Locating ligand binding sites in G-protein coupled receptors using combined information from docking and sequence conservation

Ashley R Vidad [Equal first author, 1], **Stephen Macaspac** [Equal first author, 1], **Ho Leung Ng** [Corresp. 2]

[1] Department of Chemistry, University of Hawaii at Manoa, Honolulu, Hawaii, United States

[2] Department of Biochemistry and Molecular Biophysics, Kansas State University, Manhattan, Kansas, United States of America

Corresponding Author: Ho Leung Ng
Email address: hng@ksu.edu

GPCRs (G-protein coupled receptors) are the largest family of drug targets and share a conserved structure. Binding sites are unknown for many important GPCR ligands due to the difficulties of GPCR recombinant expression, biochemistry, and crystallography. We describe our approach, ConDockSite, for predicting ligand binding sites in GPCRs using combined information from surface conservation and docking, starting from crystal structures or homology models. We demonstrate the effectiveness of ConDockSite on crystallized GPCRs such as the beta2 adrenergic and A2A adenosine receptors. We also demonstrate that ConDockSite successfully predicts ligand binding sites from high-quality homology models. Finally, we apply ConDockSite to predict ligand binding sites on a structurally uncharacterized GPCR, GPER, the G-protein coupled estrogen receptor. Most of the sites predicted by ConDockSite match those found in other independent modeling studies. ConDockSite predicts that four ligands bind to a common location on GPER at a site deep in the receptor cleft. Incorporating sequence conservation information in ConDockSite overcomes errors introduced from physics-based scoring functions and homology modeling.

TITLE: Locating ligand binding sites in G-protein coupled receptors using combined information from docking and sequence conservation

Authors: Ashley R. Vidad*[1], Stephen Macaspac*[1] & Ho Leung Ng[2#]

[1]University of Hawaii at Manoa, Department of Chemistry, Honolulu, HI. USA
[2]Kansas State University, Department of Biochemistry & Molecular Biophysics, Manhattan, KS. USA

* Authors are considered equal contributors to this paper.
[#] Corresponding author. hng@ksu.edu

Keywords: G protein-coupled receptor (GPCR), binding sites, homology modeling

Abstract

GPCRs (G-protein coupled receptors) are the largest family of drug targets and share a conserved structure. Binding sites are unknown for many important GPCR ligands due to the difficulties of GPCR recombinant expression, biochemistry, and crystallography. We describe our approach, ConDockSite, for predicting ligand binding sites in GPCRs using combined information from surface conservation and docking, starting from crystal structures or homology models. We demonstrate the effectiveness of ConDockSite on crystallized GPCRs such as the beta2 adrenergic and A2A adenosine receptors. We also demonstrate that ConDockSite successfully predicts ligand binding sites from high-quality homology models. Finally, we apply ConDockSite to predict ligand binding sites on a structurally uncharacterized GPCR, GPER, the G-protein coupled estrogen receptor. Most of the sites predicted by ConDockSite match those found in other independent modeling studies. ConDockSite predicts that four ligands bind to a common location on GPER at a site deep in the receptor cleft. Incorporating sequence conservation information in ConDockSite overcomes errors introduced from physics-based scoring functions and homology modeling.

Introduction

GPCRs (G-protein coupled receptors) are the largest family of drug targets and the targets of >30% of all drugs. Because they are membrane proteins with flexible and dynamic structures, biochemical and crystallography experiments are difficult. Only ~60 GPCRs out of ~800 in the human genome have been crystallized despite their great pharmacological importance. GPCR homology modeling remains challenging due to conformational flexibility and the abundance of flexible loops (Lai et al., 2017). Crystal structures have shown that the large majority of ligands bind in the large central, extracellular cavity of GPCRs, but the specific binding sites in the cavity can vary widely between different ligands even for the same or closely related receptors (Wacker, Stevens & Roth, 2017; Chan et al., 2019).

Various computational approaches have been used to predict ligand binding sites in G-protein coupled receptors. Traditional docking methods compute the lowest energy pose of a ligand fit to a receptor surface. Such methods are highly dependent on the form of the energy scoring function and accuracy of the receptor model structure (Katritch et al., 2010; Katritch & Abagyan, 2011; Shoichet & Kobilka, 2012; Weiss et al., 2016; Lim et al., 2018). These methods have been used to identify ligand binding sites and build pharmacophores for GPCRs (Kratochwil et al., 2011; Sanders et al., 2011; Tang et al., 2012), but the lack of diverse GPCR crystal structures presents serious challenges to using docking methods for identification of ligand binding sites. In particular, few crystal structures of non-class A GPCRs have been determined. Moreover, homology models usually cannot be used to identify ligand binding sites or for docking without extensive optimization, such as with advanced molecular dynamics sampling methods (Katritch et al., 2010; Lai et al., 2017; Zou, Ewalt & Ng, 2019). An underappreciated feature that can be used to predict ligand binding sites is surface or sequence

conservation. Binding sites for particular ligands are often conserved, and systematic sequence variation can encode ligand specificity (Capra & Singh, 2007; Kalinina, Gelfand & Russell, 2009; Wass & Sternberg, 2009). While highly conserved receptors often share similar ligand binding sites, such direct relationships often do not apply between less conserved receptors. Yet, the massive abundance of genomic data for GPCRs can provide strong constraints for possible ligand binding sites even without chemical or structural information (Madabushi et al., 2004; Sanders, 2011; Levit et al., 2012). The binding sites for synthetic, non-physiological ligands can also be identified as they often share some or even most of their binding sites with physiological ligands (Wacker, Stevens & Roth, 2017).

There has been less research on methods that combine information from chemical interactions, geometric surface analysis, and bioinformatics. Hybrid strategies, such as Concavity (Capra et al., 2009), have demonstrated superior performance in predicting ligand binding sites compared to single-mode approaches. Concavity scores binding sites by evolutionary sequence conservation, as quantified by the Jensen-Shannon divergence (Capra & Singh, 2007), and employs geometric criteria of size and shape. Here, we describe a new hybrid strategy we have developed, called ConDockSite, to predict ligand binding sites from combined information from surface conservation and docking calculations. We compare our results with those previously published using purely docking-based and other hybrid methods (Arnatt & Zhang, 2013; Méndez-Luna et al., 2015). ConDockSite is not intended to be used for docking, ie., predicting ligand binding poses, which are highly sensitive to small structural details in crystal structures. We demonstrate the effectiveness of ConDockSite for identifying ligand binding sites for the two best characterized GPCRs with known crystal structures, the β2 adrenergic and A2A adenosine receptors.

