nature chemistry

ARTICLES

PUBLISHED ONLINE: 15 DECEMBER 2013 | DOI: 10.1038/NCHEM.1821

Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways

Kai J. Kohlhoff[1,4†*], Diwakar Shukla[1,2†], Morgan Lawrenz[2†], Gregory R. Bowman[2], David E. Konerding[4], Dan Belov[4], Russ B. Altman[1,3*] and Vijay S. Pande[2*]

Simulations can provide tremendous insight into the atomistic details of biological mechanisms, but micro- to millisecond timescales are historically only accessible on dedicated supercomputers. We demonstrate that cloud computing is a viable alternative that brings long-timescale processes within reach of a broader community. We used Google's Exacycle cloud-computing platform to simulate two milliseconds of dynamics of a major drug target, the G-protein-coupled receptor β_2AR. Markov state models aggregate independent simulations into a single statistical model that is validated by previous computational and experimental results. Moreover, our models provide an atomistic description of the activation of a G-protein-coupled receptor and reveal multiple activation pathways. Agonists and inverse agonists interact differentially with these pathways, with profound implications for drug design.

G-protein-coupled receptors (GPCRs) are a family of membrane-bound α-helical proteins that regulate a large variety of physiological processes by transmitting signals from extracellular binding of diverse ligands to intracellular signalling molecules. These proteins are exceedingly prominent drug targets, responsible for at least one-third of all marketable drugs and half of the total market volume for pharmaceuticals[1]. The β_2-adrenergic receptor (β_2AR) is implicated in type 2 diabetes, obesity and asthma, and is a member of the class A, rhodopsin-like GPCRs. These proteins share a highly conserved motif of seven transmembrane helices connected by three extracellular and three intracellular loops (ICLs). β_2AR is experimentally well studied, and high-resolution X-ray structures of both the inactive[1] and several active states[2,3] have been determined in recent years. However, despite this rapid progress towards understanding of these important molecules, little is known about the mechanisms by which small molecules modulate their activity.

Molecular dynamics (MD) simulations have already begun to provide insights into the underlying dynamics and structural ensembles of GPCRs[4–8]. However, many phenomena of interest still remain out of reach. For example, one recent study used special-purpose hardware[9] to reach an unprecedented total simulation time of several hundred microseconds[5]. These results provided insights into the mechanism of deactivation, but were unable to capture activation. Moreover, it remains unclear how to make further advances, particularly for researchers without access to such specialized hardware.

To capture the mechanism of β_2AR activation, we followed an alternative approach to MD in which we extended the principles behind the volunteer-distributed computing platform Folding@home[10] to cloud computing more broadly. Specifically, we ran tens of thousands of independent simulations on Google Exacycle[11], a cloud-computing initiative that provides an interface for running compute jobs directly on Google's production infrastructure. Markov state models (MSMs) were then used to stitch together these massively parallel simulations into a single statistical model that captured rare events on timescales far longer than those reached by any individual simulation[12–14]. Our approach reproduces a variety of previous experimental and computational results, including mutual information networks of correlated residues, and we explain how key structural elements change along ligand-modulated activation pathways. Moreover, we show that the MSMs can improve our understanding of drug efficacy at GPCR receptors and can be incorporated into an effective structure-based drug-design approach.

Results

Using our cloud-based approach, we simulated 2.15 ms of β_2AR dynamics. Simulations were initiated from both an inactive (PDB 2RH1)[1] and active (PDB 3P0G)[2] crystal structure of β_2AR. We also ran simulations in the presence of two ligands (the partial inverse agonist carazolol and the full agonist BI-167107) to understand how these small molecules alter the behaviour of β_2AR. We find that activation and deactivation proceed through multiple pathways and typically visit metastable intermediate states. Our MSMs provide a human-readable view of how ligands modulate the complex conformational landscape of β_2AR and improve performance of computer-aided drug-design approaches. More generally, our cloud-based approach should be a powerful and broadly available tool for studying many biological systems.

MSMs predict ligand-specific intermediate states in activation dynamics. To elucidate the mechanism of receptor activation, we built kinetic network MSMs from our data set. MD simulations describe intrinsic receptor dynamics in atomistic detail, and an MSM provides a summarized view of the ensemble of

[1]Department of Bioengineering, Stanford University, 318 Campus Drive, Stanford, California 94305, USA, [2]Department of Chemistry, Stanford University, 318 Campus Drive, Stanford, California 94305, USA, [3]Department of Genetics, Stanford University, 318 Campus Drive, Stanford, California 94305, USA, [4]Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA, †These authors contributed equally. *e-mail: kohlhoff@google.com; russ.altman@stanford.edu; pande@stanford.edu

