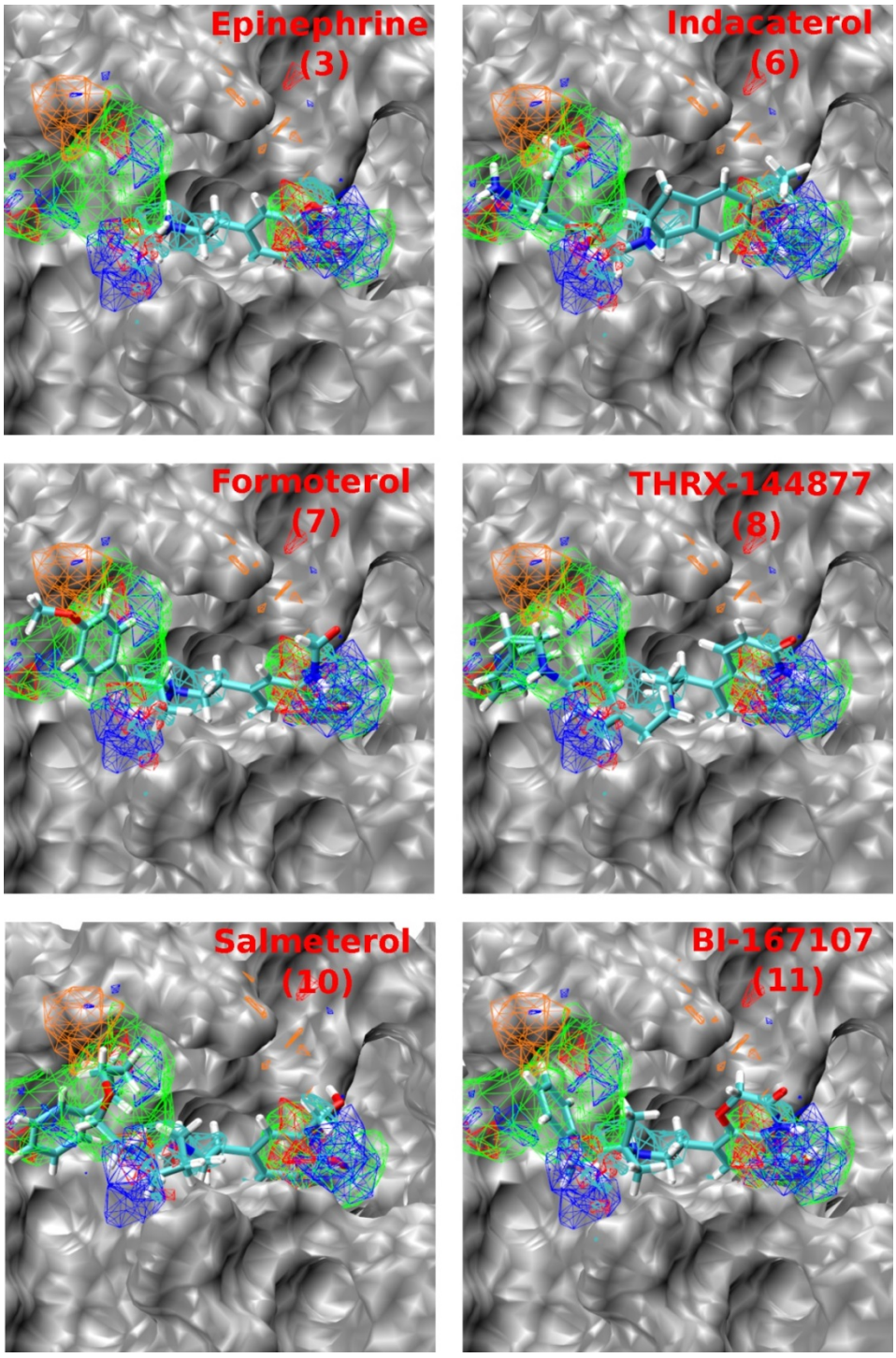

Figure S9. Binding modes of some agonist and partial agonists in the activated conformation of B2A from MC sampling of the ligands in the field of the FragMaps from the B2A SILCS-GCMC/MD.

