Integrative Computational Frameworks for GPCR Biology: From Receptor Logic to Functional Modulation

By:

Sanjay Kumar Mohanty

PhD20204

Under the Supervision of:

Dr. Gaurav Ahuja

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

May, 2025

Integrative Computational Frameworks for GPCR Biology: From Receptor Logic to Functional Modulation

By:

Sanjay Kumar Mohanty

PhD20204

A Thesis submitted

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

To the

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

May, 2025

Certificate

This is to certify that the thesis titled "**Integrative Computational Frameworks for GPCR Biology: From Receptor Logic to Functional Modulation**" being submitted by **Mr. Sanjay Kumar Mohanty** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May, 2025

Dr. Gaurav Ahuja

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

Acknowledgment

First and foremost, I extend my heartfelt gratitude to my advisor, Dr. Gaurav Ahuja, for his unwavering support, scientific mentorship, and patience throughout this Ph.D. journey. His deep insights, critical thinking, and persistent encouragement have not only shaped the direction of this thesis but also helped me grow as an independent researcher. I am particularly thankful for the constructive feedback he always provided at the right time and the many opportunities he created for me to engage in meaningful scientific discourse, both in India and abroad.

I am also thankful to IIIT-Delhi, and in particular the Department of Computational Biology, for fostering an inspiring academic environment. A special appreciation goes to the IT support team, and in particular Adarsh Sir, whose timely assistance ensured the smooth execution of compute-heavy experiments (and often saved us from last-minute panic for server hangs).

This work has greatly benefited from collaborations with researchers across disciplines. I am deeply thankful to my fellow collaborators from CSIR-Central Drug Research Institute, Lucknow, and CSIR-Institute of Microbial Technology, Chandigarh, whose contributions were vital during the experimental validation. Your expertise, timely support, and collaborative spirit significantly strengthened this research and pushed it beyond the computational realm into meaningful biological insight.

To my internal and external examiners, thank you for your time, critical review, and thoughtful feedback. Your questions challenged me to think more deeply about my work and refine its presentation with clarity and rigor.

To the Department of Science & Technology (DST) for providing me with the "Innovation in Science Pursuit for Inspired Research (INSPIRE), thank you for the research fellowship to support, as this thesis would not have been possible without the resources, travel opportunities, and academic independence that your support enabled.

A big thank-you to my fellow PhD colleagues and labmates, who made the journey not only intellectually stimulating but also emotionally manageable. I've been fortunate to share this path with a brilliant and supportive group of individuals. They transitioned from providing a helping hand to becoming a hub for fun and entertainment. I would like to extend a special mention to some interns who have been more than just work colleagues to me and continued to provide invaluable support.

A huge heartfelt thanks to my senior and "Di," who has made this journey not just bearable, but truly enjoyable. You've been with me even before the Ph.D. began, and our shared love for anime and mutual

appreciation for Asian cuisine laid the foundation for a friendship that only deepened over time. From endless discussions and late-night work marathons to shared triumphs, stressful submissions, celebratory dinners, and the inevitable frustrations of failed experiments, your presence brought light and laughter to even the most difficult phases of this journey.

To my closest friends, Sangeeta, Sukirti, Priyashree, and Rakesh, thank you for being the emotional constant when the research rollercoaster felt overwhelming. From random distractions to grounding conversations, you made sure I didn't lose myself in the chaos of deadlines and data. A special note of gratitude to Suvendu, who has been with me every step of the way, first as a classmate, then a roommate, a Ph.D. colleague, and ultimately, a brother. Your unwavering support, especially during times of illness and personal lows, has been a pillar of strength throughout this journey.

To my family, no words can truly capture my gratitude. Thank you, [Mom, Dad], for your unwavering love, trust, and sacrifice. You've been my biggest supporters, even when you didn't fully understand the science, and perhaps that made it even more meaningful.

And finally, to God, fate, or whatever mysterious force keeps the universe in balance, thank you for the sheer luck, I believe, that got me through the rough patches and brought unexpected clarity when it was most needed.

This journey has been long and intense, filled with learning, unlearning, and rediscovery. I am deeply grateful to everyone who walked even a small part of this path with me.

May, 2025

Sanjay Kumar Mohanty

Abstract

Abstract

G-protein-coupled receptors (GPCRs) are vital pharmaceutical targets, with more than one-third of FDA-approved drugs influencing their function. While central to cellular signaling and drug development, GPCR research is hindered by various challenges. This thesis introduces several innovative tools and algorithms designed to deepen our understanding of GPCR biology.

First, Reverse Cell Tracking (RCT), a novel computational framework that leverages RNA velocity embeddings to trace gene expression trajectories during cellular differentiation. Applying RCT to investigate odorant receptor (OR) gene expression during neuronal development, we uncovered insights into OR gene choice mechanisms. ORs, a subset of GPCRs traditionally associated with smell, are also expressed in non-olfactory tissues, including cancers, implicating them in processes such as migration, proliferation, and immune modulation. Their expression follows a unique "one neuron, one receptor" rule, driven by mutual exclusivity and monoallelic expression. However, recent single-cell studies have revealed co-expression of multiple ORs in immature neurons, suggesting alternative models such as winner-takes-all or stochastic selection. RCT analysis revealed a bias toward the most highly expressed OR during differentiation, offering potential breakthroughs in understanding OR expression patterns and could open up new avenues for diagnostics and therapeutic targeting outside the nose, especially in diseases like cancer, where altered GPCR signaling plays a critical role.

Second, Machine-OlF-Action (MOA), a user-friendly, open-source computational framework designed to support GPCR researchers with minimal programming experience. As GPCR signaling gains prominence, there is a growing demand for accessible tools to efficiently explore and model GPCR-ligand interactions. While machine learning-based techniques are emerging as state-of-the-art approaches in chemoinformatics, enabling selective, effective, and rapid identification of biologically relevant molecules from vast chemical databases, their broader adoption in GPCR research has been limited due to their reliance on advanced computational skills, as well as the technical complexity of existing tools. MOA bridges this gap by allowing users to input SMILES strings and known activation statuses of compounds to build reliable classification models. By simplifying complex machine learning workflows into an accessible platform, MOA enables even researchers without a deep computational background to uncover meaningful GPCR-ligand relationships and advance the field of chemosensory biology.

Third, Gcoupler, an AI-driven computational toolkit that combines *de novo* ligand design, advanced statistical approaches, Graph Neural Networks, and bioactivity-based prioritization to facilitate the unbiased identification of druggable surface cavities and the rational prediction of high-affinity ligands. While conventional GPCR-targeted therapies predominantly focus on orthosteric sites, emerging research

highlights the therapeutic potential of allosteric sites. Despite the development of synthetic allosteric modulators, endogenous intracellular modulators remain largely unexplored due to a lack of comprehensive binding and phenotypic data. This data scarcity limits the applicability of traditional machine learning approaches. Gcoupler addresses this challenge by enabling cavity-specific predictions and ligand identification even in data-scarce GPCR contexts, paving the way for more targeted and effective drug discovery.

This research introduces a suite of computational frameworks tailored to advance GPCR-targeted drug discovery by addressing key bottlenecks in modeling, data scarcity, and accessibility. These findings challenge the conventional view of OR expression and provide fresh insights into their functional roles beyond the olfactory system. By simplifying complex workflows and integrating AI-driven methods, these tools democratize computational biology for researchers with limited coding expertise. Collectively, they enhance the understanding of chemosensory GPCRs, enable unbiased ligand prioritization, and offer new strategies to tackle data-scarce targets, ultimately accelerating the development of selective and effective therapeutics.

List of Publications (link)

Thesis Related Publications (3)

1. **Mohanty, Sanjay Kumar**, Aayushi Mittal, Aakash Gaur, Subhadeep Duari, Saveena Solanki, Anmol Kumar Sharma, Sakshi Arora et al. 2025. "Deep Learning Reveals Endogenous Sterols as Allosteric Modulators of the GPCR-Gα Interface." *eLife,* May. ***(first author)***

2. **Mohanty, Sanjay Kumar**, Sidrah Maryam, Vishakha Gautam, Aayushi Mittal, Krishan Gupta, Radhika Arora, Wrik Bhadra, Tripti Mishra, Debarka Sengupta, and Gaurav Ahuja. 2022. "Transcriptional Advantage Influence Odorant Receptor Gene Choice." *Briefings in Functional Genomics*, December. ***(first author)***

3. Gupta, Anku, Mohit Choudhary, **Sanjay Kumar Mohanty**, Aayushi Mittal, Krishan Gupta, Aditya Arya, Suvendu Kumar, et al. 2021. "Machine-OlF-Action: A Unified Framework for Developing and Interpreting Machine-Learning Models for Chemosensory Research." *Bioinformatics*, January. ***(co-first author)***

Other Publications (9)

1. Arora, Sakshi, Aayushi Mittal, Subhadeep Duari, Sonam Chauhan, Nilesh Kumar Dixit, **Sanjay Kumar Mohanty**, Arushi Sharma, et al. 2024. "Discovering Geroprotectors through the Explainable Artificial Intelligence-Based Platform AgeXtend." *Nature Aging*, December. ***(contributing author; technical report)***