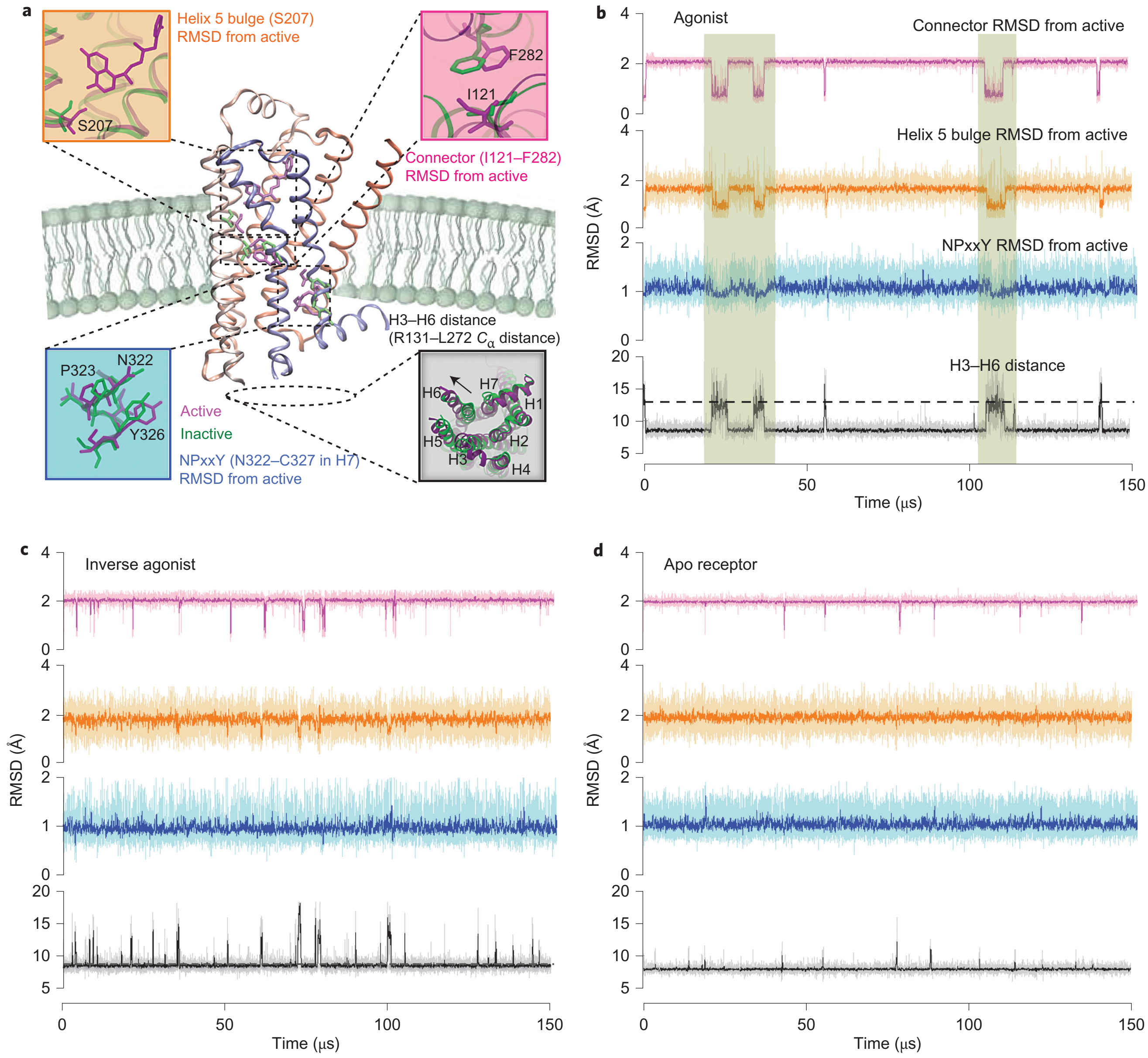

Figure 1 | MSM activation trajectories on the submillisecond timescale. a, The key differences between active and inactive structures of β_2AR. The root mean squared deviation (RMSD) of the connector region (I121–F282), the H5 bulge (S207–F208) RMSD from active crystal structure (3P0G), the distance between Helix 3 (H3) and Helix 6 (H6, measured as the R131–L272 distance) and the RMSD of NPxxY region (N322–C327) in Helix 7 (H7) from the active crystal structure (3P0G). Active crystal structure values are marked with dashed lines. **b–d**, MSM activation trajectories of 150 μs were built using the data from simulations of GPCR bound to agonist BI-167107 (**b**), inverse agonist carazolol (**c**) and the apo receptor (**d**). Agonist-bound simulations stabilize active-like conformations (highlighted with boxes) throughout the trajectory and deactivate in 2.5±0.05 μs. Inverse agonist-bound simulations deactivate from active states slightly faster, at 1.3±0.1 μs, and apo simulations deactivate in 0.67±0.02 μs (see Supplementary Figs 5 and 6). These timescales corroborate previous simulations of deactivation[5] and experiments when keeping in mind that only G protein binding can truly stabilize the active states[30].

spontaneous fluctuations exhibited by the molecule at equilibrium[15]. This helps to identify key conformational states of the receptor, and to quantify the state thermodynamic populations and the kinetics of state transitions. We built 3,000-state MSMs using clustering along four structural metrics that represent the key differences between the active[2,3] and inactive[1] crystal structures of β_2AR (see Methods) and map out the transitions between all states. Other structural metrics not highlighted in the main text, which include dynamics of the extracellular loops and movement of all transmembrane helices, are discussed in the Supplementary Information.

Our MSM provides, for the first time, a unique and novel insight into the ligand-modulated activation of β_2AR. Using the MSM, we generated 150 μs activation trajectories (see Methods). In Fig. 1b–d, these trajectories are shown as projections along the four described structural criteria for the receptor with agonist BI-167107 (Fig. 1b) and inverse agonist carazolol (Fig. 1c) bound, as well as the apo receptor (Fig. 1d). Agonist-bound simulations stabilize active-like conformations (boxed regions) for between 1.5 and 5.25 μs, but inverse agonist and apo simulations do not sample active-state conformations along all structural metrics. The differences in activation dynamics are consistent with the biological function of the ligands that we have simulated, and confirm observables from previous simulations[5].

Figure 2a shows the network representation of the 3,000-state MSM built from the simulations of the agonist-bound receptor.