92 Finally, we apply ConDockSite to predict the binding sites of four ligands to the less
93 characterized G-protein coupled estrogen receptor (GPER, formerly known as GPR30), a
94 membrane-bound estrogen receptor. GPER is proposed to mediate rapid estrogen-associated
95 effects, cAMP regeneration, and nerve growth factor expression (Kvingedal & Smeland, 1997;
96 Carmeci et al., 1997; O'Dowd et al., 1998; Filardo et al., 2002; Kanda & Watanabe, 2003).
97 GPER is known to bind estradiol and the estrogen receptor inhibitors, tamoxifen and fulvestrant,
98 that are used to treat breast cancer (Fig. S1). Recently, GPER-specific ligands G1 and G15 were
99 discovered (Bologa et al., 2006; Dennis et al., 2009). G1 and G15 are structurally similar,
100 differing by only an acetyl group. G1 is an agonist, whereas G15 is an antagonist. No crystal
101 structure of GPER is available, and details of ligand binding are unknown. We discuss how the
102 ConDockSite-predicted binding sites provide a basis for G1 and G15 binding specificity.
103 ConDockSite predictions can be tested experimentally by measuring the effects of mutagenesis
104 of predicted ligand binding sites on ligand binding. Such efforts should be straightforward given
105 our recent publication describing methods for recombinant expression and ligand binding assays
106 for GPER (Souza et al., 2019).
107
108 **Results**
109 We developed ConDockSite to predict ligand binding pockets using information from
110 surface conservation and docking calculations. ConDockSite uses a simple scoring function that
111 is the product of surface conservation scores from ConSurf (Armon, Graur & Ben-Tal, 2001) and
112 docking scores from SwissDock (Grosdidier, Zoete & Michielin, 2011a).
113 The A2A adenosine and β2 adrenergic receptors are the most heavily studied GPCRs by
114 crystallography. We used them as standards to validate the effectiveness of ConDockSite for

predicting ligand binding sites. For both receptors, we performed cross-docking of an agonist and inverse agonist against a crystal structure of the receptor bound to a different agonist or inverse agonist: ligands were cross-docked rather than self-docked into its own crystal structure. Docking was performed with SwissDock which has demonstrated high accuracy in docking ligands into receptors without prior knowledge of the binding site (also known as global or blind docking) and also includes a user-friendly web interface suitable for students (Grosdidier, Zoete & Michielin, 2011a). SwissDock docking results were then ranked by the ConDockSite scoring function (Table S1). Residues within 3.5 Å of the highest scoring predicted ligand sites were compared with the binding surfaces associated with the ligand poses in the crystal structures. In addition, we determined the distances between the centers of mass for the poses in the crystal structure and those scored highest by ConDockSite.

As a convenient proxy for the distances between predicted and experimental ligand binding sites, we use the distances between the ConDockSite-scored ligand poses and those observed in the crystal structures. The highest ConDockSite-ranked pose for adenosine within the A2A adenosine receptor was within 1.8 Å of the ligand position (distance between centers of mass) in the crystal structure. (Fig. 1A). The ConDockSite-predicted binding site had a ConSurf conservation score of 0.86 and is essentially the same as the experimental binding site. The highest ranked site for ZM241385 within the A2A adenosine receptor was within 1.0 Å of the ligand's position in the crystal structure. In this top pose, ZM241385 is found within the same binding site as that observed in the crystal structure (Fig. 1B), with a ConSurf conservation score of 0.86.

The highest ranked pose for epinephrine within the β2 adrenergic receptor was within 0.4 Å of the ligand position within the crystal structure (PDB 4ldo). This binding site for epinephrine

was again essentially the same as the observed binding pocket (Fig. 1C). The highest ranked pose for carazolol within the β2 adrenergic receptor was within 1.8 Å of the ligand's position within the crystal structure (PDB 2rh1). This binding site for carazolol was essentially the same as that in the crystal structure (Fig. 1D). This pose had a ConSurf conservation score of 0.78. The extremely accurate placement of both agonists and antagonists demonstrates ConDockSite's effectiveness when a GPCR crystal structure is available.

Unfortunately, crystal structures are not available for most GPCRs. The most valuable use of ConDockSite is to predict drug binding sites in homology models. By using surface conservation information, ConDockSite is less sensitive to homology model inaccuracies than other ligand binding site prediction methods that are based purely on geometric methods. To demonstrate the ability of ConDockSite to work with homology models, we created models of four GPCRs, the β2 adrenergic, A2A adenosine, 5HT2B serotonin, and mu opioid receptors, that excluded the known crystal structures as templates. We used I-TASSER (Yang et al., 2015) for homology modeling which does not use GPCR-specific structural constraints but allows for custom selection of templates. I-TASSER created fairly accurate models of all four receptors, with RMSDs between the models and crystal structures ranging from a best of 0.85 Å for the β2 adrenergic receptor (PDB 2rh1) to a respectable 2.1 Å for the A2A adenosine receptor (PDB 5k2a). We used ConDockSite to predict the binding sites of the β2 adrenergic receptor with carazalol, A2A adenosine receptor with ZM241385, 5HT2B serotonin receptor with methysergide, and mu opioid receptor with BU72. As expected, ConDockSite performed best with the highly accurate β2 adrenergic receptor homology model, with only 1.8 Å between the centers of mass of the predicted and crystal structure ligand poses (Fig. 2, Table S2), supporting the prediction of very similar binding pockets. Performance decreased for the other three

receptors with less reliable homology models. Models of the A2A adenosine and mu opioid receptors (PDB 5c1m) have RMSDs of 3-4 Å between the predicted and crystal structure ligand poses. In this RMSD range, most of the residues are the same between the predicted binding sites and those in the crystal structures, supporting successful ConDockSite predictions. ConDockSite performs less well with the 5HT2B serotonin receptor (PDB 6drz) for which the RMSD between the predicted and actual ligand binding sites was 7.3 Å. In the 5HT2B receptor structure, the ligand, methysergide, binds very deep in the receptor. Serious errors in homology modeling of the receptor make it difficult or impossible to dock the ligand into the deep, restricted binding site. Nevertheless, our results with ConDockSite are consistent with benchmark modeling results that show that GPCR homology models of modest accuracy from templates with low sequence identity are still useful for docking and virtual screening (Lim et al., 2018; Costanzi et al., 2019). After demonstrating the applicability of ConDockSite for homology models, we applied ConDockSite to predict the binding sites in a GPCR, GPER (G-protein coupled estrogen receptor), which has not yet been crystallized. To predict the potential ligand binding sites in GPER, we first created a homology model using GPCR-I-TASSER (Zhang et al., 2015). GPCR-I-TASSER has been shown to be among the most accurate GPCR homology modeling software package. We used the generic I-TASSER we used in our validation studies because of its fine-grained options for template selection that are lacking in GPCR-I-TASSER. Because GPCR-I-TASSER uses GPCR-specific structural constraints, it is expected to outperform the generic I-TASSER (Zhang et al., 2015). GPCR-I-TASSER identified the closest matching crystal structure to GPER to be the CCR5 chemokine receptor (PDB 4mbs) with 23% sequence identity. GPCR-I-TASSER used this crystal structure along with 9 other GPCR crystal structures as templates for homology modeling. The GPER homology model differs from chain A of the crystal structure of

CCR5 chemokine receptor with RMSD of 0.96 Å across Cα atoms (Fig. S2) and has an excellent Ramachandran plot (Fig. S3). The primary differences are in the extracellular loop between helices 4 and 5 and the intracellular loops between helices 5 and 6, and after helix 7. These two intracellular loops are predicted by ERRAT (Colovos & Yeates, 1993) to be the least reliable based on the likelihood of atom pair type interactions from high-resolution crystal structures (Fig. S4).