2. Arora, Sakshi, Shiva Satija, Aayushi Mittal, Saveena Solanki, **Sanjay Kumar Mohanty**, Vaibhav Srivastava, Debarka Sengupta, et al. 2024. "Unlocking The Mysteries of DNA Adducts with Artificial Intelligence." *Chembiochem: A European Journal of Chemical Biology* 25 (1): e202300577. ***(contributing author; review)***

3. Mittal, Aayushi, **Sanjay Kumar Mohanty**, and Mudit Gupta. 2023 "EVOLF: A MULTIMODAL DEORPHANIZATION APPROACH FOR THERAPEUTICALLY RELEVANT ODORANT RECEPTORS." *Chemical Senses*. bjad041.010 ***(second author; abstract)***

4. Mittal, Aayushi, **Sanjay Kumar Mohanty**, Vishakha Gautam, Sakshi Arora, Sheetanshu Saproo, Ria Gupta, S. Roshan, et al. 2022. "Artificial Intelligence Uncovers Carcinogenic Human Metabolites." *Nat Chem Biol* 18, 1204–1213 (2022). ***(co-first author)***

5. Gautam, Vishakha, Rahul Gupta, Deepti Gupta, Anubhav Ruhela, Aayushi Mittal, **Sanjay Kumar Mohanty**, Sakshi Arora, et al. 2022. "deepGraphh: AI-Driven Web Service for Graph-Based Quantitative Structure-activity Relationship Analysis." *Briefings in Bioinformatics* 23 (5): bbac288. ***(contributing author)***

6. Gupta, Ria, Aayushi Mittal, Vishesh Agrawal, Sushant Gupta, Krishan Gupta, Rishi Raj Jain, Prakriti Garg, **Sanjay Kumar Mohanty**, Riya Sogani, et al. 2021. “OdoriFy: A Conglomerate of Artificial Intelligence-Driven Prediction Engines for Olfactory Decoding.” *The Journal of Biological Chemistry* 297 (2): 100956. ***(contributing author)***

7. Gautam, Vishakha, Aayushi Mittal, Siddhant Kalra, **Sanjay Kumar Mohanty**, Krishan Gupta, Komal Rani, Srivatsava Naidu, Tripti Mishra, Debarka Sengupta, and Gaurav Ahuja. 2021. “EcTracker: Tracking and Elucidating Ectopic Expression Leveraging Large-Scale scRNA-Seq Studies.” *Briefings in Bioinformatics* 22 (6). ***(contributing author)***

8. Gupta, Krishan, **Sanjay Kumar Mohanty**, Aayushi Mittal, Siddhant Kalra, Suvendu Kumar, Tripti Mishra, Jatin Ahuja, Debarka Sengupta, and Gaurav Ahuja. 2021. “The Cellular Basis of Loss of Smell in 2019-nCoV-Infected Individuals.” *Briefings in Bioinformatics* 22 (2): 873–81. ***(co-first author)***

9. Gautam, Vishakha, Subhadeep Duari, Saveena Solanki, Mudit Gupta, Aayushi Mittal, Sakshi Arora, Anmol Aggarwal et al. 2025. "scCamAge: A context-aware prediction engine for cellular age, aging-associated bioactivities, and morphometrics." *Cell Reports* 44, no. 2. ***(contributing author)***

Table of Contents

List of Abbreviations

Abbreviation	Full Term
2D	2 Dimensional
3D	3 Dimensional
ADRA1A	Alpha-1A Adrenergic Receptor
AFP	Attentive FP
AI	Artificial Intelligence
AIR	Ambiguous Interaction Restraint
ANCOVA	Analysis of Covariance
AP	Average Precision
AUC	Area Under the Curve
BE	Binding Energy
BRET	Bioluminescence Resonance Energy Transfer
CoQ10	Ubiquinone 10
CoQ6	Coenzyme Q6
DGE	Differential Gene Expression
DEM	Differentially Enriched Metabolite
DOPE	Discrete Optimized Protein Energy
EC	Extracellular
EC1	Extracellular Cavity 1
ECDF	Empirical Cumulative Distribution Function
ET	Extra Tree
FBS	Fetal Bovine Serum
FRET	Förster Resonance Energy Transfer

FST	Fucosterol
GA	Genetic Algorithm
GAFF	General Amber Force Field
GAN	Generative Adversarial Network
GBM	Gradient Boosting Machine
GAT	Graph Attention Network
GBC	Globose Basal Cell
GCM	Graph Convolution Model
GCN	Graph Convolution Network
GNB	Gaussian Naive Bayes
GNN	Graph Neural Network
GPCR	G-protein-coupled Receptor
HAB	High-Affinity Binder
HAM	High-Affinity Metabolite
HBC	Horizontal Basal Cell
HK	Heat-Killed
IC	Intracellular
IC4	Intracellular Cavity 4
IC5	Intracellular Cavity 5
iOSN	Immature Olfactory Sensory Neuron
IQR	Interquartile Range
ISO	Isoproterenol
KEGG	Kyoto Encyclopedia of Genes and Genomes
kNN	k Nearest Neighbor
LAB	Low-Affinity Binder

LAM	Low-Affinity Metabolite
LBDD	Ligand-Based Drug Design
LIME	Locally Interpretable Model-Agnostic Explanation
LOOCV	Leave-One-Out Cross-Validation
LR	Logistic Regression
LST	Lanosterol
MAPK	Mitogen-Activated Protein Kinase
MCC	Matthews Correlation Coefficient
MD	Molecular Dynamics
ML	Machine Learning
MLP	Multi-Layer Perceptron
MOA	Machine-OlF-Action
MOE	Molecular Operating Environment
mOSN	Mature Olfactory Sensory Neuron
NB	Naive Bayes
ns	Nano Second
NPC	Neuronal Precursor Cell
OR	Odorant Receptor
Orco	Odorant receptor-co-receptor
ORA	Over Representation Analysis
OSN	Olfactory Sensory Neuron
PAGA	Partition-Based Graph Abstraction
PCA	Principal Component Analysis
PCD	Programmed Cell Death
PDB	Protein Data Bank

PI	Propidium Iodide
QC	Quality Control
RCT	Reverse Cell Tracking
RF	Random Forest
RMSD	Root Mean Square Deviation
RT	Room Temperature
SBDD	Structure-Based Drug Design
scRNA-seq	Single-Cell RNA Sequencing
SD	Standard Deviation
SMILES	Simplified Molecular-Input Line-Entry System
SP-LIME	Submodular Pick LIME
SVM	Support Vector Machine
TM	Transmembrane
tsOSN	Transition State Olfactory Sensory Neuron
UMAP	Uniform Manifold Approximation and Projection
WGA	Wheat Germ Agglutinin
WT	Wild Type
YMDB	Yeast Metabolome Database
YPD	Yeast Extract Peptone Dextrose
ZST	Zymosterol

List of Figures

List of Tables

Introduction

Chapter 1 - Introduction

The central dogma of molecular biology, articulated originally by Francis Crick in 1958 and refined in 1970, describes the directional flow of genetic information within biological systems, from DNA to RNA to protein (Crick 1970; Nirenberg 2004). This fundamental principle underscores the essential process through which genetic instructions encoded within DNA are transcribed into messenger RNA (mRNA), which is then translated into functional proteins (Alberts et al. 2015). While reverse transcription and RNA replication represent important exceptions to this unidirectional schema, they occur under specific circumstances and do not negate the central dogma's core tenet regarding protein synthesis (Baltimore 1970; Temin and Mizutani 1970). Proteins thus serve as the ultimate executors of genetic information, shaping cellular identity and orchestrating virtually every physiological and biochemical function (Berg et al. 2015). The intricate mechanisms governing transcription and translation are tightly regulated through multilayered control systems, ensuring appropriate spatiotemporal expression patterns essential for development and homeostasis (Levine and Tjian 2003; Maston, Evans, and Green 2006). By acting as enzymes, structural scaffolds, molecular transporters, signaling mediators, and regulators of gene expression, proteins constitute the primary operational units, directly modulating cellular activities and ensuring the adaptability and survival of cells in response to internal signals and external environmental stimuli (Hartl 2017). The complexity of post-transcriptional and post-translational modifications further expands the functional diversity of the proteome, allowing for a remarkable degree of cellular specialization despite the constraints of a finite genome (Khoury, Baliban, and Floudas 2011; Jensen 2006).

Within the extensive and varied array of proteins, cell surface receptors, especially GPCRs, hold a critically important position. As the largest and most versatile superfamily of membrane receptors, GPCRs are integral to the processes of intercellular communication and signal transduction (Robert Fredriksson et al. 2003; R. Fredriksson and Schiöth 2006; Rosenbaum, Rasmussen, and Kobilka 2009). Structurally characterized by seven transmembrane domains, GPCRs function primarily by sensing extracellular signals, ranging from photons and odorants to hormones, neurotransmitters, chemokines, and metabolites, and subsequently translating these cues into intracellular biochemical signals (Weis and Kobilka 2018; Stevens et al. 2013). This translation is achieved through the activation of heterotrimeric G proteins, composed of α, β, and γ subunits, which subsequently modulate various downstream signaling pathways, including adenylyl cyclase-cAMP, phospholipase C-inositol trisphosphate (IP3), calcium signaling, and mitogen-activated protein kinase (MAPK) cascades (Oldham and Hamm 2008; Wettschureck and Offermanns 2005; Hilger, Masureel, and Kobilka 2018). This molecular diversity underlies the remarkable versatility of GPCR signaling networks, allowing them to regulate a multitude of

physiological processes ranging from sensory perception and neurotransmission to immune function, metabolism, and development (Venkatakrishnan et al. 2013; Rockman, Koch, and Lefkowitz 2002).