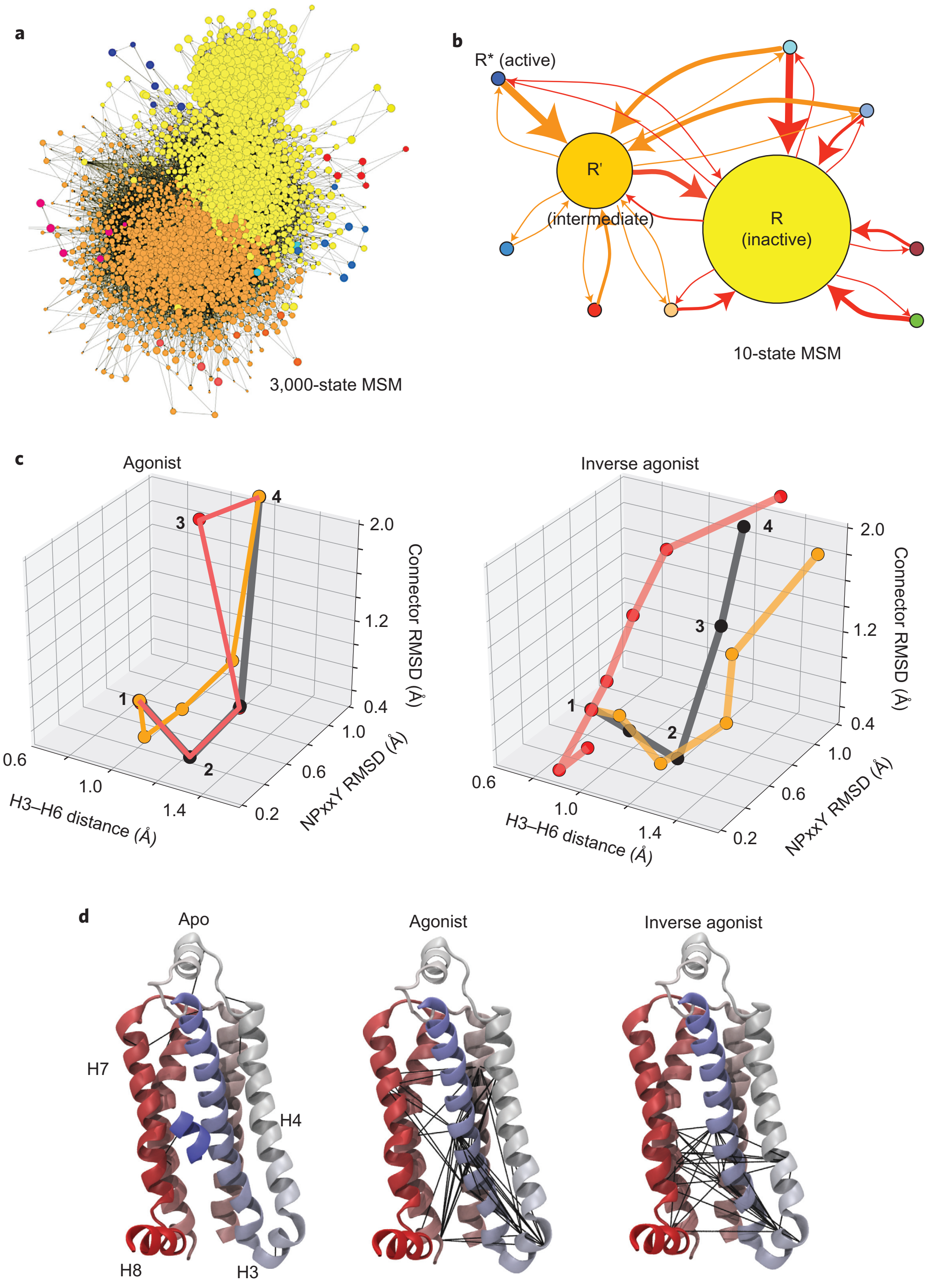

Figure 2 | MSMs and high-flux activation pathways for agonist and inverse agonist bound simulations. a, Network representation of the 3,000-state MSM built from the simulations of agonist-bound GPCR with each circle representing an individual conformational state. **b**, Ten-state MSMs built from the 3,000-state MSMs using spectral clustering methods to identify kinetically relevant states. The circles in the 3,000-state MSM are coloured according to their membership in the coarse-grained ten-state MSM. The weight of arrow indicates the transition probability between states. **c**, Pathways are shown as states (circles) connected along the three-dimension (3D) reaction coordinate used, in part, to build the MSM. Pathway connections are scaled by the path flux relative to the highest flux in black; for inverse agonist pathways, red is 61% and orange is 51% of the maximum; for agonist, red is 48% and orange is 35%. **d**, Mutual information networks of dynamically correlated residues. Black lines indicate connected residue pairs, and only Helices 3–8 are shown in the image for clarity. Agonist-bound simulations reveal a network of residues that connect the extra- and intracellular parts of the receptor to stabilize active states, whereas the inverse agonist eliminates these connections and blocks activation.

Such a detailed picture of β_2AR kinetics is useful for delineating activation pathways at atomistic resolution, which is outside the limits of temporal and spatial resolution of most experimental techniques. MSMs also provide a way to simplify this picture by discarding fast conformational dynamics to obtain a human-comprehensible model of receptor dynamics that comprises lumped, kinetically relevant states. This lumping procedure was applied to reduce the 3,000-state model to a simplified ten-state model of β_2AR dynamics (Fig. 2b). This macro state model of β_2AR reveals two highly connected states, which are identified as the inactive state (I) and the intermediate state (R′), and several states with fewer connections, including the active state (R*). The intermediate