Using the SwissDock server (Grosdidier, Zoete & Michielin, 2011a), we docked structures of the four ligands E2, G1, G15, and tamoxifen (Fig. S1) to the homology model of GPER. The docked sites from SwissDock, including those that were scored the highest, were located throughout the receptor surface and thus were considered mostly nonviable (Fig. 3). The shortcomings of a purely physics-based scoring function such as that used by SwissDock in predicting ligand binding are not surprising given the lack of an experimental crystal structure and well-known limitations of current homology modeling and docking methodology (Li, Hou & Goddard III, 2010; Merz, 2010; Wan et al., 2015; Smith et al., 2016).

We then ranked all ligand binding sites generated by SwissDock using the combined ConDockSite score. The ConDockSite score is simply the product of the ConSurf (Armon, Graur & Ben-Tal, 2001; Ashkenazy et al., 2010) binding surface sequence conservation score and the SwissDock FullFitness energy score (Grosdidier, Zoete & Michielin, 2011b). A highly negative ConDockSite score is associated with a more probable ligand binding site. For all four ligands, the ConDockSite score identified one or two ligand binding sites and poses that clearly outscored other candidates (Table S1). ConDockSite identified the same approximate binding site for all four ligands, although this was not an explicit criterion in the calculations (Fig. S5). The average ConSurf conservation score across the four ligand binding sites is 0.82 (1.0 represents complete

conservation), indicating that the site is highly but not completely conserved. The binding site is located deep in the receptor cleft, although depth was not a criterion in the prediction calculation. Given the lack of additional experimental evidence for the location of the ligand binding site, the proposed ConDockSite sites are physically reasonable.

We found two promising potential binding sites for E2 in GPER. The two sites are 4.4 Å apart, located deep in the receptor cleft (Fig. 4). E2 is oriented perpendicular to the lipid membrane and rotated about 180° between the two poses. The conservation scores for these two poses are 0.84 and 0.80. The energy scores of the two poses are similar. The amino acids contacting E2 in pose 1 are conserved in GPERs from six species, and only one residue contacting pose 2, H282, varies across species. In the top ranked pose, there is a hydrogen bond between the inward pointing D-ring hydroxyl group of E2 and the carboxyl terminal on E115. Hydrophobic interactions are present between E2 and non-polar residues L119, Y123, P303, and F314. This binding site approximately corresponds to that predicted by Lappano et al using docking (Lappano et al., 2010). In the second ranked pose, the inward pointing A-ring hydroxyl group of E2 makes a hydrogen bond with N310. This pose is in a less hydrophobic environment, contacting primarily H282 and P303.

ConDockSite predicts that G1 and G15 bind in adjacent but distinct binding sites separated by 2.3 Å. The top predicted binding site for G1 is found within the pocket bound by Y55, L119, F206, Q215, I279, P303, H307, and N310 (Fig. 5). This orientation had the highest conservation score of all predicted binding sites at 0.85. In this pose, N310 makes a long hydrogen bond with the acetyl oxygen of G1. The predicted binding site for G15 is found within the pocket bound by L119, Y123, M133, S134, L137, Q138, P192, V196, F206, C207, F208, A209, V214, E218, H307, and N310. This pose had a conservation score of 0.8. Hydrogen

bonding is not observed between GPER and G15. Hydrophobic interactions are observed with L119, Y123, F206, and V214. The ConDockSite G1 result correspond to the binding sites predicted by recent studies using docking and molecular dynamics simulations and validated by design and activity testing of new G1 derivatives (Méndez-Luna, Bello & Correa-Basurto, 2016; Martínez-Muñoz et al., 2018).

ConDockSite predicted two equally high scoring, overlapping poses for tamoxifen, near E115, L119, Y123, L137, Q138, M141, Y142, Q215, E218, W272, E275, I279, P303, G306, H307, and N310 (Fig. 6). The conservation score of this orientation is 0.81. Hydrophobic interactions are observed between tamoxifen and non-polar residues L119, Y123, Y142, P303, and F314. Notably, the amine group of tamoxifen is neutralized by E218 and E275.

We compared the GPER ligand binding sites predicted by ConDockSite to those predicted by three other software packages representing different approaches: CASTp (Dundas et al., 2006), which analyzes surface geometry, SiteHound (Hernandez, Ghersi & Sanchez, 2009), which maps surfaces with a chemical probe, and Concavity (Capra et al., 2009), which analyzes surface geometry and conservation (Fig. 7). All three methods could identify a ligand binding site very roughly matching that from ConDockSite. In comparison with the ligand binding sites predicted by traditional methods based on surface geometry and conservation (Fig. 7), the sites predicted by ConDockSite are more detailed and of higher resolution due to the information from chemical interactions from ligand docking. Moreover, prediction methods based on surface geometry and conservation cannot differentiate between binding sites for different ligands. The pocket predicted by ConDockSite is deeper than the other pockets, which while intuitively attractive, is not necessarily correct. SiteHound performed particularly poorly, with the top scoring site located on the GPER intracellular face. The site identified by SiteHound closest to

the ConDockSite site was scored third and is a shallow binding pocket near H52-G58, E275-H282, and R299-H307 (Fig. 7C). In contrast, the Concavity site was smaller and shallower than the ConDockSite site (Fig. 7D). Surprisingly, the site predicted by the simpler CASTp method best matched the ConDockSite site but is also smaller and shallower (Fig. 7B). For proteins such as GPCRs with large, concave binding pockets, geometry-based prediction methods such as Concavity and CASTp can easily identify the general, approximate location of the ligand binding site. However, such methods may have more difficulty recovering the specific, ligand-specific binding site. It is also surprising that ConDockSite more closely matched the results of the geometry-based methods given that ConDockSite does not take surface geometry into account. As described previously, the G1 and G15 binding sites predicted by ConDockSite more closely match those made using docking against very computationally expensive molecular dynamics simulations (Méndez-Luna, Bello & Correa-Basurto, 2016).

Discussion

The ConDockSite scoring method, incorporating information from both surface conservation and docking binding energy, demonstrated high accuracy in predicting ligand binding sites from the crystal structures of two class A GPCRs, the A2A adenosine and β2 adrenergic receptors. ConDockSite also successfully predicted the ligand binding sites from high-quality homology models. ConDockSite was also used to predict viable ligand binding sites for four different GPER ligands. In contrast to more typical geometry-based ligand binding site prediction methods, ConDockSite scoring takes advantage of chemistry-specific information about the ligand-receptor interface. The poor performance of SiteHound in predicting ligand binding sites on GPER suggests that a method based only on chemical interactions or docking is highly susceptible to error, most likely due to the inadequate accuracy of homology models.