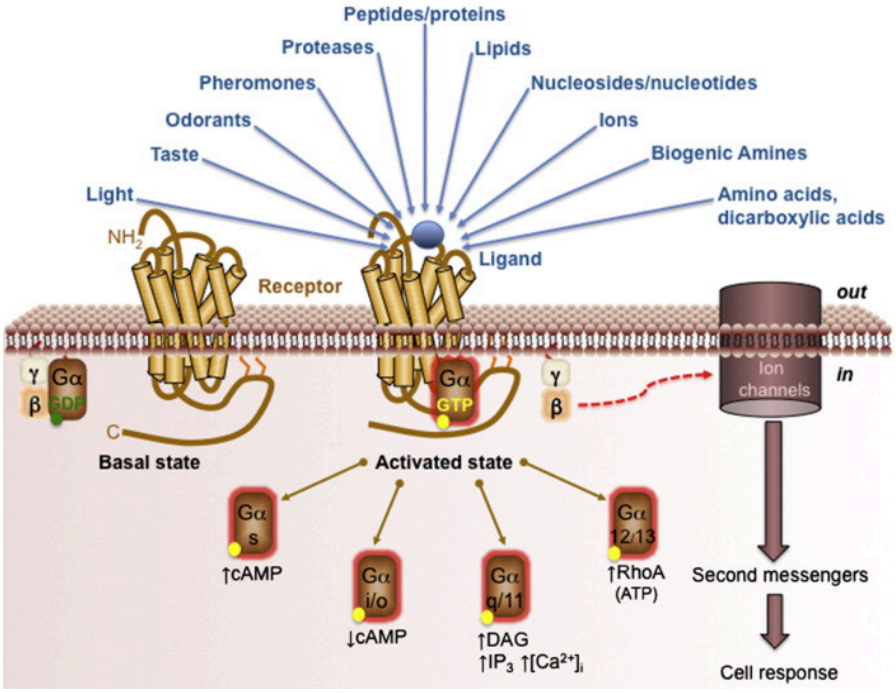


Figure 1.0: G-protein-coupled receptor signal transduction *(Heng, Aubel, and Fussenegger 2013)*
GPCRs are known to be activated by a diverse plethora of ligands and stimuli. Four major classes of G-proteINS (Gαs, Gαi/o, Gαq/11, and Gα12/13) and their associated secondary messengers constitute the canonical model of GPCR signaling. Activation of GPCRs induces conformational changes to the transmembrane and intracellular domains of the receptor, enabling them to act as guanine nucleotide exchange factors (GEFs) that catalyze exchange of GDP for GTP on the α subunit of the heterotrimeric G-protein complex. This in turn results in dissociation of the Gα and Gβγ subunits from each other, as well as the receptor. Depending on the specific subtype, the activated Gα protein subunit can in turn activate RhoGEF (Gα12/13), or effect changes in intracellular levels of cAMP (Gαs & Gαi/o), diacylglycerol (DAG), inositol-1,4,5-triphosphate (IP3) and Ca2 + (Gαq/11). The dissociated Gβγ subunit can also bind and activate other downstream effectors such as ion channels.

GPCRs thus function as molecular gatekeepers, bridging external environmental information to internal cellular responses, ensuring a coordinated and contextually appropriate cellular reaction. Their versatility and ability to interact with an extensive array of ligands make GPCRs fundamental players in physiological regulation, underpinning processes such as vision, taste, smell, immune responses, neurotransmission, cardiovascular regulation, and endocrine homeostasis (Katritch, Cherezov, and Stevens 2013; Kobilka 2013; Vassart and Costagliola 2011). The remarkable structural plasticity of these receptors enables them to adopt multiple conformational states, each capable of activating distinct

signaling pathways with varying efficacies, a phenomenon known as functional selectivity or biased signaling (Abraham et al. 2015; Wootten et al. 2018). Consequently, the precise modulation of GPCR signaling is critical for maintaining homeostatic balance across biological systems, enabling complex multicellular organisms to respond adaptively to both internal physiological demands and external environmental pressures (Hanyaloglu and von Zastrow 2008; Romero, von Zastrow, and Friedman 2011; Nieto Gutierrez and McDonald 2018). This dual capacity to sense and respond makes GPCRs central to cellular information processing networks and explains their evolutionary conservation across diverse species from unicellular organisms to mammals (Krishnan et al. 2012; de Mendoza, Sebé-Pedrós, and Ruiz-Trillo 2014).

However, the importance of GPCRs extends beyond normal physiology; dysregulation or aberrant functioning of GPCR signaling pathways is implicated in numerous pathological conditions, including cancer, neurodegenerative diseases, metabolic disorders, and cardiovascular diseases (Lappano and Maggiolini 2011; Heng, Aubel, and Fussenegger 2013; Gurevich and Gurevich 2018). Accumulating evidence demonstrates that alterations in GPCR expression, trafficking, signaling bias, or post-translational modifications can contribute to disease initiation, progression, and therapeutic resistance (Insel et al. 2018; Hauser et al. 2017; O'Hayre, Degese, and Gutkind 2014). Particularly in oncology, certain GPCRs have been identified as proto-oncogenes, while others function as tumor suppressors, highlighting the context-dependent nature of GPCR signaling in disease pathogenesis (Bar-Shavit et al. 2016; Wu et al. 2019). This widespread involvement in disease states underscores their clinical significance, positioning GPCRs as highly valuable therapeutic targets. Their extensive involvement in diverse physiological processes has established them as preeminent drug targets, with more than one-third of FDA-approved medications exerting their effects through GPCR modulation, highlighting their pivotal role in drug discovery and development (Sriram and Insel 2018; Hauser et al. 2017; Santos et al. 2017). Despite this therapeutic significance, substantial challenges persist in fully exploiting their potential.

A significant challenge is the existence of "orphan" GPCRs, which are receptors lacking identified endogenous ligands (Alexander et al. 2019, 2023). These orphan GPCRs represent a significant portion of the GPCR superfamily (Davenport et al. 2013), and their deorphanization, or the identification of their ligands, is crucial for understanding their physiological roles and therapeutic potential (Kroeze et al. 2015). Even for deorphanized GPCRs, the number of known ligands per receptor is often limited (Hauser et al. 2017), hindering our understanding of their full signaling capabilities and limiting new drug development (Hauser et al. 2017; Roth, Irwin, and Shoichet 2017).

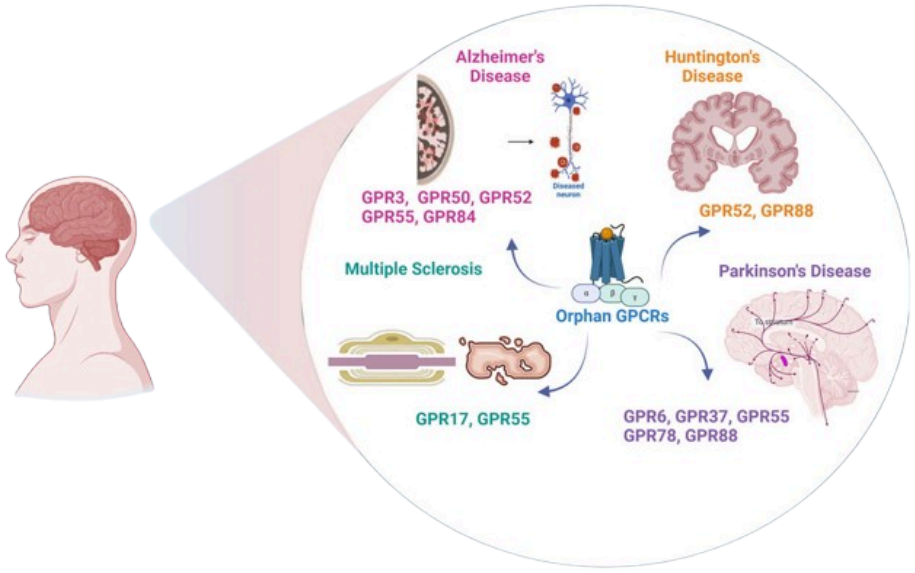


Figure 1.1: Orphan GPCRs in Neurodegenerative Disorders *(Öz-Arslan, Yavuz, and Kan 2024)*
The figure illustrates the expression and localization of orphan GPCRs in Parkinson's Disease, Alzheimer's Disease, Multiple Sclerosis, and Huntington Disease.

