`

Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

Gurdeep Singh, M. Sc.

born in: Delhi, India

Oral examination: 26.07.2021

`

A machine learning-guided framework to unravel determinants of G-protein selectivity in GPCRs

Referees:

Prof. Dr. Robert B. Russell

Prof. Dr. Peer Bork

`

`

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Singh, G., Inoue, A., Gutkind, J. S., Russell, R. B., & Raimondi, F. (2019). PRECOG: PREdicting COupling probabilities of G-protein coupled receptors. *Nucleic acids research*, *47*(W1), W395–W401. https://doi.org/10.1093/nar/gkz39

   I participated in the conceptualization of the project, data analysis, developed the machine-learning models and the web-server, and participated in the writing of the manuscript (Chapter III and Chapter IV).

2. Inoue, A., Raimondi, F., Kadji, F., Singh, G., Kishi, T., Uwamizu, A., Ono, Y., Shinjo, Y., Ishida, S., Arang, N., Kawakami, K., Gutkind, J. S., Aoki, J., & Russell, R. B. (2019). Illuminating G-Protein-Coupling Selectivity of GPCRs. *Cell*, *177*(7), 1933–1947.e25. https://doi.org/10.1016/j.cell.2019.04.044

   I participated in the data analysis, developed the machine-learning models, and contributed to the writing of the manuscript (Chapter II & Chapter IV).

3. Raimondi, F., Inoue, A., Kadji, F., Shuai, N., Gonzalez, J. C., Singh, G., de la Vega, A. A., Sotillo, R., Fischer, B., Aoki, J., Gutkind, J. S., & Russell, R. B. (2019). Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm. *Oncogene*, *38*(38), 6491–6506. https://doi.org/10.1038/s41388-019-0895-2

   I participated in the data analysis.

The abovementioned published articles (1, 2) were based on text/figures in Chapters II, III, and IV, which were originally written/drawn by me unless specified otherwise.

`

ABSTRACT

Genetic variations can have positive, negative, or neutral impacts on protein interactions, thus making it essential to understand them to obtain a mechanistic picture of biological functions and diseases. In this thesis, I studied how genetic changes affect the functions of the largest, most diverse family of cell-surface molecules involved in signal transduction: G-protein coupled receptors (GPCRs). GPCRs comprise over 2% of genes in the human genome and are the leading pharmaceutical drug target.

Analysis of one of the most comprehensive datasets quantifying GPCR/G-protein binding affinities revealed that GPCR couplings and sequence similarity are uncorrelated and that there were no clear, simple sequence changes responsible for determining which G-protein binds to a particular GPCR. While GPCRs within the same group can couple to different G-proteins, GPCRs of different groups can still couple to the same G-proteins. We used this new dataset and various protein bioinformatics tools to identify broad sequence features that are associated with specific G-protein binding events. Several of these were at or near the known GPCR/G-protein interface, but many others were not, suggesting a complex relationship between sequence and specificity.

We then applied an interpretable machine learning algorithm on the sequence- and structure-based GPCR features to develop a system and associated webserver (PRECOG) to predict and visualize GPCR/G-protein interactions. We leveraged the machine learning-guided framework to predict uncharacterized GPCRs and successfully developed the first *GNA12*-coupled designer receptor. Application of this framework to recently available binding data revealed the determinants of β-arrestin specificity in GPCRs.

Collectively, this machine learning-guided framework can be extended to other binding data to uncover sites and sequence regions that are physically or allosterically involved in determining subtype specificity. This will not only improve our understanding of protein interactions but also help us devise better chemogenetic tools and take smarter therapeutic decisions in the context of human health.

`

ZUSAMMENFASSUNG

Genetische Varianten können positive, negative oder neutrale Auswirkungen auf Proteininteraktionen haben. Daher ist es wichtig, sie zu verstehen, um ein mechanistisches Bild von biologischen Funktionen und Krankheiten zu erhalten. In dieser Arbeit habe ich untersucht, wie genetische Veränderungen die Funktionen der größten und vielfältigsten Familie von Zelloberflächenmolekülen beeinflussen, die an der Signaltransduktion beteiligt sind: G-Protein-gekoppelte Rezeptoren (GPCRs). GPCRs umfassen über 2 % der Gene im menschlichen Genom und sind das führende pharmazeutische Wirkstoffziel.

Die Analyse eines der umfangreichsten Datensätze zur Quantifizierung von GPCR/G-Protein-Bindungsaffinitäten ergab, dass GPCR-Kopplungen und Sequenzähnlichkeit unkorreliert sind und dass es keine klaren, einfachen Sequenzänderungen gibt, die dafür verantwortlich sind, welches G-Protein an einen bestimmten GPCR bindet. Während GPCRs innerhalb der gleichen Gruppe an unterschiedliche G-Proteine koppeln, können GPCRs verschiedener Gruppen immer noch an die gleichen G-Proteine koppeln. Wir haben diesen neuen Datensatz und verschiedene Protein-Bioinformatik-Werkzeuge verwendet, um breite Sequenzmerkmale zu identifizieren, die mit spezifischen G-Protein-Bindungsereignissen verbunden sind. Einige davon befanden sich an oder in der Nähe der bekannten GPCR/G-Protein-Schnittstelle, viele andere jedoch nicht, was auf eine komplexe Beziehung zwischen Sequenz und Spezifität schließen lässt.

Anschließend wendeten wir einen interpretierbaren maschinellen Lernalgorithmus auf die sequenz- und strukturbasierten GPCR-Merkmale an, um ein System und einen zugehörigen Webserver (PRECOG) zur Vorhersage sowie zur Visualisierung von GPCR/G-Protein-Interaktionen zu entwickeln. Wir nutzten das auf maschinellem Lernen basierende Framework, um uncharakterisierte GPCRs vorherzusagen und entwickelten erfolgreich den ersten *GNA12*-gekoppelten Designer-Rezeptor. Die Anwendung dieses Frameworks auf kürzlich verfügbare Bindungsdaten offenbarte die Determinanten der β-Arrestin-Spezifität in GPCRs.

`

Insgesamt kann dieses auf maschinellem Lernen basierende Framework auf andere Bindungsdaten erweitert werden, um Stellen und Sequenzregionen aufzudecken, die physikalisch oder allosterisch an der Bestimmung der Subtyp-Spezifität beteiligt sind. Dies wird nicht nur unser Verständnis von Proteininteraktionen verbessern, sondern uns auch helfen, bessere chemogenetische Werkzeuge zu entwickeln und intelligentere therapeutische Entscheidungen im Kontext der menschlichen Gesundheit zu treffen.

`

ACKNOWLEDGEMENTS

When I first came to Heidelberg for my bachelor thesis, I never thought I would do a Ph.D. There are innumerable people to thank who convinced me to go for it. I am extremely grateful to Rob for his guidance and to allow me to work in his group. His supervision gave shape, form, and clarity to my project and thesis. The best thing about Rob is that he let me do what I wished to do, including the things that were outside my Ph.D. project's gambit. Every time I step out of his office, not only I have expanded my scientific knowledge but also learned how to ideate and present my project. I also express my earnest thanks to Prof. Dr. Peer Bork who despite his busy schedule agreed to be my second supervisor and sat unabatedly throughout all my TAC meetings.

My sincere, and heartfelt thanks go to Francesco who is not just a brilliant scientist but also an excellent mentor and friend. Right from my early days in Heidelberg until today, he has provided constant, invaluable support, at and outside work, answering my silly doubts with patience and a smile. He has taught me how to think out of the box, how to interrogate data, and how to present results. My sincere appreciation to Dr. Asuka Inoue and his group (Tohoku University, Japan) who shared his valuable dataset with us and performed validation experiments leading to the development of a tool associated with this thesis.

I am fortunate to have met wonderful people in this city, who became great friends. Juan-Carlos and Torsten helped a lot with my projects and took me to PubQuizzes, game nights, and all the fun-filled activities outside work. Gaurav and Magi read the early versions of the thesis and supplied immensely valuable feedback. Yvonne diligently took care of all my admin-related stuff. I am thankful to all the former and current members of Rob's group who shared their time, conversations, and experiences during lunchtimes, making this segment of my life memorable.

A friend, teacher, and brother, Manjeet has been everything to me. Without him, this journey wouldn't have been possible. Mom, Dad, Prince, and Manisha: I shall always remain indebted to you for your constant encouragement; moral support, and boundless motivation that has helped me reach this point in my life. I can't thank you all enough for how supportive you have been during the journey.

`

`

TABLE OF CONTENTS

`

LIST OF FIGURES

`

`

LIST OF TABLES

`

ABBREVIATIONS

ATP	Adenosine Triphosphate
AUC	Area Under Curve
BLAST	Basic Local Alignment Search Tool
DREADD	Designer Receptors Exclusively Activated by Designer Drugs
FPR	False Positive Rate
GPCR	G-Protein Coupled Receptor
GtoPdb	The IUPHAR/BPS Guide to PHARMACOLOGY Database
HMM	Hidden Markov Model
ML	Machine Learning
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank
PFAM	Database of Protein families
ROC	Receiver Operating Characteristic
SDP	Specificity Determining Positions
SIFTS	Structure Integration with Function, Taxonomy, and Sequence
TPR	True Positive Rate
UniProt	Universal Protein Resource

Proteins are referred to in the text either by their common names or by their HUGO standard (Braschi et al., 2019) gene symbols.

`

`

Chapter I: Introduction

Protein interactions are crucial to biological processes. Proteins interact with each other in various conformation states to build molecular machines that steer a myriad of cellular functions including membrane transport, signal transduction, metabolism, intracellular communication, cell growth, and apoptosis. Protein sequence largely determines function and structure, and changes to this sequence can have a variety of impacts. Certain variations, seen when comparing homologous proteins, are known to determine functional specificity (Casari et al., 1995). Other variations can have positive, negative, or neutral impacts on proteins and the systems they are involved in. For instance, the binding of HIV (human immunodeficiency virus) to *CCR5* (chemokine receptor 5) triggers Acquired Immunodeficiency Syndrome (AIDS), a notorious disease that can impair the immune system of a healthy individual. However, certain *CCR5* mutations can decrease infectivity by altering its structure or level of expression (Blanpain et al., 1999). Other variations, like in that of *TP53* (tumor suppressor 53), emphatically contribute to the onset of life-threatening diseases like cancer (Ribeiro et al., 2001; Stacey et al., 2011). Some variations have almost no impact on an individual like a neutral mutation (changes amino acid sequence but almost no change in its function) or a silent mutation (does not even alter the amino acid sequence). Still elsewhere, artificially introduced variations are used to design proteins for biomedical applications (Bedbrook et al., 2019; Roth, 2016).

To explore such puzzling, and at the same time, fascinating variations in protein families is not only important to unravel their functional specificities but also to understand implications in the context of cellular pathways and human health. Here, we explore the largest family of cell-surface molecules responsible for signal transduction: G-protein coupled receptors (GPCRs). Even though GPCRs share a common structure of seven helices, they display variable specificities to their primary transducers: G-proteins. In this thesis, with the help of the most extensive dataset of GPCR/G-protein binding affinities, we have developed a machine learning-guided framework to extract the residues within GPCRs that determine their functional specificities.

`

1.1 G-PROTEIN COUPLED RECEPTORS

GPCRs share a common architecture: a single polypeptide of seven helices embedded in the cell membrane. Hence, they are also referred to as seven-transmembrane receptors. GPCRs perform a battery of functions in our body, including cell proliferation, survival, and motility. Specific external signals (like biogenic amines, light energy, lipids, sugars, proteins, and peptides) induce conformational changes in GPCRs that enable interactions, primarily with heterotrimeric G-proteins, G-protein -coupled receptor kinases (GRKs), and arrestins.

A combination of sequence and functional similarities has led to the current classification of 800 Human GPCRs into 6 classes (A-F) in GtoPdb (the IUPHAR/BPS Guide to PHARMACOLOGY) (Harding et al., 2018): a) Class A (rhodopsin-like), b) Class B (secretin receptor family), c) Class C (metabotropic glutamate), d) Class D (fungal mating pheromone receptors), e) Class E (cyclic AMP receptors), and f) Class F (frizzled/smoothened). Vertebrates lack classes D and E. With its 719 members in human accounting for almost 89% of the GPCRs, Class A is the largest subfamily. More than half of class A GPCRs encode olfactory receptors (390 out of 719 or 54%), while those remaining are either known to be activated by endogenous compounds or are classified as orphan receptors. Interestingly, even though there is little detectable sequence similarity between GPCR classes, they all share the common seven-helical architecture and transduce the external signal through a similar mechanism.

1.2 STRUCTURE AND TOPOLOGY OF A GPCR

GPCRs are integral membrane proteins. They vary considerably in sequence, but all share key common features. These are an extracellular N-terminus, seven transmembrane helices (TM) woven in and around the membrane and comprising alternating intracellular and extracellular loops (ICLs and ECLs), and a cytosolic C-terminus. Several residues of the hydrophobic TM helices are inter-connected and form functionally important motifs such as the E/DRY in TM3, WxP in TM6, and NPxxY (Palczewski et al., 2000; Rosenbaum et al., 2009) in TM7 (TMn correspond to the n^{th} transmembrane helix). Among all the helices, the TM3 helix forms the most contacts, wherein every residue is connected to either another TM helix, ligand, or G-protein/arrestin (Venkatakrishnan et al., 2013).

`

While class A GPCRs contain only a single, globular, transmembrane domain, other classes have additional extracellular domains that typically participate in ligand binding, conjointly with the TM helices (Dann et al., 2001; Grace et al., 2004; Liu et al., 2004; Rondard et al., 2006). Compared to class A, the N-terminus of classes B, C, and F are larger, have characteristic domains, and often contain disulfide bonds (Nørskov-Lauritsen et al., 2015; Perlman et al., 1995). In contrast to the extracellular partners of GPCRs, which range over thousands of ligands, intracellular partners are limited to a few key effectors: G-proteins, GRKs, and arrestins. As a result, the intracellular (IC) transmembrane region of GPCRs is structurally much more conserved compared to its solvent-exposed extracellular (EC) transmembrane region (Venkatakrishnan et al., 2014).

In GPCRs, the extracellular loops (ECLs) play a significant role in regulating the binding of ligands. While ECL1 and ECL3 are usually short and unstructured, ECL2 is longer, adopts several conformations in different GPCRs, and is often anchored to the extracellular end of the TM3 helix via disulfide bridges (Wheatley et al., 2012). The three intracellular loops (ICLs), on the other hand, mediate interactions with the cytosolic partners of GPCRs. The ICL3 shows the greatest variation in length, ranging from 5 residues in CXCR4 *(*C-X-C chemokine receptor type 4) up to over 150 residues in muscarinic acetylcholine receptors (mAChRs). The N-terminus, ICL3, and C-terminus have been predicted to lack a pre-defined structure (i.e. Intrinsically Disordered Regions or IDRs) (Venkatakrishnan et al., 2014). However, upon interaction with an effector protein, these regions likely undergo a disorder-to-order transition. For example, the fully phosphorylated C-terminus (an otherwise IDR) of *AVPR2* (V2 vasopressin receptor) transitions to β-sheet to activate β-arrestin-1 (Shukla et al., 2013).

The known features of GPCRs discussed above suggest the need for an integrated approach that captures structured and disordered regions of GPCRs to determine receptor specificity towards its cytosolic partners.

`

1.3 GPCR RESIDUE NUMBERING

In this thesis, I have used the two most common GPCR residue numbering systems: Pfam 7TM (El-Gebali et al., 2019) and Ballesteros-Weinstein (BW) (Ballesteros and Weinstein, 1995). According to the Pfam 7TM numbering system, a residue in a receptor is assigned the equivalent position in the consensus of the Pfam 7tm_1 domain. While the first character indicates the prevalent amino acid, the rest of the characters indicate the position in the consensus sequence of the domain. For example, Y268 is the name assigned to positions 678 and 288 in *TSHR* and *GPR55*, respectively. In the Ballesteros-Weinstein (BW) numbering system, residues are assigned two numbers separated by a period. The number before the period indicates the TM helix (between 1 and 7) under consideration and the number after the period indicates the position relative (upstream or downstream) to the most conserved residue (defined as 50) in the TM helix. For example, the position Y268 mentioned above corresponds to 7.53 in the BW numbering system. Here 7 refers to the seventh TM helix and 53 indicates that the residue is three positions after the most conserved residue (5.50).

1.4 CYTOSOLIC PARTNERS OF A GPCR

The three major cytosolic partners of GPCRs are heterotrimeric G-proteins, GRKs, and arrestins. The heterotrimeric G-proteins (also referred to as large G-proteins) are membrane-associated molecular switches that comprise alpha (Gα), beta (Gβ), and gamma (Gγ) subunits. Unlike monomeric G-proteins (also referred to as small G-proteins), heterotrimeric G-proteins bind directly to GPCRs. The Gα subunit moderates its switch function by binding to guanine triphosphate (GTP) (switched on) or guanine diphosphate (GDP) (off). Ligand binding induces a conformational change in GPCRs, which then act as a GEF (guanine nucleotide exchange factor) to substitute GDP by GTP in the Gα subunit. This leads to the separation of the Gα subunit and the Gβγ subunits, activating several downstream signaling pathways (Figure 1.1).

Based on sequence and functional similarity, the 16 Gα subunits can be grouped into four subfamilies: Gs, Gi/Go, G12/G13, and Gq/G11 (Figure 1.1). The Gα structure consists of two domains: a) a GTP-binding domain (G-domain), and b) a helical domain that buries the GTP within the core of the protein. The helix 5 of the G-domain

`

has been reported to play a key role in determining GPCR/G-protein coupling (Conklin et al., 1993; Inoue et al., 2012). In this study, we have used subfamily symbols (Gs, Gi/Go, G12/G13, and Gq/G11) to indicate the G-protein subfamily, and gene symbols to indicate particular G-proteins.

Two Gα subunits are responsible for regulating cyclic AMP (cAMP): $G\alpha_s$ and $G\alpha_{i/o}$. The Gs pathway is involved in the stimulation of adenylate cyclase, an enzyme that converts ATP to cAMP, which is a classical second messenger molecule. cAMP is primarily involved in the activation of protein kinase A (PKA) and cyclic-nucleotide-gated ion channels. cAMP binds to inactive PKA, dissociating it into regulatory (cAMP-bound) and activated catalytic subunits. The active catalytic subunits enter the nucleus to phosphorylate substrates that convert glycogen to glucose (Huang and Krebs, 1977; Proud et al., 1977) and stimulate transcription (Qian et al., 2017; Tremblay and Viger, 2003). In near direct contrast to Gs, the Gi/Go pathway is generally responsible for the *inhibition* of adenylate cyclase, thus suppression of the cAMP signaling.

The Gq/G11 pathway stimulates phospholipase C-β (PLC-β), which cleaves membrane-bound phosphatidylinositol 4,5-biphosphate (PIP2) into diacylglycerol (DAG) and inositol (1,4,5) triphosphate (IP3). While DAG is bound to the plasma membrane, IP3 is released into the cytosol and acts as the second messenger molecule, and binds to IP3-gated calcium channels (also called IP3 receptors) in the endoplasmic reticulum to release sequestered Ca^{+2}, which along with DAG activates protein kinase C (PKC). Besides, Ca^{+2} allosterically activates calmodulins, which are responsible for the activation of several enzymes involved in metabolism (Nishizawa et al., 1988), memory (Lledo et al., 1995), and smooth muscle contraction (Tansey et al., 1994).

The G12/G13 pathway primarily activates RhoA (Ras homolog family member A) (Hiley et al., 2006; Katoh et al., 1998), which is a regulatory factor in several cellular functions, most often related to the regulation of the cytoskeleton. These include the immune response (Girkontaite et al., 2001), angiogenesis (Offermanns et al., 1997), embryonic development (Ruppel et al., 2005), apoptosis (Althoefer et al., 1997), and platelet activation (Moers et al., 2003).

Following G-protein activation, GRKs phosphorylate the intracellular regions of liganded GPCRs and recruit arrestins, which occlude G-protein binding sites and lead

`

to receptor desensitization. Among the seven GRKs (Figure 1.1), four (GRKs 1,7,2,3) have been so far reported to phosphorylate only activated GPCRs (Gurevich and Gurevich, 2019) while the remainder were shown also to phosphorylate inactive GPCRs (Baameur et al., 2010; Li et al., 2015; Rankin et al., 2006; Tran et al., 2004). GRKs are also known to phosphorylate non-GPCR substrates such as tyrosine kinase receptors (Wu et al., 2006; Zheng et al., 2012) and regulate Gq-signaling independent of its kinase activity (Tesmer et al., 2005; Usui et al., 2005).

Humans express four arrestin subtypes (1-4) (Figure 1.1), all of which share a similar structure consisting of N- and C- terminal domains that both adopt all-beta, immunoglobulin-like structure, but perform different sub-functions. While arrestins 1 and 4 are found in photoreceptor cells, the non-visual subtypes 2 and 3 (also called β-arrestin-1 and β-arrestin-2, respectively) are ubiquitous. Besides turning off G-protein-dependent signaling, arrestins are additionally involved in G-protein-independent pathways such as β-arrestin-mediated signaling. GPCRs can activate arrestins via three modes i.e. a) the receptor C-terminus (the *tail* conformation) (Latorraca et al., 2018), b) the transmembrane core (*core* conformation) (Eichel et al., 2018; Latorraca et al., 2018); or, c) fully-engaged (Huang et al., 2020; Lee et al., 2020; Staus et al., 2020; Yin et al., 2019). While the *tail* conformation can effectively mediate receptor endocytosis and ERK signaling, and simultaneously engage with G-proteins (Nguyen et al., 2019), the *core* conformation promotes the sustained activation of arrestins following dissociation from the receptor and transduce ERK signaling from clathrin-coated endocytic structures (CCS) (Eichel et al., 2018). However, to hamper G-protein signaling, a fully engaged GPCR/arrestin complex is required (Cahill et al., 2017; Kumari et al., 2017).

`

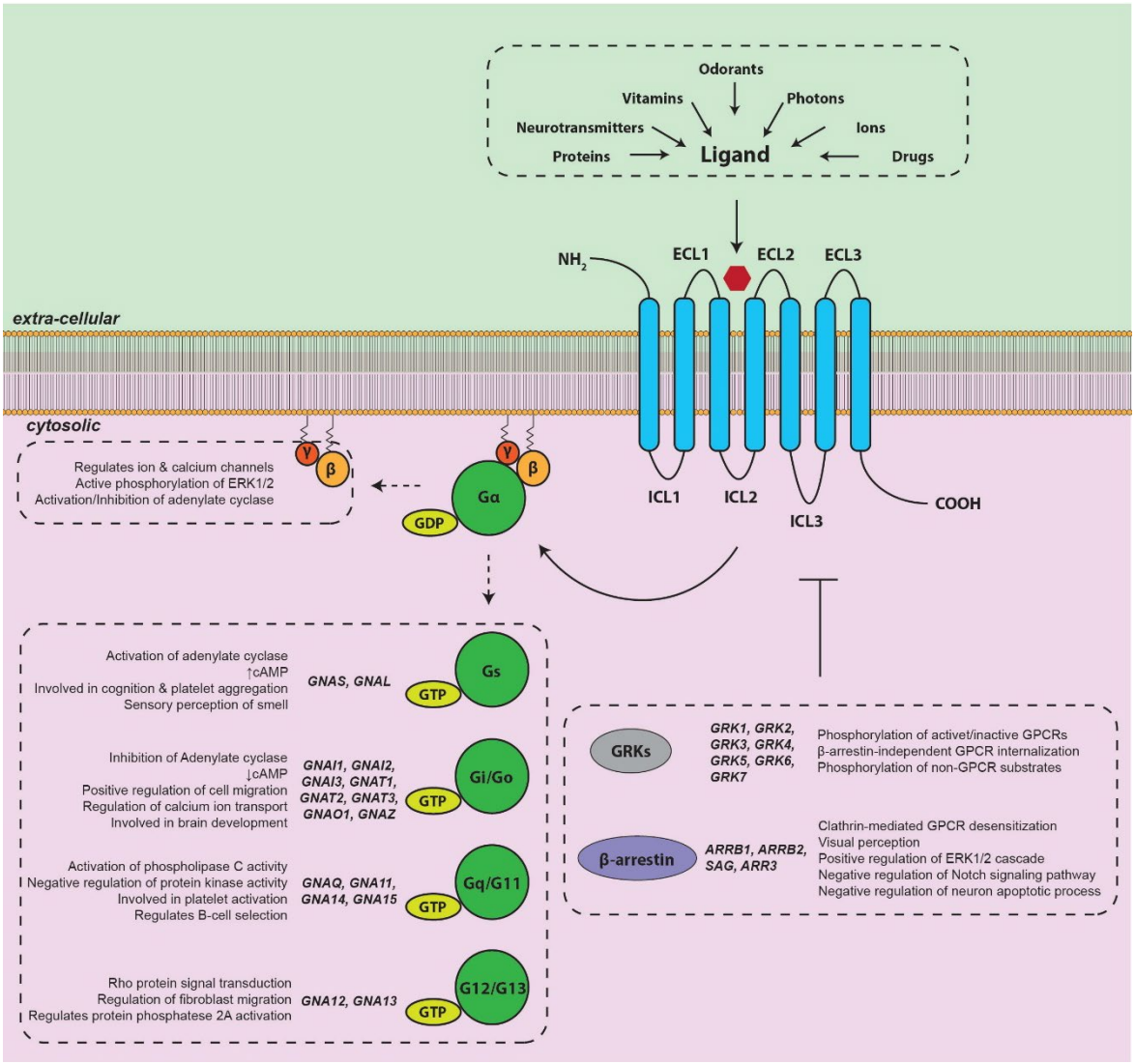


Figure 1.1: Schema of GPCR signaling. GPCRs and their cytosolic interactors.

1.5 GPCR ACTIVATION

Upon activation by an external stimulus, GPCRs undergo conformational changes that involve twisting, extension, and displacement of transmembrane helices and rearrangement of the interconnected residues (Latorraca et al., 2017) (Figure 1.2).

Previous studies (Manglik et al., 2015; Nygaard et al., 2013; Yohannan et al., 2004) have shown that a GPCR can assume several conformations, implicating various downstream signaling mechanisms. The first GPCR/G-protein crystal structure was of activated monomeric β2AR (β2-adrenergic receptor) bound to a heterotrimeric $G\alpha_s$ (*GNAS*) (Rasmussen et al., 2011). The largest conformational change observed, when

`

comparing inactive and active β2AR structures is on the intracellular side, where G-proteins interact via TM helices 5 and 6. The cytosolic end of TM6 rotates counter-clockwise (when viewed from the extracellular side) and moves outward by nearly 14Å accompanied by a smaller outward movement and an extension of the cytoplasmic side of TM5 by 7 residues and inward movement of TM7.

The GPCR/G-protein 3D complexes solved in subsequent studies revealed the motions of several residues within the transmembrane helices that transfer “cues” from the extracellular ligand-binding pocket to the cytosolic G-protein binding interfaces of GPCRs. For example, the conserved triad core Ile (BW: 3.40), Pro (BW: 5.50), and Phe (BW: 6.44) is reported to undergo structural rearrangement (moving away of Ile (3.40), separating Pro (5.50), and Phe (6.44)) upon activation of β2AR (Huang et al., 2015). In addition, the outward displacement of TM6 creates a crevice on the receptor’s intracellular side, which engages with G-proteins. At the G-protein end, there is a rotation as well inward displacement of α5-helix of *GNAS* by 6 Å into the transmembrane core of the receptor.

`

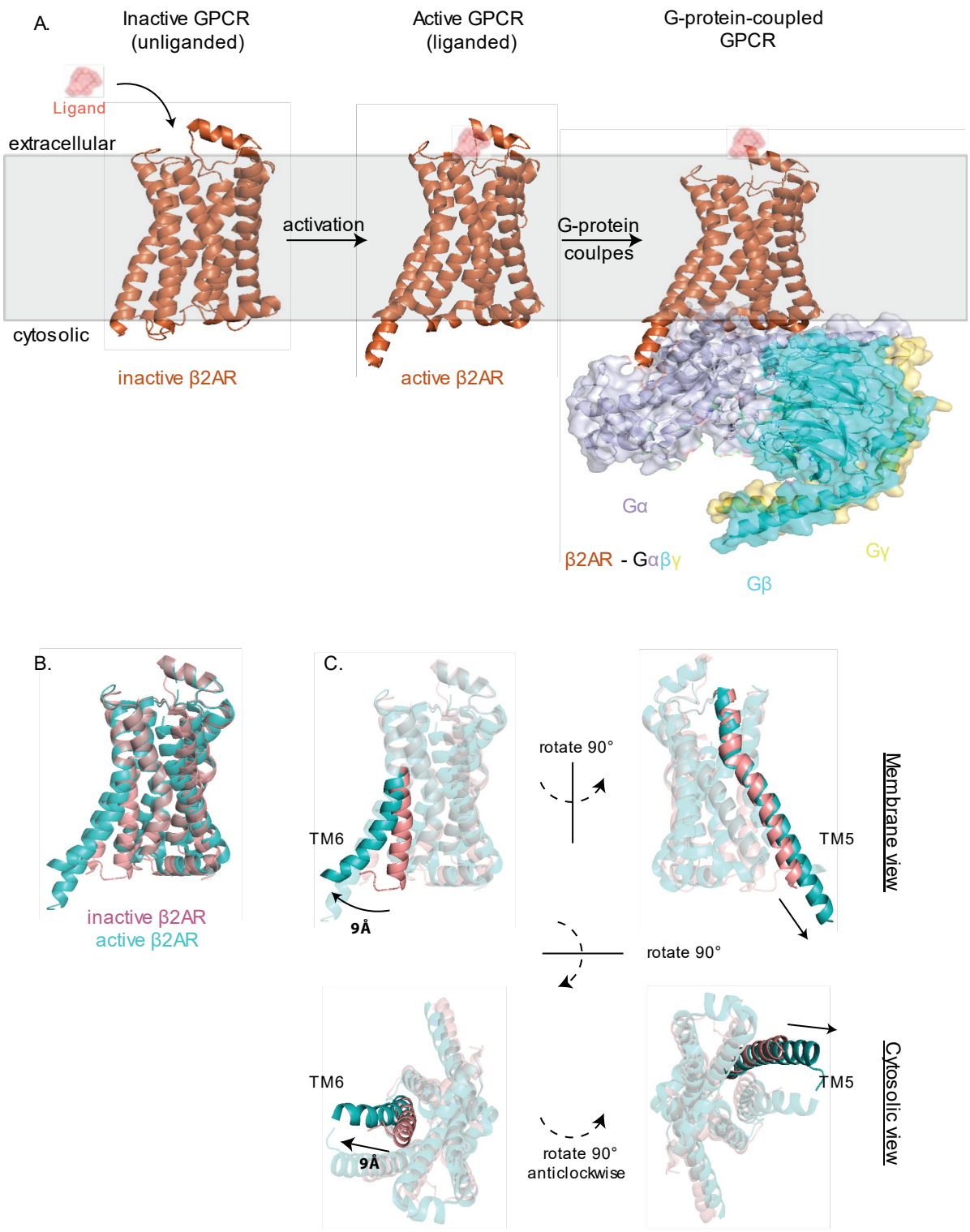


Figure 1.2: GPCR activation. A) Ligand binding induces a conformational change in GPCR, leading to the recruitment of G-protein. B) Superimposition of active β2AR (PDB ID: 3SN6) onto inactive β2AR (PDB ID: 2RH1). C) TM5 and TM6 show the largest conformational change.

A similar mechanism of activation is observed between inactive and *GNAI1*-bound μOR (μ-opioid receptor) (Koehl et al., 2018). Although most TM helices align well

`

between *GNAS*-bound β2AR (orange) and *GNAI1*-bound μOR (green), the outward displacement of the TM6 in β2AR more pronounced (by 9 Å) than that of the μOR (Figure 1.3), a trend observed in other Gs-bound vs. Gi-bound receptors (García-Nafría et al., 2018a; Kang et al., 2018). Besides, the α5-helix of *GNAI1*-bound to μOR (green) is rotated ~21° (displacement of 5 Å; towards TM7) relative to the α5-helix of *GNAS* bound to β2AR (Koehl et al., 2018). The C-terminal residues of the α5-helix are bulkier in *GNAS* than in *GNAI1*, forming a larger crevice for the Gs-specific receptors.