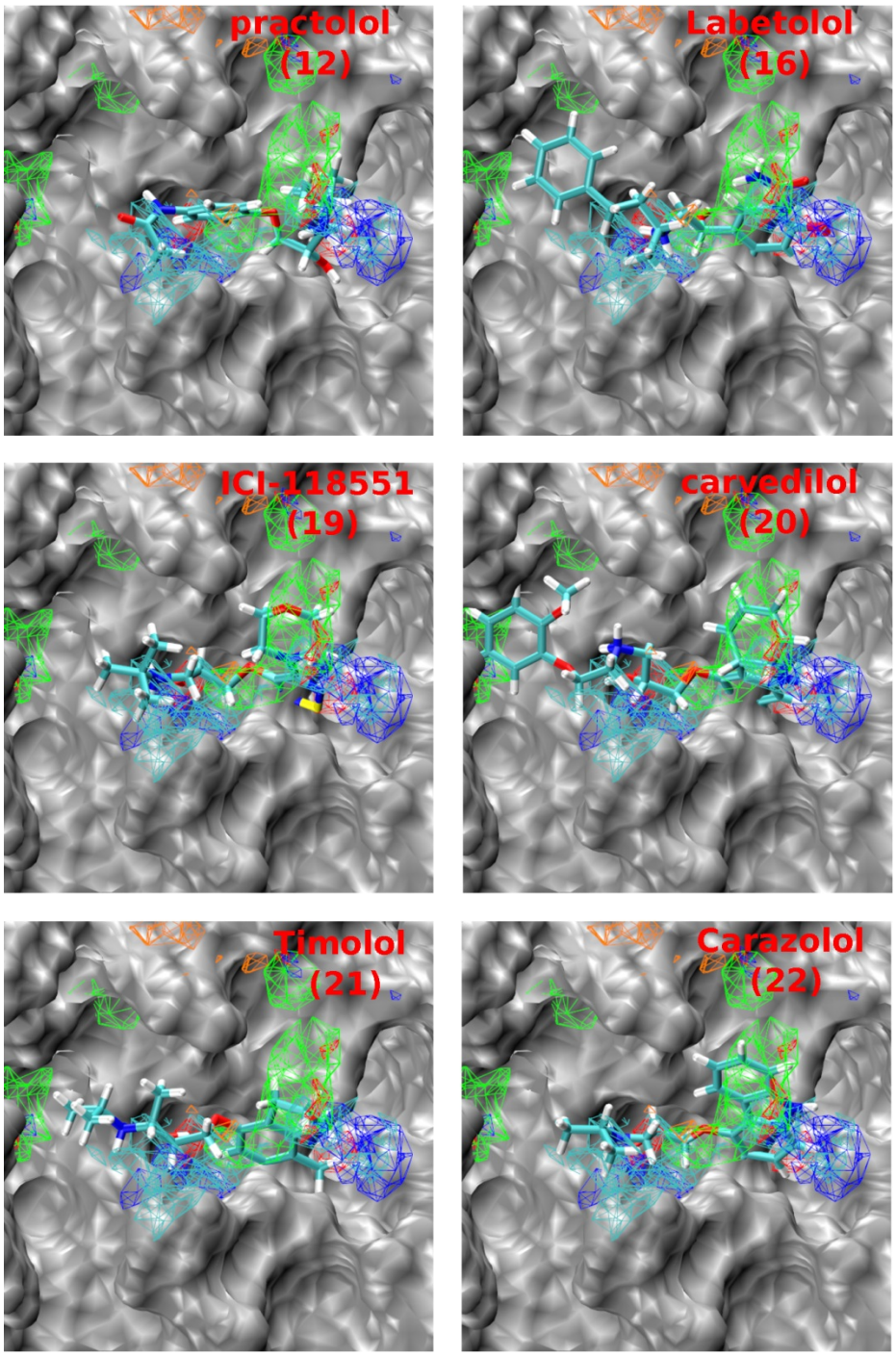

Figure S10. Binding modes of some antagonists and inverse agonists in the inactivated conformation of the β_2AR from MC sampling of the ligands in the field of the FragMaps from B2I SILCS-GCMC/MD.

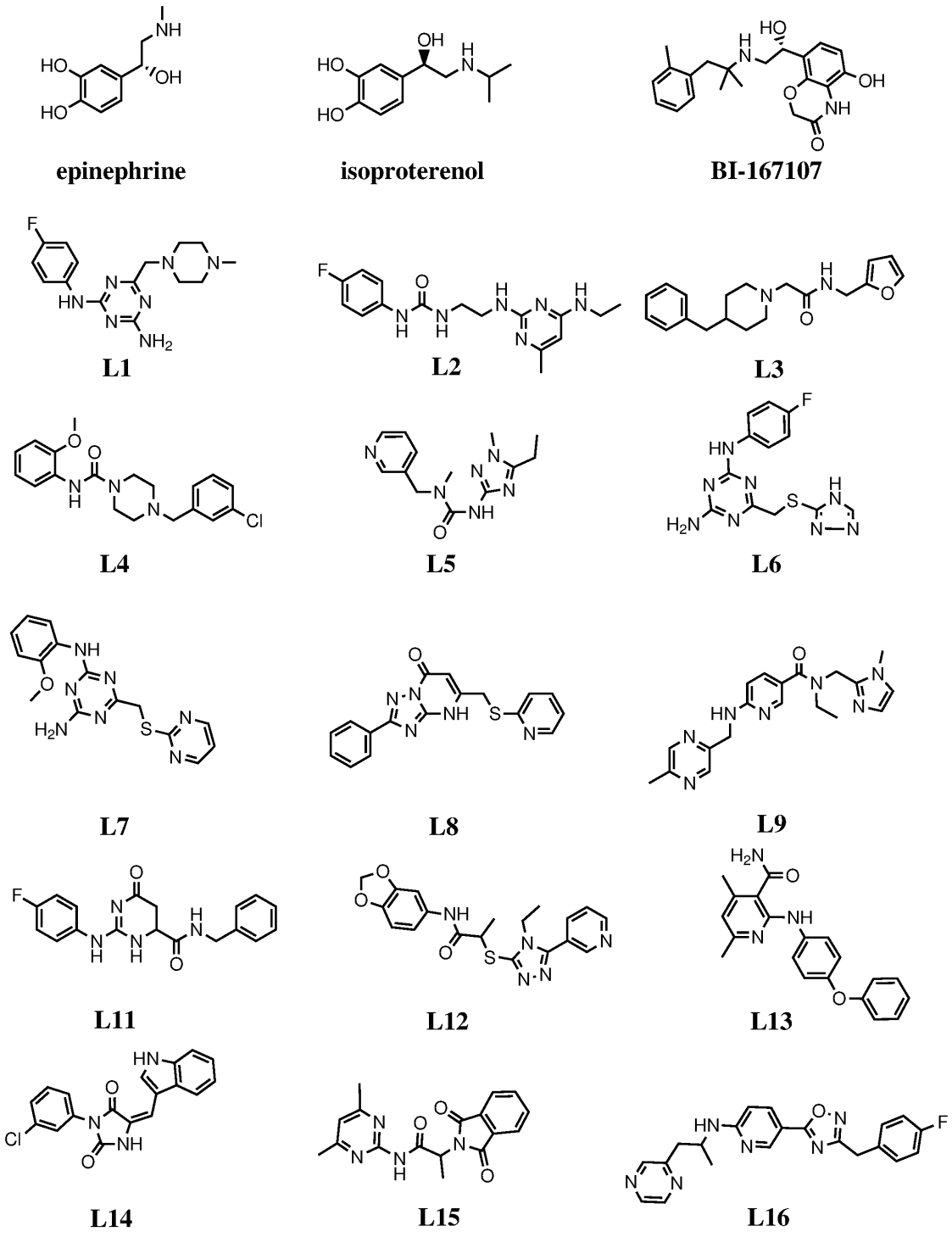

Fig. S11. Known β_2AR agonists and fifteen ligands selected for functional assessment studies from virtual screening driven by differences in FragMaps between the active and inactive states of β_2AR (L1-L16).