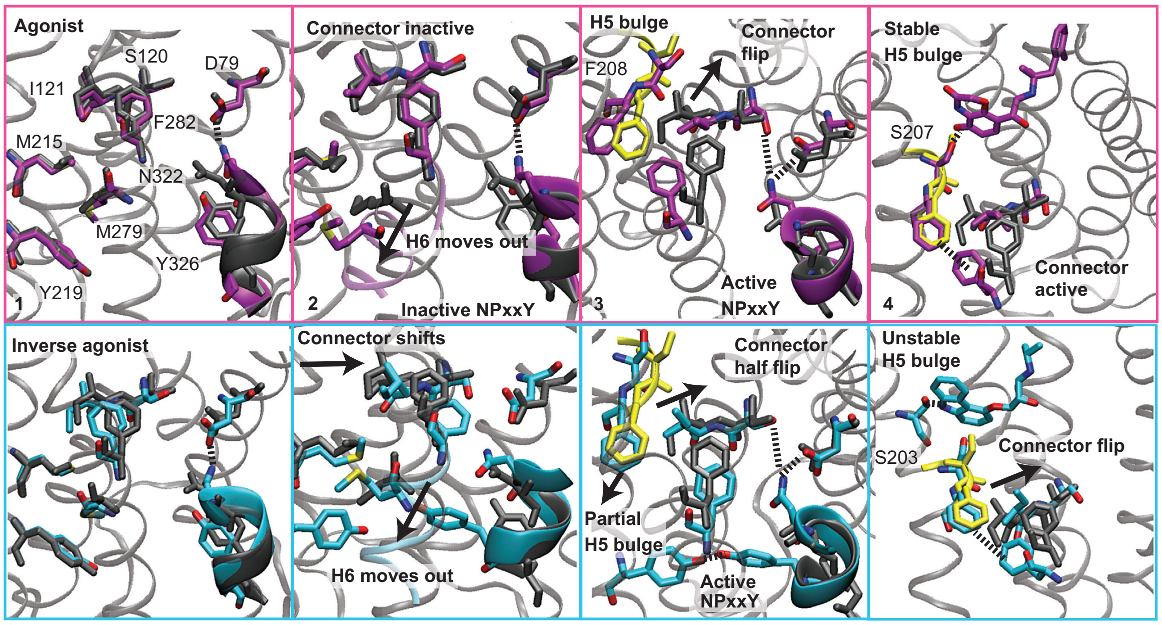

Figure 3 | Structural details of the activation pathways. Representative structures from states along the activation pathways in Fig. 2c labelled by number in four panels. The transition from inactive conformations in panel 1 proceeds via Helix 6 outwards movement (panel 2), the switching of M215 interactions from connector I121 to F282 as these residues flip conformation (panel 3) and changes in the Helix 5 region around F208, S207and S203 (panel 4) that form ligand-mediated interactions for stabilizing active conformations that can be selected by G proteins. Residues in grey are the aligned inactive crystal (2RH1) conformations; residues in yellow are from the active crystal structure (3P0G).

state (R′) is found to be different for the agonist and the inverse agonist (see Supplementary Figs 7–9). For an agonist-bound receptor, the key structural difference between the inactive and intermediate states is the flipping of the I121 and F282 residues in the connector region (shown in Fig. 1a), whereas for the inverse agonist, the twisting of the Helix 7 NPxxY region residues distinguishes the two states. The intermediate state for the apo receptor is the same as the intermediate state in the MSM of the agonist-bound receptor.

β_2AR activation occurs through many parallel pathways. Transition-path theory[16–18] analysis of the MSMs indicates a multitude of pathways with similar flux (see Supplementary Fig. 10). Projection of these pathways along three of the four criteria used to construct the MSMs and described in Fig. 1 allows the simplest view of the predominant pathways for activation. In Fig. 2c, unique high-flux pathways are shown. We employed experimentally validated mutual information analysis (see Supplementary Fig. 11 and Supplementary Tables 1–3) to find correlated residue networks that differ for simulations with different ligands and for the apo receptor (Fig. 2d). The agonist BI-167107 strengthens correlations between extracellular and intracellular residue groups, but the inverse agonist (carazolol) disconnects these groups, and the apo receptor displays only indiscriminate correlations. These different networks of correlated residues give rise to the distinct activation pathways shown in Fig. 2c. We map out the structural changes along the highest flux pathways in Fig. 3 and highlight the most general trends for activation.

Figure 3 (panel 1) shows inactive conformations for all structural metrics. The Helix 3 to Helix 6 distance is ~8 Å, with the ionic lock formed for 40–50% of the time in our simulations. The NPxxY region is in an extended α-helix, and the inactive connectors I121 and F282 are stabilized by M215 and M279, respectively. These methionines undergo significant conformational changes during activation in our simulations (Supplementary Fig. 17), as seen in crystal structures[2,3] and NMR experiments[8,19]. Shown in panel 2 (Fig. 3), activation occurs initially with Helix 6 moving away from Helix 3, and M279 moves outwards, towards its active crystal pose. For different pathways, this helix change occurs prior to other metric changes, or more gradually, concomitant with NPxxY and connector changes (Fig. 2c). In panel 3 (Fig. 3), the active NPxxY conformation is stabilized for both ligands by N322 hydrogen bonding to S120 and D79—the latter aspartic acid is known to affect binding of agonists and G protein transfer[20] (Supplementary Fig. 18). For inverse agonist simulations, the NPxxY Y326 forms stable hydrogen bonds with Y219 (Supplementary Fig. 18), which points between Helix 3 and Helix 6. Agonist-bound receptor also samples close contact of these residues, but they are rarely in hydrogen-bonding range in active states. A partial connector active transition is observed in the panel 3 snapshot (Fig. 3) from the alternate inverse agonist pathway with the slowest connector flip (red path in Fig. 2c), and the agonist-bound snapshot shows a complete connector flip along with Helix 5 bulge formation. For both ligands, the methyl group of M215 and the F208 phenyl group make close contacts with F282 in active receptor states (Supplementary Fig. 17). In panel 4 (Fig. 3), active-state ligand interactions are shown, in which agonist hydrogen bonds with S207, and the inverse agonist hydrogen bonds instead to S203, located half a helix turn above S207. Both serines are key functional residues[21], but the S207 hydrogen bond could account, in part, for agonist stabilization of active conformations, seen in Fig. 1, that can be selected by G proteins. This effect is also seen in the binding-site connections to the connector and to intracellular regions that are unique to the mutual information network for agonist-bound simulations (Fig. 2d).