Although not observed in currently known β2AR-Gs and μOR-Gi complexes, helix 8 in the receptor C-terminus is involved in Gq specificity (Qin et al., 2011).

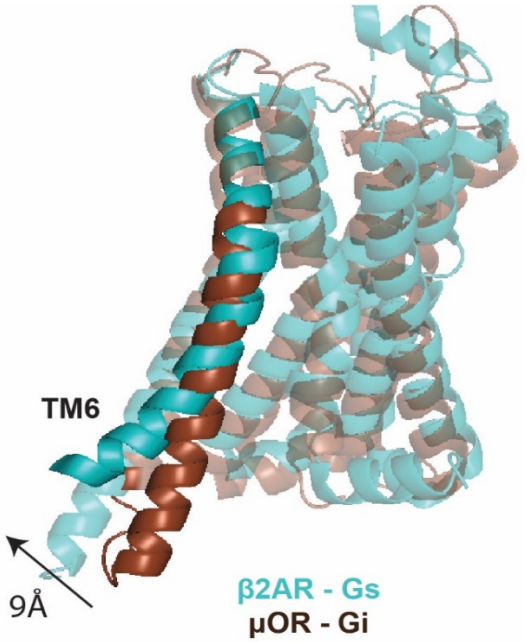


Figure 1.3: Comparison of β2AR-Gs and μOR-Gi structures. Superimposition of Gs-stabilized β2AR (PDB ID: 3SN6) onto Gi-stabilized μOR (PDB ID: 6DDE). Note that the TM6 displacement is greater in β2AR.

While the recent GPCR/G-protein 3D complexes have revealed novel insights into GPCR activation, the determinants that regulate the G-protein specificity in GPCRs remain elusive (Koehl et al., 2018). It is the contribution not just the features at the interface but also throughout the TM helices that participate in the allosteric communication to determine the G-protein selectivity (Flock et al., 2017; Wichard et al., 2011).

`

1.6 COMPUTATIONAL APPROACHES TO PREDICT GPCR/G-PROTEIN COUPLINGS

Several computational techniques have been developed to predict G-protein selectivity in GPCRs. The earliest techniques involved the use of sequence comparison methods such as BLAST (Altschul et al., 1990) and ClustalW (Thompson et al., 1994). However, these approaches lack two major aspects: a) sequences with low similarity can couple to the same G-protein, and b) sequences with high similarity can couple to different G-proteins. Later, a combination of membrane topology prediction and a pattern discovery approach in the ICLs of GPCRs (Möller et al., 2001) was used to identify unique patterns in groups of GPCR sequences, wherein each group is comprised of receptors known to couple to a G-protein subfamily. In another method (Cao et al., 2003), the authors used a naïve Bayes model to train a dataset comprising ICLs and C-terminus of GPCRs to predict their G-protein selectivity.

Other approaches used the HMM profiles (hidden Markov Model) (Eddy, 1998). In one of these studies (Sreekumar et al., 2004), the authors built the HMM profile of GPCR sequences, where each sequence is a concatenation of the predicted ICLs and the C-terminus of the receptor. Elsewhere, PRED-COUPLE (Sgourakis et al., 2005a), the authors identified a library of 25 best HMM profiles, extracted from the blocks of multiple sequence alignment (MSA) with low-entropy regions, that could discriminate for G-protein-specificity in GPCRs. In this study, only the sequence region in MSA that corresponded to the ICLs and trimmed transmembrane regions (up to 7 residues in the cell membrane) was used. The improved version, PRED-COUPLE2 (Sgourakis et al., 2005b), was trained using artificial neural networks, and the method was also extended to predict couplings with the G12/G13 subfamily, which was missing in previous approaches.

Undeniably, these studies have paved the way for use of bioinformatics techniques that can extract subtle patterns from the sequence data. However, these tools were primarily limited to only three subfamilies of G-proteins (Gi/Go, Gs, and Gq/G11) and were not designed to predict promiscuous couplings of GPCRs. Though PRED-COUPLE2 was successful at dealing with these drawbacks, like the previous approaches, its training data lacked a true negative set (i.e. a set of GPCR/G-protein

`

pairs known *not* to couple). Moreover, these tools are trained only to predict at the level of G-protein subfamilies rather than individual Gα subunits. This is critical in understanding differential G-protein couplings of GPCR isoforms (Marti-Solano et al., 2020) and members within a GPCR subfamily (Inoue et al., 2019). For example, according to the GtoPdb database, the β-adrenoreceptors couple to the Gs subfamily, while the α-adrenoreceptors couple to the Gi/Go subfamily (Altosaar et al., 2019).

Large databases such as GtoPdb and GPCRdb (G-protein coupled receptor database) (Isberg et al., 2016) have curated data for GPCRs. However, they too lack the true negative sets and Gα subunit level annotations.

Reliable prediction of GPCR/G-protein couplings remains important today. Not only are there more than 50 receptors in the GtoPdb database that lack coupling information, but it is becoming increasingly necessary to assess the impact of GPCR mutations on G-protein/β-arrestin selectivity to understand diseases. Genetic variants in GPCRs are the leading cause of several human diseases such as those related to bone development (Luo et al., 2019), Alzheimer's (Kumar et al., 2015), hypo/hyperthyroidism (Parma et al., 1997; de Roux et al., 1996), extreme obesity (Kimple et al., 2014), nephrogenic diabetes insipidus (Rosenthal et al., 1992), fertility disorders (Stoy and Gurevich, 2015) and *the emperor of all maladies*[1] – cancer (reviewed in Dorsam and Gutkind, 2007).

1.7 SPECIFICITY DETERMINING POSITIONS (SDP)

Multiple sequence alignments (MSAs) can be investigated to understand the evolution of protein families. Alignments can be used to infer functional properties of specific amino acids or positions. A long-standing view is that the positions conserved across the entire protein family participate in the functions common to all the members of it. For instance, in the context of GPCRs, the E/DRY and NPxxY motifs are conserved across all the GPCRs and known to play a role receptor function (Palczewski et al., 2000; Rosenbaum et al., 2009). On the other hand, the positions that are conserved within a subset of sequences in the MSA, participate in functions specific to the given

[1]The term was coined by Dr. Siddhartha Mukherjee in his book *The Emperor of all maladies: A Biograhpy of Cancer* (Scribner, 2010).

`

subtype (enzyme specificity) and are often referred to as Specificity Determinant Positions (SDPs) (Rausell et al., 2010). Several methods have been developed to identify functional residues within a protein subfamily and define the functional properties of the subtype.

One of the first methods to extract SDPs was a multivariate-based analysis, SequenceSpace (Casari et al., 1995). It uses principal component analysis (PCA) on aligned sequences, represented as vectors in a multi-dimensional space, to identify functional residues. The subtype information can either be predicted by SequenceSpace or provided by the user. The first three principal axes (with the largest eigenvalues) of the PCA determine the positions that are conserved across the whole superfamily while the combination of any two brings out the subtype specificity. SequenceSpace was successful at identifying experimentally known functional residues in the Ras-Rab-Rho superfamily, SH2 domains, and cyclins. A more recent approach, S3det (Rausell et al., 2010), uses multiple correspondence analysis (MCA), to link groups of proteins (subtype) to groups of residues (SDPs) in space.

rvET (Real value evolutionary trace) (Mihalek et al., 2004), a successor of ET (Lichtarge et al., 1996), is another technique that combines both evolutionary analyses (derived from phylogenetic trees) and entropic information to identify functional sites in proteins of known structure. It has been applied to several protein families such as SH2 and SH3 modular signaling domains (Lichtarge et al., 1996), the DNA binding domain of the nuclear hormone receptors (Lichtarge et al., 1996), specificity determinants in psychoactive bioamine receptors (Rodriguez et al., 2010) and to design receptors (Shenoy et al., 2006).

Unlike other methods that predict the subtype, PROUST (Hannenhalli and Russell, 2000) relies on an MSA that is already classified based on subtypes. This method constructs an HMM profile for every subtype and computes cumulative relative entropy for all the positions. The positions with Z-score > 3.0 in a given subtype are selected as functionally important for the given subtype. Besides, identifying functional residues that are either experimentally or structurally known, PROUST could highlight positions that do not lie at the interactor binding pocket, as shown for nucleotidyl cyclases, kinases, and cyclins.

`

Other SDP detection methods include SDPsite (Kalinina et al., 2009), which maps SDPs, obtained from mutual information, CPs (conserved positions), defined using Sander-Schneider conservation measure onto a structure of one of the proteins of the subtype to extract the best cluster of functional residues. kPax (Marttinen et al., 2006) simultaneously detects subtypes and associated SDPs using a Bayesian model-based approach. CEO (Combinatorial Entropy Optimization) (Reva et al., 2007) performs hierarchical clustering of possible subtypes to select the optimum one and then identifies SDPs using residue entropy. Statistical Coupling Analysis (SCA) (Lockless and Ranganathan, 1999) applied spectral decomposition to a weighted correlation matrix, obtained by combining correlation information with sequence conservation.

In this thesis, I built on these previous approaches and present a statistically-associated protocol (Chapter II), guided by machine learning (Chapter III), that exploits the HMM profiles to highlight positions and regions in an MSA that participate in determining the G-protein selectivity of GPCRs. The method is loosely based on that described in Hannenhalli & Russell, 2000.

1.8 OUTLINE OF THE REPORT

To capture and comprehend the signatures of G-protein coupling specificity in GPCRs, we analyzed and exploited one of the most extensive datasets (Inoue et al., 2019) of GPCR/G-protein couplings, created by our collaborator Dr. Asuka Inoue and his group (Tohoku University, Japan). It comprises binding affinities of a well-studied set of 144 class A and 4 class B GPCRs with 11 heterotrimeric G-proteins experimentally derived using the TGFα shedding assay technique (Inoue et al., 2012).

Chapter II presents an overview of the most extensive dataset capturing GPCR/G-protein binding affinities. We describe a novel, statistically-associated protocol to extract sequence-based determinants of G-protein coupling specificity from GPCRs. We highlight the positions in the seven-transmembrane domain of a GPCR that delineate subtle features of G-protein selectivity and also assess the influence of length and amino acid composition of ICL3 and the C-terminus.

Chapter III describes a novel, machine-learning guided framework to GPCR couplings

`

that relies on the statistically-associated protocol (Chapter II) and structure-based features of receptors. We interpret the predicted models to create a feature weight matrix that outlines the quantitative relevance of each feature. Lastly, we present the PRECOG web-server to predict the G-protein coupling specificity of any class A GPCR but also permits the user to design receptors with particular signaling properties.

In the first part of Chapter IV, I use PRECOG to a) predict the couplings of uncharacterized GPCRs, b) predict the impact of mutations on GPCR couplings, and c) develop and experimentally validate the first designer receptor that exclusively couples to GNA12 (Inoue et al., 2019). I then demonstrate the reusability of the PRECOG framework on a recently available dataset of GPCR – G-proteins/β-arrestins and explain the differences between the determinants of G-protein and β-arrestin specificity.

Chapter V summarizes the main results of Chapters II-IV and the practical implications and the outlook of the proposed machine-learning guided framework for future endeavors.

`

Chapter II: Determinants of G-protein coupling specificity in GPCRs

2.1 ABSTRACT

Upon activation, GPCRs undergo a reorganization of helices to effectively engage with their primary transducers: G-proteins. Despite several structures and databases available to date, the determinants of GPCR/G-protein coupling specificity remain elusive. Here, we investigate the most comprehensive dataset of quantified GPCR/G-protein binding affinities derived from a robust assay (TGFα shedding) to reveal the complex patterns of the receptor couplings. This dataset enabled us to develop a statistically-associated protocol to extract sequence-based determinants of G-proteins coupling specificity, encompassing both transmembrane and extra-membrane regions of the GPCRs. The proposed protocol lays the foundation for a framework that can be applied to any binding data to unravel the positions that determine the subtype specificity.

2.2 INTRODUCTION

The last decade has benefitted from the determination of several 3D complexes of GPCR/G-protein (Carpenter et al., 2016; Draper-Joyce et al., 2018; García-Nafría et al., 2018b; Koehl et al., 2018) and structural bioinformatics efforts (Flock et al., 2017; Venkatakrishnan et al., 2013, 2016) that have shed light on GPCR activation. The structural analysis of these complexes has given further insights into subtle but key differences between Gs- and Gi/Go-stabilized receptors. For example, the distribution of residues, as well the extent of displacement of TM6 of Gs and Gi/Go-coupled receptors can be used to stratify receptors based on their G-protein coupling preferences (Draper-Joyce et al., 2018; Kang et al., 2018) (see section 1.4 in Chapter I). Furthermore, recent studies have investigated the role of intrinsically disordered regions in the intra/extra-cellular loops and the N/C-termini in interactions with partner proteins (Hilger et al., 2018; Wheatley et al., 2012), by undergoing a disorder-to-order transition (Shukla et al., 2013; Venkatakrishnan et al., 2014). However, despite these

`

advances, and decades of research, the exact determinants of G-protein coupling specificity in receptors are still largely unknown (Flock et al., 2017; Koehl et al., 2018).

On the contrary, for the primary transducers of GPCRs, G-proteins, the C-terminus of the α5 helix in the Gα subunit has long been known to be a major determinant of GPCR specificity (Flock et al., 2017). The patterns of amino acids in G-proteins are recognized by GPCRs, analogous to the lock (G-proteins) and key (receptor) mechanism (Flock et al., 2017). The master keys (promiscuous receptors) can open multiple locks, while the specific keys (non-promiscuous receptors) can open just one. Studies have narrowed down the last 4-5 C-terminal residues of the α5 helix in the Gα subunits to be the determinants of GPCR coupling specificity across all G-proteins (Conklin et al., 1993; Inoue et al., 2012). This has led to the comprehensive screening of hundreds of receptors with G-proteins by altering only the last 4-6 amino acids of Gα subunits (Hsu and Luo, 2007; Kawano et al., 2016; Wang et al., 2009).

While these assay techniques have expanded our knowledge of GPCR/G-protein couplings, they suffer from major disadvantages. First, most cover only one G-protein subfamily. For example, the guanine nucleotide-binding assay for Gi/Go, cAMP assays for Gs (Thomsen et al., 2005). Second, they target G-proteins at the subfamily level, lacking the Gα subunit specificity (Inoue et al., 2012). Third, the coverage of G-proteins is heterogeneous. For instance, not all G12/G13-coupled receptors can be activated using these techniques, leaving them poorly characterized (Inoue et al., 2012); for example, they have the fewest coupling details in GPCR databases (GtoPdb/GPCRdb). The significance of the G12/G13 subfamily is evident from its implications in cardiovascular (Suzuki et al., 2009; Worzfeld et al., 2008) and metabolic diseases (Yang et al., 2020b), and its receptors in the breast (Kitayama et al., 2004), ovarian (Lee et al., 2006), and colon (Shida et al., 2003) cancers.

The Transforming Growth Factor-α (TGFα) shedding assay (Inoue et al., 2012) was used to screen and develop a dataset comprising 144 class A and 4 class B GPCRs (Inoue et al., 2019), including the poorly characterized G12/G13-receptors, against 11 chimeric G-proteins. We made use of a previously defined approach to extract specificity-determining positions from the dataset (Hannenhalli and Russell, 2000). We

`

interrogated the dataset to develop a statistically associated protocol that revealed sequence-based determinants of G-protein coupling specificity in GPCRs.

2.3 MATERIALS AND METHODS

2.3.1 Optimal binding affinity value

The data derived from the TGFα shedding assay, hereafter referred to as the coupling dataset, comprises binding affinities of GPCRs (144 class A and 4 class B) against 11 chimeric G-protein constructs (Inoue et al., 2019). These binding affinities are expressed in terms of LogRAi values, the (base 10) logarithmically transformed values of relative intrinsic activity (RAi), which range from -2 (no binding) to 0 (maximum response) (Table S1A). For detailed methods, please refer to the original article (Inoue et al., 2019). Given the small size of class B receptors, we considered only class A receptors for this analysis. We compared the coupling dataset with known coupling evidence from GtoPdb (Harding et al., 2018). While the coupling dataset is spread over a range of continuous values between -2 and 0, GtoPdb provides coupling evidence for each receptor based on published studies. For a fair comparison, we sought to binarize the LogRAi values of the coupling dataset as well the coupling evidence from GtoPdb.

A GPCR/G-protein pair was defined to be true positive if it had at least three citations in GtoPdb; true negatives were those with no citations. We then performed a Receiver Operating Characteristic (ROC) curve analysis of the receptors that overlap between the coupling dataset and GtoPdb at varying thresholds of LogRAi to calculate the true- (TPR) and false-positive rates (FPR) (Table S1B). We obtained the LogRAi value of -1.0 as the optimal threshold, where the TPR was maximum while the FPR was minimum with an Area Under the Curve (AUC) of 0.7. The above procedure was repeated by considering only primary couplings reported in GtoPdb and still obtained the same optimal threshold (LogRAi = -1.0) with an AUC of 0.78. We thus binarized all the binding affinities of the coupling dataset using this threshold. A pair was a true-coupling if LogRAi ≤ -1.0 and false otherwise (Table S1A).

`

2.3.2 LogRAi profile vs sequence identity

To determine if there is a correlation between LogRAi profile (binding affinities) and sequence identity, we performed a pairwise BLAST (Altschul et al., 1990) of each receptor in the coupling dataset against the remainder as database. Next, we drew scatter plots to compare the LogRAi profile (binding affinities) of each receptor with its most similar match by calculating their Euclidean distances.

2.3.3 Sequence-based determinants of coupling specificity

Inspired from a previously defined approach to extract determinants of coupling specificity, we built an MSA of the class A GPCR sequences by running the command-line tool - *hmmalign* - of the HMMER3 package (v3.1b2) (Eddy, 1998) using the hidden Markov models (HMM) of 7 transmembrane receptors (rhodopsin family) from Pfam 7tm_1 (Pfam accession: PF00001; 2016 release). For every chimeric G-protein, we subdivided the MSA into coupled (LogRAi ≥ –1.0) and not-coupled (LogRAi < -1.0). Lastly, the HMM profile of each sub-alignment was constructed using the *hmmbuild* tool of the HMMER3 package.

2.3.3.1 7TM1 positional features

We defined two types of positional features for each G-protein coupling group. The alignment positions that belong to the first type are present in both coupled and not-coupled HMM profiles of the given G-protein coupling group. We performed the Wilcoxon signed-rank test to compare the amino acid distribution (bit-score of each amino acid) of every position in both coupled and not-coupled HMM profiles of receptors and defined significant positions as those with p-value ≤ 0.05 (Table S1E).

We also explicitly considered insertions and deletions in one alignment relative to the other (coupled or not-coupled). Positions present only in the sub-alignment of coupled receptors were considered as insertions while those in sub-alignment of not-coupled receptors were considered as deletions. Each position was assigned its corresponding Pfam 7tm_1 position and BW numbering. The most conserved position within each helix was defined according to GPCRdb (Isberg et al., 2016) (Table S1D).

`

2.3.3.2 Extra-membrane features

The ICL3 and C-terminus are known to play a role in determining the G-protein coupling specificity of GPCRs (Flock et al., 2015; Koehl et al., 2018; Rasmussen et al., 2011). These regions are either not well aligned in the Pfam/HMMER driven alignments owing to genuine heterogeneity of the local sequences, or completely absent from them in the case of much of the C-terminus. However, as they are well established to play key roles in determining coupling specificity, we decided instead to consider the length and amino acid compositions of the ICL3 and C-terminus as potential features determining G-protein specificity. We used the nonparametric Wilcoxon rank-sums test to provide a p-value and defined significant features as those having a p-value ≤ 0.05 (Table S1F). For positions lying in extra-membrane regions (i.e. where no BW numbering is possible), we quote only the Pfam 7tm_1 position.

2.3.4 GPCR/G-protein interfaces

We extracted domain (Pfam) to chain (PDB) mappings of all 3D complexes in the SIFTS database (Structure Integration with Function, taxonomy, and sequence) (Velankar et al., 2013). We considered only the complexes with Pfam accessions PF00001 (7-transmembrane receptor) for GPCRs and PF00503 for G-protein alpha subunits. A residue of the GPCR chain was considered to belong to the interface if at least one of its atoms was ≤ 6.5 Å from at least one atom of any residue in the G-protein chain.

All the interfaces were mapped from their GPCR chains to their corresponding canonical UniProt sequences by performing a pairwise alignment of the two sequences using the command-line tool - *blastp* - of the BLAST package (Altschul et al., 1990). Every interface residue was assigned its corresponding Pfam 7tm_1 position and BW numbering (Table S1D). A residue was considered adjacent to the interface if the distance between at least one pair of atoms of the residue and the interface was ≤ 5 Å.

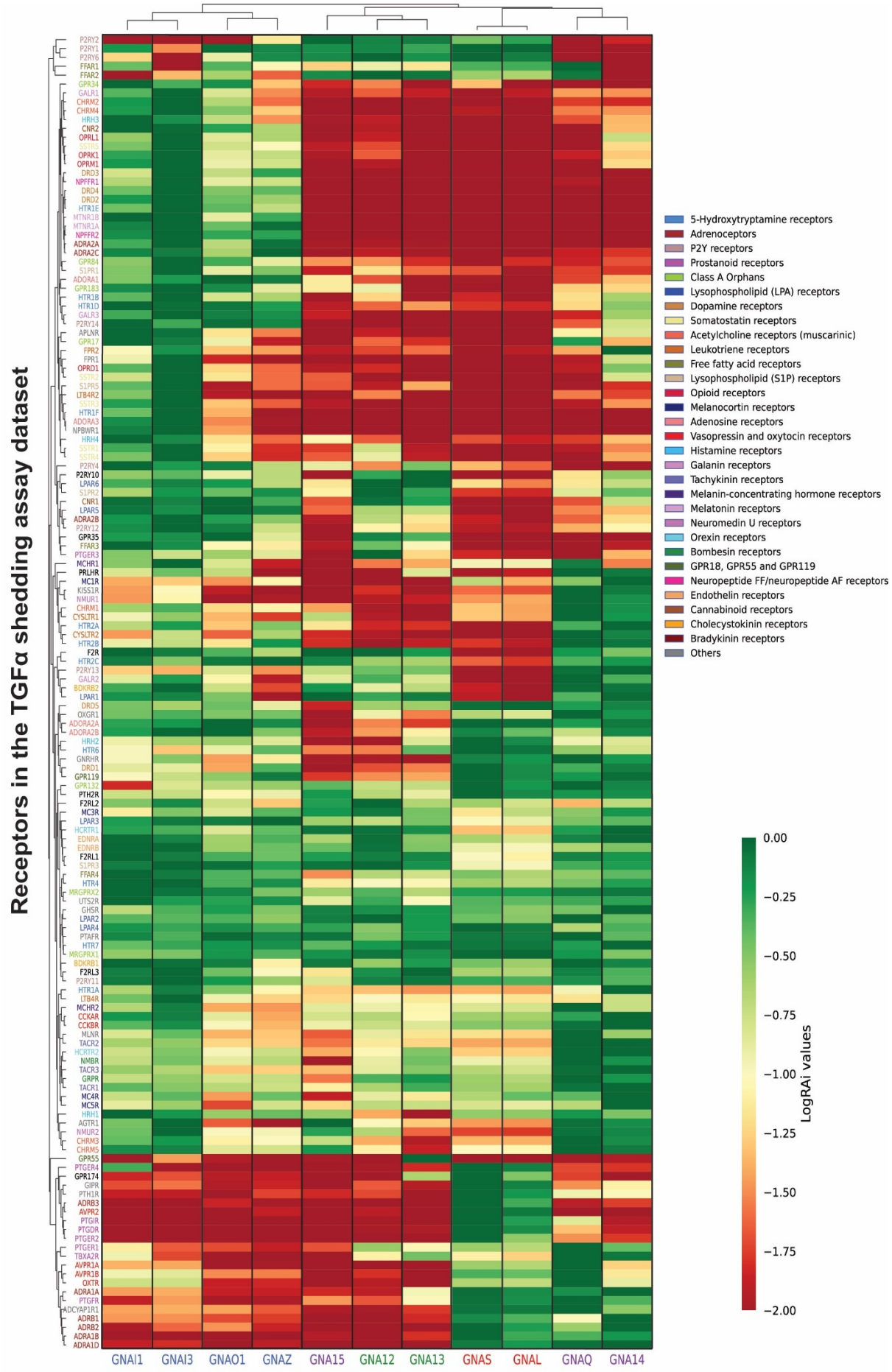
The structural analysis was performed using the Biopython library (Cock et al., 2009), and all statistical tests and Euclidean distance calculations in this section were performed using the SciPy library (Virtanen et al., 2020) with scripts written in Python

`

2.4 RESULTS

2.4.1 Overview of the coupling data-set

The dataset derived from the TGFα shedding assay provides binding affinities of 148 GPCRs (144 class A and 4 class B) against 11 chimeric G-proteins (Figure 2.1). The receptors in the dataset display a wide range of G-protein coupling profiles (Figure 2.1). Out of these 144 receptors, 11 that couple to at least one G-protein in the coupling dataset, have neither primary nor secondary couplings reported in GtoPdb (Table S1C). These include the protease-activated receptors, involved in the inflammatory response (Heuberger and Schuepbach, 2019), diet-induced obesity and metabolism (Badeanlou et al., 2011), and the purinergic receptors (*P2RY10* and *P2RY12*), which contribute to platelet aggregation, apoptosis, neurogenerative diseases and gliomas (Burnstock, 2013; Burnstock et al., 2010).


Receptors in the TGFα shedding assay dataset
GNAI1
GNAI3
GNAO1
GNAZ
GNA15
GNA12
GNA13
GNAS
GNAL
GNAQ
GNA14
5-Hydroxytryptamine receptors
Adrenoceptors
P2Y receptors
Prostanoid receptors
Class A Orphans
Lysophospholipid (LPA) receptors
Dopamine receptors
Somatostatin receptors
Acetylcholine receptors (muscarinic)
Leukotriene receptors
Free fatty acid receptors
Lysophospholipid (S1P) receptors
Opioid receptors
Melanocortin receptors
Adenosine receptors
Vasopressin and oxytocin receptors
Histamine receptors
Galanin receptors
Tachykinin receptors
Melanin-concentrating hormone receptors
Melatonin receptors
Neuromedin U receptors
Orexin receptors
Bombesin receptors
GPR18, GPR55 and GPR119
Neuropeptide FF/neuropeptide AF receptors
Endothelin receptors
Cannabinoid receptors
Cholecystokinin receptors
Bradykinin receptors
Others
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
−1.50
−1.75
−2.00
LogRAi values

`

Figure 2.1: Binding affinities of 148 GPCRs with 11 chimeric G-proteins in the coupling dataset. Binding affinities were measured in terms of LogRAi values. Cell colours vary from red (minimum response; LogRAi = -2) to green (maximum response; LogRAi = 0). Row labels are colored according to the GPCR subfamily.

2.4.2 LogRAi profile vs. Sequence similarity

We compared the LogRAi values of each receptor with the most similar match (sequence-wise) in the coupling dataset (see Methods). While we observed receptors with low sequence similarity with similar LogRAi profiles, we also noticed receptors of the same subfamily with very different LogRAi profiles. For example, *MTNR1A* (melatonin receptor type 1A) shares a low sequence similarity (~27%) with *NPFFR2* (neuropeptide FF receptor 2) but similar LogRAi profiles (Euclidean distance = 0.27; Figure 2.2), though both belong to different GPCR sub-families. In contrast, *PTH1R* and *PTH2R* (members of the parathyroid hormone receptor family) show over 50% sequence similarity but very contrasting LogRAi profiles (Euclidean distance = 3.22; Figure 2.7B).

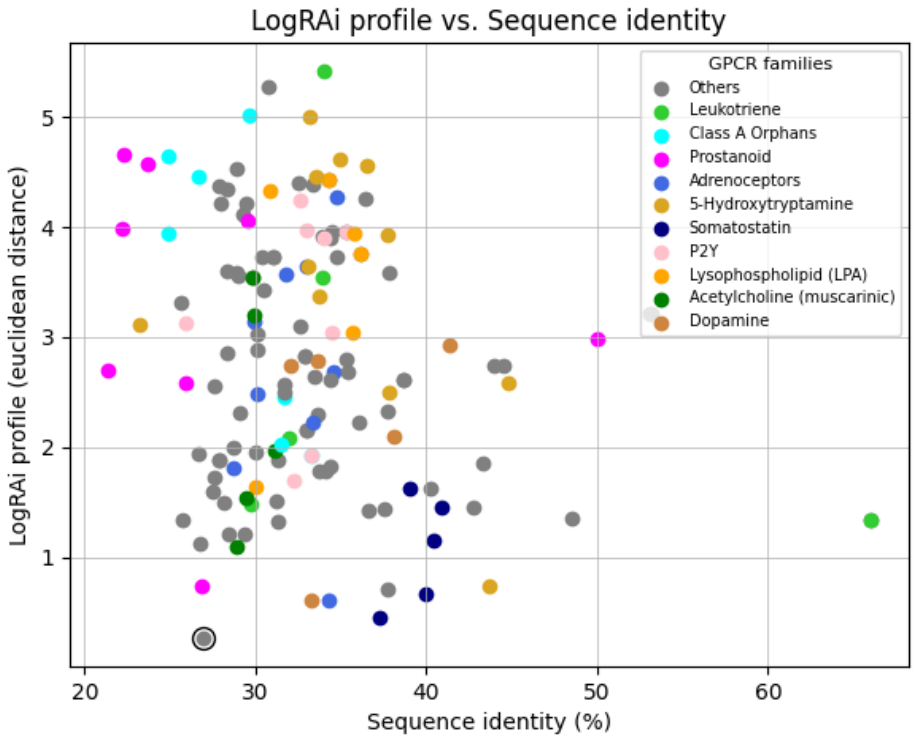


Figure 2.2: Scatter plot of LogRAi profile vs. sequence identity of receptors in the coupling dataset. Comparison of LogRAi profiles (calculated in Euclidean distance) with sequence identity (in

`

percentage). The color of bubbles corresponds to their respective GPCR subfamily. The grey-colored bubble with an edge color of black refers to *MTNR1A* (melatonin receptor type 1A) and *NPFFR2* (neuropeptide FF receptor 2) that share low sequence identity (~27%) but similar LogRAi profiles.

2.4.3 Comparison with GtoPdb

We compared the coupling preferences of the receptors in the coupling dataset with that of their known coupling evidence in GtoPdb. While the coupling dataset is a continuous range of LogRAi values, GtoPdb contains only binary true positives of known G-protein sub-families for a given receptor. We thus binarized the LogRAi values with the optimal cut-off -1.0, obtained from the ROC curve analysis (Figure 2.3; Table S1A; see Methods).

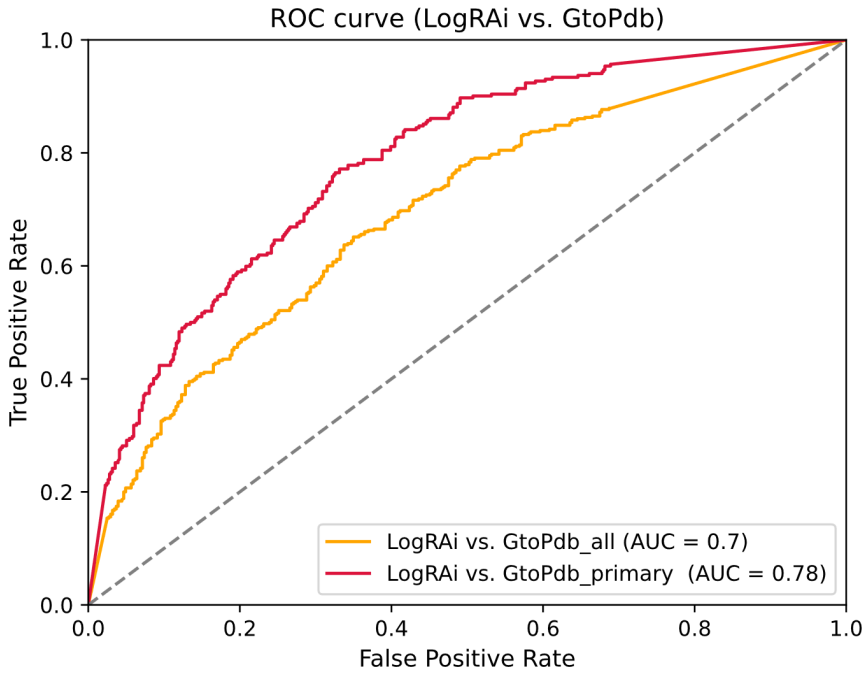


Figure 2.3: ROC curves of LogRAi values in the coupling dataset vs. coupling values known from GtoPdb. Analysis performed by considering all couplings is shown in yellow, while the one performed by considering only primary couplings is shown in red.

