

# Chapter 1

## Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding

*This chapter is adapted from the following publication:*

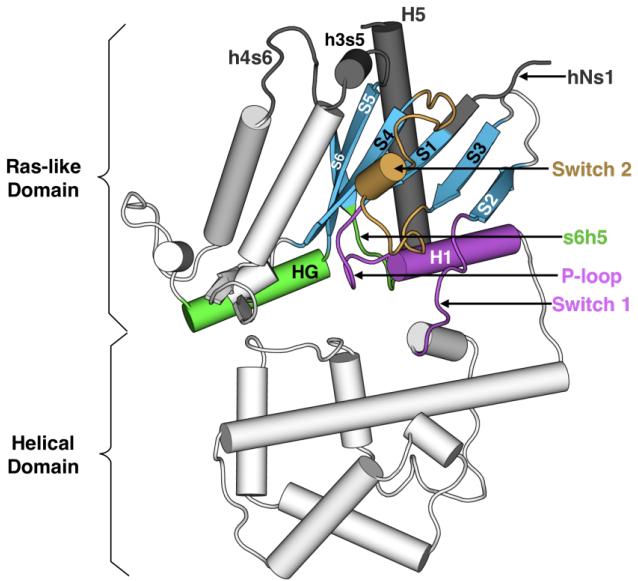
*Sun, X.\* and Singh, S.\*<sup>†</sup>, Blumer, K.J., and Bowman, G.R., Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. eLife, 7, October 2018, <https://doi.org/10.7554/eLife.38465.001> [?]*  
<sup>\*Authors contributed equally to this work</sup>

### 1.1 Abstract

Activation of heterotrimeric G proteins is a key step in many signaling cascades. However, a complete mechanism for this process, which requires allosteric communication between binding sites that are  $\sim 30 \text{ \AA}$  apart, remains elusive. We construct an atomically detailed model of G protein activation by combining three powerful computational methods: metadynamics, Markov state models (MSMs), and CARDs analysis of correlated motions. We uncover a mechanism that is consistent with a wide variety of structural and biochemical data. Surprisingly, the rate-limiting step for GDP release correlates with tilting rather than translation of the GPCR-binding helix 5.  $\beta$ -Strands 1-3 and helix 1 emerge as hubs in the allosteric network that links conformational changes in the GPCR-binding site to disordering of the distal nucleotide-binding site and consequent GDP release. Our approach and insights provide foundations for understanding disease-implicated G protein mutants, illuminating slow events in allosteric networks, and examining unbinding processes with slow off-rates.

### 1.2 Introduction

Heterotrimeric G proteins are molecular switches that play pivotal roles in signaling processes from vision to olfaction and neurotransmission [?, ?, ?]. By default, a G protein adopts an inactive state in which guanosine diphosphate (GDP) binds between the Ras-like and helical domains of the  $\alpha$ -subunit ( $G\alpha$ , Fig. 1.1). A dimer consisting of the  $\beta$ - and  $\gamma$ -subunits ( $G\beta\gamma$ ) also binds  $G\alpha$ . G protein-coupled receptors (GPCRs) trigger G protein activation by binding  $G\alpha$  and stimulating GDP release, followed by GTP binding to  $G\alpha$  and dissociation of  $G\alpha$  from  $G\beta\gamma$ .  $G\alpha$  and  $\beta\gamma$  then interact with effectors that trigger downstream cellular responses.  $G\alpha$  returns to the inactive state by hydrolyzing GTP to GDP and rebinding  $G\beta\gamma$ . Given the central role  $G\alpha$  plays, a common  $G\alpha$  numbering scheme (CGN) has been established to facilitate discussion of the activation mechanisms of different  $G\alpha$  homologs [?]. For example, the notation  $\text{Lys52}^{G.H1.1}$  indicates that Lys52 is the first residue in helix 1 (H1) of the Ras-like domain (also called the GTPase domain, or G). S6 refers to  $\beta$ -strand 6 and s6h5 refers to the loop between S6 and H5.



**Figure 1.1:** Structure of G $\alpha$ q with key secondary structure elements labeled according to the Common G $\alpha$  Numbering (CGN) system. The coloring scheme highlights the GPCR binding interface (gray), GDP phosphate-binding regions (pink), GDP nucleotide-binding regions (green),  $\beta$ -sheets (blue), and switch 2 (orange).

Strikingly, the GPCR- and nucleotide-binding sites of G $\alpha$  are  $\sim$ 30 Å apart (Fig. 1.1), but the allosteric mechanism coupling these sites to evoke GDP release remains incompletely understood [?]. Biochemical and structural studies have elucidated some key steps, but the entire process has yet to be described in atomic detail. Early studies of G $\alpha$  subunits revealed structures of the GDP- and GTP-bound states, as well as the transition state for GTP hydrolysis [?, ?, ?]. The high similarity of these structures and the binding of GDP or GTP deep in the protein's core suggests that activation occurs by adoption of other conformational states that allow GDP release [?]. One intermediate in G protein activation was suggested by the first crystal structure of a GPCR-bound G protein in which the Ras-like and helical domains of G $\alpha$  are hinged apart and GDP has dissociated [?]. Structural analysis has led to the proposal of a universal mechanism for G protein activation [?]. In this model, GPCR binding induces translation of H5 away from H1, which increases disorder in H1 and the P-loop (or Walker A motif [?]) to facilitate GDP release. However, there is evidence that additional intermediates may be involved in G $\alpha$  activation [?, ?, ?, ?], and the functional importance of this conformational ensemble has been previously suggested [?]. Furthermore, mutagenesis and nuclear magnetic resonance (NMR) studies have suggested important roles for other structural elements [?, ?, ?].

Molecular dynamics simulations promise to capture the entire mechanism of G protein activation and synthesize the wealth of experimental data on this process. Methodological advances now enable simulations to capture millisecond timescale processes for proteins with less than 100 residues [?]. For example, it is now possible to capture the binding or release of small molecules [?, ?, ?, ?, ?] and peptides [?, ?] from small proteins. Impressive simulations on the ANTON supercomputer have revealed critical conformational dynamics of G proteins in their inactive and active states, elucidating the role of domain opening in GDP unbinding [?, ?]. However, even this specialized hardware could not capture the entire process of G protein activation and GDP release due to the size of the G $\alpha$  subunit ( $>300$  residues) and the slow kinetics of GDP dissociation ( $\sim 10^{-3}$  min $^{-1}$ ) [?, ?, ?].

Here, we introduce an approach to capture rare or long-timescale events, such as GDP release, and reveal the mechanism of G $\alpha$  activation. As a test of this methodology, we apply it to G $\alpha$ q, which has one of the slowest GDP release rates [?] and is frequently mutated in uveal melanoma [?, ?]. To highlight aspects of the activation mechanism that we propose are general to all G proteins, we focus our analysis on the behavior of secondary structure elements and amino acids that are conserved across G $\alpha$  domains. Our approach first combines two powerful sampling methods, metadynamics [?] and Markov state models (MSMs) [?], to observe GDP release and identify the rate-limiting step

for this slow process. Then we use our recently developed CARDS method [?], which quantifies correlations between both the structure and disorder of different regions of a protein, to identify the allosteric network connecting the GPCR- and nucleotide-binding sites. The resulting model is consistent with a wealth of experimental data and leads to a number of predictions, described below. Taken together, our results provide the most comprehensive model of G protein activation to date. Based on this success, we expect our approach to be valuable for studying other rare events, including conformational changes and unbinding processes.