Small-molecule docking to β_2AR activation pathways enriches diverse GPCR ligand chemotypes. MSM states from high-flux activation pathways identified in this study were targeted with small-molecule docking of a database of β_2AR agonists, antagonists and decoys[22] with the program Surflex[23,24]. For both agonists and antagonists, docking to the MSM states along activation pathways gives a high area under the receiver operating

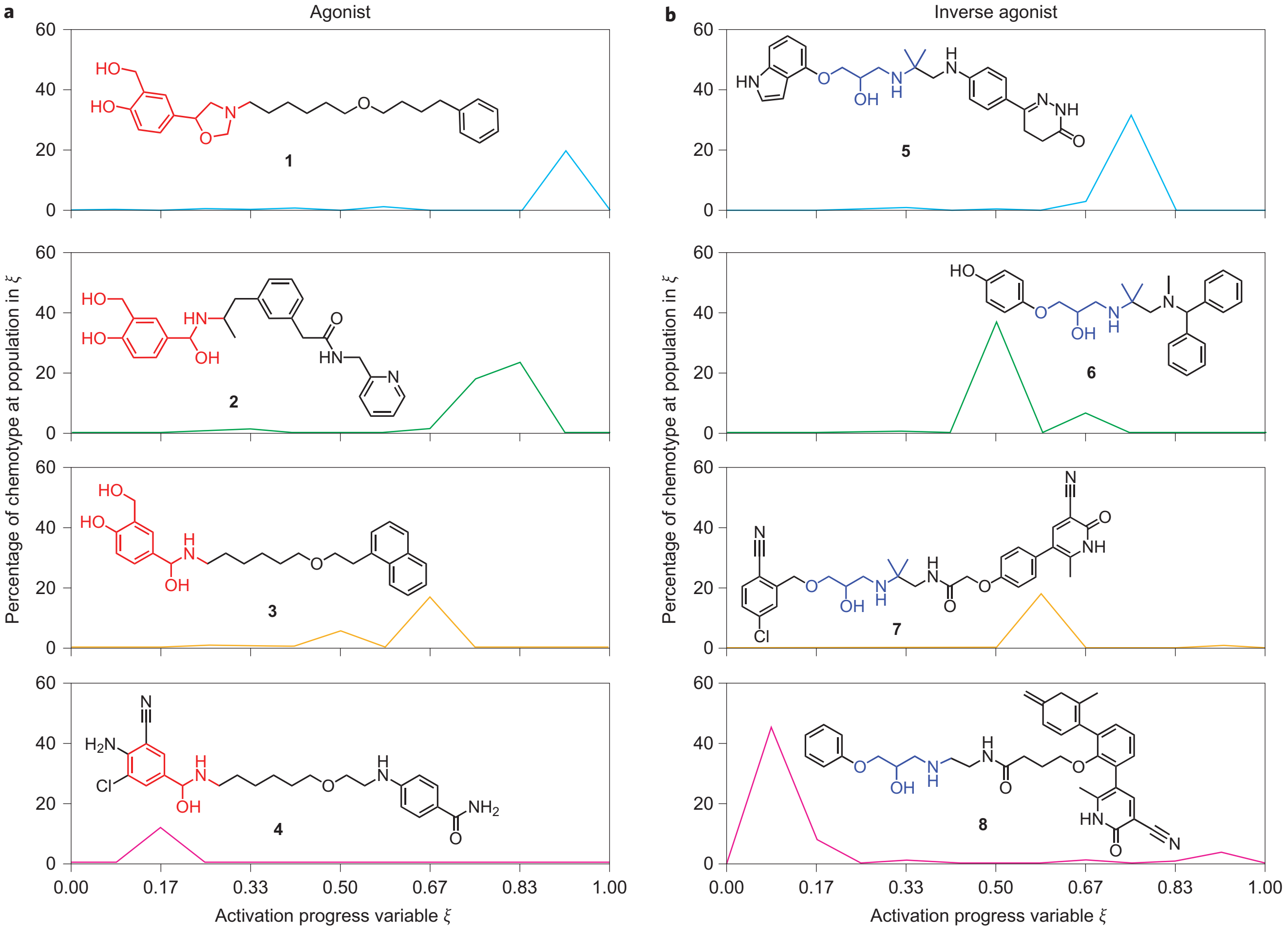

Figure 4 | Examples of GPCR ligand chemotypes enriched at MSM states along activation pathways. MSM states from high-flux activation pathways were assigned a progress score ξ based on structural metrics in Fig. 1 and range from the inactive crystal structure (2RH1) at $\xi = 0$ to the active crystal structure (3P0G) at $\xi = 1$. Top-ranking compounds from a retrospective virtual screen of known β_2AR ligands at each MSM state, and from both crystal structures, were clustered according to their 3D shape and chemistry overlap. Four examples of chemotype clusters enriched by select MSM states along the activation pathways are shown, with the percentage of a chemotype represented in the total ligands enriched at a given ξ. **a**, Example agonist chemotypes are catecholamine derivatives (1, 2 and 3) and ethanolamine (4) derivatives. **b**, Example antagonist chemotypes (5–8) share a 2-hydroxypropylamino core and include a carbostyril substituted pyridazinone (5), benzhydryl amine (6) and pyridone nitriles (7, 8). The complete distributions along ξ for all agonist and antagonist chemotype clusters are shown in Supplementary Figs 21 and 22. These results show that docking to intermediates identified by MSM TPT analysis enriches more-diverse chemotypes that could be missed by screens of only a few structures.