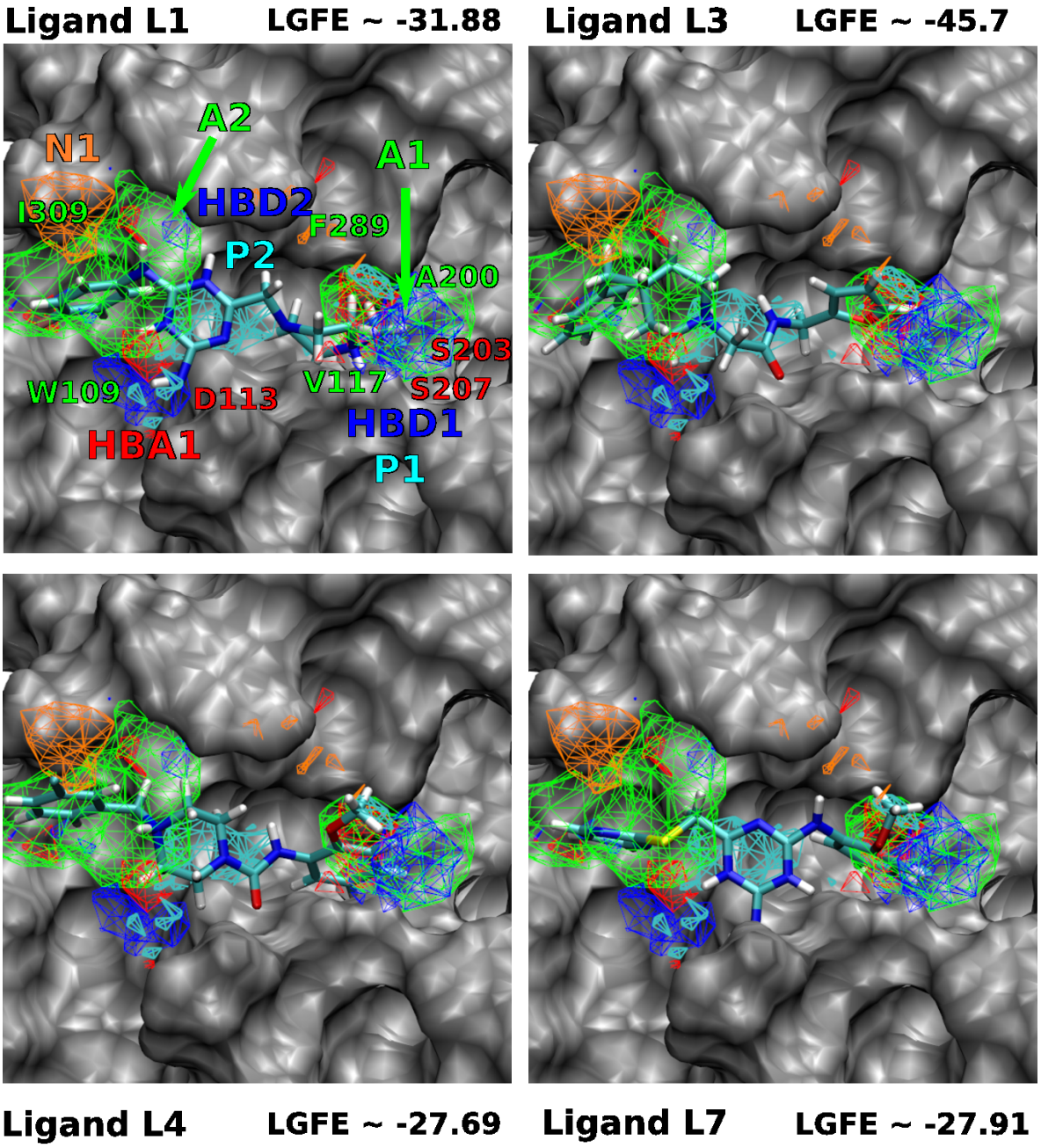

Figure S12. Docked conformations of shortlisted ligands L1, L2, L4 and L7 overlap well with the different FragMaps from B2A simulations, leading to good LGFE scores.