Surface conservation data not only provides orthogonal knowledge but also dampens the influence from the shortcomings of current computational methods in homology modeling, docking, and predicting binding affinity. How best to mathematically combine these multiple data sources has been debated (Capra & Singh, 2007; Capra et al., 2009), but we demonstrate here that a simple product scoring function is already effective. The four GPER ligands studied here differ greatly in chemical structure, but the ConDockSite scoring method predicted that all four bind to the same approximate region, deep in the extracellular cleft of the receptor. Undoubtedly, further refinement of a hybrid scoring function will lead to improved predictions.

Earlier GPER modeling studies using molecular dynamics simulations and docking identified different potential binding sites for E2, G1, and G15 near F206 and F208; the interaction with this region was described as driven primarily by π-π stacking interactions (Arnatt & Zhang, 2013; Méndez-Luna et al., 2015). Figure S6 compares the ConDockSite binding site against that predicted in the molecular dynamics simulation and docking study. The ConDockSite binding site is located deeper in the extracellular cleft; the other proposed site mostly involved surface-exposed loops. It was proposed that Q53, Q54, G58, C205, and H282 all interact with G1 and G15; however, none of these residues are conserved across the six species we analyzed. More recent studies using better homology models and computationally expensive long time-scale molecular dynamics simulations predict E2, G1, and G15 binding sites that approximately match those predicted by ConDockSite (Lappano et al., 2010; Méndez-Luna, Bello & Correa-Basurto, 2016). The ConDockSite binding site predictions can be tested experimentally by performing site-directed mutagenesis and ligand binding assays.

In summary, the simple ConDockSite hybrid scoring model predicts physically plausible ligand binding sites by combining information from ligand docking and surface conservation.

Using multiple orthogonal sources of information avoids errors introduced by modeling (Capra et al., 2009). Given a homology model of modest quality, ConDockSite can accurately predict ligand binding sites. Using this hybrid method, we identified a site in the extracellular cavity of GPER that has the potential to bind four known GPER ligands. Further optimization of hybrid scoring functions should yield significantly improved predictions. Extension of this approach may allow analysis of non-class A GPCRs.

Methods

Protein surface conservation

GPCR protein sequences were acquired from the SwissProt database (Boeckmann et al., 2005). For the A2A adenosine receptor, the protein sequences aligned were from *Homo sapiens, Canis familiaris*, *Xenopus tropicalis*, *Myotis davidii*, *Loxodonta africana*, *Gallus gallus*, *Anolis caronlinesis*, *Oncorhynchus mykiss*, *Ailuropoda melanoleuca*, and *Alligator mississippiensis*. For the β2 adrenergic receptor, the protein sequences aligned were from *Homo sapiens*, *Oncorhynchus mykiss*, *Myotis brandtii*, *Callorhinchus milii*, *Ophiophagus hannah*, *Canis familiaris*, *Loxodonta africana*, *Ailuropoda melanoleuca*, *Ficedula albicollis*, and *Xenopus laevis*. GPER protein sequences aligned were from diverse species: *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Macaca mulatta*, *Danio rerio*, and *Micropogonias undulatus*. Sequences were chosen to represent a diverse range of animal species. Multiple sequence alignment files were submitted to ConSurf (Armon, Graur & Ben-Tal, 2001; Ashkenazy et al., 2010). ConSurf assesses conservation using Bayesian reconstruction of a phylogenetic tree. Each sequence position is scored from 0-9, where 9 indicates that the amino acid was retained in all

the organisms (Fig. S7). Values from ConSurf were mapped onto the receptor surface with Chimera (Pettersen et al., 2004).

Homology modeling and docking

The crystal structures for the A2A adenosine receptor and the β2 adrenergic receptor were acquired from the RCSB protein data bank: β2 adrenergic receptor bound to epinephrine (PDB 4ldo), β2 adrenergic receptor bound to carazolol (PDB 2rh1), A2A adenosine receptor bound to adenosine (PDB 2ydo), and A2A adenosine receptor bound to ZM241385 (PDB 5k2a). The crystal structures of the mu opioid receptor and 5HT2B receptor were taken from PDB 5c1m and 6drz. Structures were prepped for docking with Chimera by removing extraneous chains and bound ligands with the DockPrep protocol. Ligands were docked into receptors with SwissDock(Grosdidier, Zoete & Michielin, 2011a). SwissDock was chosen both for its high effectiveness as well as ease of use by students. For consistency, we performed all docking studies in this paper with SwissDock although we obtained qualitatively similar docking results with AutoDock Vina, the most popular docking software, in our preliminary studies.

The crystal structure of GPER has not yet been determined. We created a homology model using GPCR I-TASSER (Iterative Threading Assembly Refinement), the most accurate homology modeling software customized for GPCRs (Zhang et al., 2015). GPCR I-TASSER modeled the GPER structure using template fragments automatically selected from the closest related GPCR crystal structures (CCR5: PDB 4mbs, sphingosine 1-phosphate: PDB 3v2y, CXCR4: PDB 3odu, delta opioid: PDB 4n6h). The homology model was validated with ERRAT (Colovos & Yeates, 1993). Coordinates for E2, G1, G15, and tamoxifen were downloaded from the ZINC ligand database (Irwin et al., 2012) and submitted to SwissDock (Grosdidier, Zoete &

346 Michielin, 2011a) for docking. SwissDock is a web interface to the EADock DSS (Grosdidier,
347 Zoete & Michielin, 2011b) engine, which performs blind, global (does not require targeting of a
348 particular surface) docking using the physics-based CHARMM22 force field (Brooks et al.,
349 2009). The "FullFitness Score" calculated by SwissDock using clustering and the FACTS
350 implicit solvent model (Haberthür & Caflisch, 2008) was used as the "Energy Score" for our
351 calculations.
352
353 **Combined analysis**
354 SwissDock poses were manually screened for those binding sites located on or near the
355 extracellular side of the protein. Ligand binding surfaces included residues with atoms within 3.5
356 Å from the docked ligand. The average conservation score of the amino acids that were
357 highlighted served as the "Conservation Score" of that specific orientation (Scheme 1). The
358 combined ConDockSite score is defined as the product of the Conservation and Energy Scores.
359 As the Energy Score is a modified free energy function, a highly negative ConDockSite score is
360 associated with a more probable ligand binding site. Binding sites predicted by ConDockSite
361 results were compared with those predicted by CASTp (Dundas et al., 2006), SiteHound
362 (Hernandez, Ghersi & Sanchez, 2009), and Concavity (Capra et al., 2009). For CASTp,
363 SiteHound, and Concavity, ligand binding pockets were defined as residues within 4 Å of the
364 selected probe/cluster.