The number of receptors that couple to a given G-protein varies across G-proteins. *GNAI1* (Gi/o subfamily) couples to the most receptors (119) and *GNA15* (Gq/G11 subfamily) to the fewest (Figure 2.4).

`

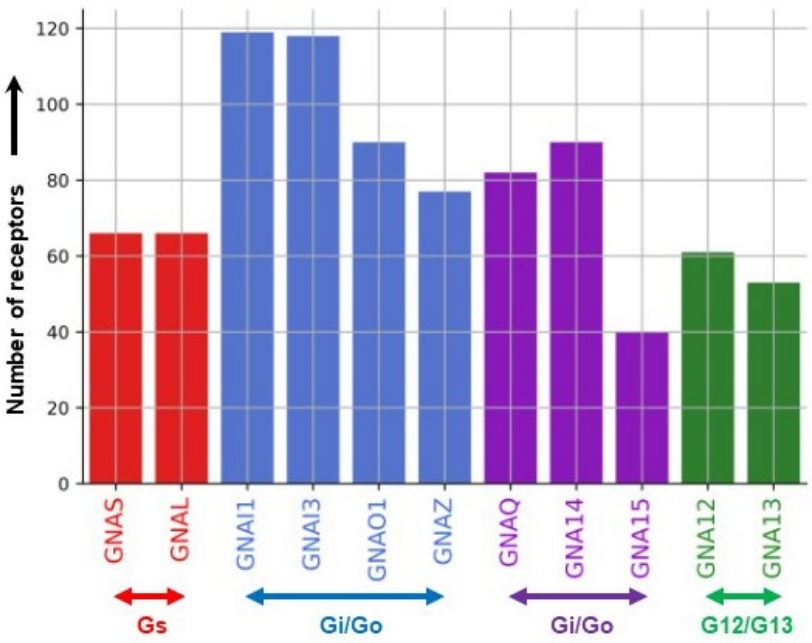


Figure 2.4: Number of receptors coupling to each G-protein in the coupling dataset. LogRAi = -1.0 was used as the optimal cut-off (see Methods).

Nearly half of the GPCR couplings (176 out of 366 or 48%) reported in the coupling dataset were not known in GtoPdb (Figures 2.5B-D) and cover all the G-protein subfamilies. The majority of the new couplings are G12/G13-coupled receptors (55 out of 176 or 31%), expected as these have to date been least studied. Of all the G12/G13 couplings, the coupling dataset accounts for 76% while GtoPdb accounts for 2% of them. Gs-coupled, Gi/Go-coupled, and Gq/G11-coupled receptors show an overlap of 35%, 58%, and 62%, respectively, between the coupling dataset and GtoPdb.

`

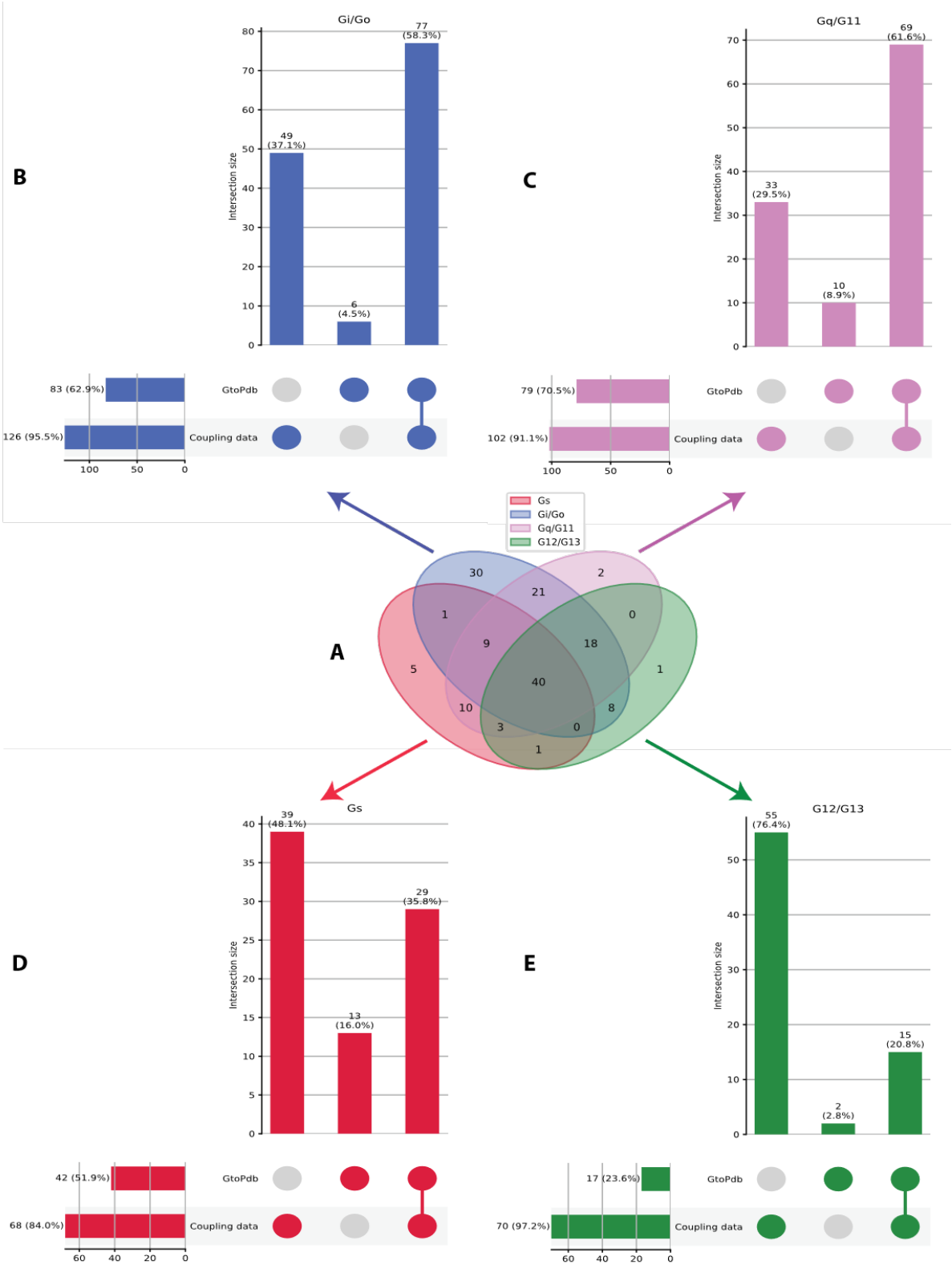


Figure 2.5: Comparison of GPCR couplings in the coupling dataset with GtoPdb. (A, center) In the center, Venn diagram of number receptors coupled to each G-protein subfamily in the TGFα shedding assay dataset. (B-E) Bar plots of receptors coupled to four G-protein subfamilies in the coupling dataset and GtoPdb.

`

2.4.4 Variable promiscuity of the receptors

We defined a GPCR to be promiscuous if it coupled to a minimum of one member each from at least two different G-protein sub-families. While G12/G13-coupled receptors are most promiscuous, Gi/Go-coupled receptors are most specific (Figures 2.5A, 2.6), later corroborated by an independent study (Avet et al., 2020).

A total of 40 receptors coupled to all G-protein subfamilies (Figure 2.5). Many receptors show coupling preference to a specific G-protein subfamily, such as the melatonin, and opioid receptors to Gi/Go subfamily; the adrenoreceptors and prostanoid receptors to the Gs subfamily (Figure 2.7A). At the other extreme, there are several receptors, such as P2Y (purinergic), lysophospholipid, and endothelin, that couple to more than one G-protein subfamily. Overall in the coupling dataset, G12/G13-coupled receptors are most promiscuous while Gi/Go-coupled receptors are most specific (Figure 2.6).

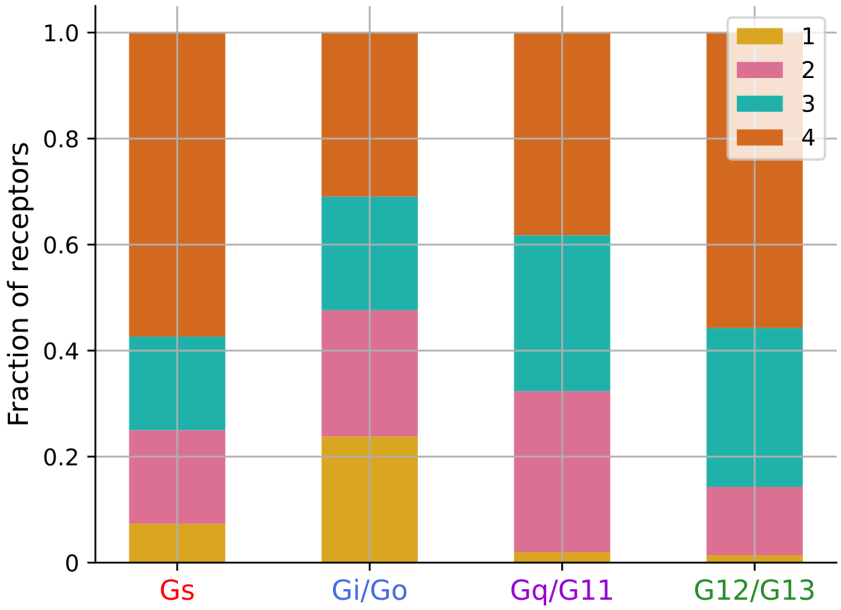


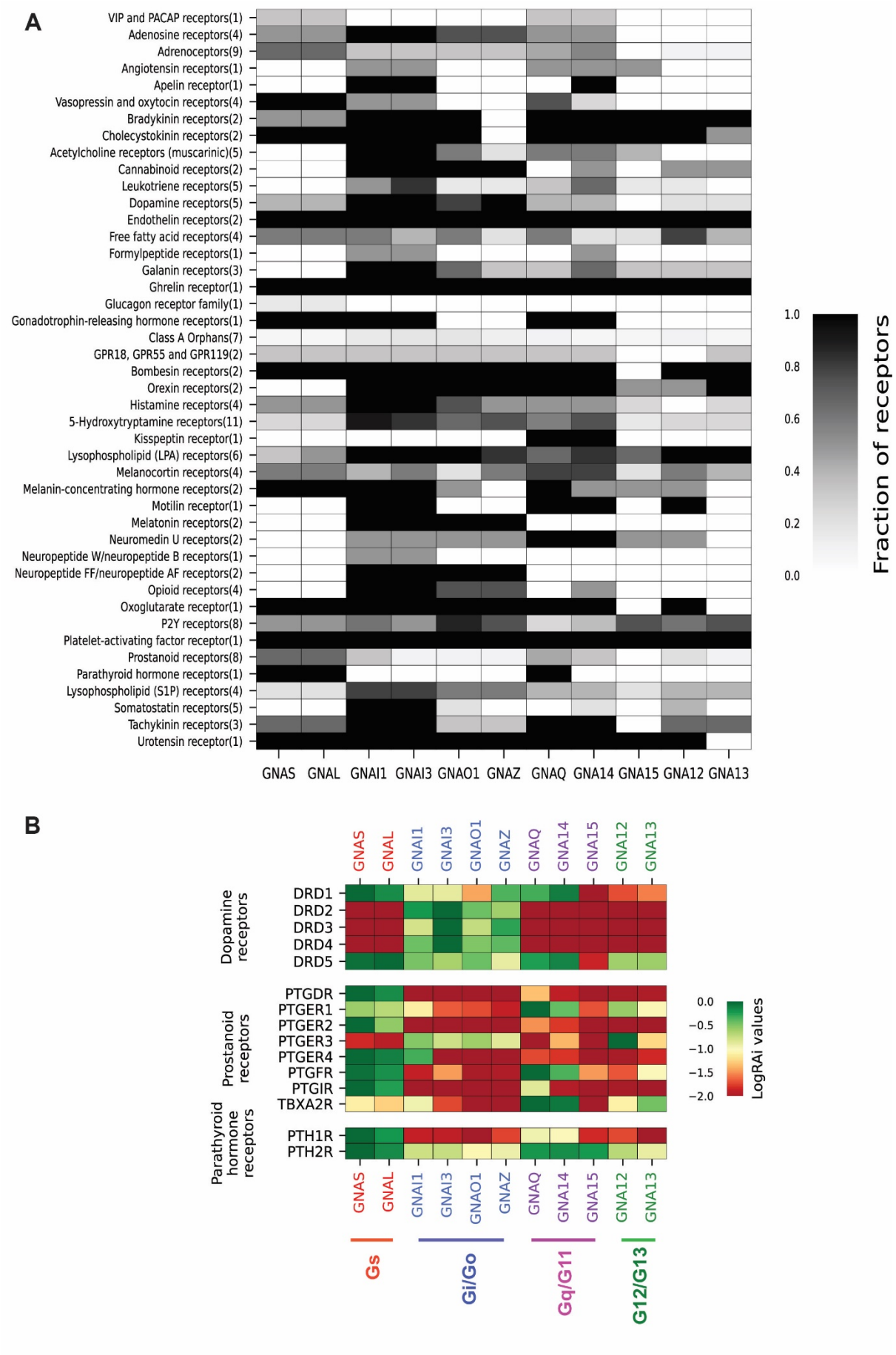
Figure 2.6: Promiscuity of the receptors in the coupling dataset. The bars represent the promiscuity of the receptors (ranging 1 to 4) that couple to the corresponding G-protein subfamily. While 1 means the receptors are specific to the given G-protein subfamily, 4 means the receptors couple to all the G-protein subfamilies.

Interestingly, there were several receptors showing specificity towards specific members of a G-protein subfamily. For instance, *GPR55* (G-protein-coupled receptor 55) is a putative cannabinoid receptor (Lauckner et al., 2008) that couples to *GNA13*,

`

but shows no coupling specificity towards *GNA12* (Figure 2.7A), though both the Gα subunits are members of the G12/G13 subfamily. Likewise, *NMBR* (neuromedin-B receptor), which plays role in the contraction of smooth muscle (Kilgore et al., 1993), neuronal responses (Mishra et al., 2012), and cell growth regulation (Matusiak et al., 2005), couples to *GNAQ* and *GNA14*, but has only a weak binding affinity towards *GNA15*, though all the three Gα subunits belong to the Gq/G11 subfamily (Figure 2.7A). Although the precise meaning of these sub-subfamily specificities is not clear, these results highlight the utility of this approach to elucidating the finer functional details of GPCR/G-protein coupling.

Similarly, the dataset also contains several examples of differential coupling specificities of receptors within the same family, a well-known characteristic of GPCRs (Shinoura et al., 2002; Williams et al., 1998). For instance, the dopamine receptors, which are involved in several neurological processes such as cognition, spatial working memory, pleasure, learning, and motor function (Girault and Greengard, 2004), are classified into two sub-types: D1 and D2. While *DRD1* and *DRD5* (D1-type) show higher binding specificity towards the Gs family, *DRD2*, *DRD3*, and *DRD4* (D2-type) couples preferentially to the Gi/Go subfamily (Figure 2.7B). Similarly, while *HRH2* of the histamine receptor family couples to the Gs subfamily, *HRH1*, *HRH3*, and *HRH4* prefer Gi/Go.


A
VIP and PACAP receptors(1)
Adenosine receptors(4)
Adrenoceptors(9)
Angiotensin receptors(1)
Apelin receptor(1)
Vasopressin and oxytocin receptors(4)
Bradykinin receptors(2)
Cholecystokinin receptors(2)
Acetylcholine receptors (muscarinic)(5)
Cannabinoid receptors(2)
Leukotriene receptors(5)
Dopamine receptors(5)
Endothelin receptors(2)
Free fatty acid receptors(4)
Formylpeptide receptors(1)
Galanin receptors(3)
Ghrelin receptor(1)
Glucagon receptor family(1)
Gonadotrophin-releasing hormone receptors(1)
Class A Orphans(7)
GPR18, GPR55 and GPR119(2)
Bombesin receptors(2)
Orexin receptors(2)
Histamine receptors(4)
5-Hydroxytryptamine receptors(11)
Kisspeptin receptor(1)
Lysophospholipid (LPA) receptors(6)
Melanocortin receptors(4)
Melanin-concentrating hormone receptors(2)
Motilin receptor(1)
Melatonin receptors(2)
Neuromedin U receptors(2)
Neuropeptide W/neuropeptide B receptors(1)
Neuropeptide FF/neuropeptide AF receptors(2)
Opioid receptors(4)
Oxoglutarate receptor(1)
P2Y receptors(8)
Platelet-activating factor receptor(1)
Prostanoid receptors(8)
Parathyroid hormone receptors(1)
Lysophospholipid (S1P) receptors(4)
Somatostatin receptors(5)
Tachykinin receptors(3)
Urotensin receptor(1)
GNAS GNAL GNAI1 GNAI3 GNAO1 GNAZ GNAQ GNA14 GNA15 GNA12 GNA13
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of receptors
B
GNAS GNAL GNAI1 GNAI3 GNAO1 GNAZ GNAQ GNA14 GNA15 GNA12 GNA13
Dopamine receptors
DRD1
DRD2
DRD3
DRD4
DRD5
Prostanoid receptors
PTGDR
PTGER1
PTGER2
PTGER3
PTGER4
PTGFR
PTGIR
TBXA2R
Parathyroid hormone receptors
PTH1R
PTH2R
0.0
−0.5
−1.0
−1.5
−2.0
LogRAi values
Gs
Gi/Go
Gq/G11
G12/G13

`

Figure 2.7: Variable LogRAi profiles in GPCR families in the coupling dataset. (A) Heatmap of differential couplings in GPCR families. The color intensity of a cell is the fraction of receptors in the corresponding GPCR family that couple to the given G-protein subfamily. The number of receptors in the family is shown in parenthesis. (B) LogRAi values of prostanoid receptors (upper panel), dopamine receptors (middle panel), and parathyroid hormone receptors (lower panel).

2.4.5 G-protein specificity determining features

We performed statistical tests to identify determinants of coupling specificity in the 7TM domain (Figure 2.8; Table S1E) and in the extra-membrane region (Table S1F) of GPCRs that is most strongly associated with binding to each of the 11 chimeric G-proteins.

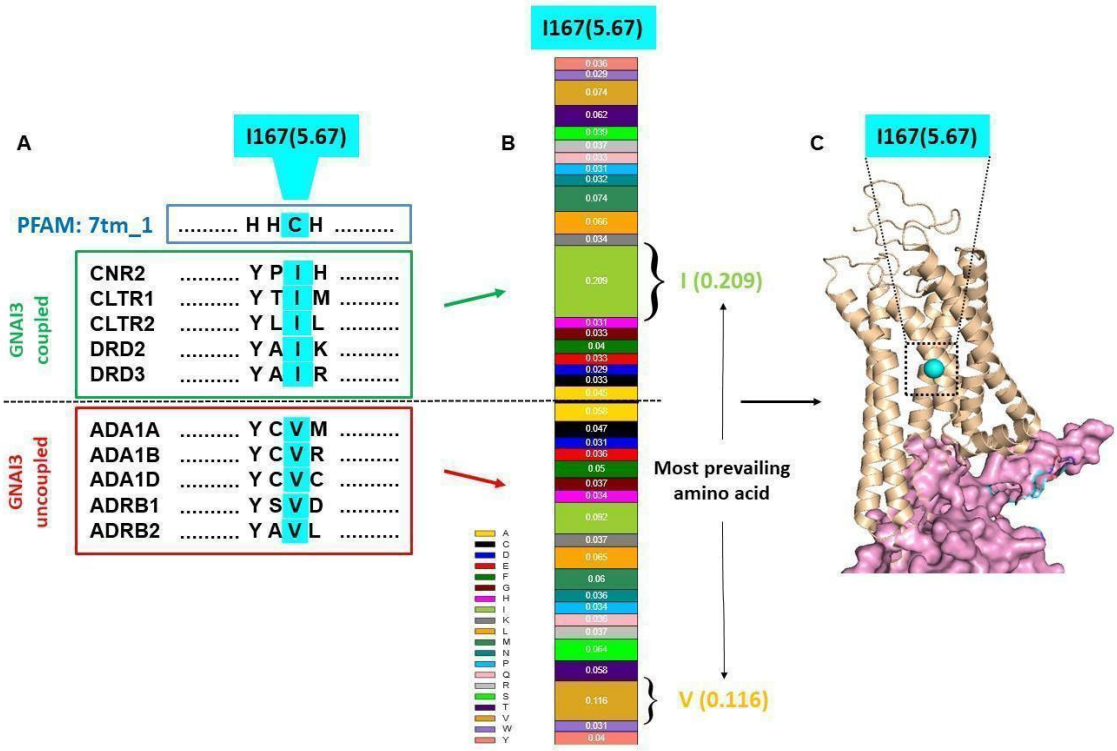


Figure 2.8: Illustrative example of a Pfam 7tm_1 position for the *GNAI3* coupling group. (A) MSA of Gi/Go- coupled/not-coupled receptors, highlighting Pfam 7tm_1 position: I167 (BW: 5.67) in cyan. (B) Amino acid distribution at the highlighted position. (C) The highlighted position is annotated as the determinant of coupling specificity (if statistically significant) and shown on structure (PDB ID: 3SN6).

The determinants of coupling specificity are more abundant in the TM helices of Gs-, Gi/Go- and Gq/G11- coupled receptors. A surprising finding was the much greater contribution from the ICL3 towards the G12/G13 specificity in the receptors (Figure 2.9).

`

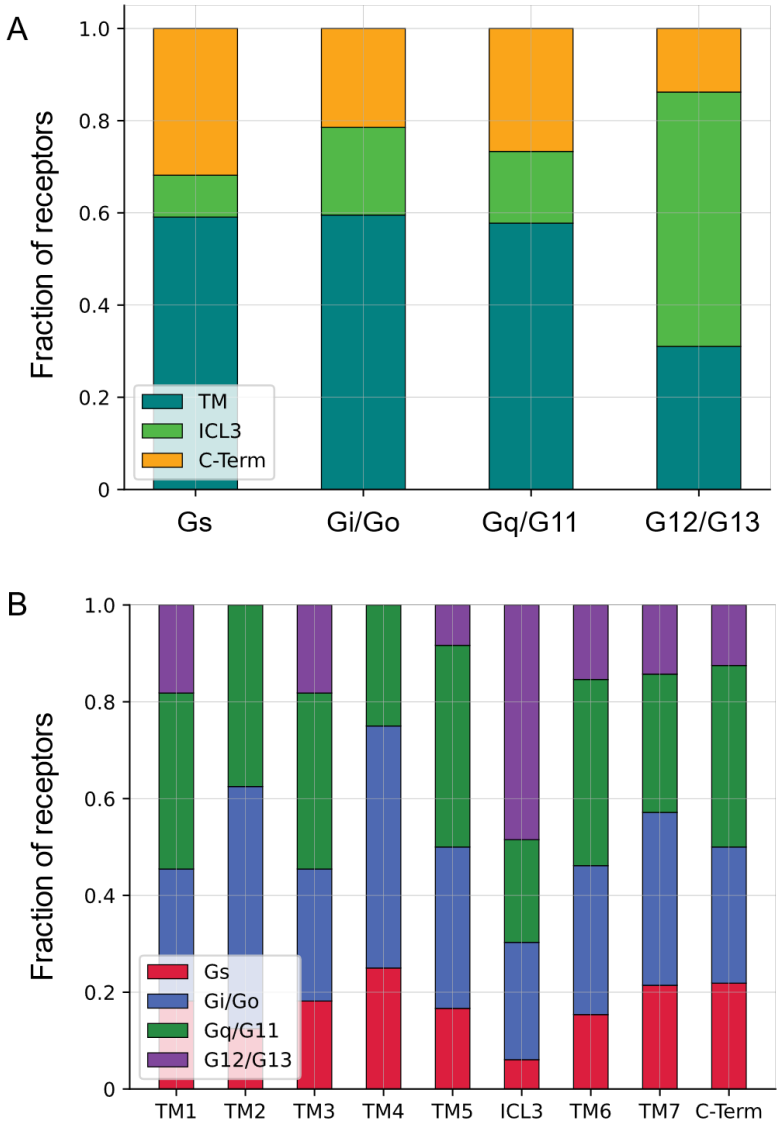


Figure 2.9: Distribution of determinants of coupling-specificity in the receptors. (A) Across G-protein subfamilies. (B) Across the topology.

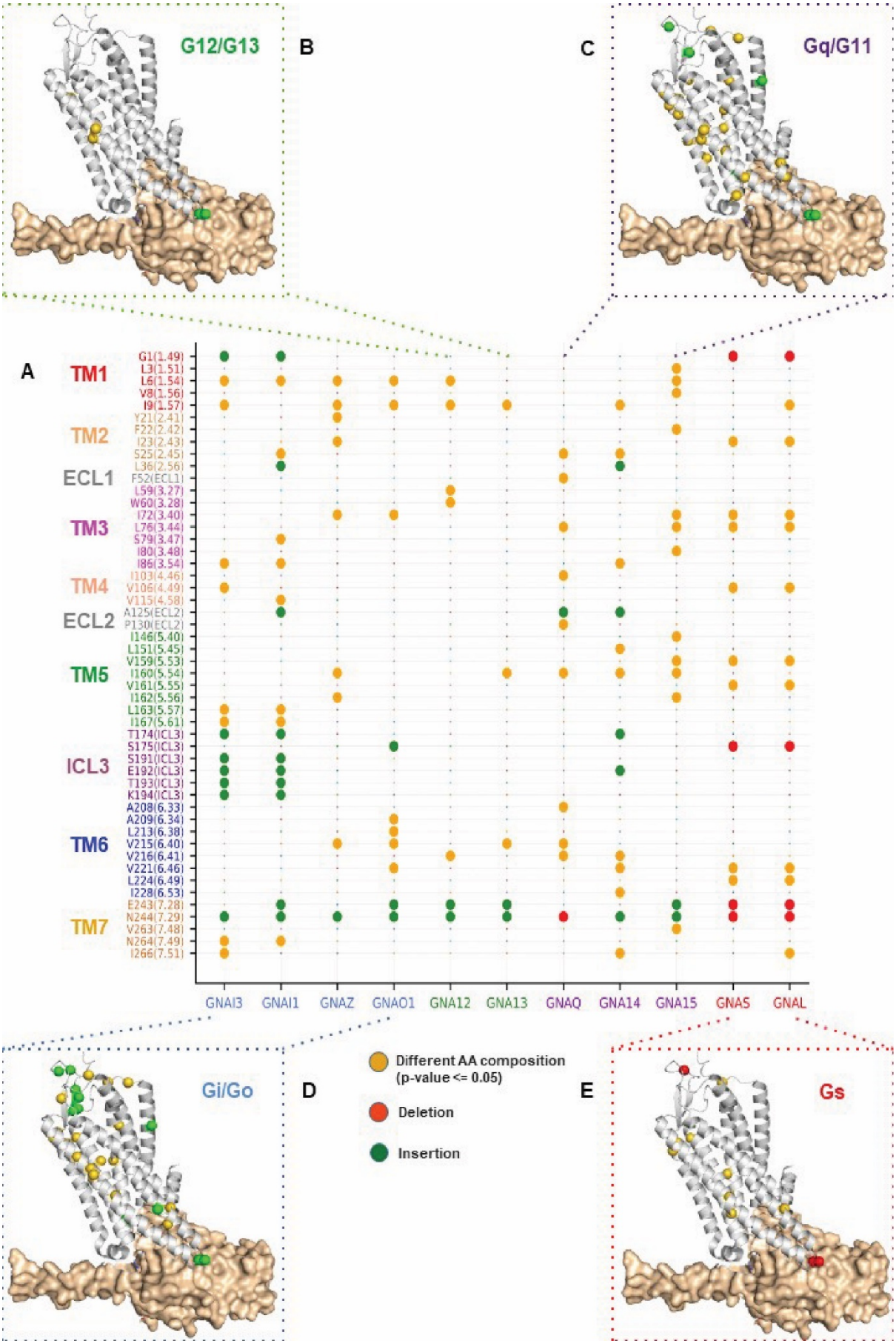
Of the 51 identified positional features (determinants of coupling specificity), only 13 (23%) lie directly at the G-protein binding sites on the cytosolic side of the ICL3 and TM helices 5 and 6, including those that are absent from Gs-coupled receptors, but present in Gi/Go-coupled receptors (Pfam 7tm_1 positions corresponding to ICL3: 174, 191-194) (Table S1E; Figure 2.10). Another 21% of the positions lie adjacent to the G-protein binding sites (see Methods) and thus likely participate in the activation mechanism (Table S1E). The remainder of positions lies either in the extracellular pockets (ligand binding sites) or within the TM helices, thus might play a role in allosteric communication during activation (Figure 2.10B).

`

The determinants of coupling specificity show a wide range of selectivity to G-protein coupling groups. Only position N244 (BW: 7.29) in the receptor is the determinant of coupling specificity in all the G-protein subfamilies (Figure 2.10A). Other positions (e.g. T193 and K194 in ICL3) only contribute to selectivity to a single (Gi/Go) G-protein subfamily (Figure 2.10A).

Specificity determining positions vary even for members of the same G-protein subfamily. For example, position L6 (BW: 1.54) is selective for *GNA12* but not *GNA13*, whereas I160 (BW: 5.54) is selective to *GNA13* but not *GNA12* (Figure 2.10A). We also identified determinants of coupling specificity that are insertions in one subfamily of G-proteins while deletions in others. For example, position G1 (BW: 1.49) of the 7TM1 domain is an insertion in *GNAI1*- and *GNAI3*- coupled receptors, but a deletion in Gs-coupled receptors (Figure 2.10A). It is noteworthy that 3 out of 4 deletions are determinants of coupling specificity to the Gs subfamily while most of the insertions are determinants of Gi/Go coupling specificity (Figure 2.10A).

`


G12/G13
B
C
Gq/G11
A
TM1
TM2
ECL1
TM3
TM4
ECL2
TM5
ICL3
TM6
TM7
G1(1.49)
L3(1.51)
L6(1.54)
V8(1.56)
I9(1.57)
Y21(2.41)
F22(2.42)
I23(2.43)
S25(2.45)
L36(2.56)
F52(ECL1)
L59(3.27)
W60(3.28)
I72(3.40)
L76(3.44)
S79(3.47)
I80(3.48)
I86(3.54)
I103(4.46)
V106(4.49)
V115(4.58)
A125(ECL2)
P130(ECL2)
I146(5.40)
L151(5.45)
V159(5.53)
I160(5.54)
V161(5.55)
I162(5.56)
L163(5.57)
I167(5.61)
T174(ICL3)
S175(ICL3)
S191(ICL3)
E192(ICL3)
T193(ICL3)
K194(ICL3)
A208(6.33)
A209(6.34)
L213(6.38)
V215(6.40)
V216(6.41)
V221(6.46)
L224(6.49)
I228(6.53)
E243(7.28)
N244(7.29)
V263(7.48)
N264(7.49)
I266(7.51)
GNAI3
GNAI1
GNAZ
GNAO1
GNA12
GNA13
GNAQ
GNA14
GNA15
GNAS
GNAL
Gi/Go
D
Different AA composition
(p-value <= 0.05)
Deletion
Insertion
E
Gs

`

Figure 2.10: Determinants of G-protein coupling specificity in the 7TM1 domain of the receptors. (A) Bubble matrix of the positional features. (B-E) Structures showing the positional features (Pfam 7tm_1 positions with BW numbering in parenthesis) in receptors specific to each G-protein subfamily. Significant positions are shown in orange (when the amino acid distribution is different in coupled and not-coupled receptors); red (when the position is found as a deletion); green (when the position is found as an insertion).

The statistical analysis of the extra-membrane region (Table S1F) reveals that, besides length, nearly every amino acid property group in ICL3 and C-terminus helps in determining the coupling specificity of G12/G13- and Gq/G11-coupled receptors (Figure 2.11). The specificity of Gi/Go-bound receptors is least influenced by ICL3/C-terminus (Figure 2.11).