## 1.3 Results and Discussion

### 1.3.1 Capturing G-protein activation and GDP release in atomic detail

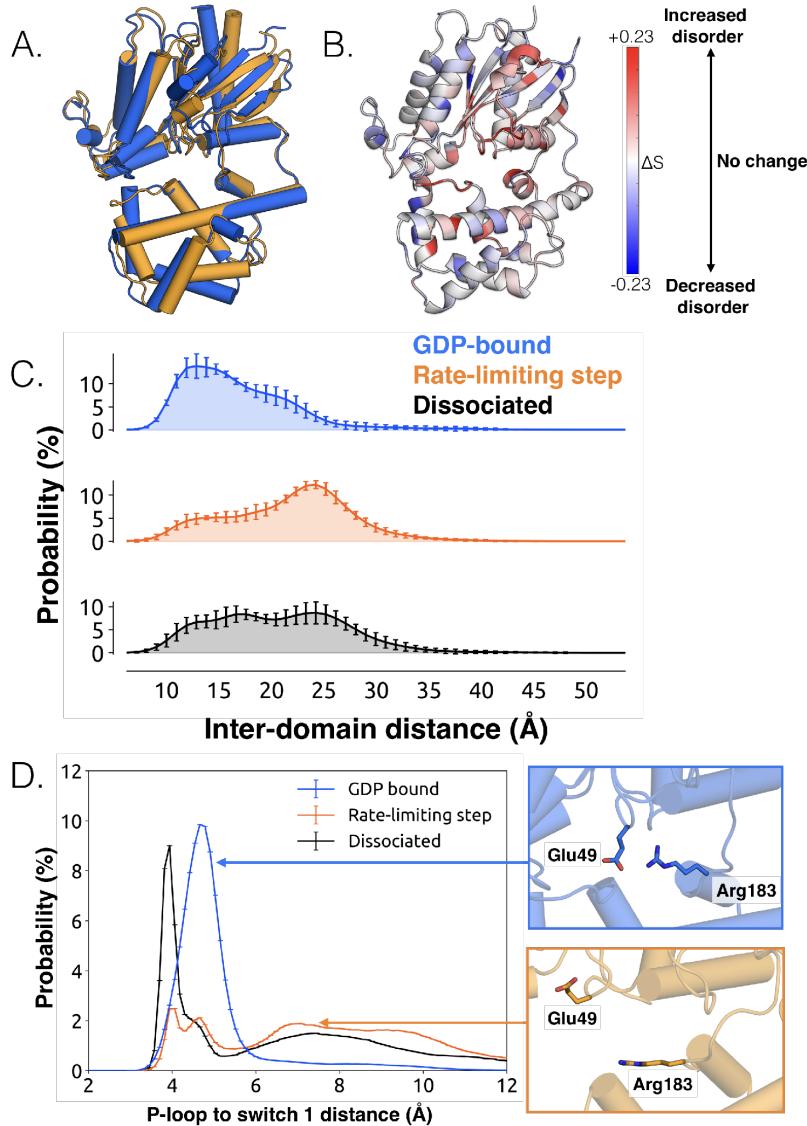
We reasoned that studying the mechanism of spontaneous GDP release from a truncated form of G $\alpha$ q would be representative of the mechanism of GPCR-mediated activation of the heterotrimeric G protein while minimizing the computational cost of our simulations. This hypothesis was inspired by previous work demonstrating that a protein's spontaneous fluctuations are representative of the conformational changes required for the protein to perform its function [?, ?, ?]. Therefore, we hypothesized that GPCRs stabilize conformational states that G proteins naturally, albeit infrequently, adopt in the absence of a receptor. We chose to focus on G $\alpha$  since it forms substantial interactions with GPCRs, compared to the relatively negligible interactions between GPCRs and G protein  $\beta$  and  $\gamma$  subunits. This view is supported by the fact that GPCRs and ‘mini’ G proteins, essentially composed of just the Ras-like domain of G $\alpha$ , recapitulate many features of GPCR-G protein interactions [?]. We also reasoned that truncating the last five residues of G $\alpha$ q would facilitate the activation process. These residues contact G $\alpha$  in some GDP-bound structures but not in GPCR-bound structures [?, ?], and removing these residues promotes GDP release due to a reduced GDP-binding affinity [?, ?]. Taken together, such evidence suggests that the last five residues of G $\alpha$ q help stabilize the inactive state and that removing them would accelerate activation. In support of this hypothesis, we find that the energetic barrier to GDP release is lower in metadynamics simulations of the truncated variant than for full-length G $\alpha$ q (Figure ??). These simulations, and those described hereafter, were initiated from an X-ray structure of the G $\alpha$ q heterotrimer bound to GDP and an inhibitor of nucleotide exchange [?]; G $\beta$  $\gamma$  and the inhibitor were excluded from all simulations.

To observe G-protein activation, we developed a variant of adaptive seeding [?] capable of capturing slow processes like ligand unbinding that are beyond reach of conventional simulation methods. First, we use metadynamics [?, ?, ?] to facilitate broad sampling of conformational space by biasing simulations to sample conformations with different distances between GDP and G $\alpha$ q. Doing so forces GDP release to occur but provides limited mechanistic information because adding a biasing force can distort the system's kinetics or cause the simulations to sample high-energy conformations that are not representative of behavior at thermal equilibrium. To overcome these limitations, we chose starting conformations along release pathways observed by metadynamics as starting points for standard molecular dynamics simulations, together yielding an aggregate simulation time of 122.6  $\mu$ s. These simulations should quickly relax away from high-energy conformations and provide more accurate kinetics. Then we use these simulations to build an MSM (Source data 1). MSMs are adept at extracting both thermodynamic and kinetic information from many standard simulations that, together, cover larger regions of conformational space than any individual simulation [?]. Related approaches have successfully captured the dynamics of small model systems [?] and RNA polymerase [?].

This protocol enabled us to capture the entire mechanism of G-protein activation, including GDP release and the rate-limiting step for this process. Identifying the rate-limiting step for this process is of great value because GDP release is the rate-limiting step for G-protein activation and downstream signaling. Therefore, the key structural and dynamical changes responsible for activation should be apparent from the rate-limiting conformational transition for this dissociation process.

To identify the rate-limiting step, we applied transition path theory [?, ?] to find the highest flux paths from bound structures resembling the GDP-bound crystal structure to fully dissociated conformations. Then, we identified the least probable steps along the 10 highest flux release pathways (Figure 1.2A and Figure ??), which represent the rate-limiting step of release. By comparing the structures before and after the rate-limiting step, we define the bound state as all conformations where the distance from the center of mass of GDP's phosphates to the center of mass of three residues on H1 that contact the GDP  $\beta$ -phosphate (Lys52<sup>G.H1.1</sup>, Ser53<sup>G.H1.2</sup>, and Thr54<sup>G.H1.3</sup>) is less than 8  $\text{\AA}$ . Consistent with this definition, this distance remains less than 8  $\text{\AA}$  throughout the entirety of 35.3  $\mu$ s of GDP-bound

simulations.



**Figure 1.2:** Structural and dynamical changes during the rate limiting step for GDP release. **A.** Overlay of representative structures before (blue) and after (orange) the rate limiting step. **B.** Change in conformational disorder (Shannon entropy) across the rate-limiting step, according to the color scale on the right. **C.** Histograms of inter-domain distances before (top, blue) and after (middle, orange) the rate limiting step, along with the inter-domain distance distribution after GDP is released (bottom, black). Inter-domain distance is measured using residues analogous to those used in DEER experiments [?], Leu97<sup>H.H.A.29</sup> in the helical domain and Glu250<sup>G.H3.4</sup> on H3. **D.** Distribution of distances between Glu49<sup>G.s1h1.4</sup> and Arg183<sup>G.hfs2.2</sup> (left) before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of the interacting residues are also shown (right).