365 $$\textbf{Combined ConDock Score} = (\text{Conservation Score}) * (\text{Energy Score})$$
366

$$\textbf{Conservation Score} = \frac{1}{10}\frac{\sum_{k=1}^{n}(\text{Amino Acid ConSurf Score})_k}{n}$$

$$\textbf{Energy Score} = \text{SwissDock FullFitness Score}$$

Scheme 1. Calculation of combined ConDockSite scores for ligand binding sites. The Conservation Score is calculated over the *n* residues in a binding site, indexed by *k*.

Crystal structure benchmarks

Crystal structures of receptors were screened for residues within 3.5 Å of their respective ligands. These residues served as a benchmark of comparison for the sites predicted by the ConDockSite scoring function.

Acknowledgments

This work was funded by the Victoria S. and Bradley L. Geist Foundation (H.L.N.), NSF CAREER Award 1833181 (H.L.N.), and the Undergraduate Research Opportunities Program at the University of Hawaii at Manoa (A.R.V., S.M.).

Author contributions

A.R.V., S.M., and H.L.N. performed the analysis and calculations. A.R.V., S.M., and H.L.N. wrote the manuscript. H.L.N. supervised the project.

Competing interests

The authors have no competing financial or non-financial interests.

390 **References**

391 Armon A, Graur D, Ben-Tal N. 2001. ConSurf: an algorithmic tool for the identification of
392 functional regions in proteins by surface mapping of phylogenetic information1. *Journal*
393 *of Molecular Biology* 307:447–463. DOI: 10.1006/jmbi.2000.4474.
394 Arnatt CK, Zhang Y. 2013. G Protein-Coupled Estrogen Receptor (GPER) Agonist Dual
395 Binding Mode Analyses Toward Understanding of Its Activation Mechanism: A
396 Comparative Homology Modeling Approach. *Molecular Informatics* 32:647–658. DOI:
397 10.1002/minf.201200136.
398 Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating
399 evolutionary conservation in sequence and structure of proteins and nucleic acids.
400 *Nucleic Acids Research* 38:W529-533. DOI: 10.1093/nar/gkq399.
401 Boeckmann B, Blatter M-C, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A. 2005.
402 Protein variety and functional diversity: Swiss-Prot annotation in its biological context.
403 *Comptes Rendus Biologies* 328:882–899. DOI: 10.1016/j.crvi.2005.06.001.
404 Bologa CG, Revankar CM, Young SM, Edwards BS, Arterburn JB, Kiselyov AS, Parker MA,
405 Tkachenko SE, Savchuck NP, Sklar LA, Oprea TI, Prossnitz ER. 2006. Virtual and
406 biomolecular screening converge on a selective agonist for GPR30. *Nature Chemical*
407 *Biology* 2:207–212. DOI: 10.1038/nchembio775.
408 Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G,
409 Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J,
410 Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW,
411 Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W,
412 York DM, Karplus M. 2009. CHARMM: the biomolecular simulation program. *Journal*

413 *of Computational Chemistry* 30:1545–1614. DOI: 10.1002/jcc.21287.
414 Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. 2009. Predicting Protein
415 Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D
416 Structure. *PLoS Comput Biol* 5:e1000585. DOI: 10.1371/journal.pcbi.1000585.
417 Capra JA, Singh M. 2007. Predicting functionally important residues from sequence
418 conservation. *Bioinformatics* 23:1875–1882. DOI: 10.1093/bioinformatics/btm270.
419 Carmeci C, Thompson DA, Ring HZ, Francke U, Weigel RJ. 1997. Identification of a Gene
420 (GPR30) with Homology to the G-Protein-Coupled Receptor Superfamily Associated
421 with Estrogen Receptor Expression in Breast Cancer. *Genomics* 45:607–617. DOI:
422 10.1006/geno.1997.4972.
423 Chan HCS, Li Y, Dahoun T, Vogel H, Yuan S. 2019. New Binding Sites, New Opportunities for
424 GPCR Drug Discovery. *Trends in Biochemical Sciences* 44:312–330. DOI:
425 10.1016/j.tibs.2018.11.011.
426 Colovos C, Yeates TO. 1993. Verification of protein structures: patterns of nonbonded atomic
427 interactions. *Protein Science: A Publication of the Protein Society* 2:1511–1519. DOI:
428 10.1002/pro.5560020916.
429 Costanzi S, Cohen A, Danfora A, Dolatmoradi M. 2019. Influence of the Structural Accuracy of
430 Homology Models on Their Applicability to Docking-Based Virtual Screening: The β_2
431 Adrenergic Receptor as a Case Study. *Journal of Chemical Information and Modeling*
432 59:3177–3190. DOI: 10.1021/acs.jcim.9b00380.
433 Dennis MK, Burai R, Ramesh C, Petrie WK, Alcon SN, Nayak TK, Bologa CG, Leitao A,
434 Brailoiu E, Deliu E, Dun NJ, Sklar LA, Hathaway HJ, Arterburn JB, Oprea TI, Prossnitz
435 ER. 2009. In vivo effects of a GPR30 antagonist. *Nature Chemical Biology* 5:421–427.

DOI: 10.1038/nchembio.168.

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. 2006. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research* 34:W116-118. DOI: 10.1093/nar/gkl282.

Filardo EJ, Quinn JA, Frackelton AR, Bland KI. 2002. Estrogen Action Via the G Protein-Coupled Receptor, GPR30: Stimulation of Adenylyl Cyclase and cAMP-Mediated Attenuation of the Epidermal Growth Factor Receptor-to-MAPK Signaling Axis. *Molecular Endocrinology* 16:70–84. DOI: 10.1210/mend.16.1.0758.

Grosdidier A, Zoete V, Michielin O. 2011a. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Research* 39:W270–W277. DOI: 10.1093/nar/gkr366.

Grosdidier A, Zoete V, Michielin O. 2011b. Fast docking using the CHARMM force field with EADock DSS. *Journal of Computational Chemistry* 32:2149–2159. DOI: 10.1002/jcc.21797.

Haberthür U, Caflisch A. 2008. FACTS: Fast analytical continuum treatment of solvation. *Journal of Computational Chemistry* 29:701–715. DOI: 10.1002/jcc.20832.

Hernandez M, Ghersi D, Sanchez R. 2009. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Research* 37:W413–W416. DOI: 10.1093/nar/gkp281.

Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. 2012. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* 52:1757–1768. DOI: 10.1021/ci3001277.