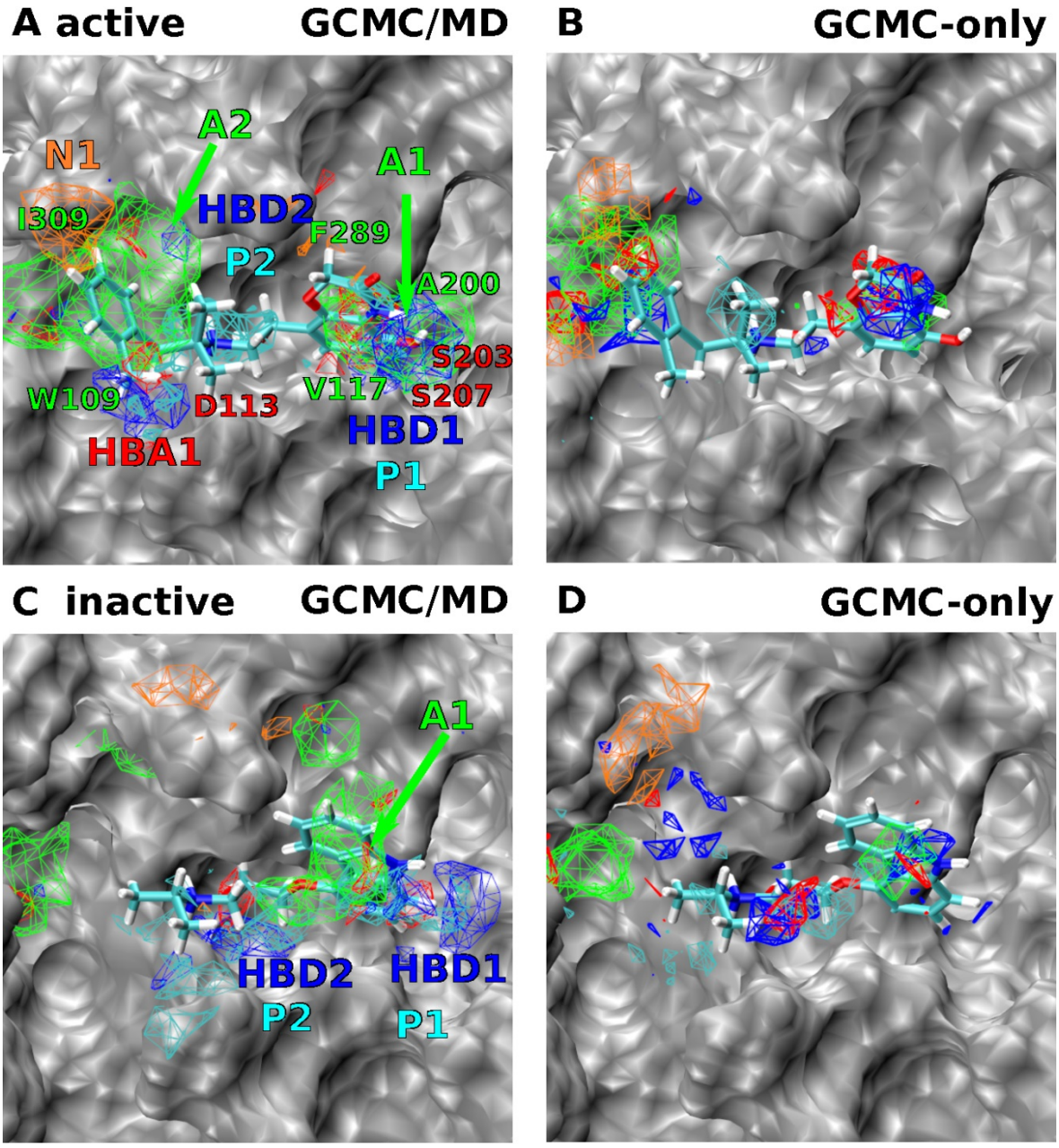

Figure S13. Comparison of FragMaps from the active (A, B) and the inactive (C, D) conformations of the β_2AR between GCMC/MD (A, C) vs. GCMC-only (B, D) simulations. In B, D all FragMap contours are displayed at -1.2 kcal/mol. In A, C HBACC and HBDON FragMaps are displayed at -0.5 kcal/mol, while APOLAR, NEG and PDON FragMaps are displayed at -1.2 kcal/mol.