The conformational changes we observe during the rate-limiting step are consistent with previous structural and biochemical work. For example, we observe that the Ras-like and helical domains separate (Figure 1.2C), as observed in DEER experiments [?] and previous simulations [?]. This finding is consistent with the intuition that these domains must separate to sterically permit GDP release, and that this separation is driven by the disruption of multiple inter-

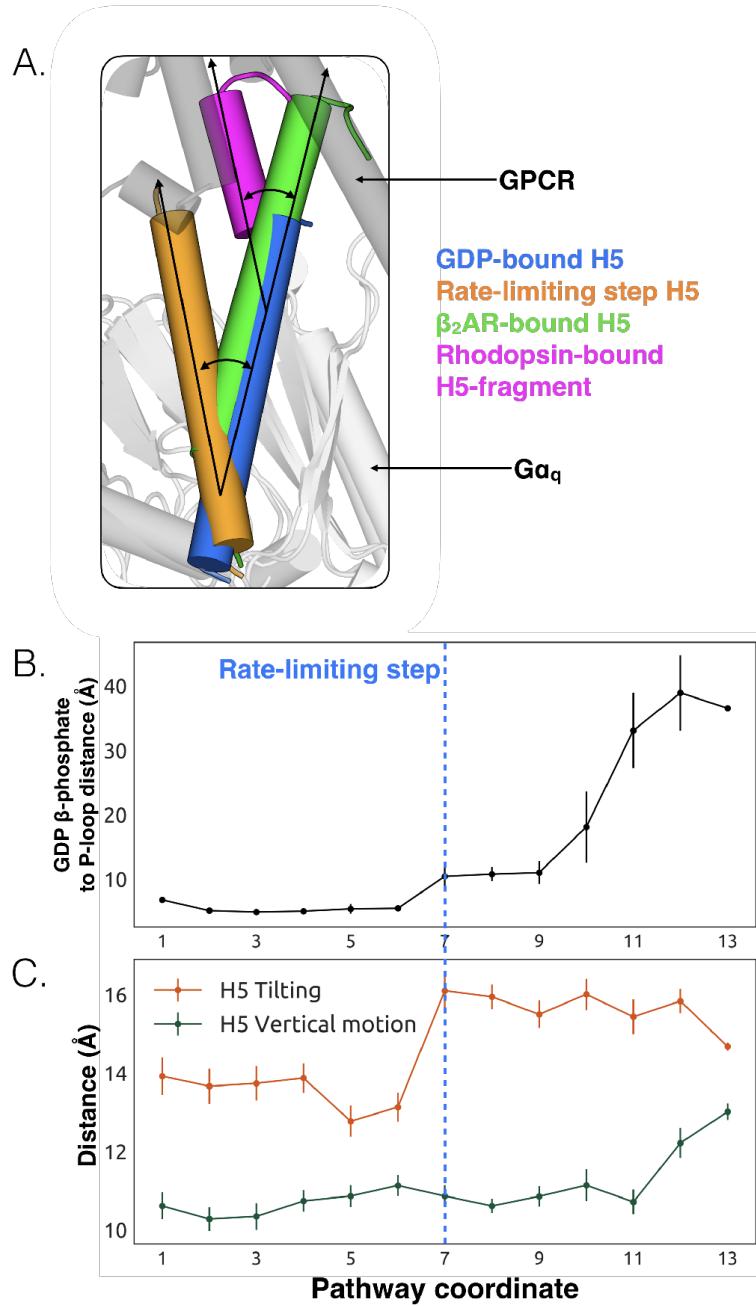
domain interactions. For example, we note a disrupted salt bridge between K275<sup>G.s5hg.1</sup> and D155<sup>H.hdhe.5</sup> (Figure ??), previously identified in structural studies [?]. Domain opening is accompanied by disruption of a key salt bridge between Glu49<sup>G.s1h1.4</sup> of the P-loop and Arg183<sup>G.hfs2.2</sup> of switch 1 (Figure 1.2D), as well as an increase in the disorder of many of the surrounding residues (Figure 1.2B and Figure ??A), consistent with the proposal that this salt bridge stabilizes the closed, GDP-bound state [?].

While domain opening is necessary for GDP release, previous simulations suggest it is insufficient for unbinding [?]. Indeed, we also see that this opening is necessary but not sufficient for GDP unbinding, as the Ras-like and helical domains often separate prior to release (Figure 1.2C). Notably, the Ras-like and helical domains only separate by  $\sim 10$  Å during the rate-limiting step. In contrast, these domains separate by 56 Å in the first structure of a GPCR-G-protein complex. This result suggests that GDP release may have occurred long before a G protein adopts the sort of widely opened conformations observed in crystallographic structures [?].

### 1.3.2 Tilting of H5 helps induce GDP release

We also observe less expected conformational changes associated with GDP release. The most striking is tilting of H5 by about 26° (Figure 1.3A, and Figure ??). We find that H5 tilting correlates strongly with the distance between GDP and G $\alpha$ q along the highest flux dissociation pathway (Figure 1.3B and ??). In particular, substantial tilting of H5 is coincident with the rate-limiting step for GDP release. This tilting contrasts with X-ray structures and the universal mechanism, in which H5 is proposed to translate along its helical axis towards the GPCR, initiating the process of GDP release (Figure 1.3A). During our simulations we also observe translation of H5, but it is not correlated with the rate-limiting step of GDP release (Figure 1.3C, Figure ??). Therefore, we are not merely missing an important role for translation due to insufficient sampling.

The potential importance of H5 tilting is supported by other structural data. For example, a crystal structure of rhodopsin [?] with a C-terminal fragment from H5 of G $\alpha$ t shows a similar degree of tilting (Figure 1.3A). Also, the tilt of H5 varies in crystal structures of the  $\beta$ 2AR-Gs complex [?], two different GLP-1 receptor-Gs complexes [?, ?], and an A2AR-mini-Gs complex [?]. The potential relevance of tilting has also been acknowledged by a number of recent works [?, ?, ?] including four recently published structures of receptor-G-protein complexes across which H5 also shows a broad range of tilting and translational motion [?, ?, ?, ?]. Interestingly, the tilting and translation we observe falls within the observed range of tilting and translational motions that H5 undergoes in available GPCR-G protein complex structures [?, ?, ?, ?, ?], providing support that conformational selection plays an important role (??). Finally, H5 is translated toward the GPCR in the A2AR-mini-Gs structure but GDP remains bound [?]. The authors of that study originally suggested that one of the mutations in mini-Gs decouples H5 translation from GDP release. However, given that we see GDP release without H5 translation in our simulations, it is also possible that amino acid substitutions required to create mini-Gs instead mitigate H5 tilting. Both of these models are consistent with the fact that some of the mutations in mini-Gs stabilize the GDP-bound state [?].



**Figure 1.3:** Tilting of H5 is correlated with GDP release but translation of H5 is not. **A.** Displacements of H5 relative to the GDP-bound crystal structure (blue). The three other orientations of H5 come from the rate-limiting step in our model (orange), the co-crystal structure of  $G\alpha_s$  and  $\beta_2$ AR (green, PDB ID 3SN6 [?]), and the co-crystal structure of a C-terminal fragment of  $G\alpha_t$  and rhodopsin (magenta, PDB ID 3PQR [?]). The black arrows highlight the change in orientation of the long axis of each helix. A representative GPCR (gray) and  $G\alpha$  (white) structure are shown for reference. **B.** GDP release distance across the highest flux pathway, defined as the distance from the GDP  $\beta$ -phosphate to the center of mass between residues Lys $52^{G.H1.1}$ , Ser $53^{G.H1.2}$ , and Thr $54^{G.H1.3}$  on H1. The state marking the rate-limiting step is highlighted by the blue dashed line. **C.** H5 motion across the highest flux pathway. The distances measured here representing H5 motion are taken from the same states as in B. H5 tilting (orange) is measured by the distance between Leu $349^{G.H5.16}$  and Tyr $325^{G.S6.2}$ . Likewise, H5 vertical motion (green) is measured by the distance between Thr $334^{G.H5.1}$  and Phe $341^{G.H5.8}$ . The rate-limiting step is marked with the blue dashed line, extended down from B.

We propose that tilting of H5 is an early step in the GDP release process, which is followed by upward translation of this helix to form a GPCR-G protein complex primed to bind GTP. This hypothesis stems from our observation that tilting of H5 is coincident with the rate-limiting step for GDP release, while translation of H5 only becomes stable after GDP dissociates (Figure 1.3C). This model is consistent with previous suggestions that G-protein activation occurs through a series of sequential interactions with a GPCR [?, ?]. Another possibility is that any displacement of H5, whether tilting or translation, may be sufficient to trigger GDP release.