459 Kalinina OV, Gelfand MS, Russell RB. 2009. Combining specificity determining and conserved
460 residues improves functional site prediction. *BMC Bioinformatics* 10:174. DOI:
461 10.1186/1471-2105-10-174.
462 Kanda N, Watanabe S. 2003. 17β-Estradiol Enhances the Production of Nerve Growth Factor in
463 THP-1-Derived Macrophages or Peripheral Blood Monocyte-Derived Macrophages.
464 *Journal of Investigative Dermatology* 121:771–780. DOI: 10.1046/j.1523-
465 1747.2003.12487.x.
466 Katritch V, Abagyan R. 2011. GPCR agonist binding revealed by modeling and crystallography.
467 *Trends in Pharmacological Sciences* 32:637–643. DOI: 10.1016/j.tips.2011.08.001.
468 Katritch V, Rueda M, Lam PC-H, Yeager M, Abagyan R. 2010. GPCR 3D homology models for
469 ligand screening: lessons learned from blind predictions of adenosine A2a receptor
470 complex. *Proteins* 78:197–211. DOI: 10.1002/prot.22507.
471 Kratochwil NA, Gatti-McArthur S, Hoener MC, Lindemann L, Christ AD, Green LG, Guba W,
472 Martin RE, Malherbe P, Porter RHP, Slack JP, Winnig M, Dehmlow H, Grether U,
473 Hertel C, Narquizian R, Panousis CG, Kolczewski S, Steward L. 2011. G protein-coupled
474 receptor transmembrane binding pockets and their applications in GPCR research and
475 drug discovery: a survey. *Current Topics in Medicinal Chemistry* 11:1902–1924.
476 Kvingedal AM, Smeland EB. 1997. A novel putative G-protein-coupled receptor expressed in
477 lung, heart and lymphoid tissue. *FEBS Letters* 407:59–62. DOI: 10.1016/S0014-
478 5793(97)00278-0.
479 Lai JK, Ambia J, Wang Y, Barth P. 2017. Enhancing Structure Prediction and Design of Soluble
480 and Membrane Proteins with Explicit Solvent-Protein Interactions. *Structure* 25:1758-
481 1770.e8. DOI: 10.1016/j.str.2017.09.002.

Lappano R, Rosano C, De Marco P, De Francesco EM, Pezzi V, Maggiolini M. 2010. Estriol acts as a GPR30 antagonist in estrogen receptor-negative breast cancer cells. *Molecular and Cellular Endocrinology* 320:162–170. DOI: 10.1016/j.mce.2010.02.006.

Levit A, Barak D, Behrens M, Meyerhof W, Niv MY. 2012. Homology model-assisted elucidation of binding sites in GPCRs. *Methods in Molecular Biology (Clifton, N.J.)* 914:179–205. DOI: 10.1007/978-1-62703-023-6_11.

Li Y, Hou T, Goddard III W. 2010. Computational Modeling of Structure-Function of G Protein-Coupled Receptors with Applications for Drug Design. *Current Medicinal Chemistry* 17:1167–1180. DOI: 10.2174/092986710790827807.

Lim VJY, Du W, Chen YZ, Fan H. 2018. A benchmarking study on virtual ligand screening against homology models of human GPCRs. *Proteins: Structure, Function, and Bioinformatics* 86:978–989. DOI: 10.1002/prot.25533.

Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O. 2004. Evolutionary Trace of G Protein-coupled Receptors Reveals Clusters of Residues That Determine Global and Class-specific Functions. *Journal of Biological Chemistry* 279:8126–8132. DOI: 10.1074/jbc.M312671200.

Martínez-Muñoz A, Prestegui-Martel B, Méndez-Luna D, Fragoso-Vázquez MJ, García-Sánchez JR, Bello M, Martínez-Archundia M, Chávez-Blanco A, Dueñas-González A, Mendoza-Lujambio I, Trujillo-Ferrara J, Correa-Basurto J. 2018. Selection of a GPER1 Ligand via Ligand-based Virtual Screening Coupled to Molecular Dynamics Simulations and Its Anti-proliferative Effects on Breast Cancer Cells. *Anti-Cancer Agents in Medicinal Chemistry* 18:1629–1638. DOI: 10.2174/1871520618666180510121431.

Méndez-Luna D, Bello M, Correa-Basurto J. 2016. Understanding the molecular basis of

505 agonist/antagonist mechanism of GPER1/GPR30 through structural and energetic
506 analyses. *The Journal of Steroid Biochemistry and Molecular Biology* 158:104–116.
507 DOI: 10.1016/j.jsbmb.2016.01.001.
508 Méndez-Luna D, Martínez-Archundia M, Maroun RC, Ceballos-Reyes G, Fragoso-Vázquez MJ,
509 González-Juárez DE, Correa-Basurto J. 2015. Deciphering the GPER/GPR30-agonist and
510 antagonists interactions using molecular modeling studies, molecular dynamics, and
511 docking simulations. *Journal of Biomolecular Structure and Dynamics* 33:2161–2172.
512 DOI: 10.1080/07391102.2014.994102.
513 Merz KM. 2010. Limits of Free Energy Computation for Protein−Ligand Interactions. *Journal of*
514 *Chemical Theory and Computation* 6:1769–1776. DOI: 10.1021/ct100102q.
515 O'Dowd BF, Nguyen T, Marchese A, Cheng R, Lynch KR, Heng HHQ, Kolakowski Jr. LF,
516 George SR. 1998. Discovery of Three Novel G-Protein-Coupled Receptor Genes.
517 *Genomics* 47:310–313. DOI: 10.1006/geno.1998.5095.
518 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004.
519 UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of*
520 *Computational Chemistry* 25:1605–1612. DOI: 10.1002/jcc.20084.
521 Sanders. 2011. ss-TEA: Entropy based identification of receptor specific ligand binding residues
522 from a multiple sequence alignment of class A GPCRs. *BMC Bioinformatics* 12:332–343.
523 DOI: 10.1186/1471-2105-12-332.
524 Sanders MPA, Verhoeven S, de Graaf C, Roumen L, Vroling B, Nabuurs SB, de Vlieg J, Klomp
525 JPG. 2011. Snooker: A Structure-Based Pharmacophore Generation Tool Applied to
526 Class A GPCRs. *Journal of Chemical Information and Modeling* 51:2277–2292. DOI:
527 10.1021/ci200088d.