characteristic curve values (Supplementary Fig. 22), which evaluates selection of true ligands from decoys. These results are a statistically significant improvement over results from docking to the active (3P0G) and inactive (2RH1) crystal structures and to randomly selected snapshots from long-timescale, agonist-bound β_2AR deactivation simulations[5]. Next, we show that docking to MSM states expands the chemical space of our docking results, an essential advantage in docking approaches[25]. Top-scoring ligands for each MSM state were clustered according to their shape and chemistry, which revealed a diversity of chemotypes that is highly ranked, or enriched, differentially by MSM states along the activation pathways. Examples of chemotypes that would have been undiscovered by virtual screen docking to the crystal structures alone or without knowledge of the full activation pathway are shown in Fig. 4, with a full description in Supplementary Figs 23 and 24. These results underline the utility of MSMs for picking functional intermediate GPCR states that have different estimated affinities for known ligand chemotypes. Knowledge of this correspondence between ligand type and receptor conformation may be fruitful for future drug-design efforts and can give testable predictions for ligands that may isolate rare intermediate conformations of receptors.

Discussion

The existence of parallel paths, comprising many intermediates, has reshaped the paradigm of protein folding, and led to the formulation of MSMs as a new scheme to conceptualize such complex phenomena. Our results indicate that GPCR conformational dynamics is also well represented within the theoretical framework of MSMs, in which the receptor exists in multiple discrete conformational states with active and inactive states connected via multiple pathways. Furthermore, we show that ligands act by modulating the receptor dynamics to prefer different pathways. In contrast, methods that reduce simulation data to a single reaction coordinate by discarding dynamics along other independent coordinates may simplify the description of the conformational change, but could miss key features of the mechanism, including alternative pathways and kinetically trapped states[26].

This study emphasizes the need to go beyond simple few-state models to capture the hidden complexity of GPCR activation. In particular, drug-discovery efforts have focused on screening available GPCR crystal structures and failed to find leads with diverse efficacy profiles beyond antagonists and inverse agonists[27]. Using MSMs to incorporate the rich structural data reported here into this process creates an opportunity to develop drugs that interact more closely with diverse receptor states[28], for overall increased efficacy and specificity.

The unprecedented millisecond simulation timescales presented here for GPCR activation require computing architectures capable of such extensive sampling. Cloud computing provides a promising new avenue to tackle these types of questions more routinely, leaving the complexity of the underlying infrastructure hidden and enabling a larger proportion of time spent on science rather than administrative tasks. Our work on Google's Exacycle platform demonstrates that large-scale exploratory analysis in the cloud can deliver new insight into biological problems. Existing commercial cloud-computing initiatives enable researchers to migrate their own computational needs to the cloud in a scalable, elastic, efficient and easily accessible computational manner. The impact of this commodity-computing approach is magnified when matched with novel algorithms, such as MSMs, that can extract the most pertinent information for researchers.

Methods

Simulation systems. Membrane-aligned crystal structures for PDB IDs 2RH1 and 3P0G were extracted from the Orientations of Proteins in Membranes database[29]. Both the inactive (PDB 2RH1) and the active (PDB 3P0G) crystal structures were simulated as ligand-free apo structures, as well as the receptor bound to the partial inverse agonist carazolol and the full agonist BI-167107 (refs 1,2). Residues in transmembrane helices and helix centres are defined as follows (with residue numbers in brackets): Helix 1 (29–60), Helix 2 (67–96), Helix 3 (103–136), Helix 4 (147–171), Helix 5 (197–229), Helix 6 (267–298) and Helix 7 (305–328), according to Rosenbaum *et al.*[30]. The crystal structures lack residues at N and C termini that were not resolved during crystallography. For 2RH1, our structures entail residues 30–342, and for 3P0G residues 23–344. Both have gaps in the sequence, where the ICL3s between Helix 5 and Helix 6 are replaced in 2RH1 and 3P0G with T4-lysozyme and a nanobody, respectively. These residues are 231–262 for 2RH1, and 228–264 for 3P0G. As β_2AR remains functional even in the absence of ICL3, most of the simulations were done without this loop. One system was added in which apo 2RH1 was complemented by a model of ICL3 obtained from SuperLooper[31] for a total of seven simulated systems.

All amino acids were protonated according to their pK_a value at neutral pH. Titratable residues D79 (2.50) and D130 (3.50) have been proposed to change protonation state during activation, and acidic pH increases β_2AR basal activity by approximately twofold[6,32,33]. Recent long-timescale simulations varied the pH of these residues and found that, although the rate of deactivation increased by twofold for charged D130, the mechanism outlined for the receptor was largely unchanged[5]. This study also focused its analysis on the neutral pH charge of β_2AR. All MD simulations were performed with the Gromacs 4.5.3 MD package[34], compiled as 64-bit Linux binaries from unmodified source code. Each system was simulated in a POPC (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine) lipid bilayer and solvated in TIP3P (transferable intermolecular potential 3P) water molecules. Comprehensive details on system setup and simulation parameters are provided in the Supplementary Information.

MD simulations on Exacycle. Many large data centres are built with sufficient redundancies to handle peaks in user activity and cluster or fibre failures and still maintain low-latency responsiveness, to leave excess cycles during normal operation. Exacycle provides a platform for running massively parallel jobs on spare capacity of Google's production infrastructure. Initially, a limited number of grants were issued for select eScience projects[11]. Exacycle was created as an alternative approach to large-scale scientific computing, to replace complex parallelization of jobs across highly interconnected, specialized computing hardware by embarrassingly parallel jobs on highly heterogeneous commodity hardware. In principle, the Exacycle approach enables the use of an arbitrary number of distributed processors. For MD, we replaced individual long-timescale trajectories by stochastic sampling across a large number of trajectories. This allowed us to simulate β_2AR, with almost 60,000 atoms, at a rate exceeding 1 ms per month with a peak throughput of 80 μs a day. For comparison, the special-purpose Anton supercomputer achieves up to 16.4 μs of chemical time per day, interpolated from Fig. 5 in Shaw *et al.*[9].