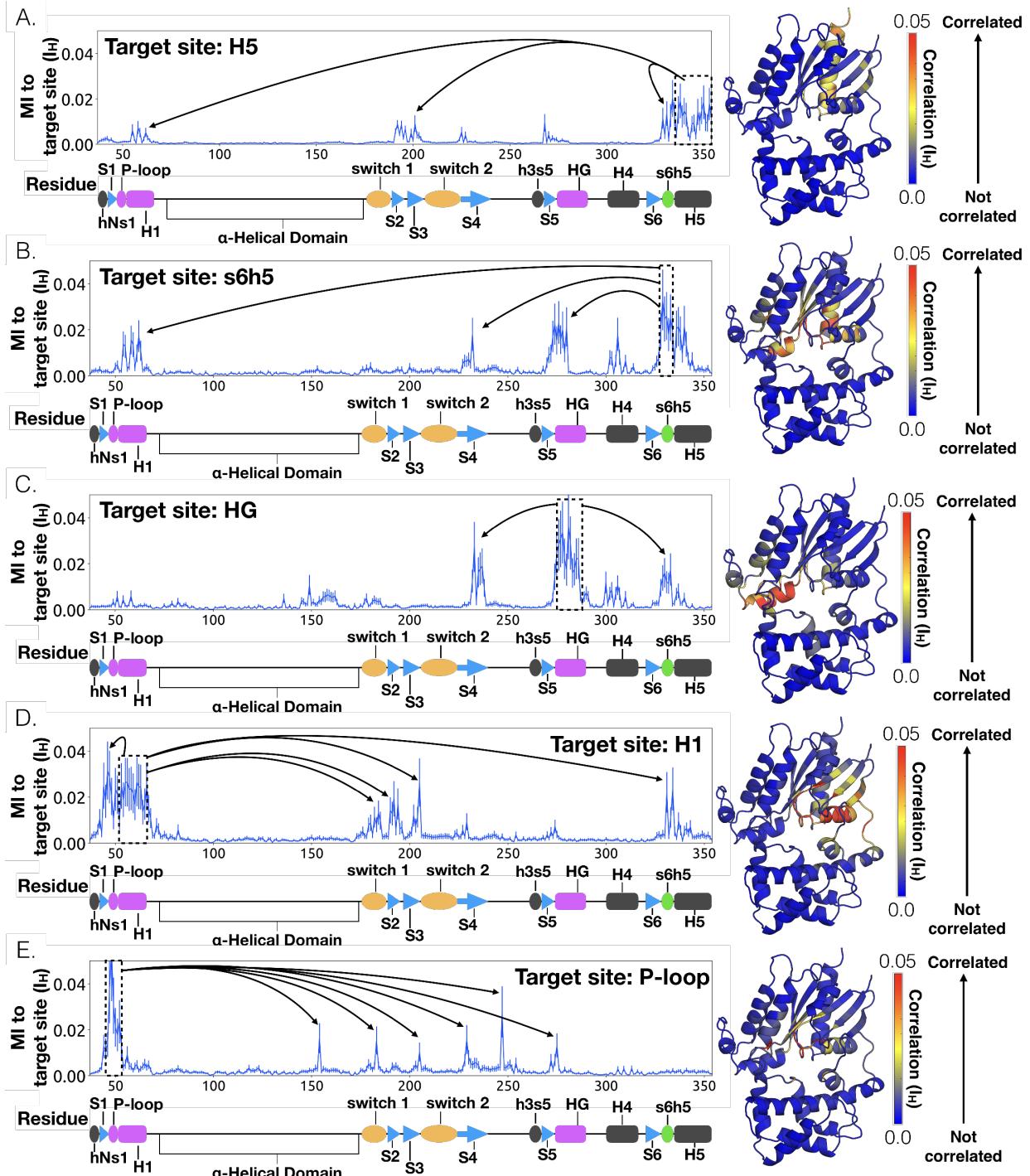
### 1.3.3 Identification of the allosteric network that triggers GDP release

While conformational changes of H5 are important for G $\alpha$  activation, other regions could also play a role [?, ?]. However, it is not straightforward to determine what other structural elements contribute to activation or their importance relative to H5. Our hypothesis that spontaneous motions of a protein encode functionally relevant conformational changes suggests that the coupling between the GPCR- and nucleotide-binding sites of G $\alpha$  should be present in simulations of the inactive protein; This provides a means to identify key elements of this allosteric network. To test this hypothesis, we ran 35.3  $\mu$ s of simulation of GDP-G $\alpha$ q. Then we applied the CARDS method [?] to detect correlations between both the structure and dynamical states of every pair of dihedral angles. Structural states are determined by assigning dihedral angles to the three dominant rotameric states for side-chains (gauche+, gauche-, and trans) and the two dominant rotameric states for backbone dihedrals (cis and trans). Dynamical states are determined by whether a dihedral angle remains in a single rotameric state (ordered) or rapidly transitions between multiple rotameric states (disordered). These pairwise correlations serve as a basis for quantifying the correlation of every residue to a target site, such as the GPCR-binding site. Combining these correlations with structural and dynamical changes in our model of GDP release provides a basis for inferring how perturbations to the GPCR-binding site are transmitted to the nucleotide-binding site. We focus our analysis on the most direct routes for communication between the GPCR- and nucleotide-binding sites by following correlated motions that emanate from structural elements that directly contact GPCRs until they reach the GDP-binding site. There are correlations between many other elements of G $\alpha$ q, including parts of the helical domain, that branch off of this allosteric network. Such correlations may be important for aspects of G $\alpha$  function besides activation, but are beyond the scope of the present study, which focuses on how GPCR-binding impacts nucleotide release.

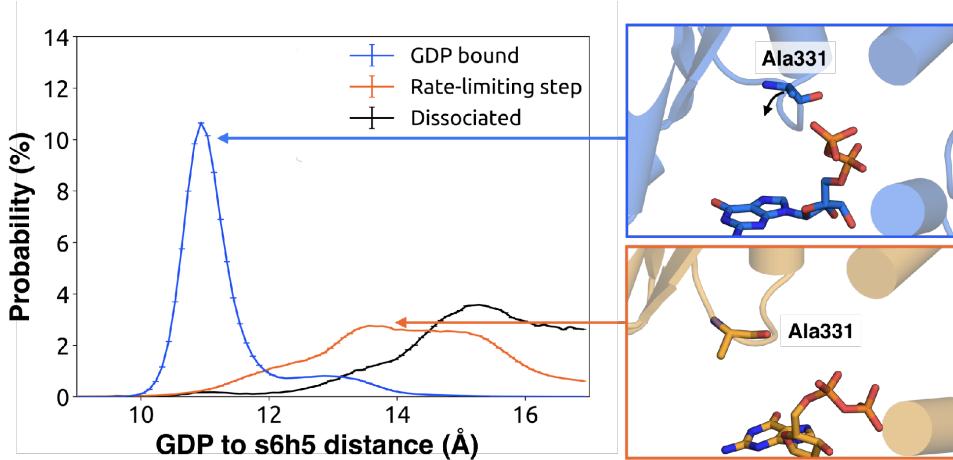
To understand how H5 tilting triggers GDP release, we first identified structural elements with strong coupling to H5 and then worked our way outward in repeated iterations until we reached the nucleotide-binding site (Figure 1.4). This analysis reveals that tilting of H5 directly communicates with and impacts the conformational preferences of the s6h5 loop, which contacts the nucleobase of GDP (Figure 1.4 and Figure 1.5). During the rate-limiting step, these contacts are broken and there is increased disorder in the s6h5 loop, particularly Ala331 of the TCAT motif (Figure 1.5 and Figure ??B). The importance of the TCAT motif in our model is consistent with its conservation and the fact that mutating it accelerates GDP release [?, ?, ?]. Based on our model, we propose these mutations accelerate release by weakening shape complementarity with GDP.