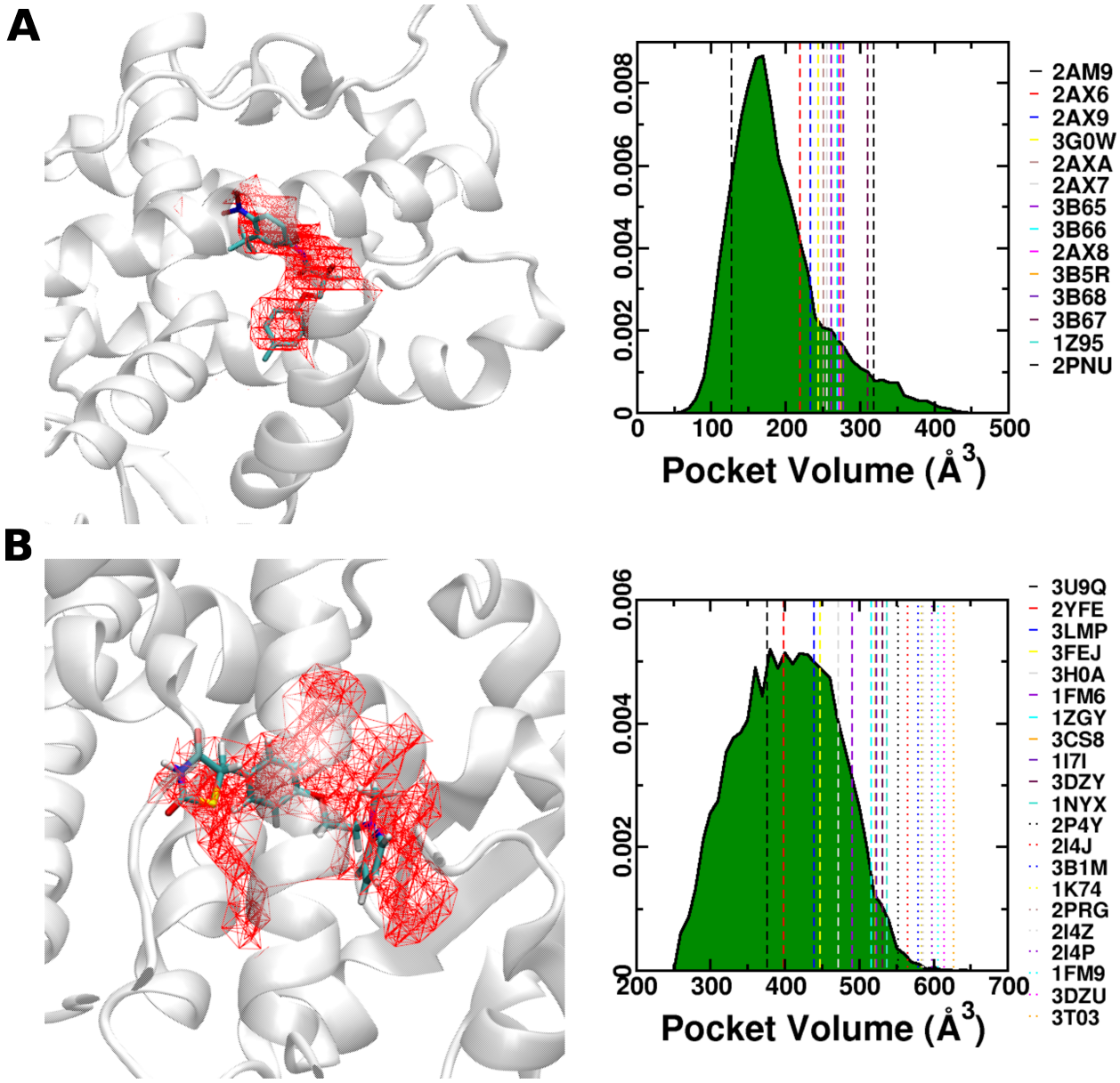

Figure S14. Volumes of the occluded binding pockets of the crystal structures of AR (A) and PPARγ (C) used to initiate the simulations was measured using POVME(red mesh) after the removal of ligand and water molecules from the crystal structures. Ligands S-1 and decanoic acid are shown in the pockets of AR and PPARγ, respectively. Distributions of the pocket volumes calculated using POVME through SILCS GCMC/MD simulations of AR (B) and PPARγ (D). The pocket volumes sampled in the simulations span through the pocket volumes of the known crystal structures bound to different ligands, marked using dashed lines.

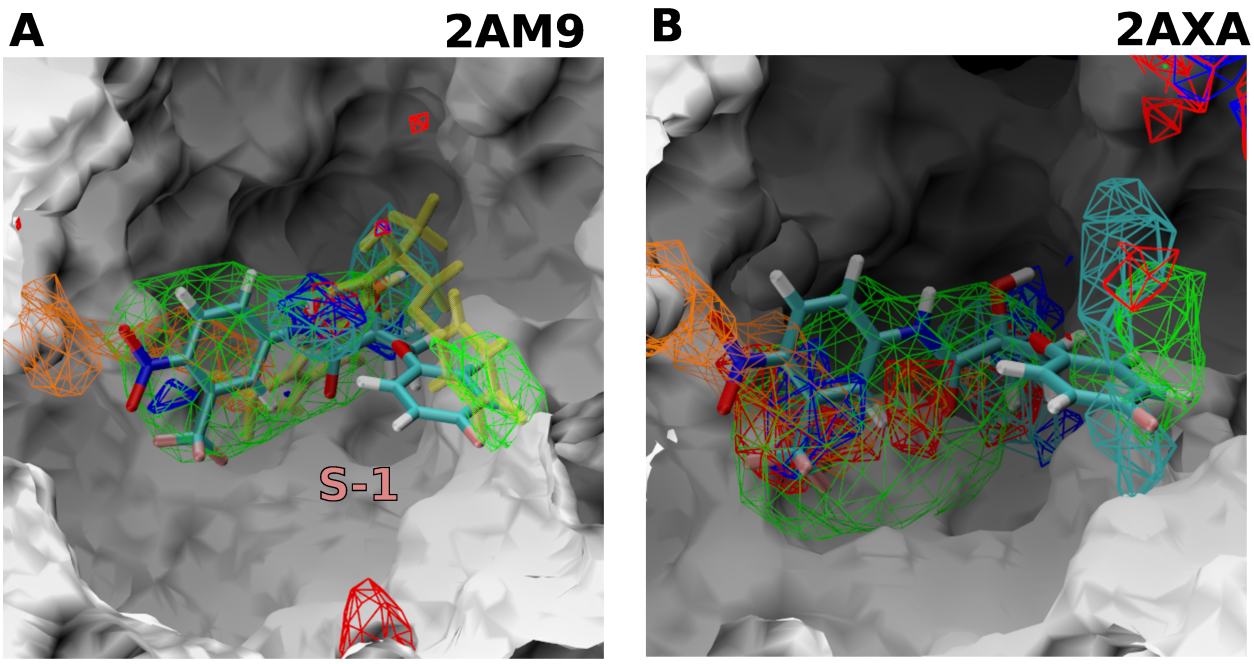

Figure S15. FragMaps of AR obtained using different initial PDB structures 2AM9 (A) and 2AXA (B) have reasonable overlap coefficients between 10x100 cycles of SILCS GCMC/MD: APOLAR : 0.42, NEG: 0.63, POS: 0.74, HBDON: 0.45 and HBACC: 0.42 indicating the FragMaps to be reasonably independent of the initial conformation of the protein.