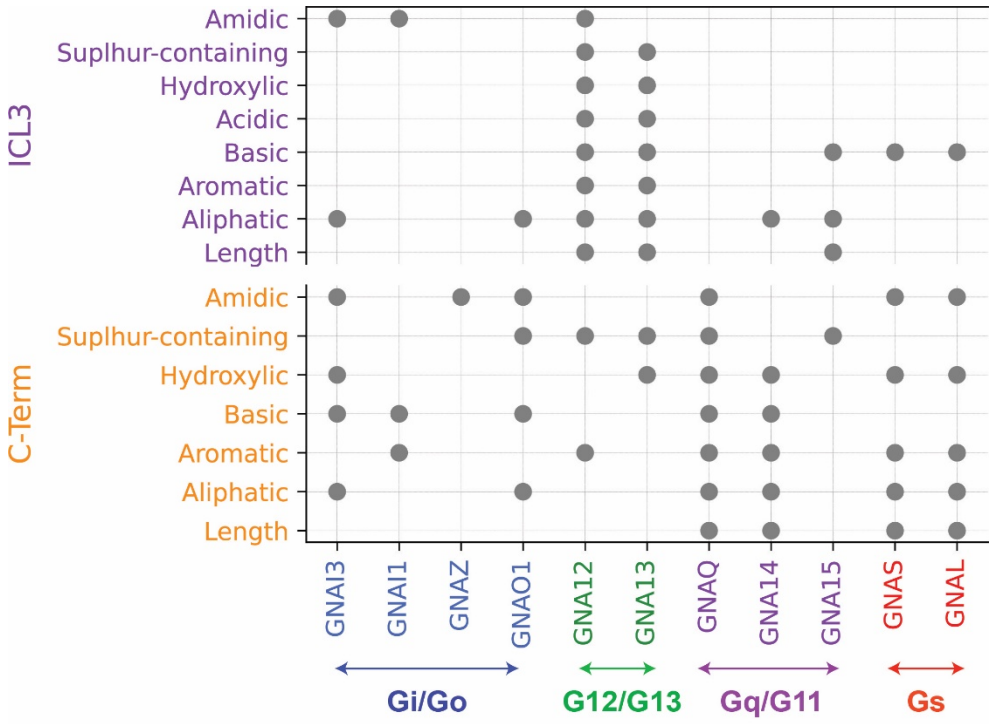


Figure 2.11: Determinants of coupling specificity in the extra-membrane region of the receptors. Length and amino acid composition (y-axis) of ICL3 and C-terminus of GPCRs that were found to be statistically significant (p-value ≤ 0.05) for a G-protein coupling group (x-axis) are shown as bubbles.

2.5 DISCUSSION

The coupling dataset is the largest and most extensive resource of GPCR/G-protein binding affinities to date. It is the first dataset that provides binding affinity at the level of individual G-protein subunits rather than their sub-families. It captures coupling

`

information of receptors that have been poorly characterized in GtoPdb (such as purinergic and protease-activated receptors), and the subtle but key differences in preferential couplings of GPCRs towards the members of the same G-protein subfamily (e.g. prostanoid receptors and dopamine receptors). The coupling dataset contains many receptors that bind to the poorly characterized G12/G13 subfamily, greatly expanding the knowledge of this coupling type. Some of the newly identified G12/G13-coupled receptors (e.g., *CNR1*, *FFAR1*, *GHSR*, *GPR35*, *HRH2*, *HTR2C*) have been implicated in type 2 diabetes, heart failure and hypoxia, inflammation, growth hormone deficiency, obesity (Addy et al., 2008; Divorty et al., 2015; Pantel et al., 2006) and are already targets for agonists approved as therapeutics (Hauser et al., 2017), suggesting possibilities for drug repurposing. Another distinctive attribute of the G12/G13-coupled receptors captured by the coupling dataset is their high promiscuity: G12/G13-coupled receptors are the most promiscuous, the Gi/Go- are the most specific, a remarkable feature also observed in a recent study (Avet et al., 2020).

Given its merits, a critical issue with this experimental dataset is the use of chimeric Gα subunits, which differ only in the last six amino acids of its C-terminus while the rest of the backbone is the same across all the G-proteins. Though recent structural studies have demonstrated the involvement of the Gα backbone in its interaction with the receptors (Carpenter et al., 2016; Draper-Joyce et al., 2018; García-Nafría et al., 2018b; Kang et al., 2018; Koehl et al., 2018), the bulk of evidence continues to support the notion that determinants of coupling specificity in G-proteins lie largely in the C-terminal α5 helix (García-Nafría et al., 2018b). Perhaps the biggest support for this comes from the fact that the coupling dataset agrees very well with known couplings (Figure 2.3). Inarguably, however, the new availability of a GPCR/G-protein coupling dataset that encompasses the native Gα sequences is sure to provide a more complete picture of the determinants of coupling specificity.

Although the determinants of GPCR selectivity in G-proteins have been identified, the determinants of G-protein selectivity in GPCRs are yet to be fully known (Flock et al., 2017). The protocol we presented here provides an approach to identify the residues that enable/disable coupling to a given G-protein. We find these determinants to be present throughout the hydrophobic regions of the 7TM bundle as well as in ICL3 and

`

C-terminus, which is partly corroborated by other studies (Carpenter et al., 2016; Flock et al., 2017; García-Nafría et al., 2018b; Kang et al., 2018; Koehl et al., 2018). This indicates that such positions play a role in allosteric regulation to control and stabilize several intermediate GPCR/G-protein 3D complexes by linking the ligand and G-protein binding sites, a mechanism demonstrated using contact network analysis by Dr. Francesco Raimondi in the paper associated with this thesis (Inoue et al., 2019), and in previous studies (Angelova et al., 2011; Venkatakrishnan et al., 2013, 2016).

The Gs-coupled receptors are predominated by deletions over insertions and the reverse is true for Gi/Go (Figure 2.9A). This can be attributed to the bulkier C-terminal side-chains in Gs compared to Gi/Go, which partly explains why Gi/Go-bound receptors, with narrower crevices, are normally unable to couple to Gs proteins (García-Nafría et al., 2018b; Kang et al., 2018). Many of the determinants we identified are specific – and different – for members of the same G-protein subfamily. This might explain why some GPCRs show diverse couplings to G-proteins that would previously have been the same. Our observation that the length and amino acid composition of the C-terminus contributes towards the Gq/G11 specificity also agrees with the findings of a previous study (Qin et al., 2011).

This chapter presents a systematic, statistically-associated protocol that can be applied to any binding assay dataset to unravel the underlying sequence-based features that regulate the molecular mechanisms of the given interacting proteins. Such a protocol can be (i) complemented with machine learning techniques (Chapter III) to predict several unknown interactions, study variants, or develop engineered proteins (Chapter IV), and (ii) developed into a framework that can be applied to other transducers of GPCRs (β-arrestins or GRKs) or any other binding data where specificity is unknown (Chapter IV).

`

Chapter III: PRECOG (PREdicting COupling probabilities of G-protein coupled receptors)

3.1 ABSTRACT

Machine learning (ML) algorithms are extensively applied to answer many biological questions. Here, we combine the method described in the previous chapter with statistically-associated structural information obtained from 3D complexes to develop a machine learning-guided framework that can predict GPCR couplings. Users can access the PRECOG webserver to (a) predict coupling probabilities of class A GPCRs; (b) visualize receptor sequence and structural features responsible for the coupling, and (c) rationally design a receptor. PRECOG can be freely accessed by academic users at precog.russelllab.org. The adaptability of the framework lends itself to the application on other binding data where the subtype information is unknown.

3.2 INTRODUCTION

Machine learning (ML) techniques have been beneficial at solving several biological problems such as prediction of post-translational modifications (PTMs) (Horn et al., 2014), detection of signal peptide cleavage sites (Almagro Armenteros et al., 2019), engineering proteins (Bedbrook et al., 2019), predicting protein tertiary structures (Senior et al., 2020) and antibiotic discovery (Stokes et al., 2020). ML-based algorithms consider raw features obtained from large, often sparsely annotated data sets, such as binding affinities or a collection of images or genomes, to uncover and exploit intricate patterns buried deep inside them to develop probabilistic models that predict the outcomes on independent datasets.

ML methods have also been applied to predict couplings of GPCRs with G-protein subfamilies using sequence-based features (Cao et al., 2003; Möller et al., 2001; Sgourakis et al., 2005a, 2005b; Yabuki et al., 2005) (see section 1.5 in Chapter I). PRED-COUPLE2 (Sgourakis et al., 2005b) is a publicly available web server that predicts GPCR/G-protein coupling specificity. It suffers from two major drawbacks. First, its predictions are limited to G-protein subfamilies and not G-protein specific. As

`

seen in the coupling dataset (see section 2.4.4 of Chapter II) and databases of known couplings, receptors display varied binding affinities even to the members of the same G-protein subfamily, implying the need to develop a predictor that can capture GPCR couplings at the level of Gα subunits. PRED-COUPLE2's other drawback is that it uses a black-box model, artificial neural networks. Although this outperforms previous predictors, it does not provide mechanistic insights by way of feature relevance. The requirement for an interpretable machine learning model is essential to understand the role played by each feature in determining the outcome (Azodi et al., 2020). This is especially useful in the context of designing receptors such as Designer Receptors Exclusively Activated by Designer Drugs (DREADDs) (Wess et al., 2013).

Protein structure provides insights into function. Though structural information was largely absent for the earlier coupling predictors, multiple structures of GPCR/G-protein 3D complexes have been solved in the last decade (Carpenter et al., 2016; García-Nafría et al., 2018b; Kang et al., 2018; Koehl et al., 2018; Rasmussen et al., 2011), providing a rich source of interaction information, that could potentially be exploited by techniques that assess how sequence changes influence interaction interface structures (Aloy and Russell, 2002; Schymkowitz et al., 2005; Yang et al., 2020a).

Here, we combine the procedure described in Chapter II and statistically significant features derived from structural interfaces to develop an ML-based predictor of GPCR/G-protein couplings (PRECOG). The PRECOG webserver significantly outperforms the previous methods on an unseen, independent dataset derived from GtoPdb (Harding et al., 2018). For a given class A GPCR, PRECOG (a) displays the determinants of coupling specificity for a given G-protein on the receptor sequence as well as structure (known or homologous); (b) predicts the impact of mutations on G-protein specificity, and; (c) suggests point mutations to aid in designing artificial GPCRs (DREADDs) with selective couplings. PRECOG is freely available for academic users at precog.russelllab.org. Lastly, the ML-guided framework holds promise for future use on other protein-protein interaction data to determine subtype specificity using the binding data.

`

3.3 MATERIALS AND METHODS

3.3.1 Dataset

The data from the TGFα shedding assay, hereafter referred to as the coupling dataset, quantifies binding affinities (in terms of LogRAi values) of 144 class A GPCR sequences with 11 chimeric G-proteins (see section 2.4.1 in Chapter II). We performed a ROC curve analysis to obtain the optimal LogRAi cutoff of -1.0 to binarize the binding affinities of the receptors to G-proteins into coupled (LogRAi value ≥ -1.0) or not-coupled (LogRAi value < -1.0) (see section 2.3.1 in Chapter II).

3.3.2 Feature generation

An ML algorithm takes a feature vector as input. The vector encodes the descriptors of the given problem statement. For GPCR/G-protein coupling, we used statistical tests to identify the sequence and structure-based features in receptors as descriptors. The sequence-based features (7TM1 position and extra-membrane) were retrieved as described before (see section 2.3.3 in Chapter II). We obtained the structure-based features using InterPreTS (Interaction prediction through tertiary prediction) (Aloy and Russell, 2002, 2003). For a given pair of putative interactors aligned to their respective homologs, InterPreTS evaluates the fitness (in terms of Z-scores) of any possible interacting pair on a given 3D complex by using the learned parameters of amino-acid pair contacts across protein interfaces (empirical potentials).

We used Pfam accessions PF00001 for GPCRs and PF00503 for G-protein alpha subunit to search for GPCR/G-protein 3D complexes in the SIFTS database (Velankar et al., 2013) (Table 3.1). For each of the six complexes retrieved, we aligned the 144 class A GPCRs to the GPCR chain and the 11 chimeric Gα subunits to the G-protein chain in the given complex structure using ClustalW (Thompson et al., 1994). The FASTA sequence of chimeric Gα subunits was constructed using only the 6 amino acids of the native Gα subunit while the remainder was replaced with that of *GNAQ.* Next, all the pairs of alignments were assessed using InterPreTS, and for every complex, a Z-score was returned for each pair of receptors and chimeric Gα subunit. Finally, we compared the Z-score distribution (Wilcoxon rank-sums test) of coupled receptors with that of not-coupled receptors for every Gα subunit in a complex. For a

`

given G-protein coupling group, only the complexes having a p-value ≤ 0.05 were selected as a suitable template to model the interaction.

PDB ID	GPCR			G-protein		
	Gene symbols	UniProt Accession	Chain	Gene symbols	UniProt Accession	Chain
6DDE	Oprm1	P42866	R	GNAI1	P63096	A
6D9H	CHRM4	P08173	R	GNAI2	P04899	A
	ADORA1	P30542	R			
6GDG	ADORA2A	P29274	A	GNAS	P63092	D
3SN6	ADRB2	P07550	R	GNAS	P04896	A
6CMO	RHO	P08100	R	GNAI1	P63096	A
6G79	HTR1B	P28222	S	GNAO1	P09471	A

Table 3.1: Known GPCR/G-protein 3D complexes (release Jan 2019). All the 3D complexes were extracted from the SIFTS database (release Jan 2019) using their PDB chain to Pfam mappings.

3.3.3 Machine learning

Logistic regression is an interpretable machine learning technique previously applied to other protein interaction problems (Dhole et al., 2014; Dou et al., 2012). We used the Logit/Log-reg classifier (logistic regression classifier) provided by the Scikit-learn library (Pedregosa et al., 2011). A logistic regression model is defined as:

$$g(x) = \sum_{i=1}^{n} w_i \, x_i$$

where x_1, x_2, x_3 ..., x_n represent the input features, w_1, w_2, w_3, ..., w_n represent the regression coefficients (also called feature weights), and n represents the number of features. Thus, the probability of input class A GPCR to couple to G-protein can be modeled in terms of logistic function as:

$$f(x) = \frac{1}{1 + e^{-g(x)}}$$

Regularization is a commonly used technique to prevent overfitting of machine learning models. Logistic regression controls overfitting using a regularization parameter called Lambda (λ), which is proportional to the penalty of finding an overfitting model. To avoid overfitting, Log-reg implements regularization in two forms:

`

L1 (adds product of *λ* and the sum of regression coefficients to the loss function) and L2 (adds product of *λ* and the sum of the squares of the regression coefficients to the loss function). In this study, we used the L2 form, which minimizes the following cost function:

$$\min\left(\frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(e^{\left(-y_i\left(X_i^T w + c\right)\right)} + 1\right)\right)$$

where $c \in R \wedge n$ is the intercept, *C* is the inverse of *λ* and *n* is the number of iterations. The lib-linear method was used as the optimization algorithm since it was shown to work well on small datasets (Fan et al., 2008) such as ours. The weights obtained through logistic regression after the training process can be used to study the importance of features (Dhole et al., 2014; Dou et al., 2012). We exploited this property of the algorithm to understand the contribution of features to every G-protein coupling group.

3.3.4 Training and test sets

For a given G-protein coupling group, a vector of three types of features (a) 7TM1 positional, (b) extra-membrane, and (c) structural features was constructed for each of the 144 class A GPCRs. First, every statistically significant 7TM1 positional feature in the input sequence encoded two bit-scores into the vector, one each from coupled and not-coupled HMM profiles of the given G-protein coupling group (see section 2.3.3 in Chapter II). If a position was an insertion or deletion (i.e. present in only one of the coupled or not-coupled HMMs), the single bit-score obtained from the respective HMM profile was encoded into the vector. If a given 7TM1 position had no amino acid present, the highest bit-score (considering both the HMM models) was encoded into the vector. Second, for a given G-protein coupling group, the extra-membrane features (statistically significant length and amino acid composition of the ICL3 and C-terminus) of GPCRs were encoded into the vector. Third, Z-scores of statistically significant complexes for every coupling group were obtained from InterPreTS (default parameters; 100 random permutations) and encoded into the vector. This way a training matrix was created for every G-protein coupling group.

Feature scaling of the training matrix aids in converging the algorithm faster and computing the feature relevance (Dou et al., 2012). All the features in the training

`

matrices were scaled in the range of 0 to 1 using the *MinMaxScaler* function of the Scikit-learn library (Pedregosa et al., 2011). The binarized coupling information (1: coupled and 0: not-coupled) for a given G-protein coupling group was added as the last column to every training matrix. Thus, the training set comprises 11 training matrices, one for each G-protein coupling group.

To compare the performance of our predictor with that of PRED-COUPLE2 (Sgourakis et al., 2005b), a publicly available web-server of GPCR/G-protein couplings, we retrieved 86 class A GPCRs that were neither included in the coupling dataset nor in that of PRED-COUPLE2. For this independent test set, 11 test matrices (one for each G-protein coupling group) were created, each containing 86 vectors encoding the features described above for the corresponding G-protein coupling group. Every test matrix was transformed using the feature-scaling parameters obtained from the corresponding training matrix of the given G-protein coupling group.

3.3.5 Cross-validation and metrics

We then performed a grid search using stratified 5-fold cross-validation (5-fold-CV) (available from the Scikit-learn library) to obtain the optimal value of *C* (inverse of *λ*). Given the imbalanced nature of the dataset, we set the *class_weight* parameter to *balanced*, which automatically adjusts the weights of the classes (in this case, coupled and not-coupled receptors) inversely proportional to their frequencies in the training matrix. Next, we randomly divided every training matrix into 5 equal stratified subsets, preserving the class ratio between the number of coupled and not-coupled receptors. During each fold, we treated one of the subsets as the validation set while the remaining four as the training set. We chose Area Under the Curve (AUC) as the metric to select the best model (hyperparameters) for every G-protein coupling group. To ensure that any random division of the training matrix will give similar results (minimum variance) during the cross-validation process, we repeated the experiment ten times. The performance of our predictor was assessed using standard metrics (Table S2A):

$$Matthews\ correlation\ coefficient\ (MCC) = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

$$Accuracy\ (ACC) = \frac{tp + tn}{tp + fp + tn + fn}$$

`

$$Precision\ (PRE) = \frac{tp}{tp + fp}$$

$$Recall\ (REC) = \frac{tp}{tp + fn}$$

$$Specificity\ (SPE) = \frac{tn}{tn + fp}$$

$$F_1 - measure\ (F1M) = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

3.3.6 Randomization test

We performed a randomization test (Salzberg, 1997), which has previously been used to assess overfitting (Murakami and Mizuguchi, 2010). For every G-protein coupling group, we randomly shuffled the original classes of the training matrix while preserving the class ratio between the number of coupled and not-coupled receptors. Next, we performed the cross-validation (see the previous section) on the randomized dataset to make 11 predictive models. The performance of the models created using the randomized dataset was assessed using the same metrics as described in the last section.

3.3.7 Workflow

To make predictions, the predictor performs steps shown in Figure 3.1A. First, the input GPCR sequence is aligned to the Pfam 7tm_1 HMM profile using the *hmmalign* tool of the HMMER package (v3.1b2) (Eddy, 1998). Using the alignment results, the determinants of coupling specificity corresponding to all the G-proteins are mapped to the input sequence and extracted. Second, InterPreTS is run in parallel to calculate the Z-scores for every input GPCR using each of the six best 3D complexes. As an input to InterPreTS, the PDB file of the 3D complex into consideration and MSA of the input receptor(s) along with that of the chimeric G-proteins are provided.

For a given G-protein, a vector consisting of corresponding sequence-based features (7TM1 positions and extra-membrane) and structure-based features (Z-score derived from InterPreTS), is constructed and its probability is calculated using its corresponding model. If mutations are given, then for each the concerned position of

`

the receptor sequence is changed before the feature matrix is constructed for a given G-protein.

3.3.8 Webserver

We developed the PRECOG web server by using the Flask web framework (Flask, 2018). The internal pipeline to execute the workflow was implemented using Python. To view positions of coupling specificity, the input sequence is aligned to all the 3D structures of class A GPCRs using BLAST (Altschul et al., 1990). The 3D structures were obtained from the PDB chain to Pfam domain mappings provided in the SIFTS database (Velankar et al., 2013). At the front end, we used several JavaScript libraries along with neXtProt sequence viewer (Gaudet et al., 2017) to view input sequence(s) and JSmol (Hanson et al., 2013) to view 3D protein structure.

3.4 RESULTS

3.4.1 3D complex information

In the last five years, the number of available GPCR/G-protein 3D complexes has gone from 1 to 6 (Carpenter et al., 2016; Draper-Joyce et al., 2018; García-Nafría et al., 2018b; Kang et al., 2018; Koehl et al., 2018; Rasmussen et al., 2011). Thus, we sought to use InterPreTS to investigate if these complex structures can be used as templates to model GPCR/G-protein couplings (Aloy and Russell, 2002, 2003) (see Methods). As expected three complexes (PDB IDs: 3SN6, 6GDG, and 6G79) are most suitable for modeling the Gα subunits in the coupling dataset that match those in the complexes (i.e. *GNAS*, *GNAL*, and *GNAI1*, respectively). Intriguingly, two additional GPCR-Gi/Go complexes (PDB IDs: 6CMO and 6DDE) are also good for modeling GPCR-G12/G13 couplings (Table 3.2).

PDB ID/G-protein	GNAS	GNAS	GNAI1	GNAI1	GNAI2	GNAO1
	3SN6	6GDG	6CMO	6DDE	6D9H	6G79
GNAI3	0,3024	0,2689	0,3346	0,8	0,5459	0,0869
GNAI1	0,1097	0,1777	0,2532	0,7891	0,6439	0,0211
GNAZ	0,761	0,4352	0,67	0,7671	0,4938	0,5142
GNAO1	0,2898	0,5427	0,2355	0,692	0,7901	0,1435

`

GNA12	0,5974	0,1736	0,0138	0,2057	0,0654	0,4283
GNA13	0,4886	0,5341	0,0023	0,0232	0,0685	0,343
GNAQ	0,1088	0,8756	0,7527	0,8346	0,7802	0,3253
GNA14	0,3056	0,9542	0,9804	0,3544	0,9837	0,0871
GNA15	0,7908	0,3556	0,0831	0,6289	0,8608	0,8187
GNAS	0,0108	0,0416	0,5957	0,2185	0,32	0,8861
GNAL	0,0125	0,1298	0,5147	0,2586	0,3726	0,5344

Table 3.2: Statistical associations of GPCR/G-protein 3D complex. Statistical significance of InterPreTS scores derived from GPCR/G-protein 3D complexes (columns) and the TGFα shedding assay couplings (rows) (scores with p-values ≤ 0.05 are highlighted in green).

3.4.2 Machine learning-based predictor

PRECOG was trained on one of the most quantified datasets of GPCR/G-protein couplings (Chapter II; Figure 3.1A, B). We created two versions of the tool: one trained only with sequence-based features, and another trained with both sequence and structure-based features. PRECOG returns G-protein coupling specificities in terms of probabilities, where a probability greater than 0.5 indicates coupling. The performances of both versions of PRECOG were compared to PRED-COUPLE2 (Sgourakis et al., 2005b) on an independent test set comprising 86 class A GPCRs reported in GtoPdb (Figure 3.1C) (see Methods) but absent from the datasets used to train both methods. For both, only the four G-protein subfamilies are considered when evaluating the predictions (since PRED-COUPLE2 only predicts these). Grouped PRECOG prediction for subfamilies was considered to be positive if at least one member was predicted to couple a given receptor. Due to the lack of a true negative set, we used recall (or sensitivity) as the metric to compare the performances of the two predictors (Figure 3.1C). PRECOG (both versions) significantly outperformed PRED-COUPLE2 on the test-set, demonstrating the relevance of the coupling dataset (Table S2B). However, the addition of structural information only marginally improved the performance of the Gs subfamily models whereas remain unaffected for the other G-protein subfamilies (Figure 3.1C). Moreover, considering both versions, PRECOG performed poorest for the Gs subfamily (Figure 3.1C).

`

We also performed a randomization test to assess overfitting by shuffling the last column that represents the coupling labels (coupling or not) and then repeated the training and cross-validation procedure (see Methods). The performance of the predictive models developed using the randomized dataset was worse than PRECOG, implying that our strategy is unbiased to the training data (Table S2C).

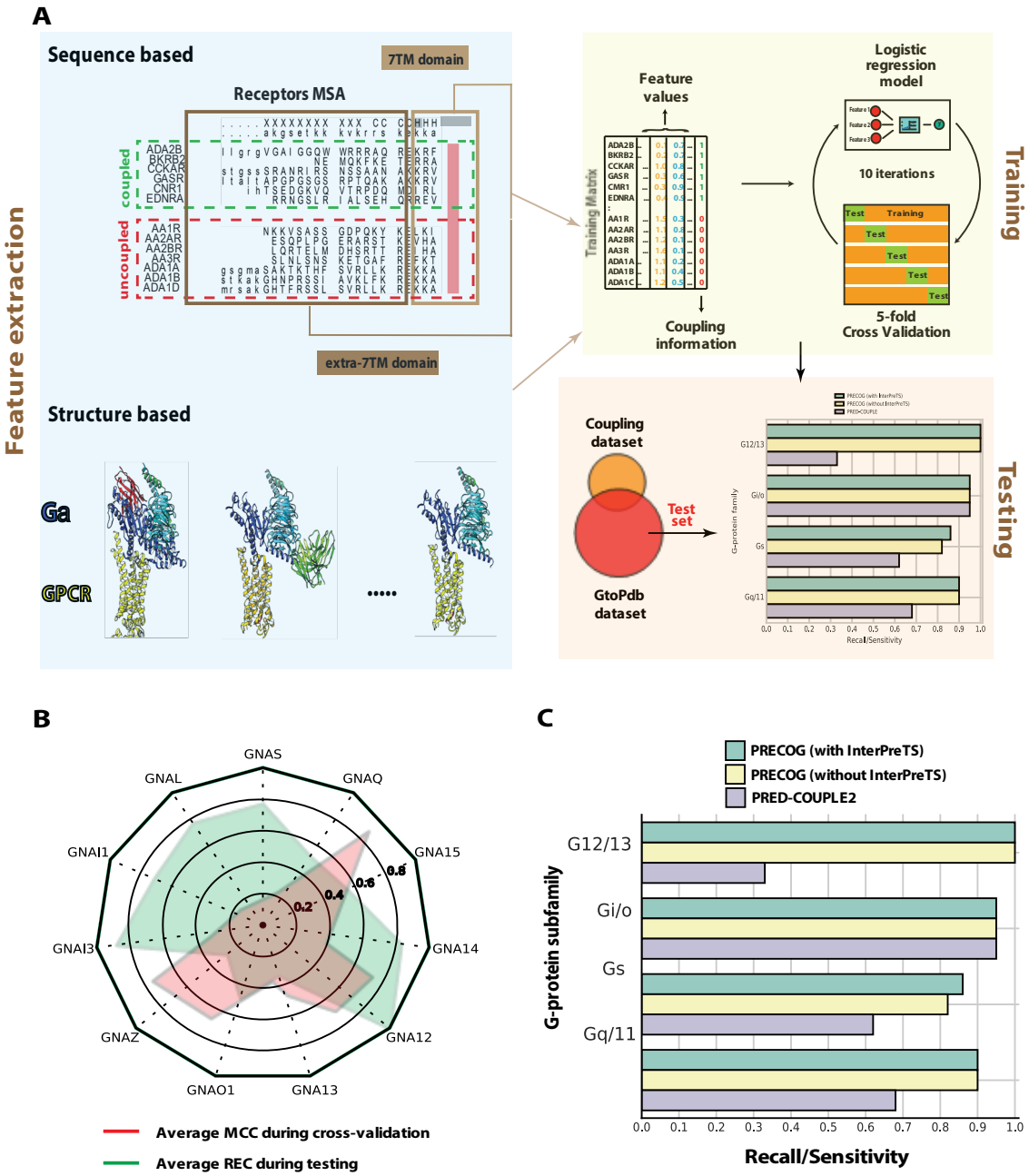


Figure 3.1: Workflow and performance of PRECOG. (A) From all the GPCR sequences in the coupling dataset, sequence- and structure-based features are extracted and used by the logistic regression algorithm to perform the training and cross-validation procedures for every G-protein

`

coupling group. The resulting predictive models are used to predict couplings of 86 unseen GPCRs that are absent from the training set of both PRECOG and PRED-COUPLE2 but present in GtoPdb. (B) Performance of PRECOG during the cross-validation and testing process. (C) Recall/Sensitivity of PRECOG over the 86 unseen GPCRs in the test set for the G-protein subfamily.

3.4.3 Importance of feature relevance

Logistic regression can be used to evaluate feature relevance (Dhole et al., 2014; Dou et al., 2012). In this study, we used the regression coefficients (also called feature weights) (see Methods) computed from the trained models of each G-protein coupling group to construct a feature weight matrix (Figure 3.2). For a given G-protein coupling group, the higher the absolute value of the feature weight, the higher the relevance of the feature. The coefficients can be either positive or negative (Figure 3.2). The feature weight matrix can be used to understand feature contribution either across or within G-protein subfamilies.

On comparing across the G-protein subfamilies, the same feature might be a determinant of coupling specificity to more than one G-protein but its feature weight might not have the same sign for all the G-proteins. For example, a 7TM1 position feature, E243 (BW: 7.28), is an insertion to Gi/Go, G12/G13, and Gq/G11 subfamilies (Figure 3.2) but the sign of its feature weight differs. The feature weights are negative for the Gi/Go subfamily, implying that a receptor with an amino acid of low feature value (i.e. bit-score in this case) is more probable to couple to the G/Go subfamily than a receptor with a higher feature value. The reverse is true for the G12/G13 and Gq/G11 subfamilies (positive feature weight; colored green in Figure 3.2). Note that the bit-score of a receptor for the concerned position is obtained from the HMM profiles of coupled/not-coupled GPCRs in a given G-protein coupling group. Another feature of similar nature is the length of ICL3. While a high feature value (i.e. length in this case) favors coupling to *GNA13* (positive feature weight; colored green in Figure 3.2), the reverse is true for *GNAQ* (negative feature weight; colored red in Figure 3.2).

On comparing within a coupling group, the features might have opposing weights. For example, two 7TM1 positional features, S191 and E192, have opposing effects (negative and positive feature weights, respectively) on probabilities of coupling to *GNAI1*. In other words, a receptor with an amino acid of low feature value (i.e. bit-

`

score in this case) at S191 is more probable to couple to the subfamily than a receptor with a higher feature value. However, the reverse is true for E192.

Therefore, the feature weights obtained from the trained models of G-protein coupling groups can be exploited to (i) unravel the underlying relevance of each feature specific to recognizing a particular G-protein; (ii) assess the impact of variants/mutations in GPCRs, and; (iii) design receptors that selectively coupled to one or more G-protein subfamilies (eg: DREADDs).