We also observe an important role for communication from H5 to H1, consistent with the universal mechanism. In particular, H1 is strongly coupled with the s6h5 loop (Figure 1.4B), which is sensitive to displacement of H5. In the rate-limiting step, s6h5 moves away from H1, contributing to an increase in disorder of H1 and the P-loop (Figure ??A and Figure ??B). Increased disorder in a set of residues that directly contact the GDP phosphates (Glu49<sup>G.s1h1.4</sup>, Ser50<sup>G.s1h1.5</sup>, Gly51<sup>G.s1h1.6</sup>, Lys52<sup>G.H1.1</sup>, and Ser53<sup>G.H1.2</sup>) likely contributes to a reduced affinity for GDP (Figure ??A). Increased disorder in these residues also contributes to disruption of the salt bridge between Glu49<sup>G.s1h1.4</sup> of the Ras-like domain and Arg183<sup>G.hfs2.2</sup> of the helical domain, facilitating inter-domain separation.

We further note that the s6h5 loop impacts the nucleotide-binding site through allosteric coupling with the HG helix, which also contacts GDP via Lys275<sup>G.s5hg.1</sup> and Asp277<sup>G.HG.2</sup> (Figures 1.4E and 1.6). The disorder of both of these residues increases during the rate-limiting step, consistent with observations of increased mobility in HG upon receptor-mediated activation (Oldham and Hamm, 2008). There are also correlations between the P-loop and Lys275<sup>G.s5hg.1</sup> on Helix G (Figure 1.4E), which result from the disruption of a key salt bridge between Lys275<sup>G.s5hg.1</sup> and Glu49<sup>G.s1h1.4</sup> on the P-loop during the rate-limiting step (Figure 1.6). Lys275<sup>G.s5hg.1</sup> is conserved across all G $\alpha$  families, suggesting it plays an important role in the stability or function of the protein. However, attempts to experimentally examine the role of this residue by mutating Lys275<sup>G.s5hg.1</sup> have resulted in aggregation [?]. Our simulations suggest Lys275<sup>G.s5hg.1</sup> plays an important role in stabilizing the GDP-bound state and that breaking

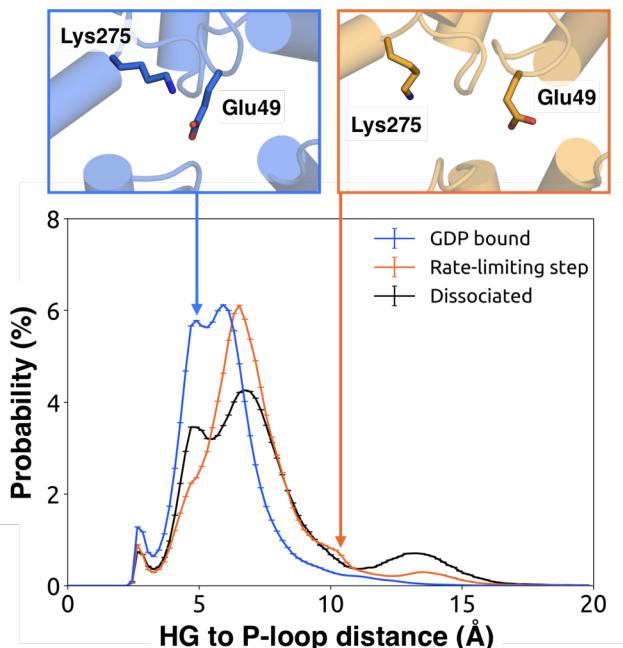


**Figure 1.4:** Allosteric network connecting H5 motion to the nucleotide binding-site via s6h5. CARDs data showing communication per residue to a target site (dashed box) is plotted (left) and mapped onto the structure of Gαq (right) for (A) H5, (B) s6h5, (C) HG, (D) H1, and (E) the P-loop. Arrows indicate regions of importance with significant communication.



**Figure 1.5:** Change in the s6h5 loop conformation across the rate-limiting step. Distribution of distances (left) from GDP's  $\beta$ -phosphate to Ala331 $G.s6h5.3$  on the s6h5 loop before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of the s6h5 loop (right) are shown for before (top right, blue) and after (bottom right, orange) the rate limiting step.

the salt bridge with Glu49 $G.s1h1.4$  facilitates GDP release. This finding demonstrates the utility of our atomistic simulations, as we can examine the role of residues that are difficult to probe experimentally.



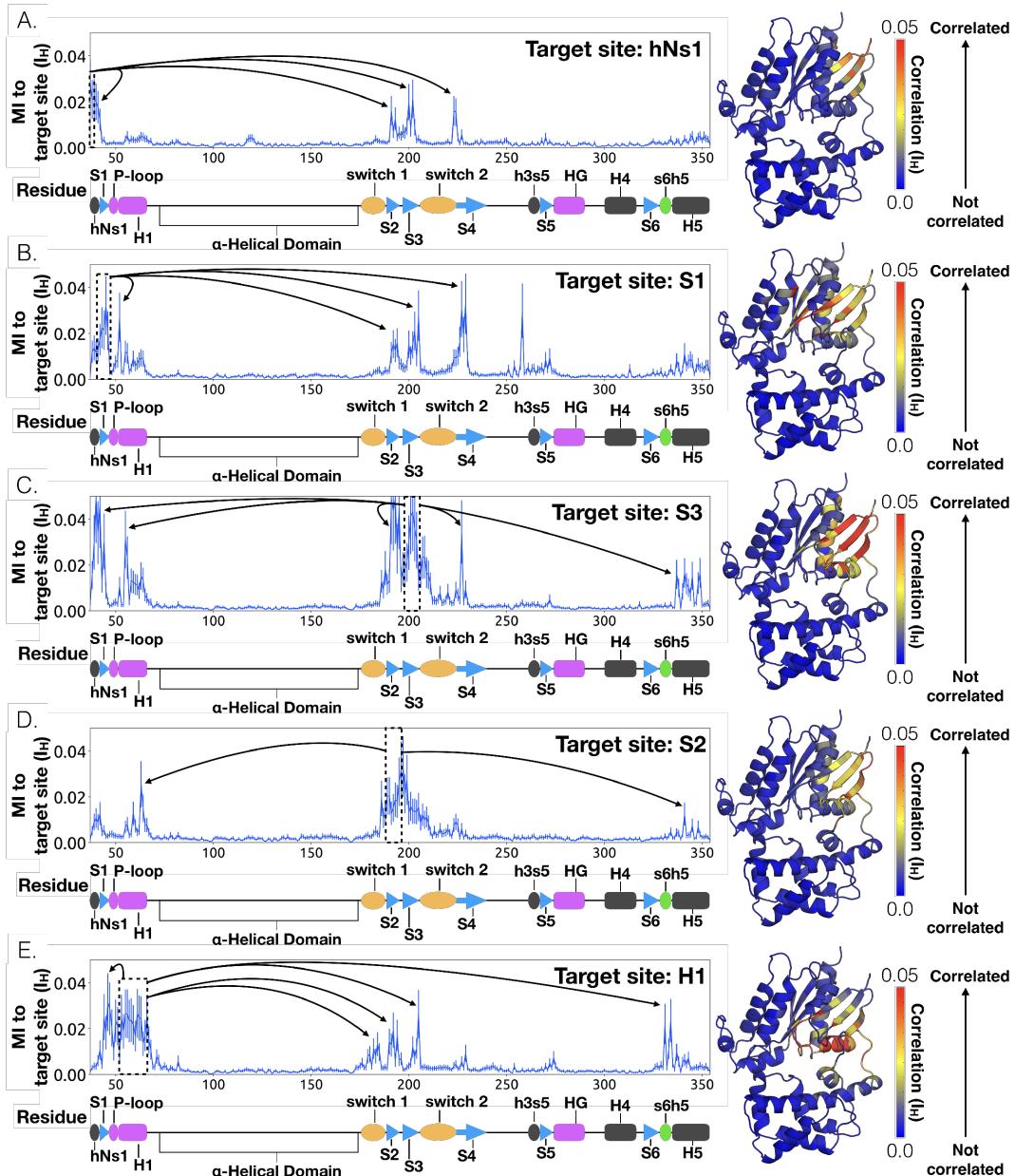
**Figure 1.6:** Probability distributions of the distance between the side-chains of Lys275 $G.s5hg.1$  and Glu49 $G.s1h1.4$ . Distributions were computed for the bound (blue), rate-limiting step (orange), and dissociated (black) states. Representative structures (above) for the bound (left, blue) and rate-limiting step (right, orange) are included with residues as sticks.