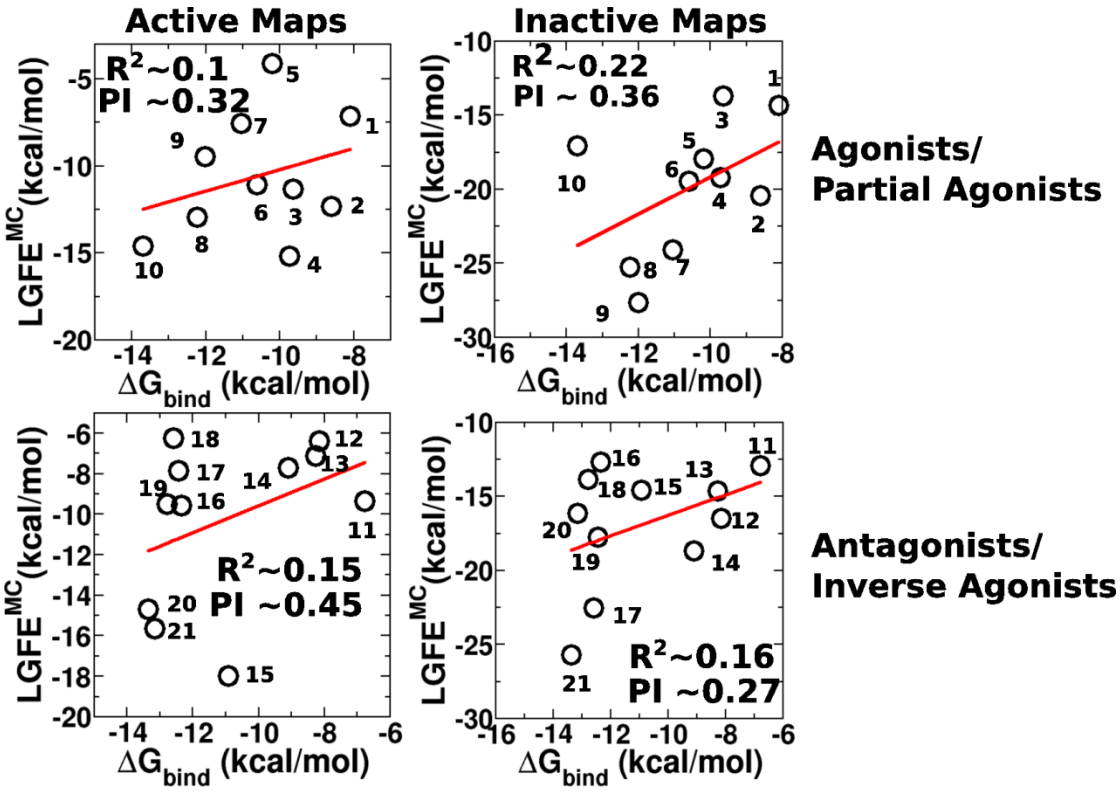

Fig. S16. LGFE versus experimental correlation analysis using GCMC-only FragMaps: LGFE scores are obtained from MC conformational ensembles of the agonists, partial agonists, antagonists and inverse agonists (Table S7) in both the B2A and B2I FragMaps from GCMC-only simulations. LGFEs and the experimental ΔG_{bind} values correlate poorly in all cases indicating the importance of the inclusion of protein flexibility in the SILCS simulations.

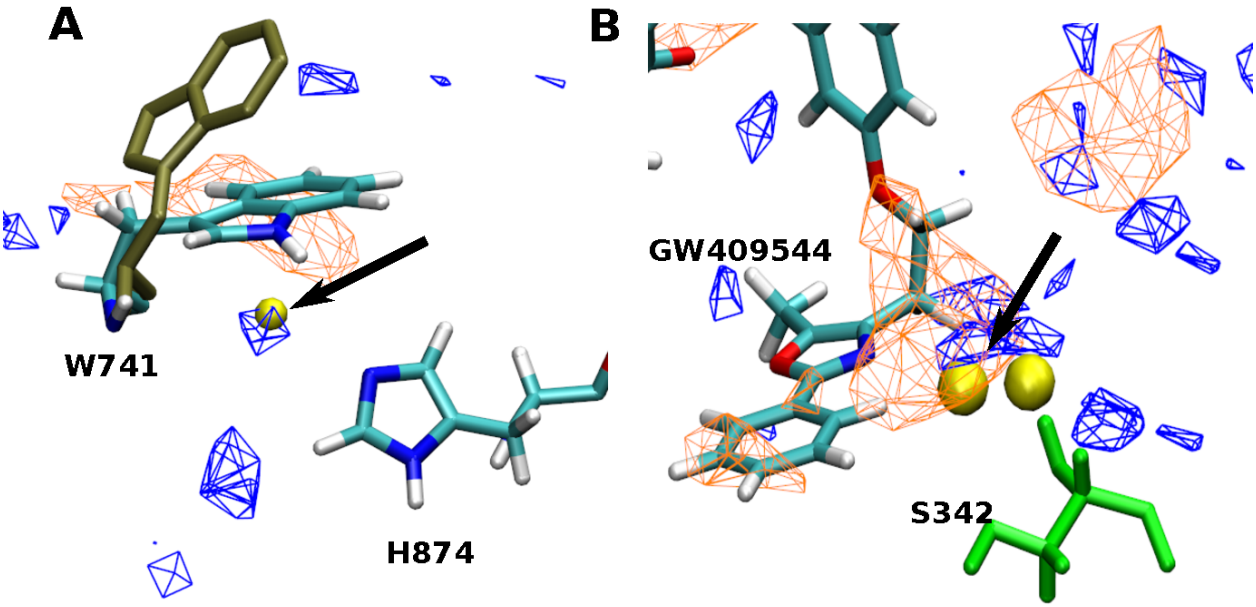

Figure S17. FragMaps tracing water oxygens (colored blue, visualized with a threshold of -0.5 kcal/mol) recapitulate crystallographic water positions (colored yellow) for the ligands in AR and PPARγ. In PPARγ, water oxygen FragMaps overlap with the NEG FragMaps (orange, visualized at a threshold of -1.2 kcal/mol), indicating that this water is displaceable.