All simulations were generated as sets of Exacycle work units. For each β_2AR system, 10,000–20,000 trajectories were simulated in three rounds with interround MSM-guided uniform sampling[12,35] to determine new starting structures. Average trajectory lengths per round correspond to 5 ns, 12.5 ns and 12.5 ns, respectively. Apo states in rounds 1 and 2 were run in a set of 20,000 trajectories, and a set of 10,000 trajectories was run for non-apo states. The first round was started from equilibrated crystal structures with a reference temperature of 300 K. Between rounds, each set of trajectories was sampled at 0.5 ns and clustered, according to the MSM criteria described below, into 1,000 microstates with MSMBuilder 2[36] from which new starting structures were chosen uniformly for the next round. All systems were heated to a reference temperature of 330 K for the third round to achieve improved sampling. The higher temperature was chosen to increase the likelihood of rare events occurring and decrease the impact of local minima. This is based on the assumption that, although we are likely to observe a shift in kinetics, the thermodynamics of the GPCRs during this last stage would be preserved because, according to our observations, changes in stabilizing enthalpic and entropic factors are dominated by subtle local conformational changes and remain small. Whereas significantly more entropically stabilized states might see a positive population shift in comparison with more enthalpically stabilized states as temperature increases, we expect this to be a negligible effect between the similarly stabilized active and inactive states. The final set of trajectories indicated convergence by comparing the conformational landscapes along the order parameters shown in Supplementary Figs 1–4. Thus, for analysis, only data from the final round were used, which equated to 125 μs aggregate MD data each, or 750 μs total, subsampled at a rate of 0.5 ns for a total of about 250,000 molecular structures per simulated system.

Small-molecule docking and chemotypes clustering. MSM states from transition pathway theory (TPT) pathways with flux >30% of the maximum flux were selected for a retrospective small-molecule docking screen. For the agonist-bound MSM, this corresponds to 20 MSM states, for the inverse agonist-bound MSM to 43 states and for the apo to 102 MSM states (to give a total of ~2.9 million docking calculations). We also dock to both the active (3P0G) and inactive (2RH1) crystal structures, and to 20 randomly selected snapshots from previously performed long-timescale agonist-bound GPCR deactivation simulations[5]. All snapshots were aligned to the same active crystal structure (3P0G) before docking. The Surflex[23,24] docking program was used to dock compounds from the GPCR Decoy and Ligand Database[22], which comprises ~200 known GPCR agonists and antagonists, and ~8,000 structurally dissimilar but property matched decoys drawn from the compound library ZINC[37]. Ligands were prepared with Surflex protonation tools, and OpenEye Omega[38] was used to enumerate stereoisomers for compounds with up to four chiral centres. Ligands were docked and scored with 20 snapshots from each MSM state, recording the highest docking score. To compute receiver operating characteristic (ROC) plots, the best-scoring stereoisomer was assigned to the ligand.

The top 10% scoring true ligands (decoys were excluded) were selected from each MSM state, which resulted in 3,300 enriched compounds for each agonist and antagonist docking set. These compounds were clustered by their chemotype with a k-centres algorithm, evaluated by Tanimoto values from unaligned (docked conformation preserved) shape and chemistry overlap computations with ROCS[39]. A cutoff for the clustering was selected as the value that separates the top 5% of all Tanimoto values computed for all compound pair overlaps, to give 935 clusters for antagonist chemotypes and 497 for agonist chemotypes (including different stereoisomers).

Analysis software. Before analysis, several million files that contained trajectory segments and related data were retrieved, merged, filtered and preprocessed on Google's infrastructure using a combination of FlumeJava[40], MapReduce[41] and Bigtable[42]. Coordinates for protein and ligand were stored in GROMACS.xtc format as well as converted into a binary format suitable for Dremel[43]. VMD[44], Gromacs[34] and Dremel were used to analyse geometric properties, such as interatomic distances and charge distributions. MSMs were generated with MSMBuilder 2[36], which was also used for TPT analysis[18,45]. Surflex 2.6[24] and OMEGA 2.4.6 were used for small-molecule docking and ROCS 3.1.2 was used to generate overlap scores for chemotype clustering.

Received 2 September 2013; accepted 11 November 2013; published online 15 December 2013

References

1. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
2. Rasmussen, S. G. F. *et al.* Structure of a nanobody-stabilized active state of the β2 adrenoceptor. *Nature* **469,** 175–180 (2011).
3. Rasmussen, S. G. F. *et al.* Crystal structure of the β2 adrenergic receptor–Gs protein complex. *Nature* **477,** 549–555 (2011).
4. Dror, R. O. *et al.* Identification of two distinct inactive conformations of the β2-adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl Acad. Sci. USA* **106,** 4689–4694 (2009).
5. Dror, R. O. *et al.* Activation mechanism of the β2-adrenergic receptor. *Proc. Natl Acad. Sci. USA* **108,** 18684–18689 (2011).
6. Vanni, S., Neri, M., Tavernelli, I. & Rothlisberger, U. Predicting novel binding modes of agonists to β adrenergic receptors using all-atom molecular dynamics simulations. *PLoS Comput. Biol.* **7,** e1001053 (2011).