This knowledge gap is particularly pronounced for certain receptor subfamilies where the endogenous ligand-receptor pairing mechanisms remain poorly characterized (Laschet, Dupuis, and Hanson 2018; Ahmad, Wojciech, and Jockers 2015). Moreover, the limited understanding of the molecular expression of certain GPCR subfamilies, particularly ORs, presents an additional layer of complexity. ORs are known to be ectopically expressed in non-olfactory tissues, suggesting previously unrecognized physiological roles beyond their traditional function in olfaction (Flegel et al. 2013; Lee, Depoortere, and Hatt 2019; Ferrer et al. 2016a). Emerging evidence indicates these GPCRs may significantly influence fundamental cellular processes, including proliferation, migration, and immune responses (Ranzani et al. 2017; Kalra et al. 2020; Masjedi, Zwiebel, and Giorgio 2019). The tissue-specific expression patterns and functional diversification of these receptors suggest potential opportunities for targeted therapeutic interventions, particularly in oncology and metabolic disorders where aberrant GPCR signaling has been implicated in disease pathogenesis (Jovancevic et al. 2017; Maßberg et al. 2016). Deciphering these non-canonical roles of ORs and other ectopically expressed GPCRs represents a frontier in receptor biology with significant implications for expanding the druggable GPCR space.

Despite their clinical importance and extensive characterization, substantial gaps remain in our understanding of GPCR biology, particularly regarding receptor dynamics, ligand specificity,

receptor-ligand interactions at atomic resolution, mechanisms of signal bias, and the identification of novel modulatory sites beyond traditional orthosteric binding regions (Latorraca, Venkatakrishnan, and Dror 2017; Congreve et al. 2020; Inoue et al. 2019). The inherent conformational flexibility of GPCRs presents significant challenges for structural biology approaches, necessitating innovative methodologies to capture their dynamic nature in physiologically relevant contexts (Weis and Kobilka 2018; Casiraghi et al. 2018). Additionally, emerging evidence suggests that GPCRs participate in complex, previously underexplored regulatory networks, including oligomerization, biased signaling, and interactions with intracellular proteins and membrane lipids, further expanding the functional versatility and therapeutic potential of this receptor family (Gahbauer and Böckmann 2016; Flock et al. 2017; Gaitonde and González-Maeso 2017). The spatiotemporal regulation of GPCR signaling through compartmentalized subcellular microdomains and the role of receptor trafficking in signaling outcomes represent additional layers of complexity that warrant further investigation (Irannejad et al. 2013; Calebiro et al. 2010). Furthermore, the increasing recognition of GPCR functional selectivity, whereby different ligands can preferentially activate distinct downstream signaling pathways through the same receptor, offers unprecedented opportunities for developing precisely targeted therapeutics with enhanced efficacy and reduced side effects (Smith, Lefkowitz, and Rajagopal 2018; Michel and Charlton 2018; Wootten et al. 2018).

Overcoming these challenges necessitates the continued development of advanced analytical techniques, enhanced bioinformatics tools, optimized computational algorithms, and standardized methodological approaches within GPCR research (Heifetz et al. 2018; Kooistra et al. 2021; Lionta et al. 2014). Equally important is the simplification and generalization of these techniques to ensure accessibility to a broader research community, thereby accelerating scientific discovery (Munk et al. 2019). The development of user-friendly computational platforms, standardized data sharing protocols, and comprehensive databases facilitates collaborative research efforts and maximizes the utility of existing experimental data (Páll et al. 2020)(Páll et al. 2020; Armstrong et al. 2020; Kooistra et al. 2021). Leveraging automation and advanced computational frameworks could further streamline research processes, fostering rapid advancements in the field and maximizing the translational potential of GPCR-related discoveries (Roth, Irwin, and Shoichet 2017). The integration of artificial intelligence and machine learning approaches with traditional molecular modeling techniques would offer unprecedented opportunities to navigate the vast chemical and conformational spaces relevant to GPCR drug discovery (Vamathevan et al. 2019; Salmaso and Moro 2018). The logical mechanistic insights, coupled with advances in computational modeling and artificial intelligence approaches, are poised to revolutionize GPCR-targeted drug discovery and expand the therapeutic applications of this versatile receptor superfamily (Kooistra et al. 2021; Lyu et al. 2019).

1.1 Central hypothesis of the Thesis

Integrative computational frameworks, when strategically aligned with mechanistic insights into receptor expression and ligand interaction, can substantially deepen our understanding of GPCR biology while simultaneously addressing critical bottlenecks in drug discovery. Specifically, it is proposed that the regulatory dynamics governing odorant receptor expression, rather than being stochastic, follow preferential mechanisms that can be captured and elucidated through computational modelling. Moreover, it is suggested that the integration of machine learning into ligand-based drug design, if implemented in an accessible and interpretable manner, can broaden the use of advanced predictive tools and accelerate the identification of biologically relevant ligands across GPCR families. Finally, it is hypothesized that AI driven methodologies, capable of leveraging structural, generative, and evolutionary information, provide a viable solution to the pervasive problem of data scarcity in GPCR research, allowing for the robust prediction of receptor–ligand interactions and the discovery of novel modulatory mechanisms.

1.2 Objective of the Thesis

This thesis aims to advance the understanding of GPCR biology through the development and application of innovative computational frameworks and integrative methodologies. The study is structured around three key objectives: first, exploring the underlying mechanism of OR, a specific subfamily of GPCRs, expression, focusing particularly on elucidating the regulatory processes governing their selective and precise expression patterns; second, simplifying the integration of machine learning techniques into ligand-based drug design (LBDD) methodologies specifically tailored for GPCRs, thereby making sophisticated computational tools accessible to researchers regardless of computational expertise; and third, addressing the critical challenge of data scarcity in GPCR research, particularly for drug discovery applications, by developing robust computational strategies that efficiently predict ligand interactions and identify novel targets for therapeutic intervention. Through these objectives, this research contributes significantly to the broader understanding of GPCR signaling mechanisms and receptor-ligand interactions and ultimately enhances the potential for targeted and effective therapeutic innovations.

This thesis hypothesizes that integrative computational methodologies, combining mechanistic modeling, interpretable machine learning, and data efficient AI, can reveal novel regulatory principles in GPCR biology, democratize ligand discovery for GPCR targets, and overcome critical limitations associated with data scarcity. By unifying these approaches, we'll gain a deeper understanding of GPCRs and find more effective ways to discover new drugs that target them.

1.3 Organization of the Chapters

This thesis is organized into five chapters, each focusing on a different aspect, including expression, ligand interaction, and generalization of AI in GPCR biology. By developing computational frameworks and advanced AI methodologies, this work aims to enhance our understanding of GPCR biology while simultaneously overcoming key bottlenecks in drug discovery. The organization of the chapters is as follows:

Chapter 1 discusses the flow of genetic information from DNA to RNA to protein, with proteins executing essential cellular functions. Among these, GPCRs stand out as key mediators of signal transduction, converting diverse extracellular stimuli into precise intracellular responses via G protein pathways. The chapter also discusses the physiological significance and disease relevance of GPCRs, underscoring their prominence as therapeutic targets. Finally, it addresses the current challenges in fully leveraging GPCRs for drug discovery, setting the stage for the objectives explored in this thesis.

Chapter 2 focuses on odorant receptors, a GPCR subset, and revisits the "one neuron, one receptor" rule. It introduces RCT, a computational method that facilitates the retroactive tracing of cellular trajectories by utilizing OR identity and cellular RNA kinetics. Using the RCT framework, it reveals that OR selection may follow a preferential, highest-expression-driven mechanism (winner-takes-all) rather than a purely random one.

Chapter 3 introduces Machine-Olf-Action, a user-friendly platform that simplifies GPCR ligand discovery using machine learning. It enables researchers, even with minimal computational expertise, to identify biologically relevant molecules efficiently. With built-in descriptors, automated preprocessing, optimized models, and interpretability tools, it ensures reliable and transparent predictions.

Chapter 4 addresses data scarcity in GPCR research using Gcoupler, an AI-driven toolkit that identifies receptor interaction sites and predicts high-affinity ligands. By combining cavity detection, generative AI, and deep learning, it streamlines GPCR-targeted drug discovery. Computational and experimental analyses revealed that intracellular metabolites can directly modulate GPCR signaling via conserved Gα-binding sites. The discovery of sterols as endogenous GPCR modulators opens new avenues in understanding stress response and cellular signaling, showcasing Gcoupler’s potential to generate insights even in data-limited scenarios.

Chapter 5 concludes the thesis by summarizing the findings, discussing their implications, and proposing future research directions. It emphasizes the integrative approach taken in this work, discusses the

associated limitations, and combines computational architecture and experimental validation to unravel the diverse aspects of GPCRs.

Objectives

Chapter 2 - Exploring The Underlying Mechanism of Odorant Receptor (Subfamily of GPCRs) Expression

2.1 Introduction

G protein-coupled receptors are essential communication centers that mediate signals from hormones, neurotransmitters, ions, photons, and various environmental stimuli, thus enabling dynamic interaction between a cell's internal and external environments. Among the many physiological roles of GPCRs, the sense of smell is mediated by a specialized class of GPCRs known as odorant receptors, expressed within the sensory region of the olfactory organ (Spehr and Munger 2009; Bettina Malnic, Godfrey, and Buck 2004; Buck and Axel 1991). The sensory function known as olfaction, one of the most evolutionarily primitive senses, allows organisms to detect, differentiate, and react to a wide variety of odor cues in their environment (Niimura, Matsui, and Touhara 2014; Niimura and Nei 2005).