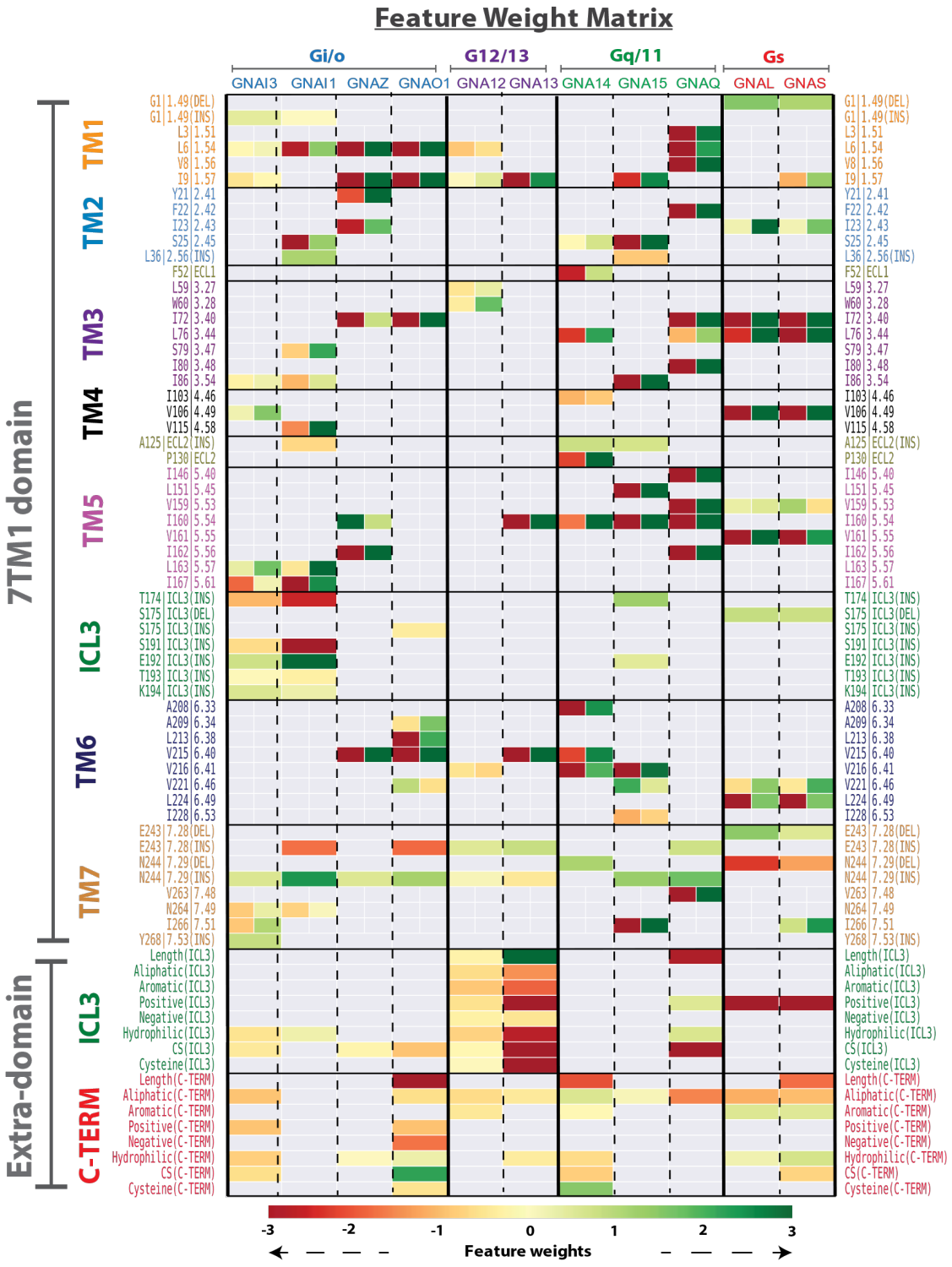


Figure 3.2: Feature weight matrix of PRECOG. The figure is taken from Singh et al., 2019. A heatmap showing contribution of statistically associated sequence-based features (x-axis) of GPCRs to at least one G-protein coupling group (y-axis). Cells are colored based on the coefficients (also called feature weights) of the given feature in the best-performing model of the corresponding interacting group (red-green scale corresponding to negative and positive weights, respectively). Color intensities of cells indicate the absolute value of the coefficients. If a significant 7TM domain position is present in both

`

coupled and not-coupled HMMs, its coefficients are shown within the same cell on left (coupled) and right (not-coupled).

3.4.4 Web-server

To aid the visualization of the determinants of G-protein coupling specificity in GPCRs, we deployed several open-source tools (see Methods) to develop the PRECOG web server. The web server consists of an input page, where the user enters GPCR sequence(s) in various formats and chooses parameters. After inputs have been processed, an output page shows a detailed summary of predictions in a tabular format along with panels to view the sequence and structures of the input or its homologs.

Input page

The input page provides the user with two options (Figure 3.3). The first is simply to make predictions of coupling for one or more GPCRs (sequence human sequence specified by UniProt identifiers, accessions, or gene symbols). Users can predict coupling probabilities of the input GPCR (wild type or mutated) to each of the 11 G-proteins from the coupling dataset. The user can also choose to display selected mutations by modifying the minimum probability difference threshold (wild type minus variant).

The other option is to design a new GPCR. Here the server will suggest the variants/mutations that are likely to alter input GPCR's coupling probabilities to one or more user-selected G-protein subfamilies. The user can select the desired G-protein subfamily (subfamilies) with help from the checkboxes provided. There are also controls to display selected variants/mutations by defining the values of P(coupling), the probability difference threshold (as above) for coupling with at least one selected G-protein, and of P(uncoupling), the same but for G-proteins other than that selected (for selectivity). The output page will display rows that satisfy either or both conditions.

`

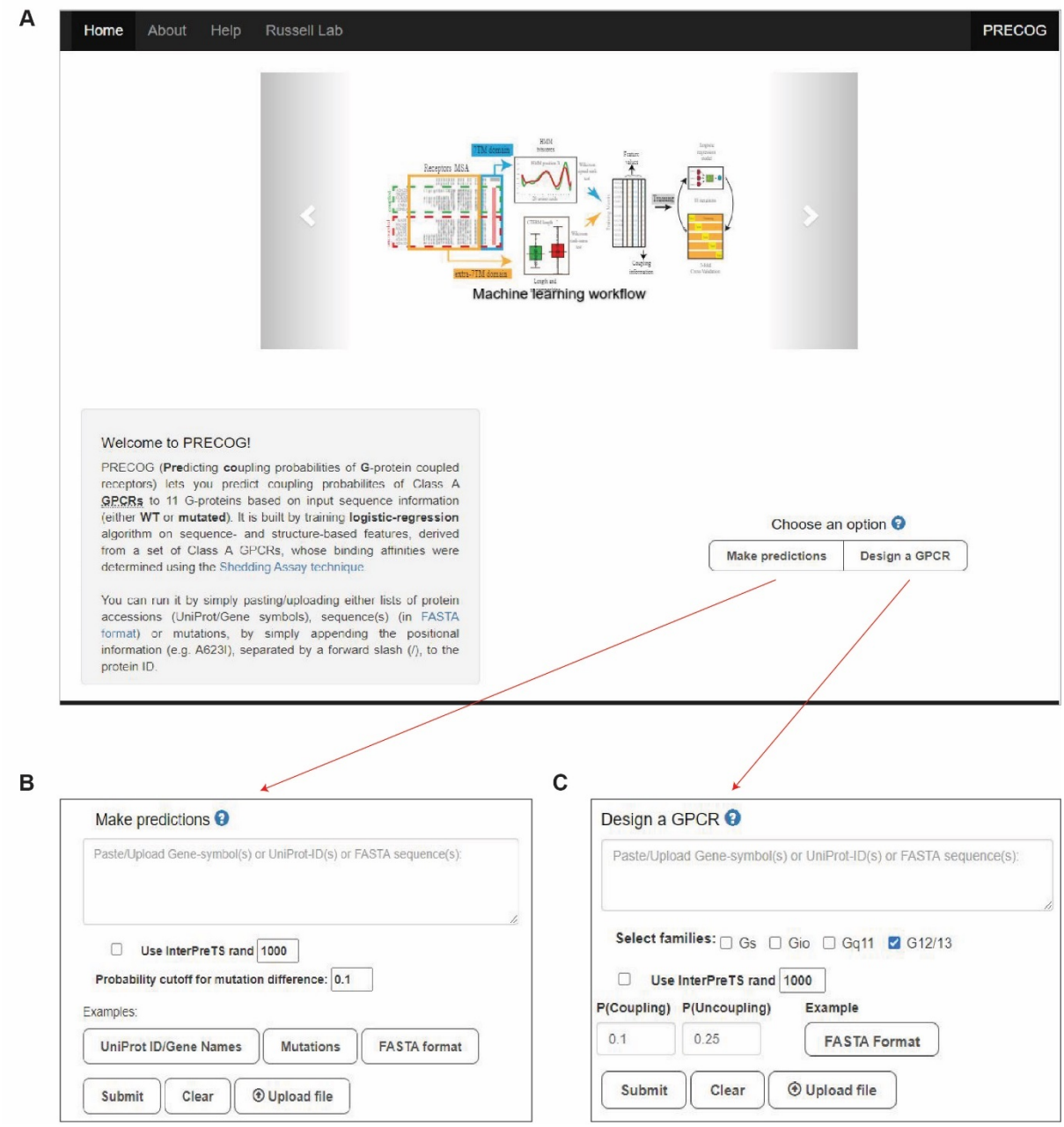


Figure 3.3: Input page of PRECOG. (A) Home page (http://precog.russelllab.org/). Available Options: (B) Make predictions and (C) Design a GPCR.

The user has the option of running either version of PRECOG (with or without InterPreTS) for both options (predicting coupling or GPCR design). The input can be UniProt accessions, gene symbols, mutations, and/or the whole sequences (FASTA formatted). Examples covering all the possible input formats are provided on the main page, which automatically fills the input boxes when selected.

`

Output page

The output page displays all the sequence-based features that are statistically associated with a given coupling. It has three panels (Figure 3.4). The Results panel shows the predicted coupling probabilities of the given input (both Wild type and/or mutations) with 11 G-proteins by PRECOG and the information known from GtoPdb and the TGFα shedding assay are displayed in a tabular format (Results panel). Each G-protein is assigned a fixed color based on their subfamily (Gi/Go: cyan; G12/G13: green; Gq/G11: yellow and; Gs: pink) and the same in uniform on the entire page. The predicted (by PRECOG) and known couplings (from GtoPdb and the TGFα shedding assay) are colored green and red if they are coupled and not coupled, respectively. Several checkboxes are provided at the top of the panel to narrow down the rows. The entire panel can be downloaded by the user as a file in TSV (tab-separated values) format.

The Sequence panel shows sequence-based determinants of coupling-specificity, and the mutations (if provided) for a selected GPCR/G-protein coupling pair (by default, the pair with the highest probability is shown). The user can further click on each of the determinants (sequence-based features) to open a bar plot that displays the distribution of amino acids at the given position, or if it is an insertion/deletion. This panel also supports regular expression (RegEx) searches.

The Structure panel shows determinants of coupling-specificity for wild type and mutants for each input GPCR (shown in the Sequence panel) mapped onto a three-dimensional structure. If the user clicks on a position (mutation or determinants of coupling-specificity) in the sequence panel, the most sequence-similar structure to the given GPCR is chosen and displayed. The user also has the option of choosing a different structure from the dropdown menu PDB ID in the panel. The values in the parenthesis represent the percentage of sequence identity of the input GPCR with the structure. GPCR/G-protein 3D complexes are highlighted with coral color in the dropdown menu. The user also has the option to toggle between determinants of coupling-specificity of more than one G-protein subfamily using the checkboxes in the panel.

`

Figure 3.4: Output page of PRECOG. (A) Results panel: For any input receptor, three rows are displayed- P(WT), IUPHAR (GtoPdb), and LogRAi. P(WT) indicates predicted coupling probabilities (values ≥ 0.5 are highlighted in green, otherwise in red), IUPHAR indicates known coupling values from GtoPdb (PC: Primary Couplings; SC: Secondary Couplings). LogRAi indicates binding affinities values known from the TGFα shedding assay (a LogRAi value ≥ –1.0 indicates coupling, otherwise not-coupling). (B) Sequence panel: Statistically-associated residues of a selected input sequence that are determinants of its coupling specificity with a given G-protein are highlighted. On clicking these residues, a bar plot with the distribution of amino acids at the given position in coupling vs non-coupling receptors is displayed. The ICL3 and/or C-terminus of the input sequence is highlighted if its length is significant for prediction. The amino acids in these stretches are underscored if their occurrence is significant for prediction. The users can click on the significant 7TM1 residues to view a bar plot with the distribution of amino acids at the given position in coupled vs not-coupled receptors and to view it on a 3D structure (or the corresponding position of the 3D structure of the closest homology). (C) Structure panel: 3D

`

structure of the input sequence (or of the closest homolog) is displayed and the significant 7TM1 positions are highlighted.

We tested the compatibility of PRECOG on different browsers (Table 3.3).

OS	Version	Chrome	Firefox	Safari
Linux	Ubuntu 16.04	not tested	Quantum 64.0	n/a
Linux	CentOS 7.2.1511	not tested	ESR 52.6.0	n/a
Windows	10	v71	not tested	n/a
macOS	Mojave 10.14.2	v71	v64.0	12.0

Table 3.3: Browser compatibility of PRECOG.

3.5 DISCUSSION

In this chapter, we present PRECOG, a machine learning-guided predictor trained on one of the most extensive GPCR/G-protein binding datasets, that can predict GPCR/G-protein selectivity, using sequence information only (Figure 3.5). The PRECOG web server has several advantages over the previous methods. It significantly outperforms the previous predictors as mentioned above and predicts coupling probabilities at the resolution of Gα subunits rather than the G-protein subfamilies. This approach is unique in that it provides easy visualization of the positions of G-protein coupling specificity on GPCR sequences and structures (known or homologous). Moreover, PRECOG is the only approach that can predict the effects of variants, which is particularly useful in the design GPCRs with desired coupling properties (DREADDs). PRECOG is freely available to academic users at precog.russelllab.org.

Integration of features derived from structural information only marginally improved the performance of PRECOG for the Gs subfamily while it remains unaffected for the remainder of G-protein subfamilies. Currently, the GPCR/G-protein 3D complexes are only available for Gi/Go and Gs subfamilies. Inclusion of GPCRs in 3D complex with Gq/G11 and G12/G13 subfamilies in the future will likely improve the performance of these subfamilies. PRECOG performed poorest in predicting Gs-coupled receptors. This can partly be attributed to a poor overlap between the coupling dataset and

`

GtoPdb (see Figure 2.5C in Chapter II). Another reason may be lack of features that are very specific to Gs-coupled receptors such as the dynamics of TM helices, especially TM6, which have been observed to undergo a distinctive outward displacement in the Gs-stabilized receptors in contrast to other complexes (Carpenter et al., 2016; Kang et al., 2018; Koehl et al., 2018; Rasmussen et al., 2011).

Interestingly, two GPCR-Gi/Go complexes (PDB IDs: 6CMO and 6DDE) are also good templates for modeling GPCR-G12/G13 couplings (Table 3.1), indicating a similar interaction topology between the two subfamilies. Additionally, in the coupling dataset, the G12/G13-coupled receptors majorly couple to the Gi/Go subfamily (see Figure 2.5A in Chapter II). This suggests that GPCR-Gi/Go complexes can also be used to model GPCR-G12/G13 interactions.

The machine learning-guided framework described here is adaptable in two significant ways. First, it can be extended to uncover specificity determining positions in other protein families where binding information is available, but sub-type specificity information is unknown or incomplete. Second, it provides the possibility to include disordered regions of the sequence that can influence the specificity but cannot be otherwise captured by MSA and are very often not seen in protein structures. Finally, the framework will enable the development of web servers to interrogate predictions and their associated mechanistic insights.

`

PRECOG

http://precog.russelllab.org/

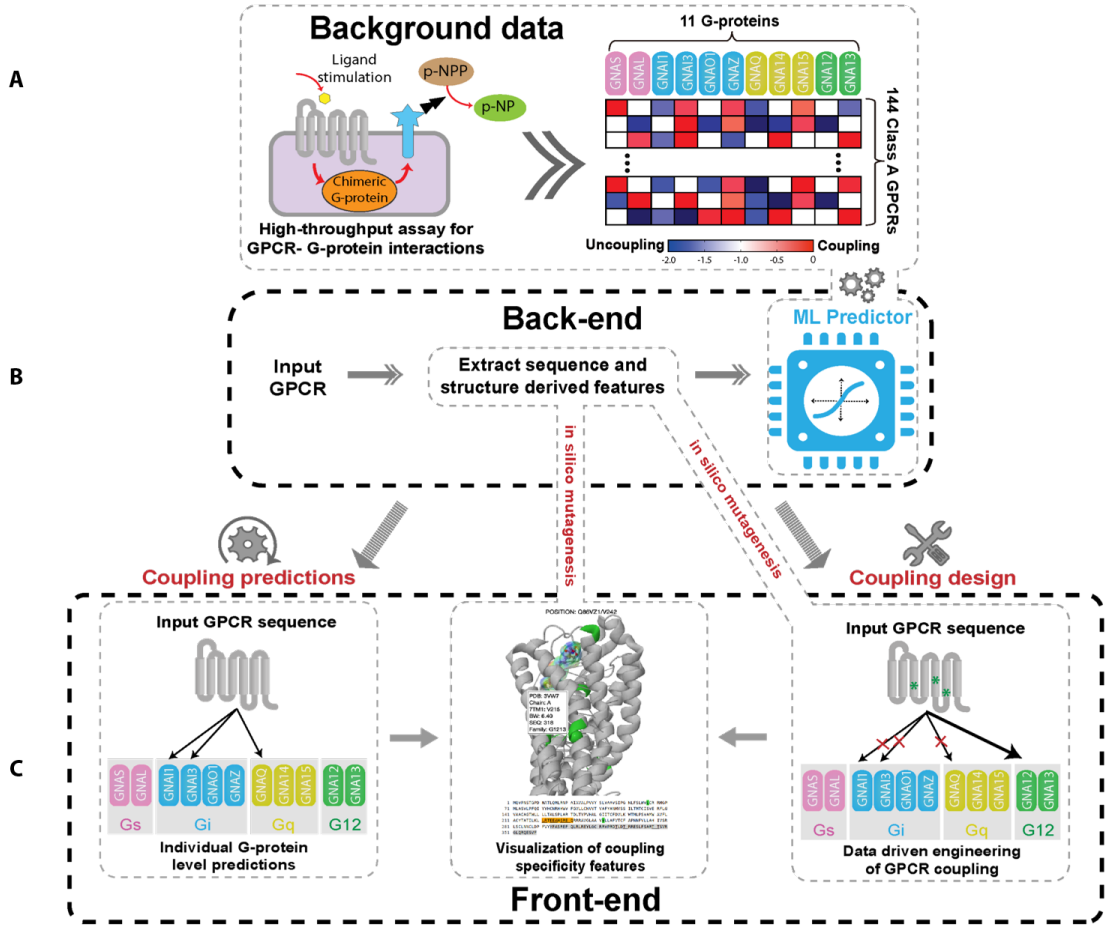


Figure 3.5: Overview of PRECOG webserver. The figure is taken from Singh et al., 2019. (A) Quantification of GPCR/G-protein binding affinities. (B) Extraction of sequence and structure-based features in GPCRs that recognize their G-protein selectivity, and development of ML-based predictor. (C) Prediction of uncharacterized GPCRs and design of artificial receptors with the aid of sequence and structure visualization tools.

`

Chapter IV: Applications of PRECOG and the machine-learning guided framework

4.1 ABSTRACT

GPCRs are of immense pharmaceutical interest. This is because they are primary points where external stimuli contact virtually every human cell, and tinkering with them with small molecules can alter a multitude of processes that can potentially treat a host of conditions. This crucial role as gatekeepers of cell signaling also provides interesting possibilities to design tools to perturb biological systems by way of synthetic biology (protein design). In this chapter, we apply PRECOG and the framework to three subjects. First, we study and predict coupling probabilities of several poorly characterized GPCRs (e.g. *P2RY8* implicated in cancer) and disease mutations in GPCRs. We then use PRECOG to design GPCR hybrid molecules to produce a new chemogenetic tool capable of specifically activating *GNA12* (and Rho signaling). Finally, we apply the entire machine learning framework to a new dataset on G-protein and β-arrestin interactions to predict specificity determinants of this important regulator of GPCR signaling.

4.2 INTRODUCTION

GPCR signaling controls a myriad of cellular pathways. Mutations in GPCRs are implicated in several diseases such as nephrogenic diabetes insipidus, cardiovascular and mental disorders, hyperthyroidism, fertility disorders, AIDS, and cancer (Insel et al., 2007). These aberrations can alter both the ligand binding or the coupling properties of receptors. Interestingly, mutations can also be introduced in GPCRs to develop chemo-genetic tools such as the Designer Receptors Exclusively Activated by Designer Drugs (DREADDs) that can hijack GPCR/G-protein couplings by allowing the use of synthetic ligands to control specific couplings in a controlled manner (Alexander et al., 2009; Armbruster et al., 2007; Farrell et al., 2013)

DREADDs were developed by screening several variants of muscarinic receptors that exhibit minimal constitutive activity and low native ligand (acetylcholine) affinity *in vitro*

`

and *in vivo* (Armbruster et al., 2007). They display high ligand affinity and specificity to synthetic ligands such as clozapine-N-oxide (CNO), which is an otherwise pharmacologically inert drug. The three most commonly used DREADDs are hM3D (activates Gq/G11 signaling) (Alexander et al., 2009), hM4D (activates Gi/Go signaling) (Armbruster et al., 2007; Stachniak et al., 2014), and rM3D (activates Gs signaling) (Farrell et al., 2013). However, a G12/13-coupled DREADD is still unavailable making it an attractive target to test the potential GPCR design tools in PRECOG.

The phosphorylation of the C-terminus of GPCRs by GPCR kinases (GRKs) recruits arrestins to the receptor. Arrestins, particularly β-arrestin-1 and β-arrestin-2 compete with G-proteins for receptor binding through steric hindrance and act as a rheostat of G-protein initiated signaling (Gutkind and Kostenis, 2018). Recent structures have revealed GPCR 'megaplex' that can simultaneously activate both G-protein and β-arrestin during receptor internalization (Nguyen et al., 2019). However, the exact determinants of β-arrestin specificity for GPCRs are still unknown.

An important utility of PRECOG is the application of its machine learning-guided framework to identify regions within the MSA of large protein families (such as GPCRs) that illuminate specificity to another interacting group of molecules (for example G-proteins or β-arrestins in the case of GPCRs) and to use these determinants to predict the specificity of uncharacterized members of the protein family. To test the general applicability of this framework, we applied it to the recently published data from the ebBRET assay (Avet et al., 2020), comprising binding affinities of 85 class A GPCRs with that of 12 G-proteins and β-arrestins1/2 (in the presence/absence of *GRK2*).

This chapter thus describes the application of PRECOG and its framework to predict coupling probabilities of uncharacterized GPCRs, to predict the effect of GPCR mutations on G-protein selectivity, to design the first *GNA12*-coupled DREADD, and to interrogate a different interaction dataset (the β-arrestin data).

`

4.3 MATERIALS AND METHODS

4.3.1 Prediction of uncharacterized and mutant receptors

To obtain a list of uncharacterized GPCRs, we considered all the class A GPCRs in GtoPdb (Harding et al., 2018) and selected those lacking primary and/or secondary coupling information (61 receptors). We then predicted the coupling probability of each of them using the *Make predictions* option of PRECOG (see section 3.4.4 of Chapter III). To analyze the results at the G-protein subfamily level, we grouped the predicted probabilities of G-proteins in each subfamily. A receptor was considered to couple to a given G-protein subfamily if it was predicted by PRECOG to couple to at least one member of the subfamily ($P \geq 0.5$, where P is the probability of G-protein coupling specificity predicted by PRECOG).

Next, we used PRECOG to predict the coupling profile of mutant GPCRs obtained from annotated disease-causing mutations in UniProt (Bateman et al., 2017). We assumed a mutation to affect GPCR coupling with any G-protein if the absolute difference between the predicted coupling probabilities of mutation and the wild type is at least 0.1 (i.e. absolute value of $P_{MUT} - P_{WT}$, where P is the probability of G-protein coupling-specificity predicted by PRECOG).

4.3.2 Prediction of *GNA12*-coupled DREADD chimeric sequences

The lack of a GPCR-G12/G13 complex as well as a DREADD to investigate the G12/G13 signaling impelled us to generate chimeric sequences that exhibit coupling specificity towards the G12/G13 subfamily. We started with the previously developed hM3D DREADD that couples to Gq/G11 (Armbruster et al., 2007). The work in previous chapters suggests that the ICL3 and C-terminus play a prominent role in determining the specificity of G12/G13 receptors (see Figure 2.11 of Chapter II). We thus constructed chimeras by swapping in ICL3 alone or combined with a swap of the C-terminus segments from other GPCRs in the coupling dataset.

Next, we aligned the sequences of hM3D, and all the receptors in the coupling dataset to the Pfam 7tm_1 HMM model using the *hmmalign* tool of the HMMER package (v3.1b2) (Eddy, 1998). We defined ICL3 as the regions between positions 173 and 205 in the Pfam alignment and the C-terminus as those after position 268. We then

`

constructed the corresponding 296 sequences with their ICL3 alone or in combination with the C-terminus swapped with those of hM3D.

We used the *Design a GPCR* option of PRECOG (see Section 3.4.4 of Chapter III) to predict the coupling probabilities of each construct with that of the G12/13 subfamily (*GNA12* and *GNA13*). The constructs were subsequently ranked according to their relative coupling probability (i.e., $\Delta P = P_{DREADD\_MUT} - P_{DREADD}$, where P refers to coupling probability predicted by PRECOG). The top 10 scoring chimeric sequences (from a total of 13 GPCRs) were then selected for experimental validation.

The validation experiments were performed by Dr. Asuka Inoue and his group (Tohoku University, Japan), and have been published elsewhere (Inoue et al., 2019). Briefly, Dr. Inoue's group performed the TGFα shedding and Nano-BiT-G assays to determine *GNA12* and *GNA13* activation. Flow cytometry was used to test the cell surface expression of the constructs.

4.3.3 Development of ebBRET assay-based predictor

The dataset from the ebBRET assay (Avet et al., 2020) comprises binding affinities of 100 G-protein coupled receptors (GPCRs) including 85, 10, and 5 class A, B, and C receptors, respectively, with 11 G-proteins and 2 β-arrestins (Table S3C). The binding affinities are measured in terms of E_{max} values, which refers to the maximum value of ligand-induced response achieved. We considered only the 85 Class A GPCRs for the development of the predictor. As described before (see Section 3.3.2 of Chapter III), we extracted the sequence- and structure-based determinants of G-protein/β-arrestin specificity in the 85 class A GPCRs. Briefly, we created an MSA with 84 class A receptors validated in the ebBRET assay. Next, we subdivided the MSA based on G-protein/β-arrestin interaction preference. For a given interacting group (G-proteins/β-arrestins), if a pair of receptor-interactor scored $E_{max} > 0$, the receptor was considered to interact with the corresponding group; if the pair scored $E_{max} = 0$, it was considered not interact. To obtain the structure-based features, we calculated the statistical association of each interacting group with their corresponding structures using InterPreTS (Aloy and Russell, 2002, 2003) as described before (see section 3.3.2 in Chapter III).

`

The statistically associated (p-value ≤ 0.05) sequence- and structure-level features were extracted to train the Logistic-regression model (see section 3.3.3 in Chapter III) for each interaction group. All the 7TM1 positional features were assigned their corresponding Pfam 7tm_1 position and BW numbering. The most conserved position within each helix was defined according to GPCRdb (Isberg et al., 2016). For positions lying in extra-membrane regions (i.e. where no BW numbering is possible), we quote only the Pfam 7tm_1 position.

We created an independent test-set of 140 class A GPCRs that are characterized in GtoPdb (Harding et al., 2018) but absent in the data from the ebBRET assay. In the case of β-arrestins, we used the STRING (combined score > 600) (Szklarczyk et al., 2019), HIPPIE (Alanis-Lobato et al., 2017), and IMEx (Orchard et al., 2012) databases to collect 57 unique class A GPCRs that interact with β-arrestins 1 or 2. 23 of these receptors were absent in the data from the ebBRET assay, and thus, used as an independent set to test the performance of the predictor. The performance of the ebBRET assay-based predictor during cross-validation and testing was evaluated using the metrics described before (see section 3.3.5 in Chapter III). To compare the performance of the ebBRET assay-based predictor with that of the coupling dataset-based predictor (PRECOG), we created an independent test-set of 71 GPCRs obtained from GtoPdb but are absent in both datasets.

To obtain the GPCR- G-protein/β-arrestin interfaces, we followed the procedure described before (see section 2.3.4 in Chapter II). For the GPCR/β-arrestin interfaces, we considered all the PDB complexes (Berman et al., 2000) that contained the Pfam accessions PF00001 (7-transmembrane receptors) for GPCRs and PF00339 (Arrestin, N-terminal domain) or PF02752 (Arrestin, C-terminal domain) for arrestins using the Pfam to PDB chain mappings from the SIFTS database (Velankar et al., 2013). A total of 31 complexes of GPCR/G-protein and 4 complexes of GPCR/β-arrestin were obtained (Table 4.1). We defined interfaces on the GPCR chains as the residues with at least one atom-atom distance ≤ 6.5Å from at least one atom in a residue from the G-protein chain in the same complex.

`

		GPCR			G-protein		
	PDB ID	Gene symbol	UniProt Accession	Chain	Gene symbol	UniProt Accession	Chain
G-proteins	6LFM	CXCR2	P25025	R	GNAI1	P63096	A
	6LFO	CXCR2	P25025	R	GNAI1	P63096	A
	7BZ2	ADRB2	P07550	R	GNAS	P63092	A
	6DDF	Oprm1	P42866	R	GNAI1	P63096	A
	6DDE	Oprm1	P42866	R	GNAI1	P63096	A
	6D9H	CHRM4	P08173	R	GNAI2	P04899	A
		ADORA1	P30542	R			
	6OS9	NTSR1	P30989	R	GNAI1	P63096	A
	6GDG	ADORA2A	P29274	A	GNAS	P63092	D
	6VMS	DRD2	P14416	R	Gnai1	P10824	A
	6OY9	RHO	P02699	R	GNAT1	P04695	A
	7JJO	ADRB1	P07700	R	GNAS	P04896	A
	6LI3	GPR52	Q9Y2T5	R	GNAS	P63092	A
	3SN6	ADRB2	P07550	R	GNAS	P04896	A
	6N4B	CNR1	P21554	R	GNAI1	P63096	A
	7D7M	PTGER4	P35408	A	GNAS	P63092	D
	6QNO	RHO	P02699	R	GNAI1	P63096	A
	6OSA	NTSR1	P30989	R	GNAI1	P63096	A
	6PT0	CNR2	P34972	R	GNAI1	P63096	A
	6NI3	ADRB2	P07550	R	GNAS	P63092	A
	7CFM	GPBAR1	Q8TDU6	R	GNAS	P63092	A
	6WWZ	CCR6	P51684	R	GNAO1	P09471	A
	7CFN	GPBAR1	Q8TDU6	R	GNAS	P63092	A
	6OMM	FPR2	P25090	R	GNAI1	P63096	A
	6KPF	CNR2	P34972	R	GNAI1	P63096	A
	6KPG	CNR1	P21554	R	GNAI1	P63096	A
	6K41	ADRA2B	P18089	R	GNAO1	P09471	A
		ADRA2A	Q28838	R			
	6K42	ADRA2A	P08913	R	GNAI1	P63097	A

`

		ADRA2B	P18089	R			
	6CMO	RHO	P08100	R	GNAI1	P63096	A
	6OIK	CHRM2	P08172	R	GNAO1	P09471	A
	6G79	HTR1B	P28222	S	GNAO1	P09471	A
	6OYA	RHO	P02699	R	GNAT1	P04695	A
β-arrestins	6U1N	CHRM2	P08172	R	Arrb1	P29066	C
		AVPR2	P30518	R			
	6TKO	ADRB1	P07700	A	ARRB1	P49407	B
	6PWC	NTSR1	P30989	R	ARRB1	P49407	A
	6UP7	NTSR1	P30989	R	ARRB1	P49407	B

Table 4.1: Known GPCR- G-protein/β-arrestin 3D complexes (release Jul 2020). All the 3D complexes were extracted from the SIFTS database (release Jul 2020) using Pfam to PDB chain mappings.

All statistical tests were performed using the SciPy library (Virtanen et al., 2020) with scripts written in python.

4.4 RESULTS

4.4.1 Prediction of couplings of uncharacterized and mutant GPCRs

We used PRECOG to predict coupling probabilities of uncharacterized receptors. Of the 61 class A GPCRs that lack coupling information in both the coupling dataset as well as in GtoPdb, PRECOG predicted 65%, 88%, 76%, and 72% to couple Gs, Gi/Go, and G12/G13 and Gq/G11 subfamilies, respectively (Figure 4.2A; Table S3A). The predicted receptors included oncogenic receptors such as *P2RY8*, which has been frequently mutated in lymphomas and displays mutual exclusivity with *GNA13* (López et al., 2019; Muppidi et al., 2014; O'Hayre et al., 2016). PRECOG predicts *P2RY8* to be a *GNA13*-coupled receptor (Figure 4.1).