References

1. Pereira de Jésus-Tran, K.; Côté, P. L.; Cantin, L.; Blanchet, J.; Labrie, F.; Breton, R. Comparison of Crystal Structures of Human Androgen Receptor Ligand-Binding Domain Complexed with Various Agonists Reveals Molecular Determinants Responsible For Binding Affinity. *Protein Sci* **2006,** 15, 987-999.
2. Malapaka, R. R.; Khoo, S.; Zhang, J.; Choi, J. H.; Zhou, X. E.; Xu, Y.; Gong, Y.; Li, J.; Yong, E.-L.; Chalmers, M. J. Identification and Mechanism of Ten-Carbon Fatty Acid as Modulating Ligand of Peroxisome Proliferator-Activated Receptors. *J Biol Chem* **2012,** 287, 183-195.
3. Fiser, A.; Do, R. K. G.; Šali, A. Modeling of Loops in Protein Structures. *Protein Sci* **2000,** 9, 1753-1773.
4. Shen, M. y.; Sali, A. Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Sci* **2006,** 15, 2507-2524.
5. Levitt, M.; Lifson, S. Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. *J Mol Biol* **1969,** 46, 269-279.
6. Andersen, H. C. Molecular Dynamics Simulations At Constant Pressure And/Or Temperature. *J Chem Phys* **1980,** 72, 2384.
7. Wu, H.; Wang, C.; Gregory, K. J.; Han, G. W.; Cho, H. P.; Xia, Y.; Niswender, C. M.; Katritch, V.; Meiler, J.; Cherezov, V. Structure of a Class C GPCR Metabotropic Glutamate Receptor 1 Bound to an Allosteric Modulator. *Science* **2014,** 344, 58-64.
8. Jo, S.; Lim, J. B.; Klauda, J. B.; Im, W. CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes. *Biophys J* **2009,** 97, 50-58.
9. Durell, S. R.; Brooks, B. R.; Ben-Naim, A. Solvent-Induced Forces Between Two Hydrophilic Groups. *J Phys Chem* **1994,** 98, 2198-2202.
10. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program For Macromolecular Energy, Minimization, And Dynamics Calculations. *J Comput Chem* **1983,** 4, 187-217.
11. Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant-Pressure Molecular-Dynamics Simulation-The Langevin Piston Method. *J Chem Phys* **1995,** 103, 4613-4621.
12. 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. *J Comput Phys* **1977,** 23, 327-341.
13. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N· log (N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1993,** 98, 10089.
14. Allen, M. P.; Tildesley, D. J.; Banavar, J. R. Computer Simulation of Liquids. *Phys Today* **1989,** 42, 105.
15. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput* **2008,** 4, 435-447.
16. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I. CHARMM General Force Field: A Force Field For Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. *J Comput Chem* **2010,** 31, 671-690.
17. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J Chem Theory Comput* **2012,** 8, 3257-3273.

18. Raman, E. P.; Yu, W.; Guvench, O.; MacKerell Jr, A. D. Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J Chem Info Model* **2011,** 51, 877-896.
19. Guvench, O.; MacKerell, A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLOS Comput Biol* **2009,** 5, e1000435.
20. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K. High-Resolution Crystal Structure of an Engineered Human β2-Adrenergic G Protein–Coupled Receptor. *Science* **2007,** 318, 1258-1265.
21. Rasmussen, S. G.; Choi, H.-J.; Fung, J. J.; Pardon, E.; Casarosa, P.; Chae, P. S.; DeVree, B. T.; Rosenbaum, D. M.; Thian, F. S.; Kobilka, T. S. Structure of a Nanobody-Stabilized Active State of the β2 Adrenoceptor. *Nature* **2011,** 469, 175-180.
22. Fiser, A.; Do, R. K. G.; Šali, A. Modeling of Loops in Protein Structures. *Protein Sci* **2008,** 9, 1753-1773.
23. Woo, H.-J.; Dinner, A. R.; Roux, B. Grand Canonical Monte Carlo Simulations of Water in Protein Environments. *J Chem Phys* **2004,** 121, 6392.
24. Mezei, M. Grand-Canonical Ensemble Monte Carlo Study of Dense Liquid: Lennard-Jones, Soft spheres and Water. *Mol Phys* **1987,** 61, 565-582.
25. Mezei, M. A cavity-biased (T, V, μ) Monte Carlo Method for the Computer Simulation of Fluids. *Mol Phys* **1980,** 40, 901-906.
26. Jayaram, B.; Beveridge, D. Grand canonical Monte Carlo Simulations on Aqueous Solutions of Sodium Chloride and Sodium DNA: Excess Chemical Potentials and Sources of Nonideality in Electrolyte and Polyelectrolyte Solutions. *J Phys Chem* **1991,** 95, 2506-2516.
27. Nosé, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol Phys* **1984,** 52, 255-268.
28. Hoover, W. G. Canonical dynamics: Equilibrium Phase-Space Distributions. *Phys Rev A* **1985,** 31, 1695.
29. Parrinello, M.; Rahman, A. Polymorphic Transitions In Single Crystals: A New Molecular Dynamics Method. *J Appl Phys* **1981,** 52, 7182-7190.
30. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997,** 18, 1463-1472.
31. Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D. Sampling of Organic Solutes in Aqueous and Heterogeneous Environments using Oscillating Excess Chemical Potentials in Grand Canonical-Like Monte Carlo-Molecular Dynamics Simulations. *J Chem Theory Comput* **2014,** 10, 2281-2290.
32. Raman, E. P.; Yu, W.; Lakkaraju, S. K.; MacKerell Jr, A. D. Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach. *J Chem Info Model* **2013,** 53, 3384-3398.
33. Pochetti, G.; Godio, C.; Mitro, N.; Caruso, D.; Galmozzi, A.; Scurati, S.; Loiodice, F.; Fracchiolla, G.; Tortorella, P.; Laghezza, A. Insights into the Mechanism of Partial Agonism Crystal Structures Of The Peroxisome Proliferator-Activated Receptor γ Ligand-Binding Domain In The Complex With Two Enantiomeric Ligands. *J Biol Chem* **2007,** 282, 17314-17324.
34. Grether, U.; Bénardeau, A.; Benz, J.; Binggeli, A.; Blum, D.; Hilpert, H.; Kuhn, B.; Märki, H. P.; Meyer, M.; Mohr, P. Design and Biological Evaluation of Novel, Balanced Dual PPARα/γ Agonists. *ChemMedChem* **2009,** 4, 951-956.
35. Henke, B. R. Peroxisome Proliferator-Activated Receptor α/γ Dual Agonists for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2004,** 47, 4118-4127.
36. Thor, M.; Beierlein, K.; Dykes, G.; Gustavsson, A.-L.; Heidrich, J.; Jendeberg, L.; Lindqvist, B.; Pegurier, C.; Roussel, P.; Slater, M. Synthesis and Pharmacological Evaluation of a New Class of Peroxisome Proliferator-Activated Receptor Modulators. *Bioorgan Med Chem Lett* **2002,** 12, 3565-3567.