7. Ivetac, A. & McCammon, J. A. Mapping the druggable allosteric space of G-protein coupled receptors: a fragment-based molecular dynamics approach. *Chem. Biol. Drug Des.* **76,** 201–217 (2010).
8. Nygaard, R. *et al.* The dynamic process of β2-adrenergic receptor activation. *Cell* **152,** 532–542 (2013).
9. Shaw, D. E. *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **51,** 91–97 (2008).
10. Shirts, M. & Pande, V. S. Screen savers of the world unite! *Science* **290,** 1903–1904 (2000).
11. Hellerstein, J. L., Kohlhoff, K. J. & Konerding, D. E. Science in the cloud: accelerating discovery in the 21st century. *IEEE Internet Comput.* **16,** 64–68 (2012).
12. Bowman, G. R., Huang, X. & Pande, V. S. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **49,** 197–201 (2009).
13. Senne, M., Trendelkamp-Schroer, B., Mey, A. S. J. S., Schütte, C. & Noé, F. EMMA: a software package for Markov model building and analysis. *J. Chem. Theory Comput.* **8,** 2223–2238 (2012).
14. Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin Struct. Biol.* **18,** 154–162 (2008).
15. Bowman, G. R. & Geissler, P. L. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl Acad. Sci. USA* **109,** 11681–11686 (2012).
16. Vanden-Eijnden, W. E, E. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.* **61,** 391–420 (2010).
17. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Mult. Mod. Sim.* **7,** 1192–1219 (2009).
18. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl Acad. Sci. USA* **106,** 19011–19016 (2009).
19. Kofuku, Y. *et al.* Efficacy of the β2-adrenergic receptor is determined by conformational equilibrium in the transmembrane region. *Nature Commun.* **3,** 1045 (2012).
20. Strader, C. D. *et al.* Conserved aspartic acid residues 79 and 113 of the beta-adrenergic receptor have different roles in receptor function. *J. Biol. Chem.* **263,** 10267–10271 (1988).
21. Liapakis, G. *et al.* The forgotten serine: a critical role for Ser-2035.42 in ligand binding to and activation of the β2 adrenergic receptor. *J. Biol. Chem.* **275,** 37779–37788 (2000).
22. Gatica, E. A. & Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model* **52,** 1–6 (2012).
23. Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* **21,** 281–306 (2007).
24. Spitzer, R. & Jain, A. N. Surflex-Dock: docking benchmarks and real-world application. *J Comput. Aided Mol. Des.* **26,** 687–699 (2012).
25. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432,** 862–865 (2004).
26. Lane, T. J., Shukla, D., Beauchamp, K. A. & Pande, V. S. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **23,** 58–65 (2013).
27. Shoichet, B. K. & Kobilka, B. K. Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **33,** 268–272 (2012).
28. Schames, J. R. *et al.* Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **47,** 1879–1881 (2004).
29. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **22,** 623–625 (2006).
30. Rosenbaum, D. M. *et al.* Structure and function of an irreversible agonist–β2 adrenoceptor complex. *Nature* **469,** 236–240 (2011).
31. Hildebrand, P. W. *et al.* SuperLooper – a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Res.* **37,** W571–W574 (2009).
32. Ballesteros, J. A. *et al.* Activation of the β2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276,** 29171–29177 (2001).
33. Ghanouni, P. *et al.* The effect of pH on β2 adrenoceptor function: evidence for protonation-dependent activation. *J. Biol. Chem.* **275,** 3121–3127 (2000).
34. 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.* **4,** 435–447 (2008).
35. Voelz, V. A., Bowman, G. R., Beauchamp, K. A. & Pande, V. S. Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1−39). *J. Am. Chem. Soc.* **132,** 1526–1528 (2010).
36. Beauchamp, K. A. *et al.* MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.* **7,** 3412–3419 (2011).
37. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **52,** 1757–1768 (2012).
38. Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A. & Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **50,** 572–584 (2010).
39. ROCS, version 3.1.2 (OpenEye Scientific Software, Santa Fe, New Mexico, 2011).
40. Chambers, C. *et al.* in *Proceedings of the 2010 ACM SIGPLAN Conference on Programming Language Design and Implementation* 363–375 (ACM, 2010).
41. Dean, J. & Ghemawat, S. MapReduce: simplified data processing on large clusters. *Commun. ACM* **51,** 107–113 (2008).
42. Chang, F. *et al.* Bigtable: a distributed storage system for structured data. *ACM Trans. Comput. Syst.* **26,** 4:1–4:26 (2008).
43. Melnik, S. *et al.* Dremel: interactive analysis of web-scale datasets. *Proc. VLDB Endow.* **3,** 330–339 (2010).
44. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
45. Berezhkovskii, A., Hummer, G. & Szabo, A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *J. Chem. Phys.* **130,** 205102 (2009).

Acknowledgements

We are grateful to M. Stumpe for his contributions to setting up the initial molecular systems and P. Kasson and J. Hellerstein for helpful advice and support. This work was funded in part by a 450M CPU core-hour donation by Google Inc. through the Exacycle eScience program, the Simbios NIH National Center on Biocomputing through the NIH Roadmap for Medical Research Grant U54 GM07297 and a Stanford School of Medicine Dean's Fellowship (K.J.K.). We also thank the users of the Folding@home distributed-computing project for donating compute time for some preliminary simulations that ensured a stable production run. Additional computations for docking and chemotype clustering were performed on the Blue Waters supercomputer at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign.

Author contributions

K.J.K., D.S. and M.L. contributed equally to this work. V.S.P., R.B.A, D.E.K. and D.B. conceived, and V.S.P. and R.B.A. supervised the project. K.J.K. and D.E.K. developed the platform for running MD simulations with Gromacs on Google Exacycle. K.J.K. set up the simulation systems. G.R.B helped with initial analysis. K.J.K. performed simulations on Google Exacycle and processed data on Google's production infrastructure. K.J.K. and G.R.B. performed preliminary simulations on Folding@home. D.S. and M.L. performed additional simulations. D.S., M.L and K.J.K. analysed the data and built MSMs. M.L. performed small-molecule docking calculations. D.S., M.L. and K.J.K. co-wrote the manuscript with inputs from G.R.B, R.B.A and V.S.P. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary information and chemical compound information are available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.J.K., R.B.A. and V.S.P.

Competing financial interests

The authors declare no competing financial interests.