`

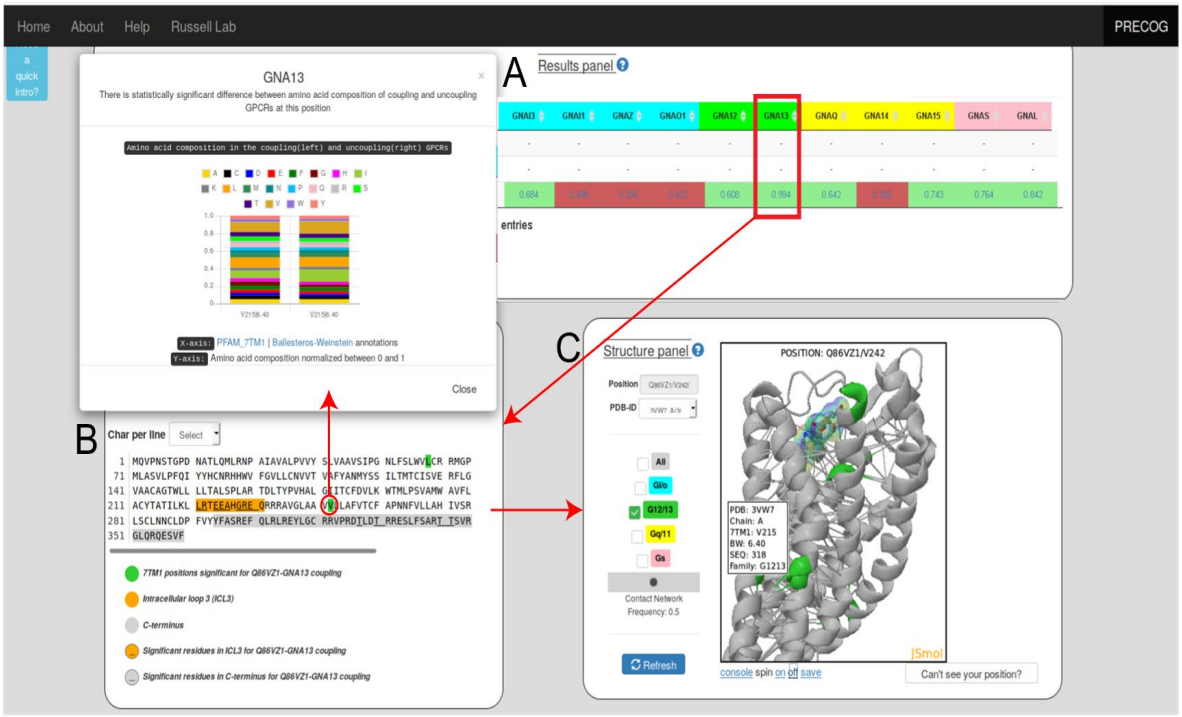


Figure 4.1: Predictions of *P2RY8* by PRECOG. The figure is taken from (Singh et al., 2019). The output page of PRECOG predicting coupling of P2Y purinoreceptor 8 (*P2RY8*) with *GNA13* is shown.

We also used PRECOG to predict the effect of GPCR mutations on G-protein selectivity (Table S3B). Of the 360 mutations across 60 class A GPCRs, we found 89 (~25%) to be a determinant of coupling specificity. 26 of the 89 mutations (29%) affect couplings with one or more G-proteins with Gi/Go being the most affected G-protein subfamily (Figure 4.2B, Table S3B; see Methods). For example, mutations in Gs-coupled *AVPR2* (vasopressin receptor 2) are responsible for nephrogenic diabetes insipidus, an X-linked recessive disease characterized by excessive urine production and thirst (reviewed in Spanakis et al., 2008) due to decreased cAMP response in kidney cells. PRECOG predicted one of these disease-causing mutations (p. Ala163Ser) (Rocha et al., 1999) to couple to *GNAI3*, which is responsible for inhibition of the cAMP-dependent pathway (Table S3B).

`

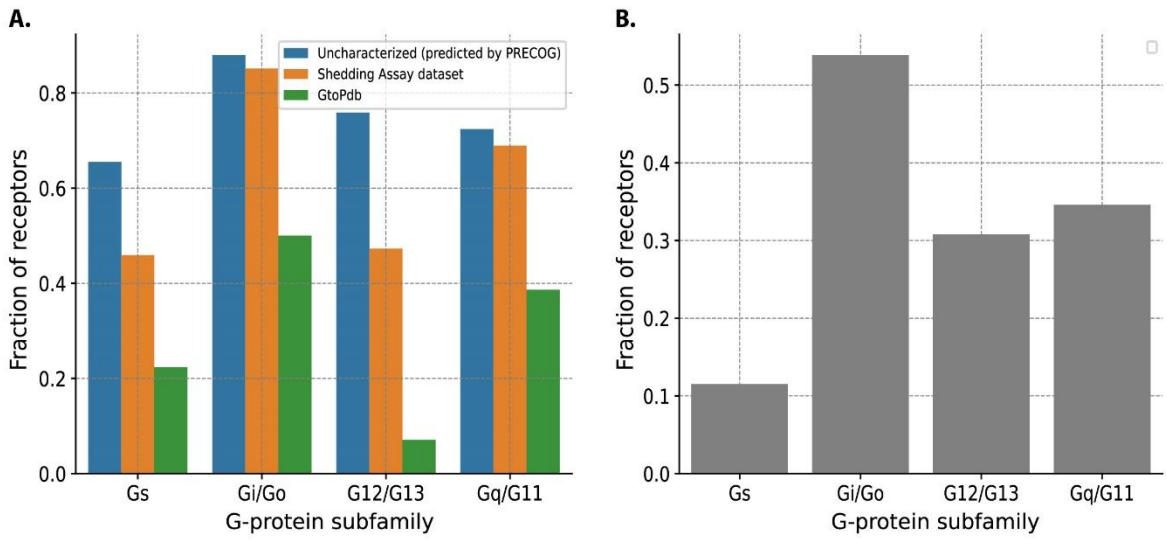


Figure 4.2: Predicted vs experimental couplings of GPCRs and predicted effect of GPCR mutations on their couplings. (A) Bar plot showing the fraction of known/uncharacterized receptor couplings (x-axis) with G-protein subfamilies (y-axis). (B) Bar plot showing the fraction of receptor couplings (y-axis) affected with the G-protein subfamilies (x-axis) because of mutations.

4.4.2 Development of the first GNA12-coupled DREADD

Several studies have highlighted the advantage of using DREADDs, especially to gain control over GPCR signaling (Farrell et al., 2013; Hu et al., 2016; Roth, 2016; Wess et al., 2013). The lack of GPCR-G12/G13 complex and poor characterization of the G12/G13 subfamily prompted us to engineer a DREADD that selectively couples to it (Inoue et al., 2012). We chose the hM3D DREADD that is experimentally known to couple to Gq/G11 and Gi/Go subfamilies (Armbruster et al., 2007), as correctly predicted by PRECOG (Figure 4.3).

`

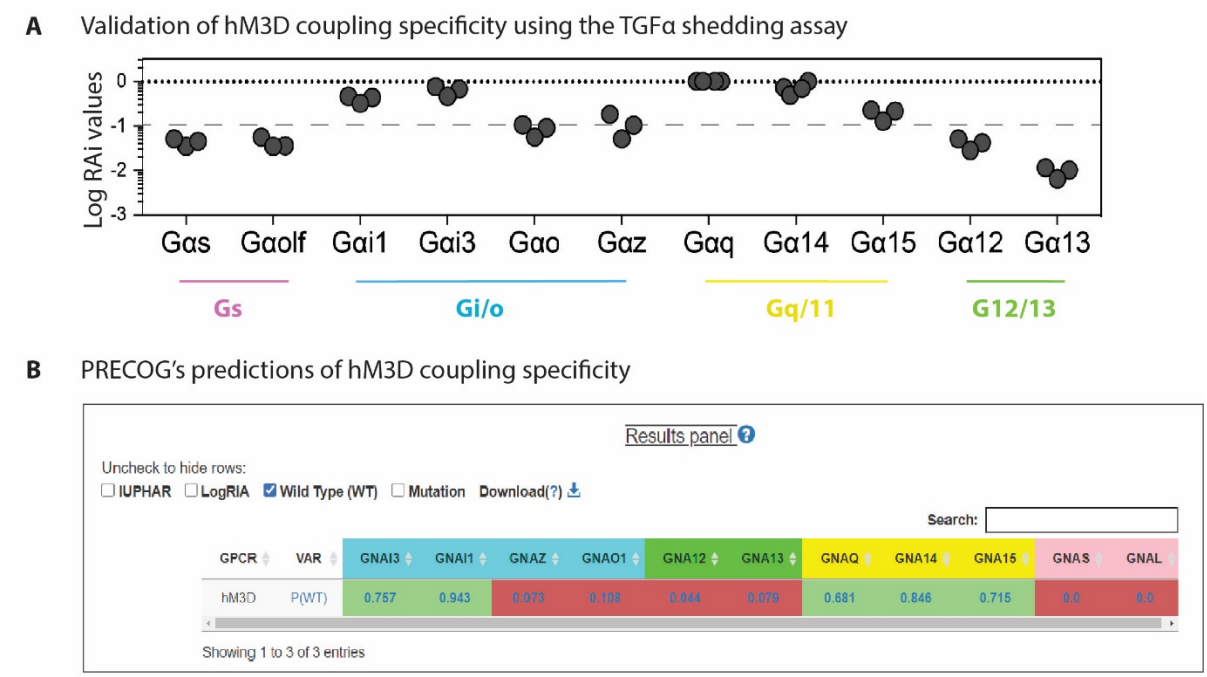


Figure 4.3: Coupling-specificity of the hM3D DREADD. (A) Experimental validation of hM3D using the TGFα shedding assay as well the generation of this panel was done by the group of Dr. Asuka Inoue (Tohoku University, Japan) (Log RAi ≥ -1.0: coupled, otherwise not-coupled) (see Chapter II). (B) Prediction of hM3D coupling probability by PRECOG.

ICL3 followed by C-terminus contribute the most towards determining the coupling probabilities of receptors with that of the G12/G13 subfamily (see Figure 2.9A of Chapter II). Thus, we sought to generate chimeric sequences of hM3D by swapping its ICL3 with or without the swap of the C-terminus with that of the GPCRs used in the coupling dataset (see Methods). Swapping of hM3D regions has been previously shown to generate Gs-coupled DREADD, known as hM3D-Gs, which involved substituting ICL2 and ICL3 of the Gq/G11-coupled hM3D with those of Gs-coupled β1AR (Guettier et al., 2009). We predicted the coupling probability of each chimeric sequence with the G12/G13 subfamily using PRECOG (Figure 4.4) and selected the 13-potential chimeric GPCRs for experimental validation (see Methods).

`

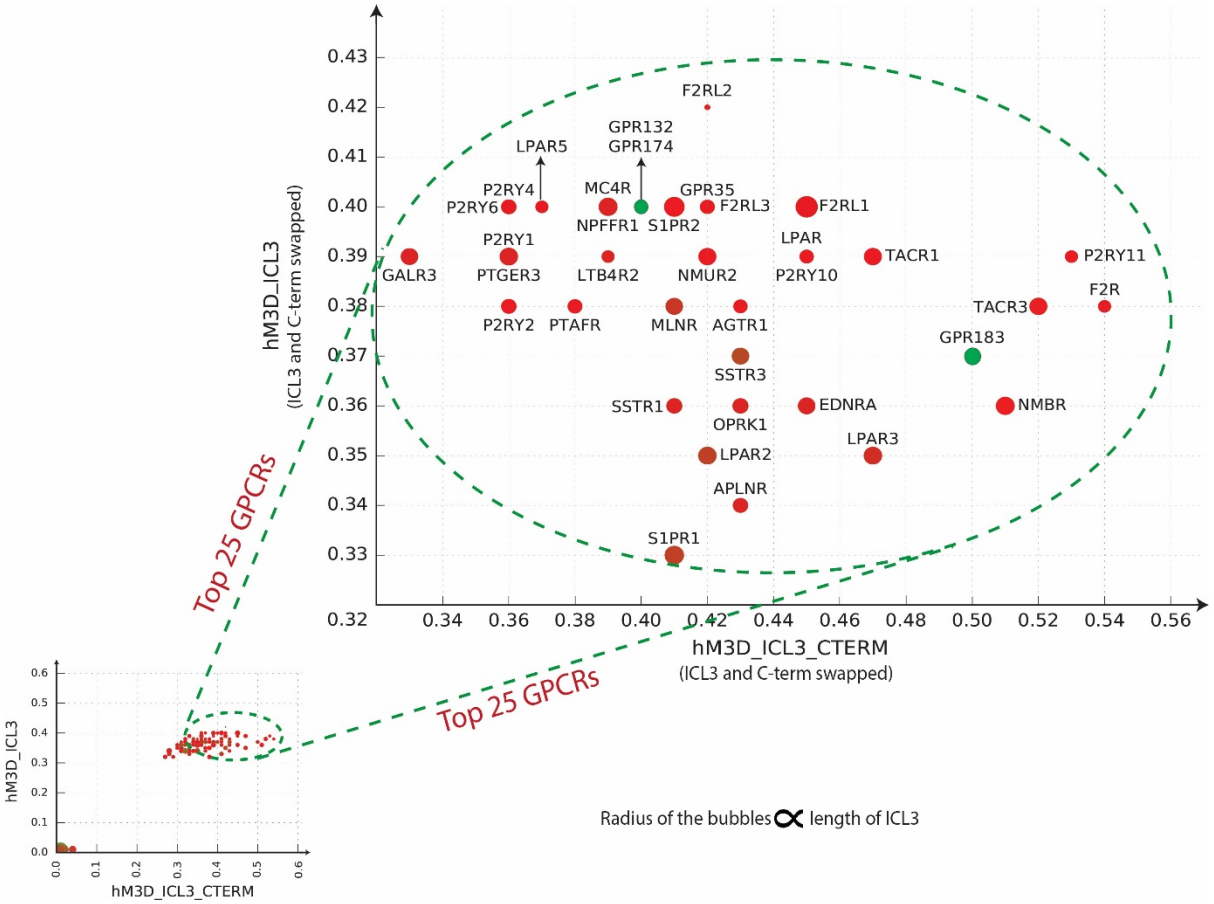


Figure 4.4: Scatter plot of predicted coupling probabilities of DREADDs with *GNA12* by PRECOG. The figure and legend are taken from Inoue et al., 2019. The smaller plot (bottom left) displays the relative coupling probabilities ($P_{DREADD} - P_{hM3D}$) of chimeric sequences obtained by substituting the stretches of hM3D by the ICL3 alone (y-axis) or in combination with the C-terminus (x-axis) of the G12/G13-coupled receptors in the TGFα shedding assay are shown. The larger plot displays the top 25 chimeric sequences obtained from both the design, including the ones derived from *GPR183* and *GPR132* (indicated with green bubbles). The size (radius) of the bubble is proportional to the length of ICL3.

Dr. Asuka Inoue and his group (Tohoku University, Japan) screened the G12-coupling activity of the best-scoring chimeric sequences using TGFα shedding and Nano-BiT assays (Figure 4.5). The detailed results of the validation experiment were published elsewhere (Inoue et al., 2019). Briefly, the TGFα shedding assay confirmed that among the 26 constructs tested, chimeras with the GPR183- and GPR132- derived ICL3 substitutions, henceforth referred to as hM3D-GPR183/ICL3 and hM3D-GPR132/ICL3, respectively, showed significant coupling towards *GNA12* (p-value ≤0.05) (Figure 4.5). As expected, *Ach* (Acetylcholine), the native ligand of muscarinic receptors, induced no *GNA12* signaling in any of the tested constructs while treatment

`

with the synthetic ligand, clozapine-N-oxide (CNO), led to the activation of the G12-coupling constructs.

An additional NanoBiT experiment (Inoue et al., 2019), performed by the same collaborators to test activation of *GNA12* and *GNAO1*, identified two additional constructs, hM3D-P2RY10/ICL3 and hM3D-NMBR/ICL3, that coupled to *GNA12* but not to *GNAO1*. A construct generated by substituting LTB4R2-derived ICL3 substitution (hM3D-LTB4R2/ICL3) induced both *GNA12* and *GNAO1* coupling with a higher dissociation signal for the latter. Another NanoBiT-G13 experiment showed that neither of the constructs induces G13 signaling. Neither of the constructs with the substitution of both ICL3 and C-terminus exhibited any coupling specificity towards *GNA12* because of their low surface expression (Figure 4.5).

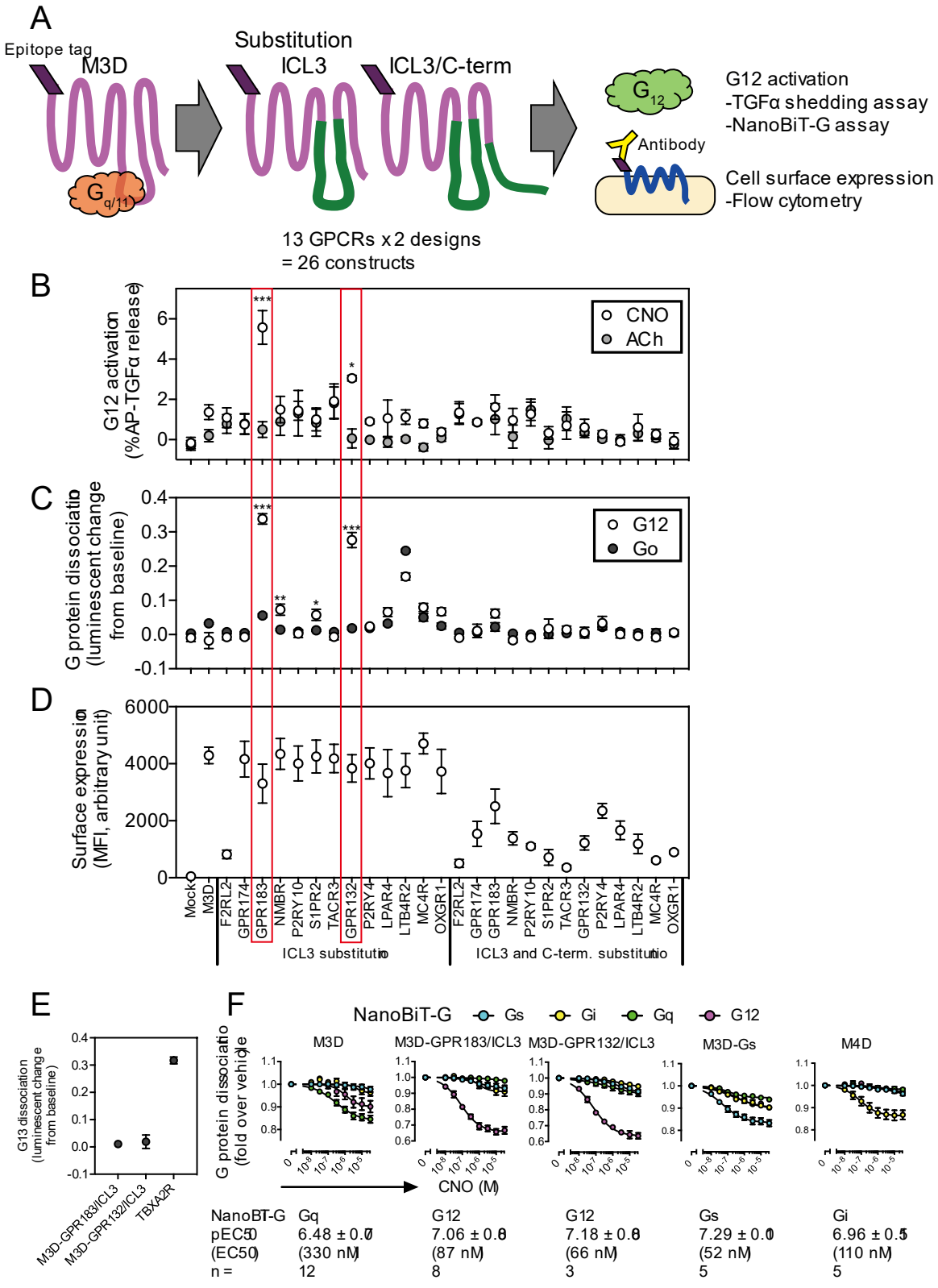


Figure 4.5: Design and validation of *GNA12*-coupled receptors. The image was taken from Inoue A et al., 2019. (A) Outline of the two types of designs used to construct hM3D-derived chimeric sequences. Based on PRECOG's predictions (without InterPreTS), the top 25 chimeric sequences

`

(corresponding to 13 GPCRs) were selected from each type of design. 26 chimeric sequences were constructed for these 13 GPCRs. (B) The TGFα shedding assay response of chimeric sequences was measured to assess *GNA12* (G12) signaling. (C) The NanoBiT-G-protein dissociation assay response of chimeric sequences was measured to assess *GNA12* (G12) and *GNAO1* (Go) activation. (D) The surface expression of the 26 hM3D-derived chimeric sequences was measured using flow cytometry. The newly designed G12/G13 DREADDs are outlined in red. (E) The NanoBiT-G13 protein dissociation assay response of chimeric sequences was measured to assess *GNA13* (G13) activation. (F) The NanoBiT-G protein dissociation assay response of the two newly constructed G12/G13-coupled DREADDs (hM3D-GPR183/ICL3 and hM3D-GPR132/ICL3) as well as of the previously established DREADDs (Gq/G11-coupled hM3D, Gi/Go-coupled hM4D, and Gs-coupled hM3D) was used to assess G-protein activation. Bubbles and error bars represent mean and standard deviation, respectively. Experimental validation of using the TGFα shedding assay as well the generation of this figure was done by the group of Dr. Asuka Inoue (Tohoku University, Japan). For detailed legend, please refer to the original article (Inoue et al., 2019).

4.4.3 Application of the framework on the ebBRET Assay

The comparison of the datasets from the ebBRET and TGFα shedding assays reveals an overlap of 71 receptors (Figure 4.6). The order of decreasing agreement of receptor couplings to G-protein subfamilies between the two datasets is Gi/Go > Gq/G11 > G12/G13 > Gs (Figure 4.6B). For more details, please refer to the original article (Avet et al., 2020) on the data from the ebBRET assay.

`

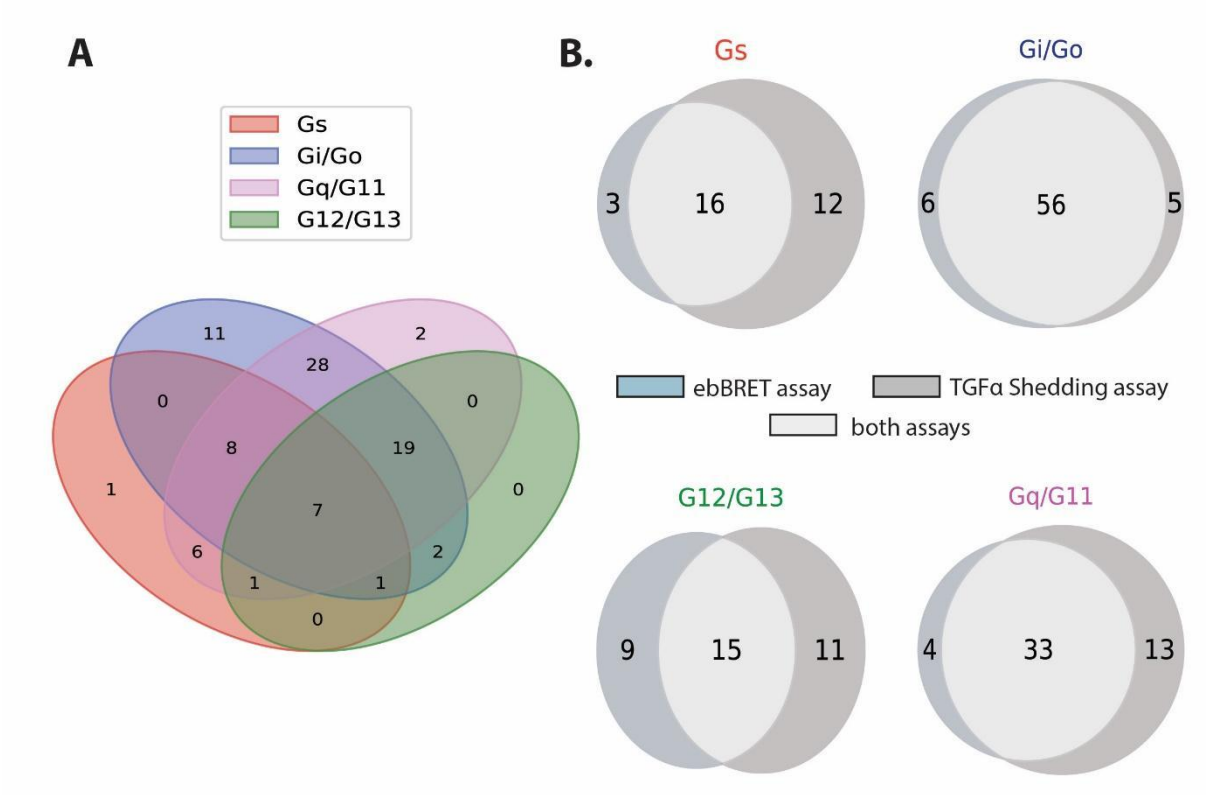


Figure 4.6: Comparison between the datasets from the ebBRET and TGFα shedding assays. (A) Venn diagrams depicting the number of GPCRs coupled to each G-protein subfamily (E_{max}>0). (B) Venn diagrams depicting receptor couplings to the two datasets.

We applied the PRECOG framework to obtain the determinants of G-protein/β-arrestin specificity in receptors of the data from the ebBRET assay (Tables S3D, S3E). We identified 53 positional features (or determinants of interaction specificity) in the 7TM1 domain of the receptors that are statistically associated with each of the 12 G-proteins (49 positions) and 2 β-arrestins (in presence or absence of *GRK2*) (25 positions) (Figure 4.7D; Table S3D). There is an intersection of 21 (of 53) positional features between the two interacting groups (Figure 4.7E; Table S3D). Intriguingly, like observed in the case of the data from the TGFα shedding assay (see section 2.4.5 in Chapter II), only 16 of 53 (or 30%) positional features lie on the known GPCR/G-protein interfaces (14 of 49 or 28%) or GPCR/β-arrestin interfaces (7 of 14 or 28%) (see Methods; Table S3D).

As observed in the TGFα shedding assay, ebBRET assay-derived determinants of both G-proteins (Figures 4.7A, B) and β-arrestins (Figures 4.7A, B, 4.7D) are also distributed throughout the TM helices and extracellular as well cytosolic regions of

`

receptors. Regarding β-arrestins in the data from the ebBRET assay, positional features followed by C-terminus were observed to contribute the most towards determining their specificity in receptors (Figures 4.7A, B).

Next, we compared the determinants of G-protein coupling specificity extracted from the datasets from the ebBRET and the TGFα shedding assays. A significant number of positional features (34 of 49 or 69%) determining G-protein specificity in the data from the ebBRET assay were found to overlap with those obtained from the TGFα shedding assay (Figure 4.7E). Though the distribution of determinants at the level of G-protein subfamilies is very similar in the two datasets (Figure 4.7A), we observed certain differences. First, the structure-based determinants (obtained from InterPreTS; see Methods) extracted from the data from the ebBRET assay contribute more towards the specificity of Gs and Gi/Go subfamilies than those derived from the data from the TGFα shedding assay (Figure 4.7A). Second, while positional features (sequence-based features in the 7TM1 domain) extracted from the data from the ebBRET assay provide a major share in determining G12/G13-specific couplings in GPCRs, the ICL3 influences G12/G13 specificity greater in the determinants derived from the data from the TGFα shedding assay (Figures 4.7A, B, C).

Among the determinants derived from both the datasets, TM3, TM5, and TM6 (Figure 4.7D) are the major contributors of G-protein and β-arrestins specificities.

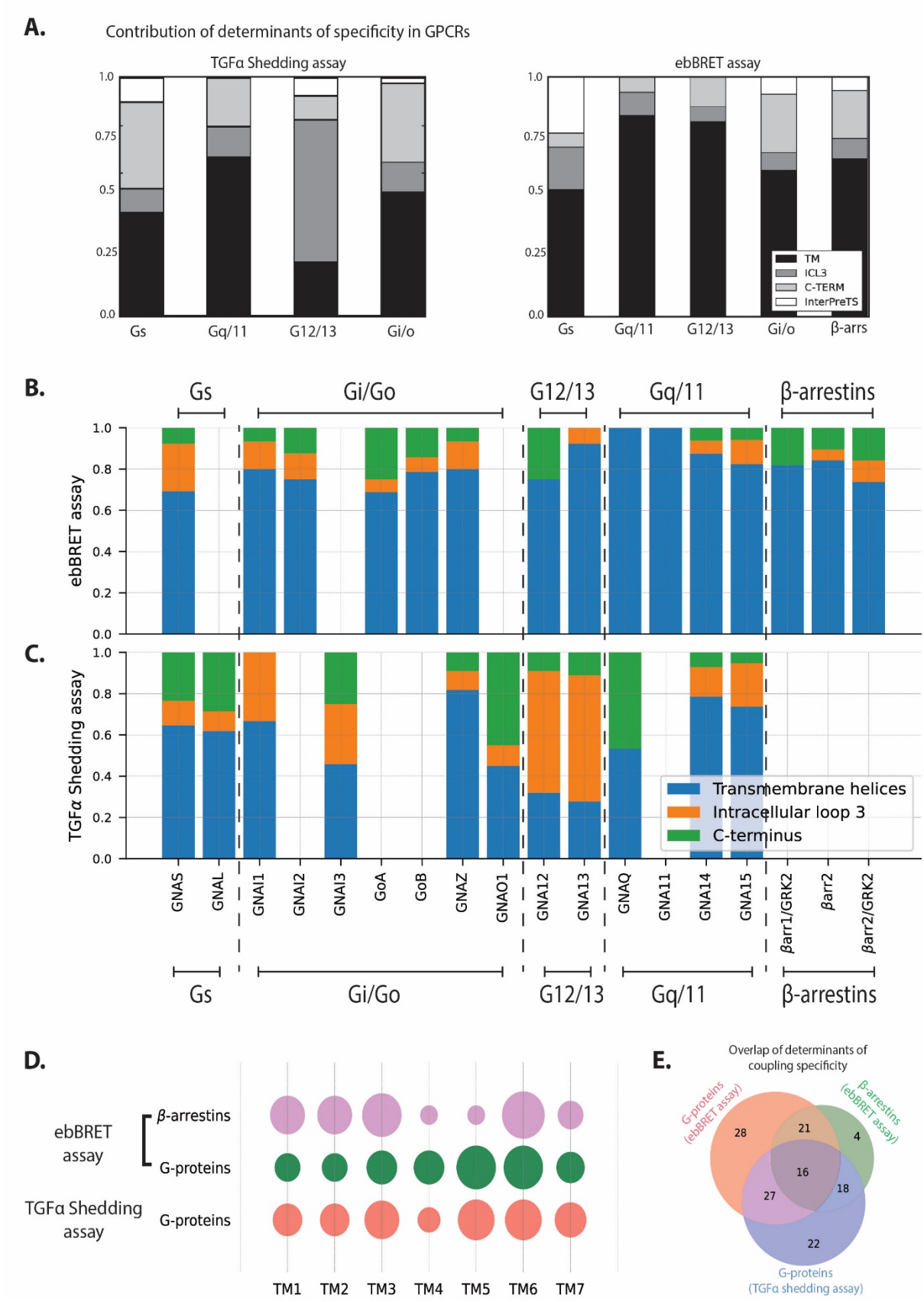


Figure 4.7: Comparison of determinants of G-protein/β-arrestin specificity obtained from the two datasets. (A) Contribution of G-protein/β-arrestin specificity in receptors in the two datasets. (B) Bar-

`

plots showing contribution of TM helices, ICL3, and C-terminus as determinants of interaction specificity for individual interacting groups in the data from the ebBRET assay. (C) Bar-plots showing contribution of TM helices, ICL3, and C-terminus as determinants of interaction specificity for individual interacting groups in the dataset derived from TGFα shedding assay. (D) Scatter plot showing the contribution of TM helices as a determinant of specificity of interacting groups in the two datasets. (E) Venn diagrams depicting the overlap of determinants (positional features) of coupling specificity in the datasets derived from the ebBRET (for G-proteins and β-arrestins) and the TGFα shedding assays (G-proteins).