37. Parmenon, C.; Guillard, J.; Caignard, D.-H.; Hennuyer, N.; Staels, B.; Audinot-Bouchez, V.; Boutin, J.-A.; Dacquet, C.; Ktorza, A.; Viaud-Massuard, M.-C. 4, 4-Dimethyl-1, 2, 3, 4-tetrahydroquinoline-based PPARα/γ Agonists. Part I: Synthesis and Pharmacological Evaluation. *Bioorgan Med Chem Lett* **2008,** 18, 1617-1622.
38. Amato, A. A.; Rajagopalan, S.; Lin, J. Z.; Carvalho, B. M.; Figueira, A. C.; Lu, J.; Ayers, S. D.; Mottin, M.; Silveira, R. L.; Souza, P. C. GQ-16, A Novel Peroxisome Proliferator-Activated Receptor γ (PPARγ) Ligand, Promotes Insulin Sensitization Without Weight Gain. *J Biol Chem* **2012,** 287, 28169-28179.
39. Liu, K. G.; Lambert, M. H.; Ayscue, A. H.; Henke, B. R.; Leesnitzer, L. M.; Oliver Jr, W. R.; Plunket, K. D.; Xu, H. E.; Sternbach, D. D.; Willson, T. M. Synthesis and Biological Activity of l-Tyrosine-based PPARγ Agonists with Reduced Molecular Weight. *Bioorgan Med Chem Lett* **2001,** 11, 3111-3113.
40. Connors, R. V.; Wang, Z.; Harrison, M.; Zhang, A.; Wanska, M.; Hiscock, S.; Fox, B.; Dore, M.; Labelle, M.; Sudom, A. Identification of a PPARδ Agonist with Partial Agonistic Activity on PPARγ. *Bioorgan Med Chem Lett* **2009,** 19, 3550-3554.
41. Furukawa, A.; Arita, T.; Satoh, S.; Wakabayashi, K.; Hayashi, S.; Matsui, Y.; Araki, K.; Kuroha, M.; Ohsumi, J. Discovery of a Novel Selective PPARγ Modulator From (−)-Cercosporamide Derivatives. *Bioorgan Med Chem Lett* **2010,** 20, 2095-2098.
42. Weidner, C.; de Groot, J. C.; Prasad, A.; Freiwald, A.; Quedenau, C.; Kliem, M.; Witzke, A.; Kodelja, V.; Han, C.-T.; Giegold, S. Amorfrutins Are Potent Antidiabetic Dietary Natural Products. *Proc Natl Acad Sci USA* **2012,** 109, 7257-7262.
43. Einstein, M.; Akiyama, T. E.; Castriota, G. A.; Wang, C. F.; McKeever, B.; Mosley, R. T.; Becker, J. W.; Moller, D. E.; Meinke, P. T.; Wood, H. B. The Differential Interactions Of Peroxisome Proliferator-Activated Receptor γ Ligands with Tyr473 is a Physical Basis for Their Unique Biological Activities. *Mol Pharmacol* **2008,** 73, 62-74.
44. Wakabayashi, K.; Hayashi, S.; Matsui, Y.; Matsumoto, T.; Furukawa, A.; Kuroha, M.; Tanaka, N.; Inaba, T.; Kanda, S.; Tanaka, J. Pharmacology and in Vitro Profiling of a Novel Peroxisome Proliferator-Activated Receptor. GAMMA. Ligand, Cerco-A. *Biological and Pharmaceutical Bulletin* **2011,** 34, 1094-1104.
45. Milbank, J. B.; Knauer, C. S.; Augelli-Szafran, C. E.; Sakkab-Tan, A. T.; Lin, K. K.; Yamagata, K.; Hoffman, J. K.; Zhuang, N.; Thomas, J.; Galatsis, P. Rational Design of Seven-Arylquinolines As Non-Competitive Metabotropic Glutamate Receptor Subtype 5 Antagonists. *Bioorgan Med Chem Lett* **2007,** 17, 4415-4418.
46. Larsson, I. A Comparison Between Clenbuterol, Salbutamol and Terbutaline in Relation To Receptor Binding And In Vitro Relaxation of Equine Tracheal Muscle. *J Vet Pharmacol Therapeut* **1998,** 21, 388-392.
47. Toews, M. L.; Harden, T. K.; Perkins, J. P. High-Affinity Binding of Agonists to Beta-Adrenergic Receptors on Intact Cells. *Proc Natl Acad Sci USA* **1983,** 80, 3553-3557.
48. Eimerl, S.; Schramm, M.; Lok, S.; Goodman, M.; Khan, M.; Melmon, K. The Four Stereoisomers of a High Potency Congener of Isoproterenol: Biological Activity and the Relationship Between the Native and the Chemically Inserted Asymmetric Carbon. *Biochem Pharmacol* **1987,** 36, 3523-3527.
49. Düringer, C.; Grundström, G.; Gürcan, E.; Dainty, I. A.; Lawson, M.; Korn, S. H.; Jerre, A.; Håkansson, H. F.; Wieslander, E.; Fredriksson, K. Agonist-Specific Patterns of β2-Adrenoceptor Responses in Human Airway Cells During Prolonged Exposure. *Brit J Pharmacol* **2009,** 158, 169-179.
50. Baker, J. G. The Selectivity of β-Adrenoceptor Antagonists at the Human β1, β2 and β3 Adrenoceptors. *Brit J Pharmacol* **2005,** 144, 317-322.
51. Chidiac, P.; Hebert, T. E.; Valiquette, M.; Dennis, M.; Bouvier, M. Inverse Agonist Activity of Beta-Adrenergic Antagonists. *Mol Pharmacol* **1994,** 45, 490-499.

52. Innis, R. B.; Corrêa, F.; Snyder, S. H. Carazolol, an extremely potent β-adrenergic blocker: Binding to β-receptors in brain membranes. *Life sciences* **1979,** 24, 2255-2264.
53. Eyrisch, S.; Helms, V. Transient Pockets On Protein Surfaces Involved in Protein-Protein Interaction. *J. Med. Chem.* **2007,** 50, 3457-3464.