### 1.3.4 H1 and $\beta$ -sheets are communication hubs

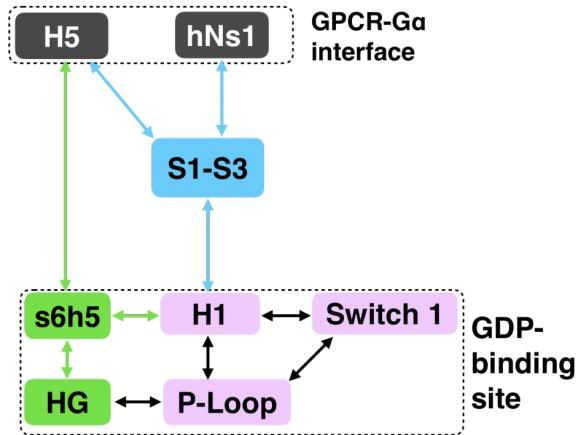
To identify other important structural elements in the allosteric network underlying G protein activation, we followed correlated motions emanating from other sites that are known to interact directly with GPCRs, including the hNs1 loop, the h3s5 loop, and the h4s6 loop [?]. We find that h3s5 and h4s6 are largely isolated, suggesting they play a role in forming a stable GPCR-G protein complex but not in the allosteric mechanism that triggers GDP release. This finding is consistent with sequence analysis suggesting these structural elements are important determinants of the specificity of GPCR-G $\alpha$  interactions [?]. In contrast, the hNs1 loop has strong correlations with  $\beta$ -strands S1-S3 (Figure 1.7). These strands, in turn, communicate with H1, switch 1, and the P-loop.

Integrating our correlation analysis with structural insight from the rate-limiting step described above suggests an important role for S1-S3 in a complex allosteric network that couples the GPCR- and nucleotide-binding sites (Figures 1.7 and 1.8, Figure ??). S2 and S3 twist relative to S1 and away from H1 (Figures 1.2A and 1.9, Figure ??). This twisting disrupts stacking between Phe194<sup>G,S2.6</sup> on S2 and His63<sup>G,H1.12</sup> on H1 and increases disorder of side-chains in H1 (Figures 1.2B and 1.9, Figure ??C). Increased disorder in H1 is also a crucial component of the proposed universal mechanism, but in that model translation of H5 is the key trigger for changes in H1. The role for the  $\beta$ -sheets in our model is consistent with previous work identifying interactions between S2 and H1 [?], NMR experiments showing chemical exchange in the methyls of S1-S3 upon receptor binding [?], and mutational data. In particular, Flock et al. have previously noted the important interaction between residues Phe194<sup>G,S2.6</sup> and His63<sup>G,H1.12</sup> [?].

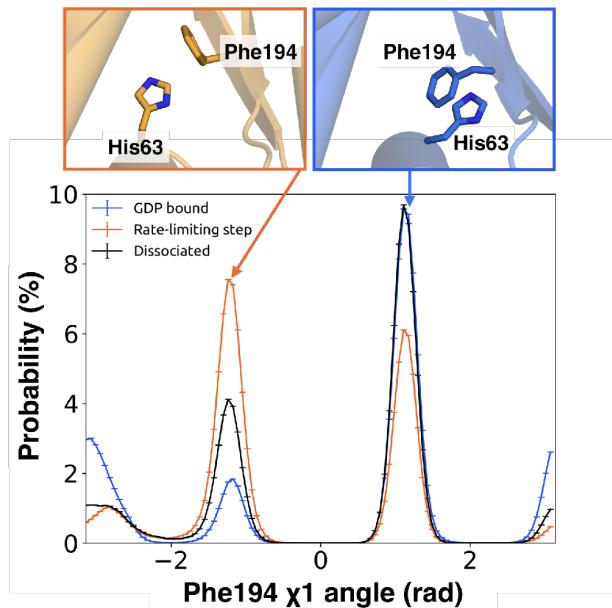
The importance of H1 and  $\beta$ -strands S1-S3 is underscored by mapping the global communication of every residue onto a structure of G $\alpha$  (Figure ??). The global communication of a residue is the sum of its correlations to every other residue and is a useful metric for identifying residues that are important players in allosteric networks [?]. Interestingly, these  $\beta$ -strands and H1 have higher global communication than H5 and the s6h5 loop. This suggests that H1 and the  $\beta$ -sheets integrate conformational information from multiple sources, including the hNs1 loop, and not just H5. The importance of the  $\beta$ -sheets and H1 for allosteric communication is consistent with their conservation [?], which may not simply reflect the role they play in protein folding and stability, as had been suggested previously [?, ?].



**Figure 1.7:** CARDs data showing communication per residue to a target site (dashed box) is plotted (left) and mapped onto the structure of G $\alpha$ q (right) for (A) hNs1, (B) S1, (C) S3, (D) S2, and (E) H1. Arrows indicate regions of importance with significant communication to the target site.



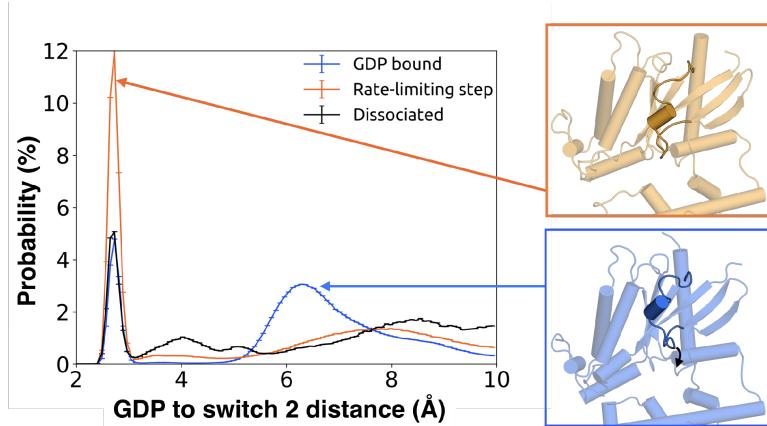
**Figure 1.8:** Allosteric network connecting the GPCR- and nucleotide-binding interfaces. The coloring scheme corresponds to that used in 1.1, highlighting the GPCR binding interface (gray), GDP phosphate-binding regions (pink), GDP nucleotide-binding regions (green), and the  $\beta$ -sheets (blue).



**Figure 1.9:**  $\pi$ -stacking between S2 and H1 is disrupted during the rate-limiting step. Distribution of the  $\chi_1$  angle (bottom) of Phe194<sup>G,S2.6</sup> on S2 before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of Phe194<sup>G,S2.6</sup> and His63<sup>G,H1.12</sup> (top) corresponding to before and after the rate-limiting step are also shown.

### 1.3.5 GDP release alters the structure and dynamics of the G $\beta$ -binding site

We also find that switch 2 has strong correlations with the nucleotide-binding site, especially switch 1 (Figure ??). Given that switch 2 is a major component of the interface between G $\alpha$  and G $\beta$ , this communication could enable GDP release to trigger dissociation of G $\alpha$  from G $\beta\gamma$ . Examining the rate-limiting step for GDP release reveals that switch 2 shifts towards the nucleotide-binding site (Figure 1.10) and exhibits increased conformational disorder (Figure 1.2B and Figure ??). These findings are consistent with previous kinetic studies postulating that switch 2 dynamics are impacted prior to GDP release [?].



**Figure 1.10:** Switch 2 moves towards GDP across the rate-limiting step. Distance distribution (left) of Gly207<sup>G.s3h2.1</sup> to GDP before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of Switch 2 motion are shown (right).