We then applied the logistic regression algorithm to train and develop the ebBRET assay-based predictor and compared its performance with that of the TGFα shedding assay-based predictor (i.e. PRECOG) (Figure 4.8B; Tables S3F, G). In terms of individual G-protein coupling groups, the performance of the ebBRET assay-based predictor is better than PRECOG for *GNAQ*, *GNA15*, *GNA13*, *GNAZ,* and *GNAI1*-specificities with the reverse being the case for the others (Figure 4.8B; Table S3H). The inclusion of structure-based features does not significantly improve the performance of the ebBRET assay-based predictor (Figure 4.8C; Tables S3F, G). However, the major addition of the ebBRET assay-based predictor (first ever) is the prediction of β-arrestin specificity.

`

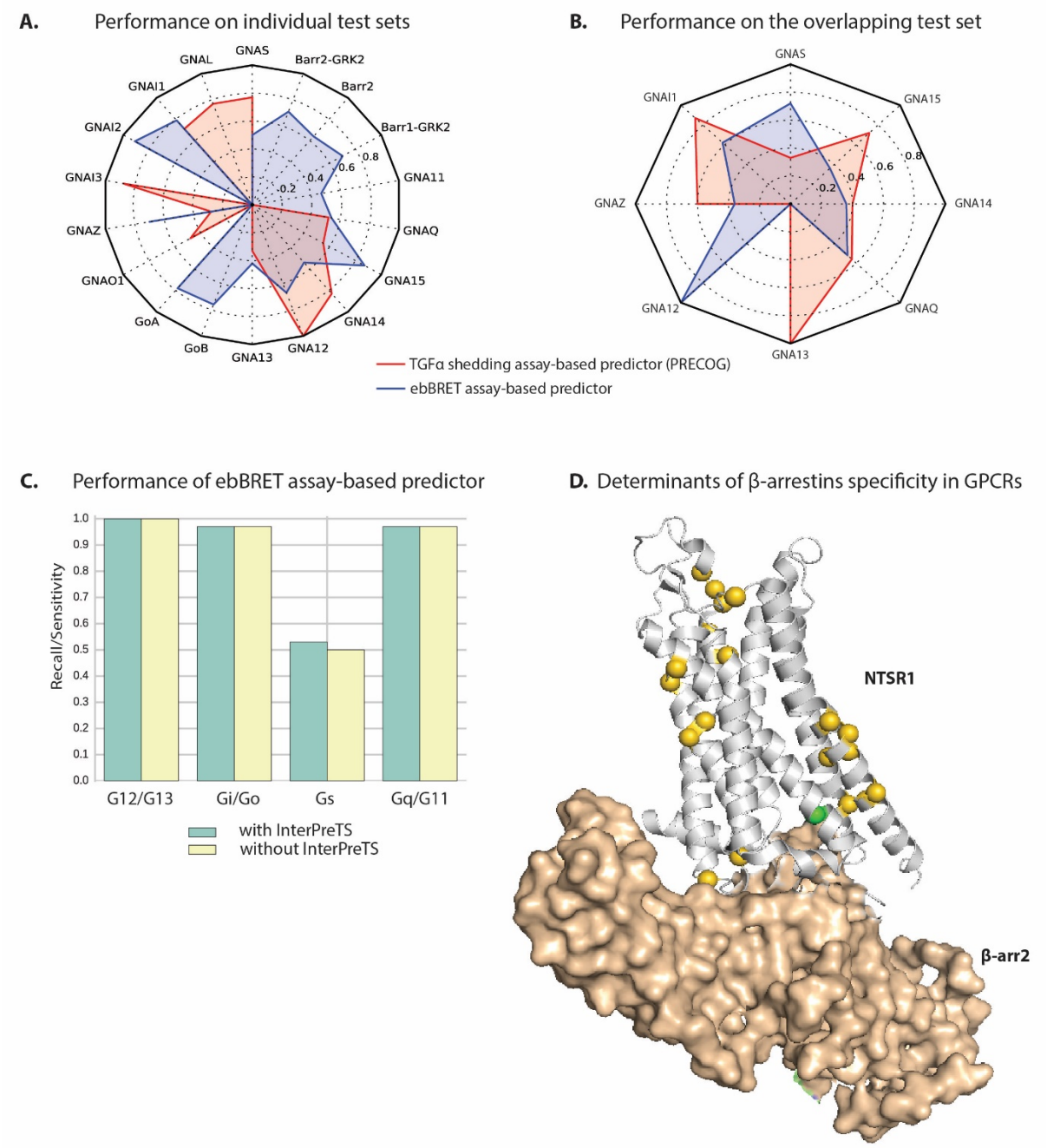


Figure 4.8: Performance of the ebBRET assay-based predictor. (A) Recall/Sensitivity of the ebBRET assay-based predictor and (B) the TGFα shedding assay-based predictor (PRECOG) on their test sets. (C) Comparison of performance (recall/sensitivity) of the ebBRET assay-based predictors (with or without InterPreTS) at the level of G-protein subfamily. (D) Determinants of β-arrestin-1/2 specificity mapped onto a structure (PDB ID: 6UP7) (Positions with a difference in amino acid distribution between the interacting and non-interacting receptors are shown in yellow, residues with insertions are shown in green, and residues with deletions are shown in red).

We computed the feature weights from the trained models of each interacting group to construct a feature weight matrix (Figure 4.9), which has been previously used to

`

study feature relevance (Dhole et al., 2014; Dou et al., 2012). The colors of the cells in the matrix can aid to deduce the directionality of the effect of changes in it (see section 3.4.2 in Chapter III). For example, the length of the C-terminus, an extra-membrane feature with positive feature weights (green-colored cell highlighted by a dotted circle in Figure 4.9), is computed to affect *GNA12*-coupled receptors. In other words, an increase in its length will favor coupling to *GNA12*.

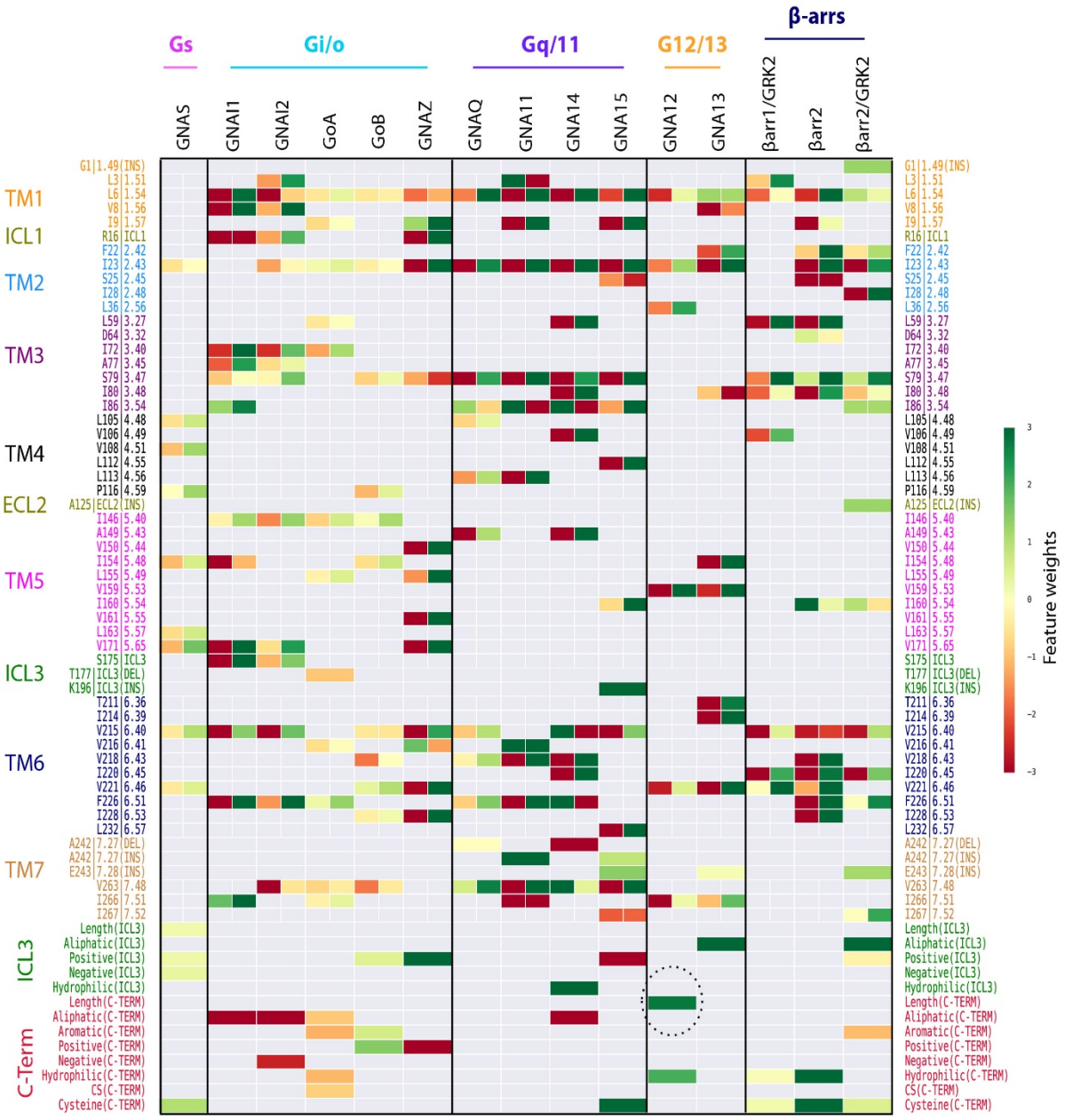


Figure 4.9: Feature weight matrix of the ebBRET-based predictor. A heatmap showing contribution of statistically associated sequence-based features (x-axis) in GPCRs to at least one interacting group (y-axis). Cells are colored based on the coefficients (also called feature weights) of the given feature in the best-performing model of the corresponding interacting group (red-green scale corresponding to negative and positive weights, respectively). Color intensities of cells indicate the absolute value of the

`

coefficients. If a significant 7TM domain position is present in both coupled and not-coupled HMMs, its coefficients are shown within the same cell on left (coupled) and right (not-coupled). The cell highlighted with a dotted circle is discussed in section 4.4.3.

4.5 DISCUSSION

PRECOG and the application of its framework show strong potential. We demonstrated the first application of PRECOG: to predict couplings of several uncharacterized receptors such as *P2RY8* that has been reported to have an oncogenic potential (López et al., 2019; Muppidi et al., 2014; O'Hayre et al., 2016). We then described the second application in identifying variants that could alter G-protein specificity. The third application of PRECOG involved the design of artificial GPCRs (DREADDs) that hold much promise in studying cell functions. Unlike TM helices in receptors of other G-protein subfamilies, ICL3 of receptors contributes towards the G12/G13-coupling specificity (Figure 4.7A). We exploited this property to successfully develop the first *GNA12*-specific DREADDs: hM3D-GPR183/ICL3 and hM3D-GPR132/ICL3.

A key application of PRECOG lies in the employability of its framework on other binding assay datasets. We applied the framework on recently available data from the ebBRET assay to train and develop predictors of G-protein and importantly, β-arrestin, specificity. As observed with the coupling dataset, only 30% of the positional features lie on the interfaces while the rest are distributed across the TM helices and extra-membrane region, corroborating the known role of allosteric mechanisms (Angelova et al., 2011; Flock et al., 2017; Venkatakrishnan et al., 2013, 2016; Wichard et al., 2011). Another key observation was the major contribution of TM helices 5 and 6 as determinants of specificity, which has also been described in previous studies (García-Nafría et al., 2018b; Kang et al., 2018; Koehl et al., 2018). For β-arrestins, the determinants of specificity (such as TM6, ICL3, and C-terminus) extracted here have also been previously reported to play a role in GPCR- β-arrestin interactions (Ranjan et al., 2017; Shukla et al., 2013). A significant overlap of sequence-based determinants of coupling specificity in β-arrestin and G-protein groups is logical as β-arrestins mediate steric hindrance of G-proteins, leading to receptor desensitization (Shukla et al., 2011). It is known that phosphorylation of GPCRs in the C-terminus (Sente et al., 2018) and ICL3 (Kumari et al., 2017) is essential for the recruitment of

`

β-arrestin. Thus, features that encode phosphosite information can be included to obtain a wider view of GPCR- β-arrestin couplings and to further improve the performance of the predictor.

The data from the ebBRET assay complements the data from the TGFα shedding assay for specific G-proteins as they are better covered by the former and hence show better sensitivity than PRECOG (Figure 4.8A). This can be useful in the development of a *GNA13*-specific receptor, which is still unavailable. At the level of the subfamily, the performances of all groups are fairly the same. However, it is noteworthy that the performance of the Gs subfamily is lower than the other subfamilies in both predictors. As described before (see section 3.5 in Chapter III), this can be due to the lack of features that cover the dynamics of the Gs subfamily in the training matrices, such as the outward movement of the TM6 helix, which is more pronounced in the Gs subfamily than in the Gi/Go subfamily (Kang et al., 2018; Koehl et al., 2018). The inclusion of ICL/ECLs and the N-terminus as features could also improve the performance of Gs-specificity.

A significant overlap of determinants of G-protein coupling specificity derived from the two datasets (69% of determinants derived from the ebBRET assay and 66% of determinants derived from the TGFα shedding assay intersect) highlights the strength of the PRECOG framework, which can be employed on binding datasets of large protein families (such as GPCRs) and to perform a systematic analysis to extract and analyze signatures of molecular mechanisms from their MSA.

`

Chapter V: General conclusions and discussion

Proteins undergo different conformational states to interact with each other, leading to disruption and the creation of several (typically non-covalent) interactions between residues. Aberrations in any of these residues can lead to dysregulation of physiological mechanisms, most of which are implicated in human diseases. Additionally, unraveling such residues is beneficial in protein engineering to unravel associated downstream signaling pathways. In this study, we have developed a method to extract statistically-associated positions and sequence regions within GPCRs that determine coupling-specificity towards the heterotrimeric G-proteins (Chapter 2). We then implemented a machine learning approach to learn the sequence and structure-based features to develop a predictor of G-protein coupling specificity: PRECOG (Chapter III). Finally, we demonstrated the applications of this machine learning-guided framework to predict couplings of uncharacterized and mutated GPCRs, to design the first G12-coupled receptors, and to use the PRECOG framework to a different class of interactions between GPCRs and β-arrestins.

5.1 MAIN RESULTS

5.1.1 Determinants of G-protein coupling specificity in GPCRs (Chapter II)

We identified statistically-associated sequence-based features that include both the positions across the 7TM1 domain and the intrinsically disordered ICL3 and C-terminus for each G-protein group. Intriguingly only 23% of the identified 7TM1 positions lie on the GPCR/G-protein interfaces with most being spread across the transmembrane helices, supporting the notion that GPCR activation is a complex interplay of residues involving allosteric communication between ligand and G-protein binding pockets (Huang et al., 2015; Koehl et al., 2018). Of the seven helices, we found the TM6 helix, which has been shown to undergo an outward movement in activated GPCR structures (García-Nafría et al., 2018b; Koehl et al., 2018; Lin et al., 2020; Nojima et al., 2020; Rasmussen et al., 2011), contributes the most positional features. The length and charge distribution of the extra-membrane regions - the ICL3 and C-terminus - also play a role in determining G-protein specificity, especially for the G12/G13 subfamily. Structural comparison of GPCR/G-protein 3D complexes

`

revealed a pronounced outward movement of the TM6 helix in Gs-coupled vs Gi/Go-coupled receptors, which can be attributed to bulkier side chains of C-terminus of the α5 in *GNAS*. The amino acid deletion in the TM5-ICL3-TM6 regions, that we found statistically associated with Gs-coupled receptors, might be instrumental in creating a large crevice to accommodate bulky side chains of *GNAS*.

5.1.2 A machine learning-guided framework (Chapter III)

In addition to sequence-based descriptors, we derived a set of structure-based properties by identifying GPCR/G-protein 3D complexes better suited to model the interactions for every G-protein coupling group using InterPreTS (Aloy and Russell, 2002, 2003), that evaluated the fit of all the given receptor-G-protein pairs onto these complexes. We then applied an interpretable, logistic regression algorithm to learn from the sequence- and structure-based features and to develop predictive models (called PRECOG) of GPCR/G-protein couplings. The PRECOG web-server is a first predictor of GPCR coupling that predicts at the level of individual G-protein rather than the subfamily. For a given receptor sequence or its UniProt identifier, the PRECOG web server can quickly predict the interacting G-protein(s), assess the impact of mutations, and design receptors that show coupling to a specific receptor.

5.1.3 Applications of the ML-guided framework (Chapter IV)

Amongst other uncharacterized mutated GPCRs, PRECOG readily predicted the uncharacterized purinergic receptor *P2RY8*, a protein with oncogenic potential, to couple *GNA13*, which is in line with growing evidence (López et al., 2019; Muppidi et al., 2014; O'Hayre et al., 2016). The predictive models of PRECOG were successful at developing the first *GNA12*-coupled designer receptors: hM3D-GPR183/ICL3 and hM3D-GPR132/ICL3. Finally, we demonstrated the versatility of the PRECOG framework on new binding data to unravel the determinants of β-arrestin specificity in GPCRs. As observed for G-proteins, the β-arrestin specificity determining positions in receptors also lie at the intracellular binding interface as well as throughout the hydrophobic regions of the 7TM1 domain and extra-membrane regions. Further, a significant overlap of G-protein specificity positions obtained from the two binding datasets highlights the reusability of the PRECOG framework.

`

5.2 PRACTICAL IMPLICATIONS

5.2.1 An adaptable framework

The framework lends itself to binding data of other protein-protein interactions that lack the sub-type specificity information. Unlike most SDP methods (see section 1.6 of Chapter I) that consider only positions within an MSA, our framework offers flexibility to also include the statistically significant disordered regions of proteins (that can often not be meaningfully aligned) by assessing their length and charge distributions (for example the ICL3 and C-terminus of GPCRs in the current study). Other physical properties of a protein such as hydrophobicity and predicted backbone-dynamics or secondary structure or post-translational modifications information can also be added as additional descriptors. Finally, the framework also provides the flexibility of employing other *interpretable* machine learning algorithms (for example support vector machines or random forests) depending on the nature, size, and quality of data.

5.2.2 ML-guided protein designing

The significant strength of the framework comes from the use of logistic regression, an *interpretable* machine-learning algorithm, that generates data-driven predictors based on input information. These predictors assess the contribution of features towards determining the subtype and this property can be exploited to engineer proteins, as we demonstrated with the development of the first G12-coupled receptors. This significantly reduces the immense cost and time required to screen hundreds or thousands of sequences. ML-guided protein engineering has recently shown promising results. For example, Gaussian process models were used to design minimally-invasive channelrhodopsins of high light sensitivity (Bedbrook et al., 2019) and to engineer cytochrome P450s with increased thermostability (Romero et al., 2013). In another study, a Boltzmann machine-learning-based method was applied to develop an artificial protein that mimics chorismate mutase, an enzyme essential to the biosynthesis of aromatic amino acids (Russ et al., 2020). As we continue to map sequence to function relationships using various experimental assays, machine-learning guided frameworks, such as the one presented here, will be beneficial tools to design as well develop artificial proteins.

`

5.3 OUTLOOK

5.3.1 Expanding the feature set

Over 1000 ligands bind on the extracellular sites of the receptor, which in turn recruit one of the 16 human G-proteins and/or one of the four β-arrestins that bind to intracellular pockets of the receptor. While our framework is effective at mining subtype-specific positions in receptors, a much more comprehensive feature set that encompasses other properties of not just GPCRs but also their extracellular and cytosolic partners can provide a complete picture.

Descriptors such as SMILES strings (Weininger, 1988) or PubChem fingerprints (Kim et al., 2021) can be used to include the structural and chemical nature of ligands. Though such inclusion may not be possible in the current scope as each receptor was tested by only one agonist in the data from the TGFα shedding assay, this information can be supplemented by other publicly available datasets such as of PRESTO-Tango (Kroeze et al., 2015) that have screened hundreds of compounds against GPCRs.

The fifth helix of the Gα subunit, particularly the last 6 amino acids, is involved in the interaction with GPCRs (Carpenter et al., 2016). A recent bioinformatics study identified additional patterns of amino acids in the G-proteins that determine GPCR selectivity (Flock et al., 2017). Thus, a vector encoding the G-protein properties can be appended into the feature set to enhance predictions.

The inclusion of other extra-membrane regions of the receptors such as the ICL1 (García-Nafría et al., 2018b; Nojima et al., 2020), ICL2 (García-Nafría et al., 2018b; Kang et al., 2018; Koehl et al., 2018; Nojima et al., 2020), and ECL2 (Lin et al., 2020) that have been shown to participate in cytosolic interactions, might provide supplementary subtype information. Upon the availability of more data in the future, other receptor features such as conformational dynamics, which captures the variable outward movement receptors (Koehl et al., 2018), can further fine-tune the G-protein specificity predictions.

`

5.3.2 Include contextual information

A GPCR can express more than one isoform, and each of them can have different downstream outcomes (Kendall and Senogles, 2011; Smith et al., 2017). While in the current framework, PRECOG can be used to predict the G-protein/β-arrestin specificity of each of these isoforms, it is the different combinations of these isoforms co-expressed in tissues that dictate the signaling outcome (Marti-Solano et al., 2020). Thus, an advanced framework would involve the calculation of co-expression scores of receptor isoforms and G-proteins in tissues and combine them with sequence and/or structural-based predictors such as PRECOG using statistical or cognitive models (Lee and Danileiko, 2014) to estimate the overall probability of the given coupling in a tissue.

5.3.3 Building regression models

In the current study, we have built a classification model of each coupling group by binarizing the LogRAi values into 1 (high binding affinity) or 0 (low binding affinity) (see section 2.3.1 in Chapter II). While the current classification model (logistic regression) predicts the coupling probability of any given GPCR/G-protein pair, these probabilities do not reflect the LogRAi value of the interaction. Thus, an improved framework would involve the building of regression models that will make predictions that resemble the actual target values (LogRAi values). The availability of bigger and more detailed datasets will enable such a predictor.

5.4 EPILOGUE

In this study, we present a powerful, machine-learning guided framework that uses GPCR binding data to extract statistically associated sequence and structural features to develop predictive models of G-protein subtype specificity. We apply the framework to predict the couplings of uncharacterized receptors, assess the impact of mutations, and design two novel G12-specific receptors. Collectively, the application of this framework to other binding data, as we showed with β-arrestins, can uncover novel allosteric sites involved in subtype activation; improve our understanding of human diseases; help us devise better chemogenetic tools and diagnostic techniques, and ultimately make smarter therapeutic decisions.

`

REFERENCES

Addy, C., Wright, H., Van Laere, K., Gantz, I., Erondu, N., Musser, B.J., Lu, K., Yuan, J., Sanabria-Bohórquez, S.M., Stoch, A., et al. (2008). The Acyclic CB1R Inverse Agonist Taranabant Mediates Weight Loss by Increasing Energy Expenditure and Decreasing Caloric Intake. Cell Metab. *7*, 68–78.

Alanis-Lobato, G., Andrade-Navarro, M.A., and Schaefer, M.H. (2017). HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res.

Alexander, G.M., Rogan, S.C., Abbas, A.I., Armbruster, B.N., Pei, Y., Allen, J.A., Nonneman, R.J., Hartmann, J., Moy, S.S., Nicolelis, M.A., et al. (2009). Remote control of neuronal activity in transgenic mice expressing evolved G protein-coupled receptors. Neuron *63*, 27–39.

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. *37*, 420–423.

Aloy, P., and Russell, R.B. (2002). Interrogating protein interaction networks through structural biology. Proc. Natl. Acad. Sci. U. S. A. *99*, 5896–5901.

Aloy, P., and Russell, R.B. (2003). InterPreTS: protein Interaction Prediction through Tertiary Structure. Bioinformatics *19*, 161–162.

Althoefer, H., Eversole-Cire, P., and Simon, M.I. (1997). Constitutively active Gαq and Gα13 trigger apoptosis through different pathways. J. Biol. Chem. *272*, 24380–24386.

Altosaar, K., Balaji, P., Bond, R.A., Bylund, D.B., Cotecchia, S., Devost, D., Doze, V.A., Eikenburg, D.C., Gora, S., Goupil, E., et al. (2019). Adrenoceptors (version 2019.4) in the IUPHAR/BPS Guide to Pharmacology Database. IUPHAR/BPS Guid. to Pharmacol. CITE *2019*.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol.

Angelova, K., Felline, A., Lee, M., Patel, M., Puett, D., and Fanelli, F. (2011). Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor. Cell. Mol. Life Sci. *68*, 1227–1239.

Armbruster, B.N., Li, X., Pausch, M.H., Herlitze, S., and Roth, B.L. (2007). Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. Proc. Natl. Acad. Sci. U. S. A. *104*, 5163–5168.

Avet, C., Mancini, A., Breton, B., Le Gouill, C., Hauser, A., Normand, C., Kobayashi, H., Gross, F., Hogue, M., Lukasheva, V., et al. (2020). Selectivity Landscape of 100 Therapeutically Relevant GPCR Profiled by an Effector Translocation-Based BRET Platform. SSRN Electron. J. 2020.04.20.052027.

Azodi, C.B., Tang, J., and Shiu, S.H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends Genet.

Baameur, F., Morgan, D.H., Yao, H., Tran, T.M., Hammitt, R.A., Sabui, S., McMurray, J.S., Lichtarge, O., and Clark, R.B. (2010). Role for the regulator of G-protein signaling homology domain of G protein-coupled receptor kinases 5 and 6 in β2-adrenergic receptor and rhodopsin phosphorylation. Mol. Pharmacol. *77*, 405–415.

Badeanlou, L., Furlan-Freguia, C., Yang, G., Ruf, W., and Samad, F. (2011). Tissue factor-protease-activated receptor 2 signaling promotes diet-induced obesity and adipose inflammation. Nat. Med. *17*, 1490–1497.

Ballesteros, J.A., and Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods Neurosci.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al. (2017). UniProt: The universal protein knowledgebase. Nucleic Acids Res.

Bedbrook, C.N., Yang, K.K., Robinson, J.E., Mackey, E.D., Gradinaru, V., and Arnold, F.H. (2019).

`

Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. Nat. Methods *16*, 1176–1184.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res.

Blanpain, C., Lee, B., Vakili, J., Doranz, B.J., Govaerts, C., Migeotte, I., Sharron, M., Dupriez, V., Vassart, G., Doms, R.W., et al. (1999). Extracellular cysteines of CCR5 are required for chemokine binding, but dispensable for HIV-1 coreceptor activity. J. Biol. Chem.

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., and Bruford, E. (2019). Genenames.org: The HGNC and VGNC resources in 2019. Nucleic Acids Res. *47*, D786–D792.

Burnstock, G. (2013). Introduction to Purinergic Signalling in the Brain. (Springer, Dordrecht), pp. 1–12.

Burnstock, G., Fredholm, B.B., North, R.A., and Verkhratsky, A. (2010). The birth and postnatal development of purinergic signalling. Acta Physiol. *199*, 93–147.

Cahill, T.J., Thomsen, A.R.B., Tarrasch, J.T., Plouffe, B., Nguyen, A.H., Yang, F., Huang, L.Y., Kahsai, A.W., Bassoni, D.L., Gavino, B.J., et al. (2017). Distinct conformations of GPCR-β-arrestin complexes mediate desensitization, signaling, and endocytosis. Proc. Natl. Acad. Sci. U. S. A.

Cao, J., Panetta, R., Yue, S., Steyaert, A., Young-Bellido, M., and Ahmad, S. (2003). A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics *19*, 234–240.

Carpenter, B., Nehmé, R., Warne, T., Leslie, A.G.W., and Tate, C.G. (2016). Structure of the adenosine A2A receptor bound to an engineered G protein. Nature *536*, 104–107.

Casari, G., Sander, C., and Valencia, A. (1995). A method to predict functional residues in proteins. Nat. Struct. Biol.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Conklin, B.R., Farfel, Z., Lustig, K.D., Julius, D., and Bourne, H.R. (1993). Substitution of three amino acids switches receptor specificity of G qα to that of Giα. Nature.

Dann, C.E., Hsieh, J.C., Rattner, A., Sharma, D., Nathans, J., and Leahy, D.J. (2001). Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. Nature *412*, 86–90.

Dhole, K., Singh, G., Pai, P.P., and Mondal, S. (2014). Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier. J. Theor. Biol. *348*, 47–54.

Divorty, N., Mackenzie, A.E., Nicklin, S.A., and Milligan, G. (2015). G protein-coupled receptor 35: An emerging target in inflammatory and cardiovascular disease. Front. Pharmacol. *6*.

Dorsam, R.T., and Gutkind, J.S. (2007). G-protein-coupled receptors and cancer. Nat. Rev. Cancer *7*, 79–94.

Dou, Y., Wang, J., Yang, J., and Zhang, C. (2012). L1pred: A sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier. PLoS One.

Draper-Joyce, C.J., Khoshouei, M., Thal, D.M., Liang, Y.L., Nguyen, A.T.N., Furness, S.G.B., Venugopal, H., Baltos, J.A., Plitzko, J.M., Danev, R., et al. (2018). Structure of the adenosine-bound human adenosine A1 receptor-Gi complex. Nature *558*, 559–565.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Eichel, K., Jullié, D., Barsi-Rhyne, B., Latorraca, N.R., Masureel, M., Sibarita, J.B., Dror, R.O., and Von Zastrow, M. (2018). Catalytic activation of β-Arrestin by GPCRs. Nature.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res.

`

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., and Lin, C.J. (2008). LIBLINEAR: A library for large linear classification. J. Mach. Learn. Res.

Farrell, M.S., Pei, Y., Wan, Y., Yadav, P.N., Daigle, T.L., Urban, D.J., Lee, H.M., Sciaky, N., Simmons, A., Nonneman, R.J., et al. (2013). A Gα;s DREADD mouse for selective modulation of cAMP production in striatopallidal neurons. Neuropsychopharmacology *38*, 854–862.

Flask (2018). Flask Web Framework.

Flock, T., Ravarani, C.N.J., Sun, D., Venkatakrishnan, A.J., Kayikci, M., Tate, C.G., Veprintsev, D.B., and Babu, M.M. (2015). Universal allosteric mechanism for Gα activation by GPCRs. Nature.

Flock, T., Hauser, A.S., Lund, N., Gloriam, D.E., Balaji, S., and Babu, M.M. (2017). Selectivity determinants of GPCR-G-protein binding. Nature *545*, 317–322.

García-Nafría, J., Nehmé, R., Edwards, P.C., and Tate, C.G. (2018a). Cryo-EM structure of the serotonin 5-HT1B receptor coupled to heterotrimeric Go. Nature *558*, 620–623.

García-Nafría, J., Lee, Y., Bai, X., Carpenter, B., and Tate, C.G. (2018b). Cryo-EM structure of the adenosine A2A receptor coupled to an engineered heterotrimeric G protein. Elife *7*.

Gaudet, P., Michel, P.A., Zahn-Zabal, M., Britan, A., Cusin, I., Domagalski, M., Duek, P.D., Gateau, A., Gleizes, A., Hinard, V., et al. (2017). The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res.

Girault, J.A., and Greengard, P. (2004). The Neurobiology of Dopamine Signaling. In Archives of Neurology, (Arch Neurol), pp. 641–644.

Girkontaite, I., Missy, K., Sakk, V., Harenberg, A., Tedford, K., Pötzel, T., Pfeffer, K., and Fischer, K.D. (2001). Lsc is required for marginal zone B cells, regulation of lymphocyte motility and immune responses. Nat. Immunol. *2*, 855–862.

Grace, C.R.R., Perrin, M.H., DiGruccio, M.R., Miller, C.L., Rivier, J.E., Vale, W.W., and Riek, R. (2004). NMR structure and peptide hormone binding site of the first extracellular domain of a type B1 G protein-coupled receptor. Proc. Natl. Acad. Sci. U. S. A. *101*, 12836–12841.