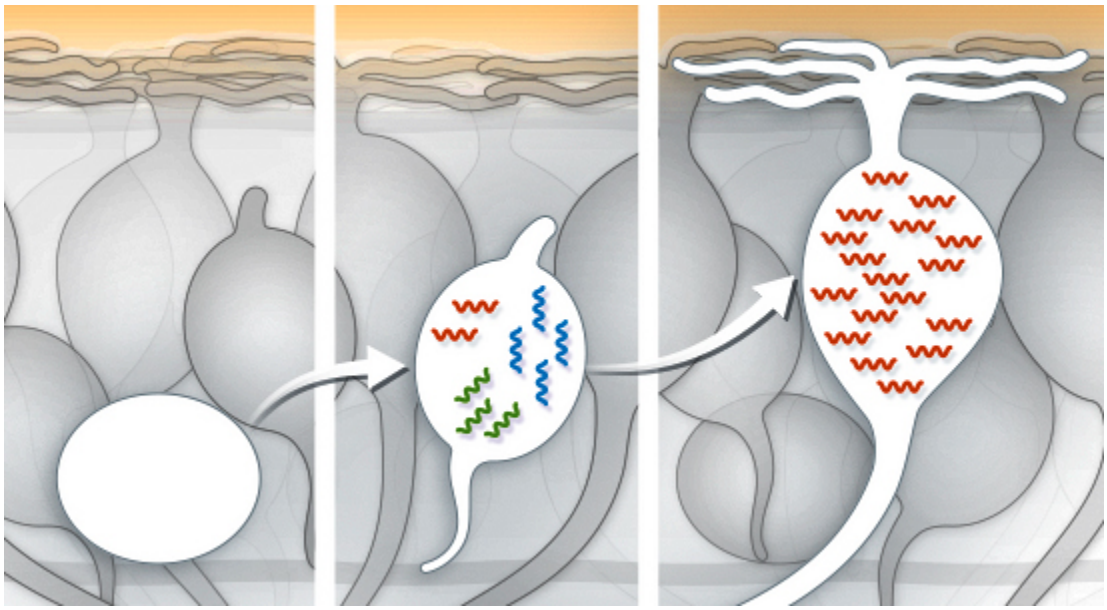


Figure 2.0: one-neuron-one-receptor rule of olfaction *(L. Tan, Li, and Sunney Xie 2015)*
Olfactory sensory neurons co-express multiple olfactory receptors at an immature stage, while mature neurons express only one olfactory receptor.

A fundamental regulatory characteristic of the olfactory system is that each mature sensory neuron expresses a single functional odorant receptor gene from a pool of hundreds, thereby upholding the "one-neuron-one-receptor" principle. For instance, while mice possess over a thousand functional OR genes, each mOSN expresses only one (Buck and Axel 1991). The initial experimental validation of this rule was derived from investigations employing the exponential cDNA amplification method on RNA extracted from individual mouse olfactory sensory neurons (OSNs) (B. Malnic et al. 1999; Chess et al.

1994). The feasibility of these findings was subsequently evaluated and confirmed on multiple other vertebrate models (Mombaerts 2004; Bystrova and Kolesnikov 2021). Furthermore, a comprehensive mechanistic investigation was conducted to elucidate the molecular mechanisms governing OR selection in the mOSNs. Despite the comprehensive investigations that have improved our comprehension of the one-neuron-one-receptor principle, there remains an absence of consensus regarding the model that might elucidate this OR gene selection (Serizawa, Miyamichi, and Sakano 2004, 2005; Chess et al. 1994; B. Malnic et al. 1999).

A previously accepted model, known as silence-all-and-activate-one, posits that OR selection in mature olfactory sensory neurons (mOSNs) is a sequential process predominantly governed at the epigenetic level (Monahan et al. 2017; Bashkirova and Lomvardas 2019; Lomvardas et al. 2006). In summary, this model posits that differentiating OSNs selectively express one OR gene from among the numerous ORs within the same extensive family. The selected OR is translated into a functional protein, triggering the transcriptional suppression of excluded OR genes via a negative feedback mechanism. The feedback mechanism entails epigenetic silencing in the early immature olfactory sensory neuron (iOSN) state, resulting from a transcriptionally repressive environment (H3K9me3) (Lyons et al. 2013; Bashkirova and Lomvardas 2019). As the iOSNs diverge, the de-repression of an OR gene locus transpires through the low-level activation of lysine-specific demethylase LSD1. This selective expression of a single OR triggers a feedback mechanism involving the unfolded protein response that inhibits LSD1 activity and prevents the de-repression of additional OR gene loci (Lyons et al. 2013, 2014). Although mOSNs were meticulously examined for their monogenic OR expression, recent independent studies employing single-cell RNA sequencing (scRNA-seq) in olfaction revealed the expression of multiple ORs in iOSNs and, in one study, also in Neuronal Precursor Cells (NPCs) (Hanchate et al. 2015; L. Tan, Li, and Sunney Xie 2015; Saraiva, Ibarra-Soria, et al. 2015; Fletcher et al. 2017). These results elucidated the molecular intricacy of the fundamental mechanism driving the one-receptor-one-neuron rule and severely challenged the feasibility of the silence-all-and-activate-one model.

Although the expression of a singular OR in mOSNs is widely recognized, recent findings from single-cell RNA profiling of olfactory cell types indicate the significant influence of intermediate cell types on OR gene selection during olfactory sensory neuron differentiation. These findings resulted in the development of two distinct models: winner-takes-all and stochastic selection, as originally suggested by Hanchate and associates (Hanchate et al. 2015). Furthermore, it indicates that the fundamental mechanism of OR selection is more complex than previously expected (Hanchate et al. 2015; L. Tan, Li, and Sunney Xie 2015; Saraiva, Ibarra-Soria, et al. 2015). The winner-take-all model illustrates the unfair selection advantage for the transcriptionally advanced (highest expressed) OR in the iOSN state. This preference is

referred to as transcriptional resource bias. The stochastic selection model posits that OR selection occurs randomly, resulting in the selection of any one OR from ~1000 potential options (expressed and non-expressed) in mice. Although both models appear plausible, substantial evidence supporting these possibilities is largely absent, likely due to technical constraints, including the lack of computational methods capable of retracing differentiation trajectories and complications related to cellular viability and vitality associated with in vivo labeling of multiple olfactory receptors and their monitoring during olfactory sensory neuron differentiation in the olfactory mucosa (Bystrova and Kolesnikov 2021). In this study, we established an innovative computational framework, RCT, which facilitates the retroactive tracing of cellular trajectories by utilizing OR identity and cellular RNA kinetics (Bergen et al. 2020; La Manno et al. 2018). We assumed that the RCT framework, by using RNA velocity and single-cell transcriptomics, can accurately and quantitatively reconstruct the developmental path of cells. This computational reconstruction is expected to retrace the lineage of OR expression, thereby validating the observed preference for the receptors and revealing any transitional cell states that are crucial for the receptor selection process. Using this method, we were able to identify evidence that supported the selection bias of the OR that was expressed the highest during the iOSNs stage. A further finding from the analysis of neuronal trajectory was the identification of an intermediate neuronal state. In this state, the neurons have mixed identities of both iOSNs and mOSNs, and the relative abundance of these is controlled by the nonsense-mediated mRNA decay factor *Upf3b*.

2.2 Materials and Methods

2.2.1 Data Pre-processing

The raw expression matrices were acquired from NCBI (https://www.ncbi.nlm.nih.gov/) under accession numbers GSE146043 and GSE157119. Raw FASTQ files were obtained from EMBL-EBI (https://www.ebi.ac.uk/) using project numbers PRJNA609124 and PRJNA660170 for these datasets. The CellRanger v4.0.0 (Zheng et al. 2017) software was utilized to produce BAM files from FASTQ files using default parameters (Zheng et al. 2017). Notably, GRCm38 (mm10) served as the reference genome. The BAM files were sorted by coordinates utilizing the Samtools sort function. Following the sorting, the sorted BAM files are converted into loom files using the Velocyto v0.17.17 (La Manno et al. 2018) run10x function with standard parameters. In the case of the GSE146043 dataset, individual loom files of wild-type and *Upf3b(-/-)* datasets were generated and combined respectively using the loompy.merge() function. Following this, scVelo (Bergen et al. 2020) was used for downstream analysis. The wild-type merged loom file contained 20303 cells and 55487 genes/features ,and the *Upf3b(-/-)* merged loom file contained 20868 cells and 55487 genes/features. In the GSE157119 dataset, the loom file with the

olfactory-stimulated condition consisted of 7920 cells and 32285 genes/features, and the olfactory-deprived condition loom file consisted of 4480 cells and 32285 genes/features. For the GSE146043 dataset, we used the authors' cell annotations and segregated the GBCs, HBCs, iOSNs, and mOSNs. This sub-filtering resulted in 5184 cells and 55487 genes in the wild-type sub-dataset while 6472 cells and 55487 genes in the *Upf3b(-/-)* sub-dataset. For the GSE157119 dataset, the clusters expressing gene markers of the aforementioned cell types were sub-clustered. We further filtered the cells based on the dropouts (minimum 200 genes expressed). Moreover, we also filtered the genes based on their expression (less than 3 cells) or source (mitochondrial genome-encoded genes).