## 1.4 Conclusion

We have succeeded in simulating G protein activation, including both the allosteric coupling between the GPCR- and nucleotide-binding sites of G $\alpha$ q and consequent unbinding of GDP. Our results reveal a previously unobserved intermediate that defines the rate-limiting step for GDP release and, ultimately, G protein activation. Our model synthesizes a wealth of experimental data and previous analyses. For example, we identify an important role for coupling from H5 to the s6h5 loop and H1 that is consistent with a previously proposed universal mechanism for G protein activation. However, we also find that this allosteric network incorporates the hNs1 loop,  $\beta$ -strands S1-S3, and the HG helix. Strands S1-S3 and H1 serve as hubs in this network, simultaneously integrating information from both H5 and the hNs1 loop. Our observation is consistent with previous postulates that information flows from H5 and hNs1 to H1 [?]. It is important to note that our model was extracted using simulations of G $\alpha$ q, and so some correlations or changes in conformation and dynamics may apply only to G $\alpha$ q. However, by focusing our analysis on secondary structure elements and residues that are shared across all G $\alpha$  homologs, our model likely captures a universal ‘skeleton’ of changes involved in G $\alpha$  activation, expanding upon a previously proposed universal mechanism for G $\alpha$  activation [?]. The consistency of our model with a wide variety of structural and biochemical data suggests that it is a promising foundation for future efforts to understand the determinants of GPCR-G $\alpha$  interaction specificity, how mutations cause aberrant signaling and disease, and how small molecule inhibitors modulate G $\alpha$  activation. Our model also adds weight to the growing appreciation for the fact that a protein’s spontaneous fluctuations encode considerable information about its functional dynamics [?, ?, ?, ?]. Construction of our model was enabled by a powerful combination of simulation methods, namely metadynamics and MSMs. In the future, we expect this methodology will prove valuable for understanding other slow conformational changes and unbinding processes.

## 1.5 Materials and Methods

### 1.5.1 Molecular dynamics simulations of GDP unbinding

#### System preparation

We used the crystal structure of G $\alpha$  from the heterotrimeric G $\alpha$ q protein (PDB entry 3AH8) as the initial structure to set up our simulations [?]. The first 36 residues were removed prior to simulation, as they come from G $\alpha$ i, along with the G $\beta$  and G $\gamma$  subunits which were removed to minimize the system size and maximize our chances of observing GDP release [?].

The protein structure was solvated in a dodecahedron box of TIP3P water [?] that extended 1 nm beyond the protein in every dimension. A single Mg $^{2+}$  ion was added coordinating the phosphate groups of GDP as its presence accelerates GDP release [?]. Thereafter, Na $^+$  and Cl $-$  were added to produce a neutral system at 0.15 M NaCl. The final system consists of one GDP, one G $\alpha$ q, 57 Cl $^{-1}$ , 64 Na $^{+1}$ , one Mg $^{2+}$ , and 18696 TIP3P water molecules, for a total of 61,893 atoms.

Molecular dynamics (MD) simulations were carried out using Gromacs [?, ?] and the AMBER03 force field [?]. The force field parameters of GDP were obtained from the AMBER Parameter Database (<http://research.bmh.manchester.ac.uk/bryce/amber/>) [?]. The system was energy minimized with the steepest descent algorithm for 50,000 steps until the maximum force fell below 100 kJ/mol/nm using a step size of 0.02 nm and a cut-off distance of 0.9 nm for the neighbor list, electrostatic interactions and van der Waals interactions. Afterward, the solvent was relaxed by a NVT simulation of 100 ps with the constraint of 1000 kJ/mol/nm applied to the protein heavy atoms and 2 fs as the time step. In this relaxation simulation, the temperature of the system is constrained to 300 K using V-rescale thermostat (with a time constant of 0.1 ps) [?]. A cut-off distance of 1 nm was used for the van der Waals, short-range electrostatic interactions. Periodic boundary conditions are applied to x, y and z directions. The Particle-Mesh-Ewald method is employed to recover the long-range electrostatic interactions with 0.16 nm as the grid spacing and with a fourth order spline [?]. All the bonds connecting to hydrogens are constrained using LINCS algorithm [?]. After the NVT relaxation, the system further underwent an NPT simulation for one ns with the time step of 2 fs for equilibration. The simulation parameters here are kept the same as the NVT relaxation except for the application of Parrinello-Rahman barostat for pressure coupling [?]. After these equilibration runs, the constraint on heavy atoms were removed for all subsequent production runs. Virtual sites were used to allow for a 4 fs time-step [?]. We then applied a three-step protocol to simulate GDP release.

#### Step one: MD simulations of the GDP-bound state

We performed 100 parallel simulations of the GDP bound state on the Folding@home [?] distributed computing environment for an aggregate simulation time of 35.3  $\mu$ s.

#### Step two: Metadynamics simulations

We subsequently ran metadynamics simulations [?, ?] initiated from conformations generated in step one to actively promote GDP release. Starting conformations were selected from step one by clustering protein conformations into 625 states using a hybrid K-center/K-medoids method [?] with a 2 Å cutoff. The 10 most populated states included conformations with large inter-domain separation distances between the Ras-like and AH domains as measured by the angle formed between the  $\alpha$ -carbon atoms of Leu97<sup>H.HA.29</sup>, Asn82<sup>H.HA.14</sup>, Ile258<sup>G.H3.12</sup>, and Glu250<sup>G.H3.4</sup>. This inter-domain angle ranged from  $-30^\circ$  to  $0^\circ$ . From these states, three representative structures were chosen with inter-domain angles of  $-6^\circ$  (Conf. 1),  $-10^\circ$  (Conf. 2), and  $-26.6^\circ$  (Conf. 3) as starting conformations for metadynamics simulations, which were run on PLUMED [?]. We defined two collective variables for our metadynamics simulations: 1) the distance between GDP's phosphate groups and the backbone of Lys52<sup>G.H1.1</sup>-Thr54<sup>G.H1.3</sup> in G $\alpha$ q subunit (CV1), and 2) the RMSD of GDP to the starting conformation (CV2).

In metadynamics, a history-dependent biased potential VG(S,t) is added to the two selected CVs.

$$V_G(S, t) = \int_0^t (dt' \omega \exp(-\sum_{i=1}^d \frac{(S_i(R) - S_i(R(t')))^2}{2\sigma_i^2})) \quad (1.1)$$

**Table 1.1:** Details of metadynamics simulations

Starting conformation	Width of CV1 (Å)	Width of CV2 (Å)	Number of conformations selected
<b>1</b>	0.1	0.1	171
<b>1</b>	0.08	0.03	132
<b>1</b>	0.03	0.01	504
<b>2</b>	0.1	0.1	145
<b>2</b>	0.08	0.03	141
<b>2</b>	0.03	0.01	320
<b>3</b>	0.1	0.1	198
<b>3</b>	0.08	0.03	186
<b>3</b>	0.03	0.01	288

where  $t$  represents time,  $S$  are collective variables,  $\omega$  is the energy rate and  $\sigma_i$  controls the width of the Gaussian for the  $i$ th collective variable. Summing up the Gaussians allows us to obtain the biased potential  $V_G$ . The free energy  $-F(S)$  is derived by the assumption,

$$\lim_{t \rightarrow \infty} (S, t) \sim -F(S) \quad (1.2)$$

The metadynamics simulations were repeated three times for each selected representative structure using different Gaussian widths (1.1). We set the Gaussian height to 1.5 kJ/mol.