528 Shoichet BK, Kobilka BK. 2012. Structure-based drug screening for G-protein-coupled
529 receptors. *Trends in Pharmacological Sciences* 33:268–272. DOI:
530 10.1016/j.tips.2012.03.007.
531 Smith RD, Damm-Ganamet KL, Dunbar JB, Ahmed A, Chinnaswamy K, Delproposto JE,
532 Kubish GM, Tinberg CE, Khare SD, Dou J, Doyle L, Stuckey JA, Baker D, Carlson HA.
533 2016. CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined
534 Computational Protein Design, Docking, and Scoring/Ranking Challenge. *Journal of*
535 *Chemical Information and Modeling* 56:1022–1031. DOI: 10.1021/acs.jcim.5b00387.
536 Souza SA, Kurohara DT, Dabalos CL, Ng HL. 2019. G Protein–Coupled Estrogen Receptor
537 Production Using an Escherichia coli Cell-Free Expression System. *Current Protocols in*
538 *Protein Science* 97:e88. DOI: 10.1002/cpps.88.
539 Tang H, Wang XS, Hsieh J-H, Tropsha A. 2012. Do crystal structures obviate the need for
540 theoretical models of GPCRs for structure-based virtual screening? *Proteins: Structure,*
541 *Function, and Bioinformatics* 80:1503–1521. DOI: 10.1002/prot.24035.
542 Wacker D, Stevens RC, Roth BL. 2017. How Ligands Illuminate GPCR Molecular
543 Pharmacology. *Cell* 170:414–427. DOI: 10.1016/j.cell.2017.07.009.
544 Wan S, Knapp B, Wright DW, Deane CM, Coveney PV. 2015. Rapid, Precise, and Reproducible
545 Prediction of Peptide–MHC Binding Affinities from Molecular Dynamics That Correlate
546 Well with Experiment. *Journal of Chemical Theory and Computation* 11:3346–3356.
547 DOI: 10.1021/acs.jctc.5b00179.
548 Wass MN, Sternberg MJE. 2009. Prediction of ligand binding sites using homologous structures
549 and conservation at CASP8. *Proteins: Structure, Function, and Bioinformatics* 77:147–
550 151. DOI: 10.1002/prot.22513.

551 Weiss DR, Bortolato A, Tehan B, Mason JS. 2016. GPCR-Bench: A Benchmarking Set and
552 Practitioners' Guide for G Protein-Coupled Receptor Docking. *Journal of Chemical*
553 *Information and Modeling* 56:642–651. DOI: 10.1021/acs.jcim.5b00660.
554 Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure
555 and function prediction. *Nature Methods* 12:7–8. DOI: 10.1038/nmeth.3213.
556 Zhang J, Yang J, Jang R, Zhang Y. 2015. GPCR-I-TASSER: A Hybrid Approach to G Protein-
557 Coupled Receptor Structure Modeling and the Application to the Human Genome.
558 *Structure* 23:1538–1549. DOI: 10.1016/j.str.2015.06.007.
559 Zou, Ewalt, Ng. 2019. Recent Insights from Molecular Dynamics Simulations for G Protein-
560 Coupled Receptor Drug Discovery. *International Journal of Molecular Sciences* 20:4237.
561 DOI: 10.3390/ijms20174237.
562

Figure 1

Predicted and experimental ligand binding sites in A2A adenosine and β2 adrenergic receptors.

Superposition of crystal structure with ligand bound (red) with ConDockSite predicted pose (blue). A) Adenosine with A2A receptor. B) ZM241385 with A2A receptor. C) Epinephrine with β2 adrenergic receptor. D) Carazolol with β2 adrenergic receptor.

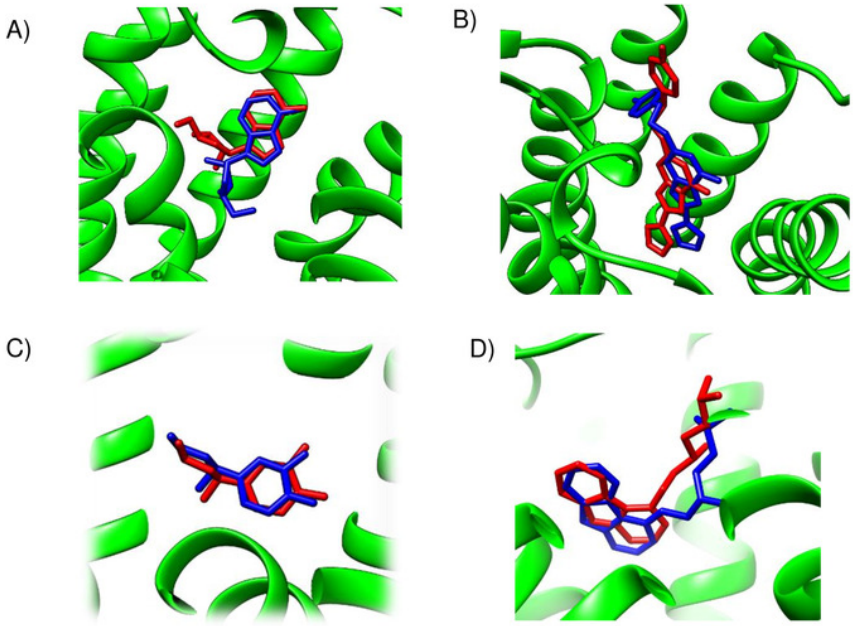

Figure 1. Predicted and experimental ligand binding sites in A2A adenosine and β2 adrenergic receptors. Superposition of crystal structure with ligand bound (red) with ConDockSite predicted pose (blue). A) Adenosine with A2A receptor. B) ZM241385 with A2A receptor. C) Epinephrine with β2 adrenergic receptor. D) Carazolol with β2 adrenergic receptor.

Figure 2

Predicted and experimental ligand binding sites for homology models of four GPCRs.

Superposition of crystal structure with ligand bound (red) with ConDockSite predicted pose (blue). A) ZM241385 with A2A adenosine receptor. B) Carazolol with β2 adrenergic receptor. C) BU72 with mu opioid receptor. D) Methysergide with 5HT2B serotonin receptor.

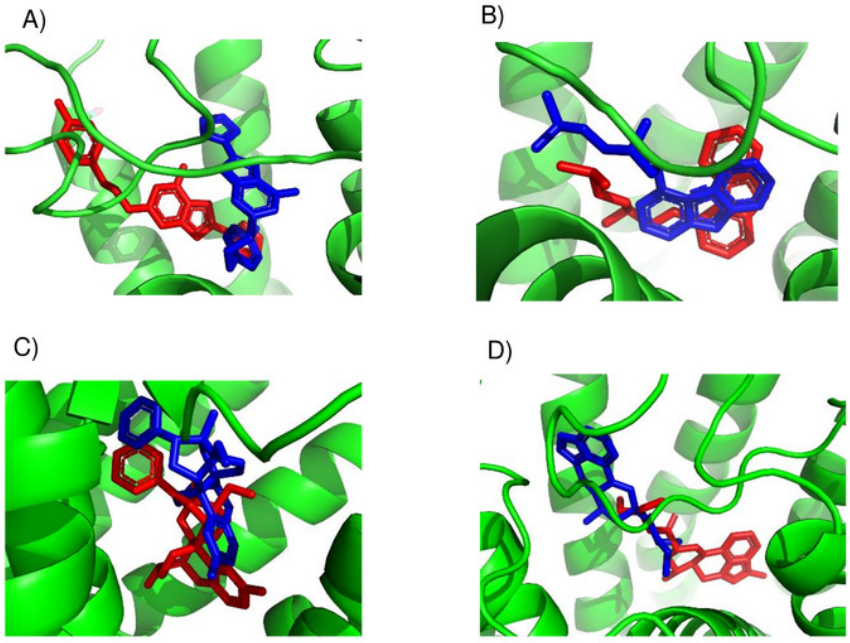

Figure 2. Predicted and experimental ligand binding sites for homology models of four GPCRs. Superposition of crystal structure with ligand bound (red) with ConDockSite predicted pose (blue). A) ZM241385 with A2A adenosine receptor. B) Carazolol with β2 adrenergic receptor. C) BU72 with mu opioid receptor. D) Methysergide with 5HT2B serotonin receptor.