2.2.2 Cellular differentiation trajectory inference using scVelo

In the scVelo workflow, we filtered the top 2000 influential genes (default parameter) and normalized the data using the default settings. Following normalization, we computed the moments and neighbors for each cell. Moments are utilized for velocity estimation, whereas neighbors construct neighborhood graphs. The unsupervised clustering utilized the Louvain algorithm. Cellular identities were confirmed using the known bona fide markers of HBCs, GBCs, NPCs, iOSNs, and mOSNs. In summary, significant clusters were identified comprising HBCs (418), GBCs (103), NPCs (35), iOSNs (520), tsOSNs (309), and mOSNs (2277) in the *Upf3b* wild-type; HBCs (547), GBCs (117), NPCs (76), iOSNs (607), tsOSNs (175), and mOSNs (682) in the open nostril dataset; and HBCs (335), GBCs (32), NPCs (48), iOSNs (199), tsOSNs (39), and mOSNs (438) in the closed nostril dataset. In the *Upf3b* knockout data, only five clusters were identified: HBCs (319), GBCs (98), NPCs (127), iOSNs (288), and mOSNs (2516). To ascertain the presence of tsOSNs in *Upf3b* knockout mice, we amalgamated OSN expression profiles from wild-type and *Upf3b* knockout (GSE146043) under the premise that tsOSNs from both cohorts would exhibit analogous transcriptomic signatures, subsequently conducting clustering on the integrated data. In accordance with this methodology, the revised clusters in the *Upf3b* knockout condition comprised HBCs (319), GBCs (98), NPCs (127), iOSNs (288), tsOSNs (96), and mOSNs (2420). scVelo employs the velocity length and confidence of each cell to determine the rate of cell differentiation. We employed CellRank (Lange et al. 2022) to compute the velocity pseudotime for each cell utilizing the default parameters. Pseudotime quantifies the number of transitions a cell undergoes while traversing the graph from the root cell. The trajectory inference was conducted using Partition-based Graph Abstraction (PAGA) (Wolf et al. 2019), which establishes the directionality of the clusters. Finally, for the downstream analysis, we extracted the cellular Uniform Manifold Approximation and Projection (UMAP) coordinates of all cells while maintaining their cluster identities. Differential gene expression analysis to identify authentic markers for transition state olfactory sensory neurons was conducted utilizing the scVelo workflow.

2.2.3 Doublet Analysis

We employed Scrublet (Single-Cell Remover of Doublets) v0.2.3 (Wolock, Lopez, and Klein 2019), a Python module, to detect doublets in the single-cell expression datasets. For each individual tsOSN, a doublet score and a boolean value (True or False) indicating doublet status were computed using the scrub.scrub_doublets() function.

2.2.4 Reverse Cell Tracking (RCT)

The raw expression matrix for various genes across all cells was subjected to 'LogNormalization' using Seurat software (Satija et al. 2015). The LogNormalized matrix was subsequently employed to compute the z-score. Cluster-specific coordinates of cells were obtained from the scVelo pipeline for the RCT algorithm, and the distance between each pair of cells was computed. In brief, every mature OSN is selected recursively, and the expressed OR, say ORx (mostly 1 OR per mOSN), is noted. Next, the RCT1 algorithm scans every tsOSN that expresses ORx, allowing the pairing of one mOSN to multiple tsOSNs based on the common OR expressed (ORx). The last and most important step involves the ranking of selected tsOSNs based on two criteria: (1) the distance between the selected mOSN and tsOSNs must be within the selected threshold (distance percentiles), and (2) the ORx must possess the highest expression rank in that tsOSN. The positive hits refer to unique tsOSN-mOSN pairs, where the OR under investigation is the OR expressed by the mOSN and possesses the maximum expression among other ORs in the tsOSN. Different percentile cut-offs were applied for the distance matrix with increasing distances between selected mOSN and tsOSNs. A similar analysis was performed with RCT2 and RCT3 algorithms allowing the pairing of tsOSNs to iOSNs and further iOSNs to NPCs, respectively. We hypothesize that if the winner-takes-all model is true, then the observed enrichment pattern along different cutoffs must be exclusive to that particular OR (ORx); therefore, as a control, we randomly selected an alternative OR (not ORx) and computed the normalized hit rate across given thresholds. Analysis of covariance (ANCOVA) was calculated using PAST (version 4.10) and computed the statistical significance for line plots. The Chi-Square test was performed for the test of significance between the enrichment pattern and random control at each percentile distance cutoff. The cut‑off used for statistical significance is < 0.05. *, **, *** and **** in the figure refer to P‑values < 0.05, < 0.01, < 0.001 and < 0.0001, respectively. The statistically non-significant comparisons are annotated as ns. Notably, since earlier studies have shown that the relative abundance of cell types could influence their spatial positions in the low dimensional space, e.g., 2D-UMAP; therefore, for the accurate estimation of the trajectories, we did not perform any cell elimination cutoffs/criteria except for the ones that failed in the Quality Check (QC) step (minimal number of genes>= 200). Importantly, only a smaller fraction of mOSNs were positive for multiple ORs (17.4%); among them, 12.5% expressed only 2 ORs (wild type), and within this, only one OR was highly

expressed, while other co-expressed OR was marginally expressed. We performed RCT on all mOSNs, where for a fraction of them that carry multiple ORs, highly expressed OR was used for the RCT calculations.

2.2.5 Data and Code Availability

This research is entirely computational and employs publicly accessible single-cell RNA sequencing datasets. Raw FASTQ sequencing files were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with the accession IDs: GSE146043 and GSE157119. An end-to-end workflow for the entire analysis is provided on GitHub: https://github.com/the-ahuja-lab/OR-selection-model and Zenodo: https://zenodo.org/badge/latestdoi/340450648.

2.3 Results

2.3.1 Addressing Cell-Type Annotation Ambiguities via Integrated Marker Curation

The mammalian neuroepithelium comprises various functionally distinct cell types that collectively contribute to the differentiation of olfactory sensory neurons (Moran, Rowley, and Jafek 1982; K. Gupta et al. 2020). Recent advancements in computational trajectory inference methods have facilitated nearly precise tracking of differentiation time courses by utilizing RNA expression kinetics data (Trapnell et al. 2014; Bergen et al. 2020; La Manno et al. 2018). We employed scVelo (Bergen et al. 2020), an RNA Velocity-based (La Manno et al. 2018) cell fate estimator, to reconstruct the lineage mapping of various cell types within the mouse olfactory epithelium, utilizing the RNA splicing kinetics of individual cells. We deduced the neuronal differentiation trajectory from two independent single-cell datasets (GSE146043 and GSE157119) that encompass expression profiles of four biologically distinct conditions. The initial dataset (GSE146043) comprises single-cell expression profiles of horizontal basal cells (HBCs), globose basal cells (GBCs), iOSNs, and mOSNs from both wild-type and *Upf3b* knockout mice (K. Tan et al. 2020), totaling 3662 and 3348 cells, respectively **(Figure 2.1A)**. The second dataset (GSE157119) comprises expression profiles of the specified cell types from olfactory-deprived (closed nostrils) and olfactory-stimulated (open nostrils) mice (van der Linden et al. 2020), totaling 1091 and 2204 cells, respectively **(Figure 2.1A)**. Significantly, in all these datasets, mOSNs represent the predominant cell type based on their absolute quantity **(Figure 2.1A)**. A confounding factor in scRNA-seq is cellular annotations, primarily accomplished through established cell-type markers. Previous studies examined scRNA profiles of olfactory cell types utilizing a diverse, though slightly overlapping, array of markers for each cell type. For instance, Neurog1 and Neurod1 were utilized to label NPCs by Hanchate et al.

(Hanchate et al. 2015), whereas, Neurod1, Ascl1, and Lgr5 were employed to label GBCs by Tan et al. (K. Tan et al. 2020). Additionally, Saraiva et al. (Saraiva, Ibarra-Soria, et al. 2015) employed Neurog1 and Krt5 to identify GBCs.