We observed the release of GDP from  $G\alpha$  in the metadynamics simulations and obtained the free-energy landscape for the release. We then applied the string method [?] to detect potential release pathways and use these conformations as the seeds for simulations in step 3. We can use the function  $\chi(\alpha)$  to represent  $\chi(C1, C2)$  which showing the minimum free-energy path. We thus have

$$\frac{dZ(\alpha)}{d\alpha} = \sum_{k=1}^n \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{d\chi_k}{d\alpha} \quad (1.3)$$

which is parallel to

$$\sum_{k=1}^n \frac{\partial \theta(\chi(\alpha))}{\partial \chi_k} \frac{dF(\chi_k)}{d\chi_k} = \sum_{j,k=1}^n \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{\partial \theta_j(\chi(\alpha))}{\partial \chi_k} \frac{dF(z(\alpha))}{dZ(\alpha)} \quad (1.4)$$

The average of the tensor

$$\sum_{j,k=1}^n \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{\partial \theta_j(\chi(\alpha))}{\partial \chi_k} \quad (1.5)$$

can be represented as

$$M_{ij}(z) = \Omega^{-1} e^{\beta F(z)} \int_{\mathbb{R}^N} \sum_{k=1}^n \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{\partial \theta_j(\chi(\alpha))}{\partial \chi_k} e^{-\beta F(z)} \quad (1.6)$$

$$\prod_{v=1}^N \delta(z_v - \theta_v(\chi)) d\chi \quad (1.7)$$

The points determining the minimum free-energy path along the surface satisfy

$$0 = (M_{ij}(z) \Delta F(z(\alpha)))^\perp \quad (1.8)$$

We applied this method [?, ?] to obtain a minimum free energy path, extracting 2085 conformations along potential GDP release pathways. Notably, we only observed the transition from the bound state and the unbound state for one time in a single metadynamics simulation. This implies that the predicted free-energy surface from metadynamics cannot be used to describe the release events accurately. In spite of this, we can still use the conformations along the release pathway explored by metadynamics to seed unbiased parallel MD simulations.

### Step three: Metadynamics-seeded MD simulation of GDP release

Lastly, we carried out unbiased MD simulations initiated from the 2085 conformations obtained from metadynamics using the Folding@home platform [?]. A total of  $122.6 \mu\text{s}$  of simulation was generated in this step. All the following analyses are based on unbiased MD simulations.

### 1.5.2 Identifying the allosteric network with CARDS

To determine how the GPCR- and GDP-binding regions communicate with one another, we applied the CARDS [?] methodology to simulations of the GDP-bound state of G $\alpha$ q. CARDS measures communication between every pair of dihedrals via both correlated changes in structural motions and dynamical behavior. Structural states are captured by discretizing backbone  $\phi$  and  $\psi$  dihedrals into two structural states (cis and trans), while side-chain  $\chi$  angles are placed into three states (gauche+, gauche-, and trans). Every dihedral is also parsed into dynamical states, capturing whether the dihedral is stable in a single state (ordered), or rapidly transitioning between multiple states (disordered). These dynamical states are identified using two kinetic signatures of dihedral motion: the average time a dihedral persists in a structural state (an ordered timescale), and the typical timescale for transitions between structural states (a disordered timescale). Parsing into dynamical states utilizes a two-step process by (i) calculating the distribution of ordered and disordered times from the simulations and (ii) assigning each period of time between two consecutive transitions into ordered and disordered states based on which distribution the time between two transitions is most consistent with.

From these states, a holistic communication ( $I_H(X, Y)$ ) is computed for every pair of dihedrals  $X$  and  $Y$ :

$$I_H(X, Y) = \overline{I_{ss}(X, Y)} + \overline{I_{dd}(X, Y)} + \overline{I_{ds}(X, Y)} + \overline{I_{sd}(X, Y)} \quad (1.9)$$

where  $\overline{I_{ss}(X, Y)}$  is the normalized mutual information between the structure (i.e., rotameric state) of dihedral  $X$  and the structure of dihedral  $Y$ ,  $\overline{I_{ds}(X, Y)}$  is the normalized mutual information between the structure of dihedral  $X$  and the dynamical state of dihedral  $Y$ ,  $\overline{I_{sd}(X, Y)}$  is the normalized mutual information between the dynamical state of dihedral  $X$  and the structure of dihedral  $Y$ , and  $\overline{I_{dd}(X, Y)}$  is the normalized mutual information between the dynamical state of dihedral  $X$  and the dynamical state of dihedral  $Y$ . The Mutual Information ( $I$ ) is

$$I(X, Y) = - \sum_{x \in X} \sum_{y \in Y} \frac{P(x, y)}{P(x)P(y)} \quad (1.10)$$

where  $x \in X$  refers to the set of possible states that dihedral  $X$  can adopt,  $p(x)$  is the probability that dihedral  $X$  adopts state  $x$ , and  $p(x, y)$  is the joint probability that dihedral  $X$  adopts state  $x$  and dihedral  $Y$  adopts state  $y$ . Normalized mutual information is computed using the maximum possible mutual information for any specific mode of communication.

From the pairwise correlation for every dihedral-pair, we extracted how much each individual residue communicates with a target site of interest via bootstrapping with 10 random samples with replacement. After locating the group of residues communicating most strongly with a specific target site, we set this newly identified group as the new target site; The iteration of this process allows us to identify a pathway of communication from one region of interest to another. Here, we set the GPCR contact sites as our initial target sites. We then iteratively used this approach to identify pathways connecting these contact sites with the GDP-binding site of G $\alpha$ q.

### 1.5.3 Markov state model construction

We clustered G $\alpha$ q conformations and GDP binding states separately and combined the assignments to build a Markov State Model using MSMbuilder [?, ?] and enspara [?]. First, we clustered protein conformations into 5040 states using a hybrid k-center/k-medoids method with 1.8 Å cutoff. Then we clustered the GDP-binding state into 321 states using the automatic partitioning algorithm (APM) [?] with a residence time of 2 ns. By combining the assignments from protein conformations and the GDP-binding states, we obtained a total of 221,965 states. The implied timescales of this MSM show Markovian behavior with a lag time of 5 ns [?]. (Figure ??). Analyses of distances, angles, and dihedrals of interest were carried out using bootstrapping with ten random samples, with replacement. Results were insensitive to varying the number of bootstrapped samples between 5 and 30. Histograms were generated using 100 bins.

### 1.5.4 Quantifying conformational disorder

The disorder of every residue was measured by computing Shannon entropy [?] of each dihedral as they are natural degrees of freedom for describing protein dynamics. Shannon entropy ( $H$ ) is defined as

$$H(X) = - \sum_{x \in X} p(x) \log(p(x)) \quad (1.11)$$

where  $x \in X$  refers to the set of possible states that dihedral  $X$  can adopt and  $p(x)$  is the probability that dihedral  $X$  adopts state  $x$ . Dihedral angles were calculated using MDTraj (McGibbon et al., 2015) and assigned to discrete rotameric states as described above using CARDS. The entropy of a single residue was computed by summing up the entropies of its dihedrals, and normalized by the residue's maximum possible Shannon entropy. This maximum possible Shannon entropy, using a flat distribution for the appropriate number of bins, is referred to as the 'channel capacity' and has been used to normalize other information-theoretic metrics [?]. Summing entropies within a residue establishes an upper bound on the degree of motion for a single residue, while ignoring intra-residue correlations between dihedrals.

### 1.5.5 Identification of the rate-limiting step for GDP release

We used transition path theory (TPT) [?, ?] to find the highest flux paths from the bound state to the unbound state [?]. The bound state was defined as all clusters that satisfied two criteria: (i) GDP is within 6 Å of the backbone atoms of Lys52<sup>G,H1,1</sup>-Thr54<sup>G,H1,1</sup>, and (ii) GDP has an RMSD < 0.5 Å to its crystallographic conformation. The unbound state was defined as all clusters with GDP > 55 Å from the binding pocket. The rate-limiting step was identified by finding the bottleneck in the highest flux paths. To obtain this, we first calculate the flux between states  $i$  and  $j$  along any possible unbinding path using

$$f(x) = \begin{cases} \pi_i q_i^- T_{ij} q_j^+, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}$$

where  $q^+$  is the committer probability from the bound to the unbound state, and  $q_i^-$  is  $1 - q_i^+$ ;  $\pi_i$  is the weighted probability, and  $T_{ij}$  is the transition matrix. The highest flux paths can be identified by maximizing the fluxes between the bound states and the unbound states using

$$c(w) = \min (f_{i_l i_{l+1}}^+ | l = 1 \dots n_w - 1) \quad (1.12)$$

where  $i_l$  are intermediate states. From this, the slowest step was extracted as the minimum flux step of the highest flux release pathway.

## 1.6 Acknowledgements

We thank TE Frederick and TD Todd for their helpful discussion and insight. We are grateful to the Folding@home users for computing resources.