Figure 3

E2 binding sites calculated by SwissDock.

E2 poses are in blue. The top of the figure corresponds to the extracellular face of GPER.

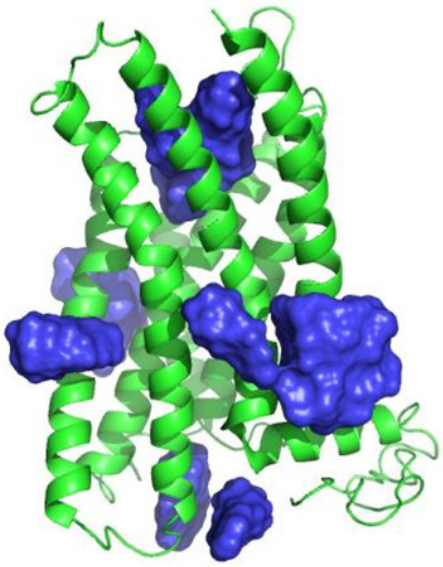

Figure 3. E2 binding sites calculated by SwissDock. E2 poses are in blue. The top of the figure corresponds to the extracellular face of GPER.

Figure 4

Predicted E2 binding sites in GPER

A) The two highest scoring docking poses for E2. B) Receptor-ligand interactions for E2 pose 1. C) Receptor-ligand interactions for E2 pose 2.

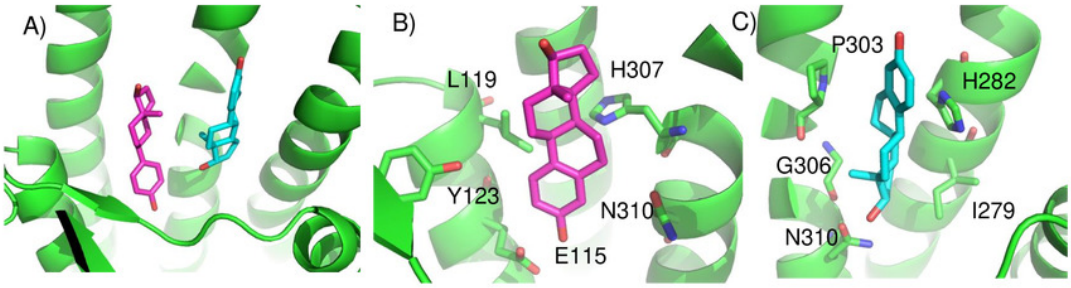

Figure 4. Predicted E2 binding sites in GPER. A) The two highest scoring docking poses for E2. B) Receptor-ligand interactions for E2 pose 1. C) Receptor-ligand interactions for E2 pose 2.

Figure 5

Predicted G1 and G15 binding sites in GPER

A) The highest scoring docking poses for G1 (maroon) and G15 (cyan). B) Receptor-ligand interactions for G1. C) Receptor-ligand interactions for G15.

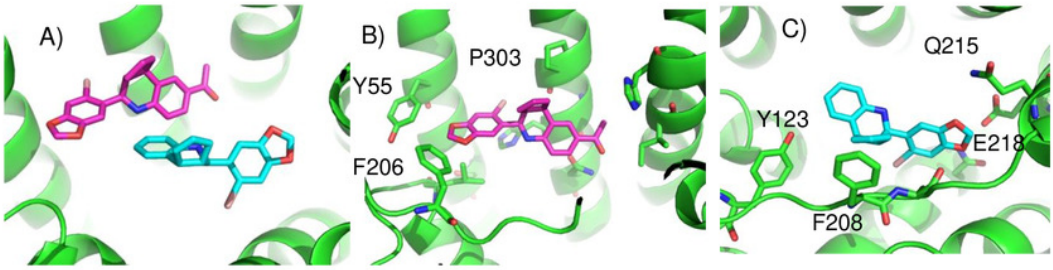

Figure 5. Predicted G1 and G15 binding sites in GPER. A) The highest scoring docking poses for G1 (maroon) and G15 (cyan). B) Receptor-ligand interactions for G1. C) Receptor-ligand interactions for G15.

Figure 6

Predicted tamoxifen binding sites in GPER

A) The highest scoring docking poses for tamoxifen, pose 1 (maroon) and pose 2 (cyan).

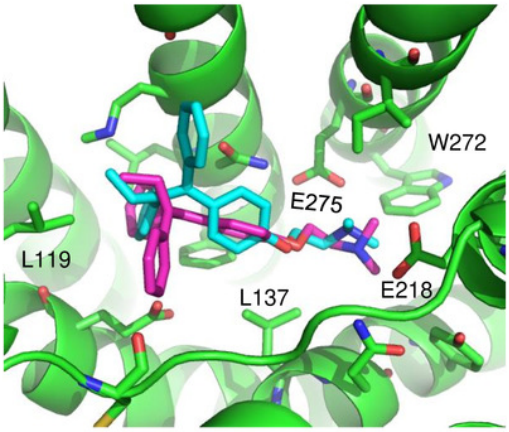

Figure 6. Predicted tamoxifen binding sites in GPER. A) The highest scoring docking poses for tamoxifen, pose 1 (maroon) and pose 2 (cyan).

Figure 7

Predicted E2 binding sites by ConDockSite, CASTp, SiteHound, Concavity.

Ligand binding sites are colored, predicted by A) ConDockSite, B) CASTp, C) SiteHound, D) Concavity.

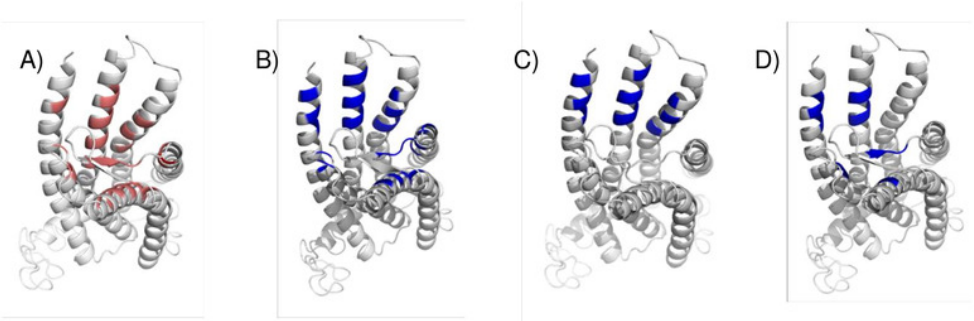

Figure 7. Predicted E2 binding sites by ConDockSite, CASTp, SiteHound, Concavity. Ligand binding sites are colored, predicted by A) ConDockSite, B) CASTp, C) SiteHound, D) Concavity.