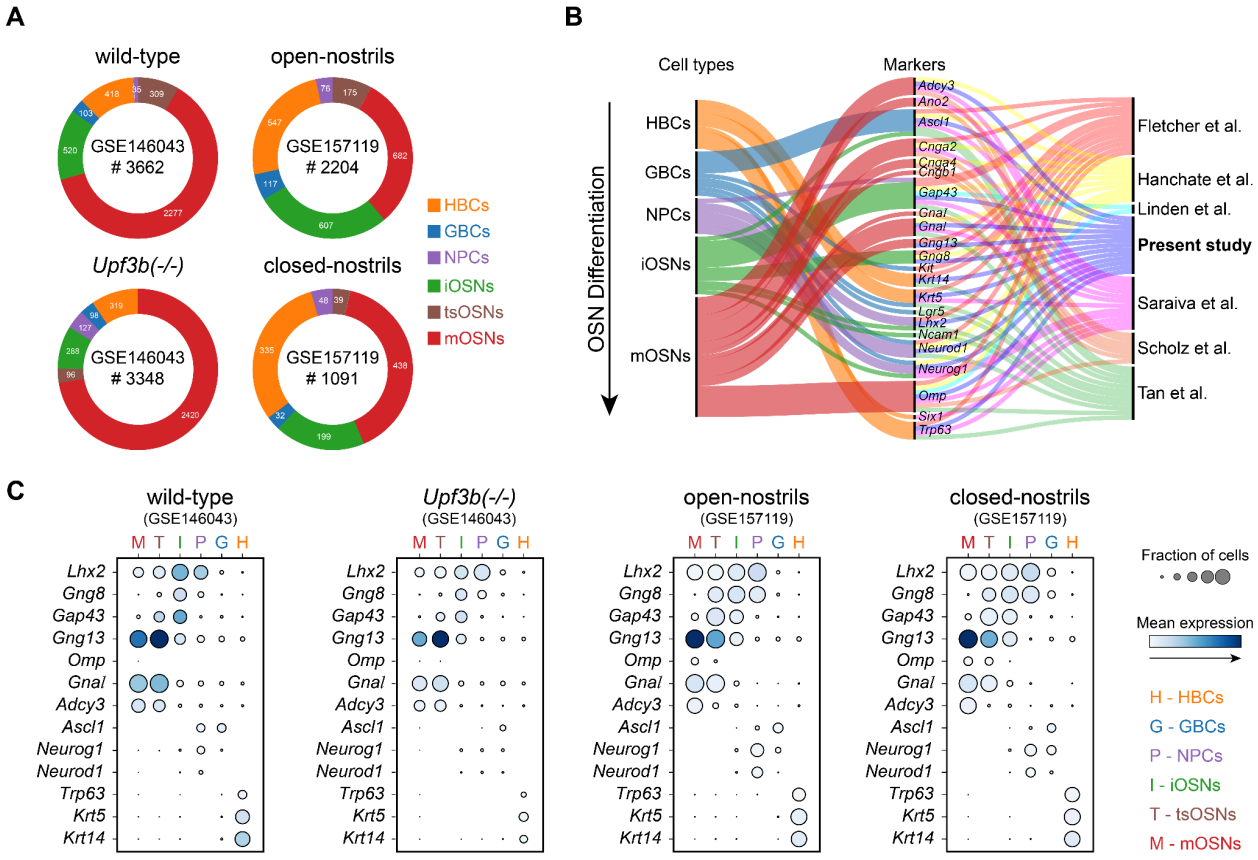


Figure 2.1 Olfactory Cell Annotation: Cell Counts and Marker Profiles

(A) Donut plots depicting the number of the indicated cell types in each dataset used in this study. Accession IDs and the total number of cells are mentioned inside each donut. (B) Alluvial plot indicating expression markers of the indicated cell types used in the present and other studies. (C) Dot plots depicting the relative expression of the indicated markers in the mature olfactory sensory neurons, transition-state olfactory sensory neurons, immature olfactory sensory neurons, neuronal precursor cells, globose basal cell, and horizontal basal cell in the indicated datasets.

To prevent misinterpretation of results stemming from cell annotations, we aggregated known cell-type-specific markers from various scRNA-seq datasets, cross-validated them with the datasets utilized in this study, and subsequently selected a subset for downstream analysis (Saraiva, Ibarra-Soria, et al. 2015; Hanchate et al. 2015; van der Linden et al. 2020; Scholz et al. 2016; Fletcher et al. 2017; K. Tan et al. 2020) **(Figure 2.1B-C)**.

2.3.2 Mixed-Identity Cell Cluster Suggests Putative Transitional Subpopulation

We conducted scVelo, PAGA, and CellRank analyses (Wolf et al. 2019) on these distinct datasets **(Figure 2.2A)**. RNA dynamics-based trajectory analysis demonstrated consistent lineage trajectories originating from stem cells and progressing towards mOSNs **(Figure 2.2A)**.

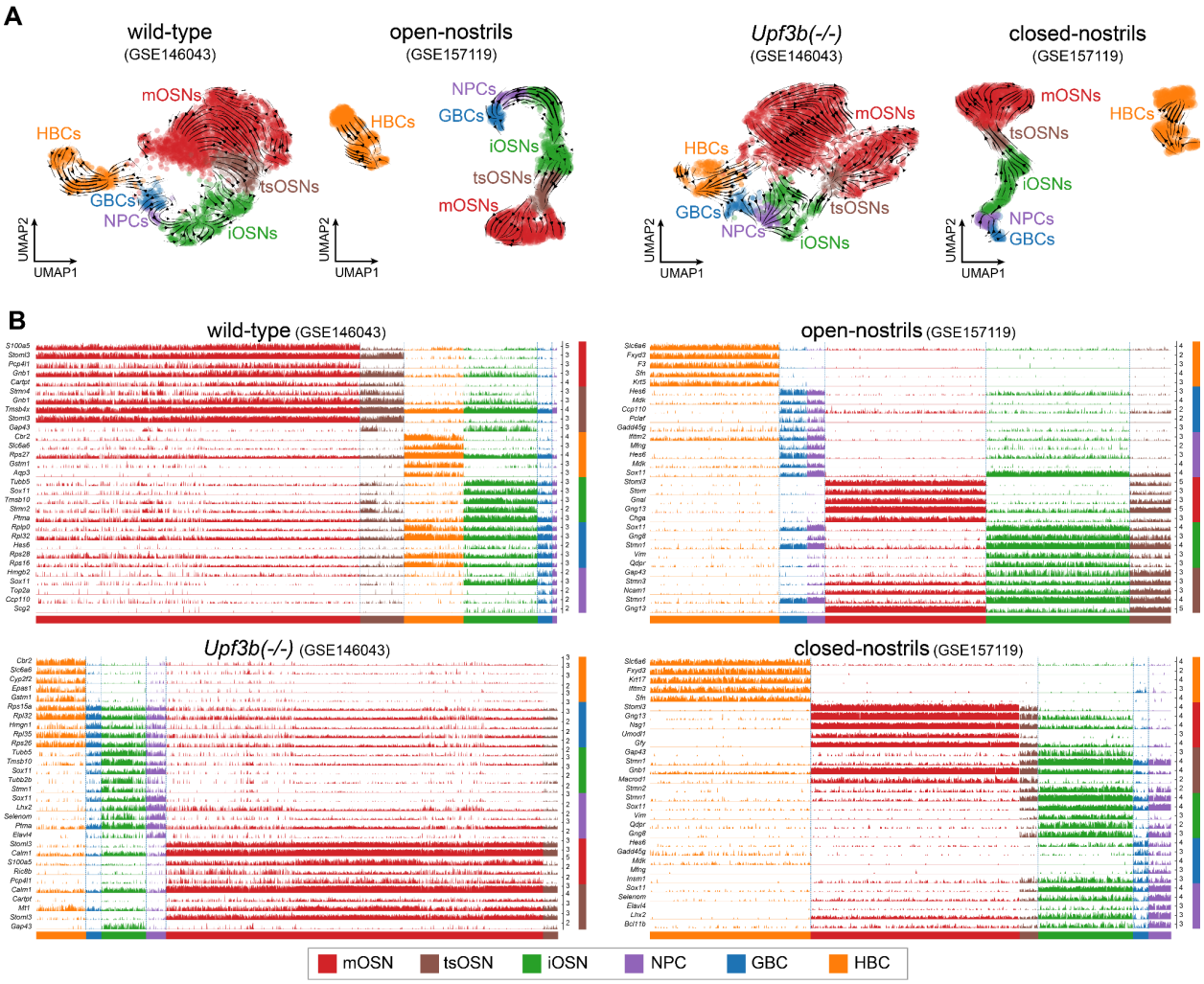


Figure 2.2 Cell Differentiation and Marker Gene Expression Across Conditionss

(A) Stream graphs depicting the differentiation trajectories of the olfactory cell types in the four biological conditions (wild-type, Upf3b knockout (Upf3b(-/-)), open-nostril, closed-nostril). (B) Track plots depicting the cluster-wise differentially expressed genes in the four indicated datasets.

Significantly, unsupervised clustering of wild-type cells from the GSE146043 and GSE157119 datasets, utilizing the default parameters of the scVelo workflow, identified a minor cluster situated between the iOSNs and mOSNs **(Figure 2.2A)**. A detailed examination of this cluster indicated that it contains mixed signatures of both iOSNs and mOSNs, likely representing a transitional cellular state **(Figure 2.1C; Figure 2.2A-B)**. To obtain enhanced insights, we conducted differential gene expression (DGE) analysis utilizing the scVelo workflow to identify potential markers for this intermediate cellular state. The DGE

analysis failed to identify any specific marker, which we hypothesize may be attributed to their overlapping transcriptomic identities **(Figure 2.1C; Figure 2.2B)**. Transition state cells frequently exhibit mixed markers, as exemplified by the extensively researched Epithelial to Mesenchymal Transition (Hybrid EMT), wherein the proposed markers for these transitional cell populations are characterized by both Epithelial and Mesenchymal markers (Sinha et al. 2020). We designated this transient cell state as tsOSNs (transition state olfactory sensory neurons) due to their mixed identities and intermediate position during differentiation events. Despite earlier single-cell investigations indicating the presence of late immature neurons (Hanchate et al. 2015), we classify them as tsOSNs owing to their heterogeneous molecular characteristics.