Guettier, J.M., Gautam, D., Scarselli, M., De Azua, I.R., Li, J.H., Rosemond, E., Ma, X., Gonzalez, F.J., Armbruster, B.N., Lu, H., et al. (2009). A chemical-genetic approach to study G protein regulation of β cell function in vivo. Proc. Natl. Acad. Sci. U. S. A. *106*, 19197–19202.

Gurevich, V. V., and Gurevich, E. V. (2019). GPCR signaling regulation: The role of GRKs and arrestins. Front. Pharmacol. *10*, 125.

Gutkind, J.S., and Kostenis, E. (2018). Arrestins as rheostats of GPCR signalling. Nat. Rev. Mol. Cell Biol. *19*, 615–616.

Hannenhalli, S.S., and Russell, R.B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. J. Mol. Biol. *303*, 61–76.

Hanson, R.M., Prilusky, J., Renjian, Z., Nakane, T., and Sussman, J.L. (2013). JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. Isr. J. Chem.

Harding, S.D., Sharman, J.L., Faccenda, E., Southan, C., Pawson, A.J., Ireland, S., Gray, A.J.G., Bruce, L., Alexander, S.P.H., Anderton, S., et al. (2018). The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: Updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Res. *46*, D1091–D1106.

Hauser, A.S., Attwood, M.M., Rask-Andersen, M., Schiöth, H.B., and Gloriam, D.E. (2017). Trends in GPCR drug discovery: New agents, targets and indications. Nat. Rev. Drug Discov.

Heuberger, D.M., and Schuepbach, R.A. (2019). Protease-activated receptors (PARs): Mechanisms of action and potential therapeutic modulators in PAR-driven inflammatory diseases. Thromb. J. *17*.

Hiley, E., McMullan, R., and Nurrish, S.J. (2006). The Gα12-RGS RhoGEF-RhoA signalling pathway regulates neurotransmitter release in C. elegans. EMBO J. *25*, 5884–5895.

`

Hilger, D., Masureel, M., and Kobilka, B.K. (2018). Structure and dynamics of GPCR signaling complexes. Nat. Struct. Mol. Biol.

Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (2014). KinomeXplorer: An integrated platform for kinome biology studies. Nat. Methods *11*, 603–604.

Hsu, S.H., and Luo, C.W. (2007). Molecular dissection of G protein preference using Gsα chimeras reveals novel ligand signaling of GPCRs. Am. J. Physiol. - Endocrinol. Metab. *293*.

Hu, J., Stern, M., Gimenez, L.E., Wanka, L., Zhu, L., Rossi, M., Meister, J., Inoue, A., Beck-Sickinger, A.G., Gurevich, V. V., et al. (2016). A G protein-biased designer G protein-coupled receptor useful for studying the physiological relevance of Gq/11-dependent signaling pathways. J. Biol. Chem. *291*, 7809–7820.

Huang, T.S., and Krebs, E.G. (1977). Amino acid sequence of a phosphorylation site in skeletal muscle glycogen synthetase. Biochem. Biophys. Res. Commun. *75*, 643–650.

Huang, W., Manglik, A., Venkatakrishnan, A.J., Laeremans, T., Feinberg, E.N., Sanborn, A.L., Kato, H.E., Livingston, K.E., Thorsen, T.S., Kling, R.C., et al. (2015). Structural insights into μ-opioid receptor activation. Nature.

Huang, W., Masureel, M., Qu, Q., Janetzko, J., Inoue, A., Kato, H.E., Robertson, M.J., Nguyen, K.C., Glenn, J.S., Skiniotis, G., et al. (2020). Structure of the neurotensin receptor 1 in complex with β-arrestin 1. Nature *579*, 303–308.

Inoue, A., Ishiguro, J., Kitamura, H., Arima, N., Okutani, M., Shuto, A., Higashiyama, S., Ohwada, T., Arai, H., Makide, K., et al. (2012). TGFα shedding assay: An accurate and versatile method for detecting GPCR activation. Nat. Methods *9*, 1021–1029.

Inoue, A., Raimondi, F., Marie, F., Kadji, N., Gutkind, J.S., Aoki, J., and Russell, R.B. (2019). Illuminating G-Protein-Coupling Selectivity of GPCRs Graphical Abstract Highlights d Large datasets of quantitative coupling between 148 human GPCRs and 11 G proteins d Identification of GPCR sequence-encoded features underlying G-protein selectivity d A . Cell *177*, 1933-1947.e25.

Insel, P.A., Tang, C.M., Hahntow, I., and Michel, M.C. (2007). Impact of GPCRs in clinical medicine: Monogenic diseases, genetic variants and drug targets. Biochim. Biophys. Acta - Biomembr.

Isberg, V., Mordalski, S., Munk, C., Rataj, K., Harpsøe, K., Hauser, A.S., Vroling, B., Bojarski, A.J., Vriend, G., and Gloriam, D.E. (2016). GPCRdb: An information system for G protein-coupled receptors. Nucleic Acids Res. *44*, D356–D364.

Kalinina, O. V., Gelfand, M.S., and Russell, R.B. (2009). Combining specificity determining and conserved residues improves functional site prediction. BMC Bioinformatics *10*, 174.

Kang, Y., Kuybeda, O., De Waal, P.W., Mukherjee, S., Van Eps, N., Dutka, P., Zhou, X.E., Bartesaghi, A., Erramilli, S., Morizumi, T., et al. (2018). Cryo-EM structure of human rhodopsin bound to an inhibitory G protein. Nature *558*, 553–558.

Katoh, H., Aoki, J., Yamaguchi, Y., Kitano, Y., Ichikawa, A., and Negishi, M. (1998). Constitutively active Gα12, Gα13, and Gα(q) induce rho-dependent neurite retraction through different signaling pathways. J. Biol. Chem. *273*, 28700–28707.

Kawano, T., Baki, A., Xiang, G., and Logothetis, D.E. (2016). Construction of G Alpha-16 Chimeras for Detection of GPCR Activation. Biophys. J. *110*, 425a.

Kendall, R.T., and Senogles, S.E. (2011). Isoform-specific uncoupling of the D 2 dopamine receptors subtypes. Neuropharmacology *60*, 336–342.

Kilgore, W.R., Mantyh, P.W., Mantyh, C.R., McVey, D.C., and Vigna, S.R. (1993). Bombesin/GRP-preferring and neuromedin B-preferring receptors in the rat urogenital system. Neuropeptides *24*, 43–52.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res.

`

Kimple, M.E., Neuman, J.C., Linnemann, A.K., and Casey, P.J. (2014). Inhibitory G proteins and their receptors: Emerging therapeutic targets for obesity and diabetes. Exp. Mol. Med.

Kitayama, J., Shida, D., Sako, A., Ishikawa, M., Hama, K., Aoki, J., Arai, H., and Nagawa, H. (2004). Over-expression of lysophosphatidic acid receptor-2 in human invasive ductal carcinoma. Breast Cancer Res. *6*, R640.

Koehl, A., Hu, H., Maeda, S., Zhang, Y., Qu, Q., Paggi, J.M., Latorraca, N.R., Hilger, D., Dawson, R., Matile, H., et al. (2018). Structure of the μ-opioid receptor-Gi protein complex. Nature *558*, 547–552.

Kroeze, W.K., Sassano, M.F., Huang, X.P., Lansu, K., McCorvy, J.D., Giguère, P.M., Sciaky, N., and Roth, B.L. (2015). PRESTO-Tango as an open-source resource for interrogation of the druggable human GPCRome. Nat. Struct. Mol. Biol. *22*, 362–369.

Kumar, A., Dhull, D.K., and Mishra, P.S. (2015). Therapeutic potential of mGluR5 targeting in Alzheimer's disease. Front. Neurosci. *9*, 215.

Kumari, P., Srivastava, A., Ghosh, E., Ranjan, R., Dogra, S., Yadav, P.N., and Shukla, A.K. (2017). Core engagement with β-arrestin is dispensable for agonist-induced vasopressin receptor endocytosis and ERK activation. Mol. Biol. Cell.

Latorraca, N.R., Venkatakrishnan, A.J., and Dror, R.O. (2017). GPCR dynamics: Structures in motion. Chem. Rev. *117*, 139–155.

Latorraca, N.R., Wang, J.K., Bauer, B., Townshend, R.J.L., Hollingsworth, S.A., Olivieri, J.E., Xu, H.E., Sommer, M.E., and Dror, R.O. (2018). Molecular mechanism of GPCR-mediated arrestin activation. Nature.

Lauckner, J.E., Jensen, J.B., Chen, H.-Y., Lu, H.-C., Hille, B., and Mackie, K. (2008). GPR55 is a cannabinoid receptor that increases intracellular calcium and inhibits M current.

Lee, M.D., and Danileiko, I. (2014). Using cognitive models to combine probability estimates.

Lee, Y., Warne, T., Nehmé, R., Pandey, S., Dwivedi-Agnihotri, H., Chaturvedi, M., Edwards, P.C., García-Nafría, J., Leslie, A.G.W., Shukla, A.K., et al. (2020). Molecular basis of β-arrestin coupling to formoterol-bound β1-adrenoceptor. Nature *583*, 862–866.

Lee, Z., Swaby, R.F., Liang, Y., Yu, S., Liu, S., Lu, K.H., Bast, R.C., Mills, G.B., and Fang, X. (2006). Lysophosphatidic acid is a major regulator of growth-regulated oncogene α in ovarian cancer. Cancer Res.

Li, L., Homan, K.T., Vishnivetskiy, S.A., Manglik, A., Tesmer, J.J.G., Gurevich, V. V., and Gurevich, E. V. (2015). G protein-coupled receptor kinases of the GRK4 protein subfamily phosphorylate inactive G protein-coupled receptors (GPCRs). J. Biol. Chem. *290*, 10775–10790.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol.

Lin, X., Li, M., Wang, N., Wu, Y., Luo, Z., Guo, S., Han, G.W., Li, S., Yue, Y., Wei, X., et al. (2020). Structural basis of ligand recognition and self-activation of orphan GPR52. Nature *579*, 152–157.

Liu, X., He, Q., Studholme, D.J., Wu, Q., Liang, S., and Yu, L. (2004). NCD3G: A novel nine-cysteine domain in family 3 GPCRs. Trends Biochem. Sci. *29*, 458–461.

Lledo, P.M., Hjelmstad, G.O., Mukherji, S., Soderling, T.R., Malenka, R.C., and Nicoll, R.A. (1995). Calcium/calmodulin-dependent kinase II and long-term potentiation enhance synaptic transmission by the same mechanism. Proc. Natl. Acad. Sci. U. S. A. *92*, 11175–11179.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science (80-.).

López, C., Kleinheinz, K., Aukema, S.M., Rohde, M., Bernhart, S.H., Hübschmann, D., Wagener, R., Toprak, U.H., Raimondi, F., Kreuz, M., et al. (2019). Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. Nat. Commun.

Luo, J., Sun, P., Siwko, S., Liu, M., and Xiao, J. (2019). The role of GPCRs in bone diseases and

dysfunctions. Bone Res.

Manglik, A., Kim, T.H., Masureel, M., Altenbach, C., Yang, Z., Hilger, D., Lerch, M.T., Kobilka, T.S., Thian, F.S., Hubbell, W.L., et al. (2015). Structural insights into the dynamic process of β2-adrenergic receptor signaling. Cell *161*, 1101–1111.

Marti-Solano, M., Crilly, S.E., Malinverni, D., Munk, C., Harris, M., Pearce, A., Quon, T., Mackenzie, A.E., Wang, X., Peng, J., et al. (2020). Combinatorial expression of GPCR isoforms affects signalling and drug responses. Nature *587*, 650–656.

Marttinen, P., Corander, J., Tö Rö Nen, P., and Holm, L. (2006). Bayesian search of functionally divergent protein subgroups and their function specific residues. *22*, 2466–2474.

Matusiak, D., Glover, S., Nathaniel, R., Matkowskyj, K., Yang, J., and Benya, R. V. (2005). Neuromedin B and its receptor are mitogens in both normal and malignant epithelial cells lining the colon. Am. J. Physiol. - Gastrointest. Liver Physiol. *288*.

Mihalek, I., Reš, I., and Lichtarge, O. (2004). A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance. J. Mol. Biol. *336*, 1265–1282.

Mishra, S.K., Holzman, S., and Hoon, M.A. (2012). A nociceptive signaling role for neuromedin B. J. Neurosci. *32*, 8686–8695.

Moers, A., Nieswandt, B., Massberg, S., Wettschureck, N., Grüner, S., Konrad, I., Schulte, V., Aktas, B., Gratacap, M.P., Simon, M.I., et al. (2003). G13 is an essential mediator of platelet activation in hemostasis and thrombosis. Nat. Med. *9*, 1418–1422.

Möller, S., Vilo, J., and Croning, M.D.R. (2001). Prediction of the coupling specificity of G protein coupled receptors to their G proteins. Bioinformatics *17*, 1–8.

Muppidi, J.R., Schmitz, R., Green, J.A., Xiao, W., Larsen, A.B., Braun, S.E., An, J., Xu, Y., Rosenwald, A., Ott, G., et al. (2014). Loss of signalling via G α 13 in germinal centre B-cell-derived lymphoma. Nature.

Nguyen, A.H., Thomsen, A.R.B., Cahill, T.J., Huang, R., Huang, L.Y., Marcink, T., Clarke, O.B., Heissel, S., Masoudi, A., Ben-Hail, D., et al. (2019). Structure of an endosomal signaling GPCR–G protein–β-arrestin megacomplex. Nat. Struct. Mol. Biol.

Nishizawa, Y., Okui, Y., Inaba, M., Okuno, S., Yukioka, K., Miki, T., Watanabe, Y., and Morii, H. (1988). Calcium/calmodulin-mediated action of calcitonin on lipid metabolism in rats. J. Clin. Invest. *82*, 1165–1172.

Nojima, S., Fujita, Y., Kimura, K.T., Nomura, N., Suno, R., Morimoto, K., Yamamoto, M., Noda, T., Iwata, S., Shigematsu, H., et al. (2020). Cryo-EM Structure of the Prostaglandin E Receptor EP4 Coupled to G Protein. Structure.

Nørskov-Lauritsen, L., Jørgensen, S., and Bräuner-Osborne, H. (2015). N-glycosylation and disulfide bonding affects GPRC6A receptor expression, function, and dimerization. FEBS Lett. *589*, 588–597.

Nygaard, R., Zou, Y., Dror, R.O., Mildorf, T.J., Arlow, D.H., Manglik, A., Pan, A.C., Liu, C.W., Fung, J.J., Bokoch, M.P., et al. (2013). The dynamic process of β2-adrenergic receptor activation. Cell.

O'Hayre, M., Inoue, A., Kufareva, I., Wang, Z., Mikelis, C.M., Drummond, R.A., Avino, S., Finkel, K., Kalim, K.W., Dipasquale, G., et al. (2016). Inactivating mutations in GNA13 and RHOA in Burkitt's lymphoma and diffuse large B-cell lymphoma: A tumor suppressor function for the Gα13/RhoA axis in B cells. Oncogene.

Offermanns, S., Mancino, V., Revel, J.P., and Simon, M.I. (1997). Vascular system defects and impaired cell chemokinesis as a result of Gα13 deficiency. Science (80-.). *275*, 533–536.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F., Cesareni, G., et al. (2012). Protein interaction data curation: The International Molecular Exchange (IMEx) consortium. Nat. Methods.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., et al. (2000). Crystal structure of rhodopsin: A G protein-coupled

receptor. Science (80-.). *289*, 739–745.

Pantel, J., Legendre, M., Cabrol, S., Hilal, L., Hajaji, Y., Morisset, S., Nivot, S., Vie-Luton, M.P., Grouselle, D., De Kerdanet, M., et al. (2006). Loss of constitutive activity of the growth hormone secretagogue receptor in familial short stature. J. Clin. Invest. *116*, 760–768.

Parma, J., Duprez, L., Van Sande, J., Hermans, J., Rocmans, P., Van Vliet, G., Costagliola, S., Rodien, P., Dumont, J.E., and Vassart, G. (1997). Diversity and Prevalence of Somatic Mutations in the Thyrotropin Receptor and G s α Genes as a Cause of Toxic Thyroid Adenomas 1 . J. Clin. Endocrinol. Metab.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Perlman, J.H., Wang, W., Nussenzveig, D.R., and Gershengorn, M.C. (1995). A disulfide bond between conserved extracellular cysteines in the thyrotropin-releasing hormone receptor is critical for binding. J. Biol. Chem. *270*, 24682–24685.

Proud, C.G., Rylatt, D.B., Yeaman, S.J., and Cohen, P. (1977). Amino acid sequences at the two sites on glycogen synthetase phosphorylated by cyclic AMP-dependent protein kinase and their dephosphorylation by protein phosphatase-III. FEBS Lett. *80*, 435–442.

Qian, W., Gang, X., Zhang, T., Wei, L., Yang, X., Li, Z., Yang, Y., Song, L., Wang, P., Peng, J., et al. (2017). Protein kinase A-mediated phosphorylation of the Broad-Complex transcription factor in silkworm suppresses its transcriptional activity. J. Biol. Chem. *292*, 12460–12470.

Qin, K., Dong, C., Wu, G., and Lambert, N.A. (2011). Inactive-state preassembly of Gq-coupled receptors and G q heterotrimers. Nat. Chem. Biol. *7*, 740–747.

Ranjan, R., Dwivedi, H., Baidya, M., Kumar, M., and Shukla, A.K. (2017). Novel Structural Insights into GPCR–β-Arrestin Interaction and Signaling. Trends Cell Biol. *27*, 851–862.

Rankin, M.L., Marinec, P.S., Cabrera, D.M., Wang, Z., Jose, P.A., and Sibley, D.R. (2006). The D1 dopamine receptor is constitutively phosphorylated by G protein-coupled receptor kinase 4. Mol. Pharmacol.

Rasmussen, S.G.F., Devree, B.T., Zou, Y., Kruse, A.C., Chung, K.Y., Kobilka, T.S., Thian, F.S., Chae, P.S., Pardon, E., Calinski, D., et al. (2011). Crystal structure of the β 2 adrenergic receptor-Gs protein complex. Nature *477*, 549–557.

Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand binding: From protein subfamilies to functional specificity. Proc. Natl. Acad. Sci. U. S. A. *107*, 1995–2000.

Reva, B., Antipin, Y., and Sander, C. (2007). Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. *8*, R232.

Ribeiro, R.C., Sandrini, F., Figueiredo, B., Zambetti, G.P., Michalkiewicz, E., Lafferty, A.R., DeLacerda, L., Rabin, M., Cadwell, C., Sampaio, G., et al. (2001). An inherited p53 mutation that contributes in a tissue-specific manner to pediatric adrenal cortical carcinoma. Proc. Natl. Acad. Sci. U. S. A.

Rocha, J.L., Friedman, E., Boson, W., Moreira, A., Figueiredo, B., Liberman, B., De Lacerda, L., Sandrini, R., Graf, H., Martins, S., et al. (1999). Molecular analyses of the vasopressin type 2 receptor and aquaporin-2 genes in brazilian kindreds with nephrogenic diabetes insipidus. Hum. Mutat. *14*, 233–239.

Rodriguez, G.J., Yao, R., Lichtarge, O., and Wensel, T.G. (2010). Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. Proc. Natl. Acad. Sci. U. S. A. *107*, 7787–7792.

Romero, P.A., Krause, A., and Arnold, F.H. (2013). Navigating the protein fitness landscape with Gaussian processes. Proc. Natl. Acad. Sci. U. S. A. *110*, E193–E201.

Rondard, P., Liu, J., Huang, S., Malhaire, F., Vol, C., Pinault, A., Labesse, G., and Pin, J.P. (2006). Coupling of agonist binding to effector domain activation in metabotropic glutamate-like receptors. J.

Biol. Chem. *281*, 24653–24661.

Rosenbaum, D.M., Rasmussen, S.G.F., and Kobilka, B.K. (2009). The structure and function of G-protein-coupled receptors. Nature *459*, 356–363.

Rosenthal, W., Seibold, A., Antaramian, A., Lonergan, M., Arthus, M.F., Hendy, G.N., Birnbaumer, M., and Bichet, D.G. (1992). Molecular identification of the gene responsible for congenital nephrogenic diabetes insipidus. Nature *359*, 233–235.

Roth, B.L. (2016). DREADDs for Neuroscientists. Neuron *89*, 683–694.

de Roux, N., Misrahi, M., Brauner, R., Houang, M., Carel, J.C., Granier, M., Le Bouc, Y., Ghinea, N., Boumedienne, A., Toublanc, J.E., et al. (1996). Four families with loss of function mutations of the thyrotropin receptor. J. Clin. Endocrinol. Metab. *81*, 4229–4235.

Ruppel, K.M., Willison, D., Kataoka, H., Wang, A., Zheng, Y.W., Cornelissen, I., Yin, L., Mei Xu, S., and Coughlin, S.R. (2005). Essential role for Gα13 in endothelial cells during embryonic development. Proc. Natl. Acad. Sci. U. S. A. *102*, 8281–8286.

Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. (2020). An evolution-based model for designing chorismate mutase enzymes. Science (80-.). *369*, 440–445.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: An online force field. Nucleic Acids Res. *33*.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710.

Sente, A., Peer, R., Srivastava, A., Baidya, M., Lesk, A.M., Balaji, S., Shukla, A.K., Babu, M.M., and Flock, T. (2018). Molecular mechanism of modulating arrestin conformation by GPCR phosphorylation. Nat. Struct. Mol. Biol. *25*, 538–545.

Sgourakis, N.G., Bagos, P.G., Papasaikas, P.K., and Hamodrakas, S.J. (2005a). A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models. BMC Bioinformatics *6*, 104.

Sgourakis, N.G., Bagos, P.G., and Hamodrakas, S.J. (2005b). Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. Bioinformatics *21*, 4101–4106.

Shenoy, S.K., Drake, M.T., Nelson, C.D., Houtz, D.A., Xiao, K., Madabushi, S., Reiter, E., Premont, R.T., Lichtarge, O., and Lefkowitz, R.J. (2006). β-arrestin-dependent, G protein-independent ERK1/2 activation by the β2 adrenergic receptor. J. Biol. Chem. *281*, 1261–1273.

Shida, D., Kitayama, J., Yamaguchi, H., Okaji, Y., Tsuno, N.H., Watanabe, T., Takuwa, Y., and Nagawa, H. (2003). Lysophosphatidic Acid (LPA) Enhances the Metastatic Potential of Human Colon Carcinoma DLD1 Cells through LPA1. Cancer Res. *63*.

Shinoura, H., Shibata, K., Hirasawa, A., Tanoue, A., Hashimoto, K., and Tsujimoto, G. (2002). Key amino acids for differential coupling of α1-adrenergic receptor subtypes to Gs. Biochem. Biophys. Res. Commun. *299*, 142–147.

Shukla, A.K., Xiao, K., and Lefkowitz, R.J. (2011). Emerging paradigms of β-arrestin-dependent seven transmembrane receptor signaling. Trends Biochem. Sci. *36*, 457–469.

Shukla, A.K., Manglik, A., Kruse, A.C., Xiao, K., Reis, R.I., Tseng, W.C., Staus, D.P., Hilger, D., Uysal, S., Huang, L.Y., et al. (2013). Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide. Nature *497*, 137–141.

Singh, G., Inoue, A., Gutkind, J.S., Russell, R.B., and Raimondi, F. (2019). PRECOG: PREdicting COupling probabilities of G-protein coupled receptors. Nucleic Acids Res. *47*, W395–W401.

Smith, J.S., Alagesan, P., Desai, N.K., Pack, T.F., Wu, J.H., Inoue, A., Freedman, N.J., and Rajagopal, S. (2017). C-X-C Motif Chemokine Receptor 3 Splice Variants Differentially Activate Beta-Arrestins to

`

Regulate Downstream Signaling Pathways s. Mol. Pharmacol.

Spanakis, E., Milord, E., and Gragnoli, C. (2008). AVPR2 variants and mutations in nephrogenic diabetes insipidus: Review and missense mutation significance. J. Cell. Physiol. *217*, 605–617.

Sreekumar, K.R., Huang, Y., Pausch, M.H., and Gulukota, K. (2004). Predicting GPCR-G-protein coupling using hidden Markov models. Bioinformatics *20*, 3490–3499.

Stacey, S.N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D.F., Magnusson, O.T., Gudjonsson, S.A., Sigurgeirsson, B., Thorisdottir, K., et al. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nat. Genet.

Stachniak, T.J., Ghosh, A., and Sternson, S.M. (2014). Chemogenetic Synaptic Silencing of Neural Circuits Localizes a Hypothalamus→Midbrain Pathway for Feeding Behavior. Neuron *82*, 797–808.

Staus, D.P., Hu, H., Robertson, M.J., Kleinhenz, A.L.W., Wingler, L.M., Capel, W.D., Latorraca, N.R., Lefkowitz, R.J., and Skiniotis, G. (2020). Structure of the M2 muscarinic receptor–β-arrestin complex in a lipid nanodisc. Nature *579*, 297–302.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackerman, Z., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. Cell.

Stoy, H., and Gurevich, V. V. (2015). How genetic errors in GPCRs affect their function: Possible therapeutic strategies. Genes Dis.

Suzuki, N., Hajicek, N., and Kozasa, T. (2009). Regulation and physiological functions of G12/13-mediated signaling pathways. NeuroSignals *17*, 55–70.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res.

Tansey, M.G., Luby-Phelps, K., Kamm, K.E., and Stull, J.T. (1994). Ca2+-dependent phosphorylation of myosin light chain kinase decreases the Ca2+ sensitivity of light chain phosphorylation within smooth muscle cells. J. Biol. Chem.

Tesmer, V.M., Kawano, T., Shankaranarayanan, A., Kozasa, T., and Tesmer, J.J.G. (2005). Snapshot of Activated G Proteins at the Membrane: The Gαq-GRK2-Gßγ Complex. Science (80-.).

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Thomsen, W., Frazer, J., and Unett, D. (2005). Functional assays for screening GPCR targets. Curr. Opin. Biotechnol. *16*, 655–665.

Tran, T.M., Friedman, J., Qunaibi, E., Baameur, F., Moore, R.H., and Clark, R.B. (2004). Characterization of Agonist Stimulation of cAMP-Dependent Protein Kinase and G Protein-Coupled Receptor Kinase Phosphorylation of β 2-Adrenergic Receptor Using Phosphoserine-Specific Antibodies. Mol. Pharmacol. *65*, 196–206.

Tremblay, J.J., and Viger, R.S. (2003). Transcription factor GATA-4 is activated by phosphorylation of serine 261 via the cAMP/protein kinase A signaling pathway in gonadal cells. J. Biol. Chem. *278*, 22128–22135.

Usui, I., Imamura, T., Babendure, J.L., Satoh, H., Lu, J.C., Hupfeld, C.J., and Olefsky, J.M. (2005). G protein-coupled receptor kinase 2 mediates endothelin-1-induced insulin resistance via the inhibition of both Gαq/11 and insulin receptor substrate-1 pathways in 3T3-L1 adipocytes. Mol. Endocrinol.

Velankar, S., Dana, J.M., Jacobsen, J., Van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res.

Venkatakrishnan, A.J., Deupi, X., Lebon, G., Tate, C.G., Schertler, G.F., and Madan Babu, M. (2013).

Molecular signatures of G-protein-coupled receptors. Nature *494*, 185–194.

Venkatakrishnan, A.J., Flock, T., Prado, D.E., Oates, M.E., Gough, J., and Madan Babu, M. (2014). Structured and disordered facets of the GPCR fold. Curr. Opin. Struct. Biol. *27*, 129–137.

Venkatakrishnan, A.J., Deupi, X., Lebon, G., Heydenreich, F.M., Flock, T., Miljus, T., Balaji, S., Bouvier, M., Veprintsev, D.B., Tate, C.G., et al. (2016). Diverse activation pathways in class A GPCRs converge near the G-protein-coupling region. Nature *536*, 484–487.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods.

Wang, C.J., Hsu, S.H., Hung, W.T., and Luo, C.W. (2009). Establishment of a chimeric reporting system for the universal detection and high-throughput screening of G protein-coupled receptors. Biosens. Bioelectron. *24*, 2298–2304.

Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci.

Wess, J., Nakajima, K., and Jain, S. (2013). Novel designer receptors to probe GPCR signaling and physiology. Trends Pharmacol. Sci.

Wheatley, M., Wootten, D., Conner, M.T., Simms, J., Kendrick, R., Logan, R.T., Poyner, D.R., and Barwell, J. (2012). Lifting the lid on GPCRs: The role of extracellular loops. Br. J. Pharmacol. *165*, 1688–1703.

Wichard, J.D., ter Laak, A., Krause, G., Heinrich, N., Kühne, R., and Kleinau, G. (2011). Chemogenomic analysis of G-protein coupled receptors and their ligands deciphers locks and keys governing diverse aspects of signalling. PLoS One.

Williams, N.G., Zhong, H., and Minneman, K.P. (1998). Differential coupling of α1-, α2-, and β-adrenergic receptors to mitogen-activated protein kinase pathways and differentiation in transfected PC12 cells. J. Biol. Chem. *273*, 24624–24632.

Worzfeld, T., Wettschureck, N., and Offermanns, S. (2008). G12/G13-mediated signalling in mammalian physiology and disease. Trends Pharmacol. Sci.

Wu, J.H., Goswami, R., Cai, X., Exum, S.T., Huang, X., Zhang, L., Brian, L., Premont, R.T., Peppel, K., and Freedman, N.J. (2006). Regulation of the platelet-derived growth factor receptor-β by G protein-coupled receptor kinase-5 in vascular smooth muscle cells involves the phosphatase Shp2. J. Biol. Chem.

Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H., and Suwa, M. (2005). GRIFFIN: A system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. Nucleic Acids Res. *33*.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020a). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A. *117*, 1496–1503.

Yang, Y.M., Kuen, D., and Chung, Y. (2020b). G α 12 / 13 signaling in metabolic diseases. Exp. Mol. Med.

Yin, W., Li, Z., Jin, M., Yin, Y.L., de Waal, P.W., Pal, K., Yin, Y., Gao, X., He, Y., Gao, J., et al. (2019). A complex structure of arrestin-2 bound to a G protein-coupled receptor. Cell Res. *29*, 971–983.

Yohannan, S., Faham, S., Yang, D., Whitelegge, J.P., and Bowie, J.U. (2004). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc. Natl. Acad. Sci. U. S. A.

Zheng, H., Worrall, C., Shen, H., Issad, T., Seregard, S., Girnita, A., and Girnita, L. (2012). Selective recruitment of G protein-coupled receptor kinases (GRKs) controls signaling of the insulin-like growth factor 1 receptor. Proc. Natl. Acad. Sci. U. S. A.

`

SUPPLEMENTARY INFORMATION

The supplementary information can be downloaded by at the following links:

Table S1 (http://russelllab.org/GS_thesis/Table_S1.xlsx)

Table S1A: Binding affinities of GPCR with G-proteins derived from the TGFα shedding assay

Table S1B: ROC curve analysis of the data from the TGFα shedding assay with GtoPdb

Table S1C: GPCR couplings in the data from the TGFα shedding assay and GtoPdb

Table S1D: Correspondences between Pfam 7tm_1 positions and BW numbering

Table S1E: Functional annotation of 7TM coupling features

Table S1F: Significance of extra-membrane region of GPCRs in determining G-protein coupling specificity

Table S2 (http://russelllab.org/GS_thesis/Table_S2.xlsx)

Table S2A: Cross-validation and test results of PRECOG

Table S2B: Test results PRECOG and PRED-COUPLE2 at the G-protein subfamily level

Table S2C: Comparison of 5-fold-CV between randomized and original dataset

Table S3 (http://russelllab.org/GS_thesis/Table_S3.xlsx)

Table S3A: Predicted probabilities of uncharacterized GPCRs

Table S3B: Predicted probabilities of mutated class A GPCRs reported in UniProt

Table S3C: Binding affinities of GPCR - G-proteins/β-arrestins in the data from the ebBRET assay

Table S3D: Statistical association of 7TM1 positional features with G-protein coupling group and known interfaces

Table S3E: Statistical association of 3D complexes with G-proteins/β-arrestins

Table S3F: Performance of the ebBRET assay-based predictor (without InterPreTS)

Table S3G: Performance of the ebBRET assay-based predictor (with InterPreTS)

Table S3H: Comparison of PRECOG with ebBRET assay-based predictors