2.3.3 *Upf3b*-Dependent Disruption of Olfactory Sensory Neuron Lineage Trajectories

A recent study demonstrated that *Upf3b* affects the cellular composition of the olfactory epithelium (K. Tan et al. 2020), with notable alterations in the frequencies of HBCs and mOSNs, indicating defects in differentiation. *Upf3b* is a pivotal element of nonsense-mediated RNA decay (NMD), recognized for its regulation of numerous cellular processes and its association with various diseases (Lykke-Andersen and Jensen 2015); however, its function in olfaction has only recently been assessed (K. Tan et al. 2020). Given that *Upf3b* knockout mice exhibit impaired mOSNs, we hypothesize that this differentiation defect is facilitated through tsOSNs. To investigate this, we initially examined the scRNA profiles of olfactory cell types from *Upf3b* knockout mice utilizing the standard clustering parameters of the scVelo workflow, and did not identify any tsOSNs-specific cluster **(Figure 2.3A)**. We subsequently inquired whether the lack of a tsOSN cluster might result from the constrained resolution of the clustering algorithm. To counter this, we combined the OSNs expression profiles from wild-type and *Upf3b* knockout mice (GSE146043) and conducted clustering on the integrated dataset **(Figure 2.3B-D)**.

In this consolidated dataset, we identified a singular cluster between iOSNs and mOSNs, which we designated as tsOSNs due to their common markers. Subsequently, we differentiated mixed tsOSNs into wild-type and *Upf3b* knockout categories, noting a significant reduction in the tsOSNs within the *Upf3b* knockout mice (4.2% of all OSNs) compared to the wild-type (9.4% of all OSNs). We hypothesize that *Upf3b* modulates tsOSNs, thereby inducing differentiation anomalies in the *Upf3b* knockout models **(Figure 2.3E)**. These findings build upon the work of Tan and colleagues' who propose that *Upf3b* may play a significant role in the differentiation of OSNs (K. Tan et al. 2020). To eliminate the potential for mixed signatures arising from doublets in this minor neuronal cluster (tsOSNs), we conducted a doublet analysis utilizing the widely adopted Scrublet software (Wolock, Lopez, and Klein 2019) and found that the majority of the tsOSNs were in a singlet state **(Figure 2.3F)**.

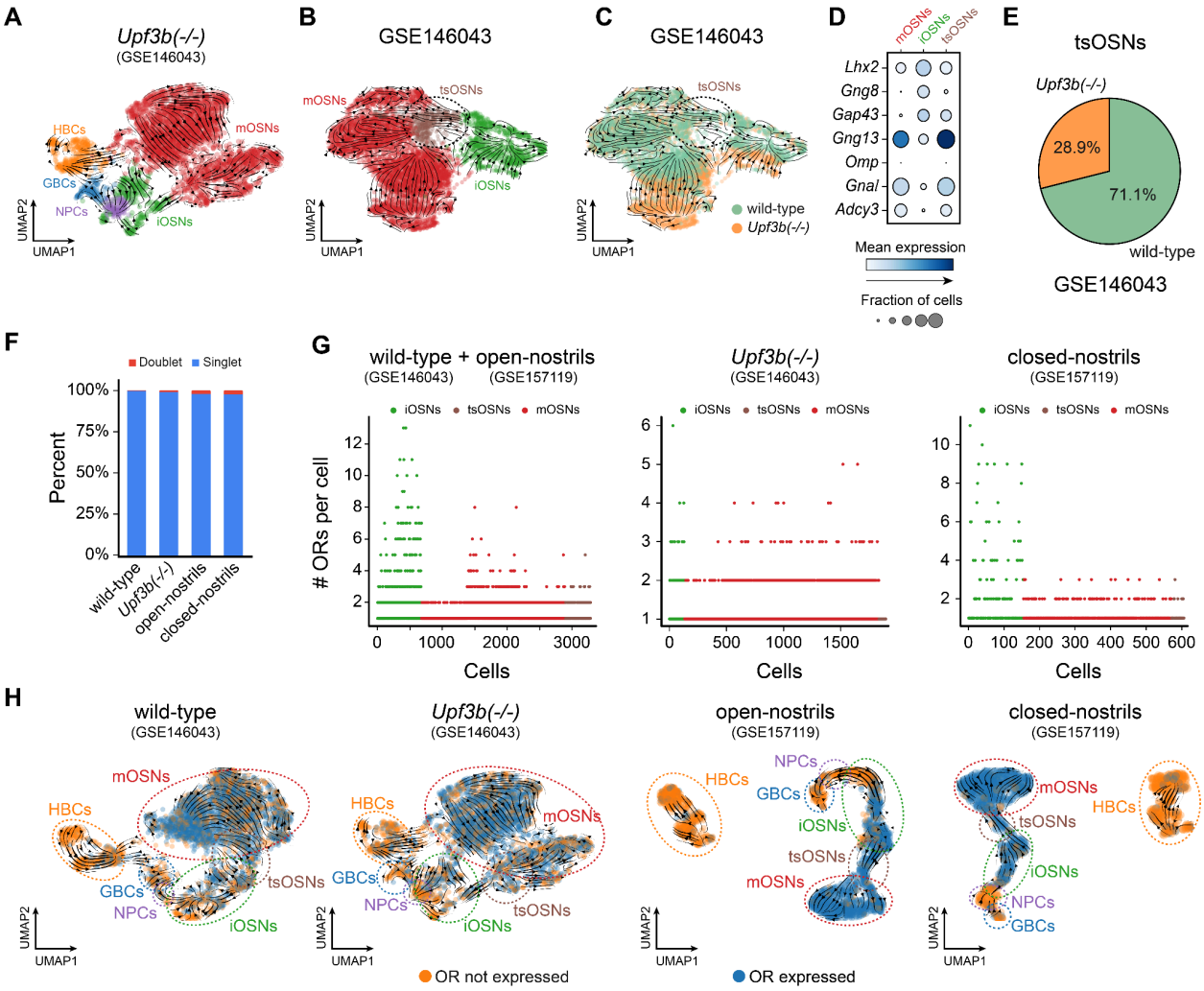


Figure 2.3 Impaired tsOSN Resolution and OR Dynamics in Upf3b-Deficient Conditions

(A) Stream graph depicting the differentiation trajectories of the olfactory cell types in the Upf3b knockout (Upf3b(-/-)) conditions, obtained using the default parameters of the scVelo workflow. Of note, default parameters could not generate a distinct cluster for tsOSNs in the case of the Upf3b(-/-) dataset. (B) Stream graphs depicting the direction of olfactory sensory neuron differentiation on the merged dataset combining olfactory sensory neurons from wild-type and Upf3b(-/-) cells (GSE146043). (C) UMAP reveals the identities of single cells from wild-type (green) and Upf3b(-/-) cells (orange) (GSE146043). (D) Dot plot depicting the relative expression of the indicated markers of the olfactory sensory neurons (mature and immature) on the merged dataset. (E) Pie chart depicting the fractions of transition state olfactory sensory neurons in the merged dataset. (F) Bar plot depicting the percentage of singlets/doublets identified in the transition state olfactory sensory neurons in the indicated conditions. (G) Scatterplots depicting the number of ORs expressed in the individual iOSNs, tsOSN, and mOSNs olfactory sensory neurons of the wild-type, Upf3b knockout (Upf3b(-/-)) and closed-nostrils conditions. (H) UMAPs indicating single cells expressing (blue) or not expressing (orange) at least one or more odorant receptor (OR).

Notably, the Scrublet software produced a binary classification of tsOSNs into singlet or doublet across the four datasets utilized in this study. After confirming the cluster annotations and cellular identities, we conducted PAGA analysis to reconstruct the lineage relationships among clusters. PAGA results demonstrated cell fate transitions from GBCs to mOSNs across all datasets; however, in *Upf3b* knockout cells, the lineage relationships generated by PAGA were distorted **(Figure 2.4)**.

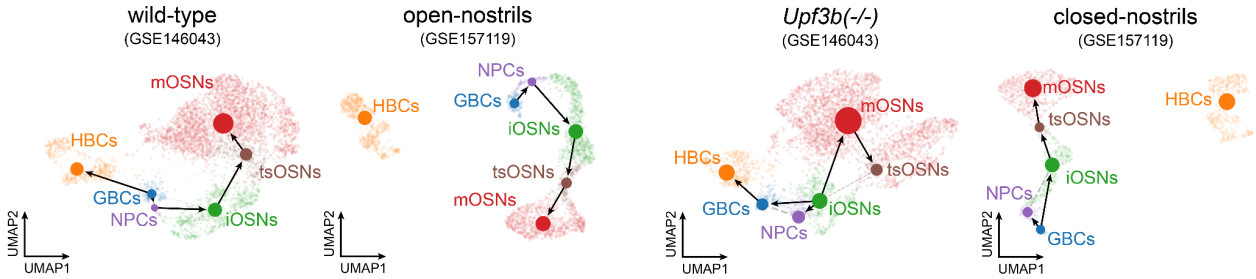


Figure 2.4 RNA Velocity-Based Mapping of Differentiation Trajectories

Trajectory inference graph depicting the directionality of cluster transition using RNA Velocities in the indicated datasets across four biological conditions.

Subsequently, we delineated the expression of OR in these cell types across all datasets, as recent single-cell profiling of olfactory-associated cells demonstrated the presence of OR transcripts in NPCs, iOSNs, and mOSNs (Fletcher et al. 2017; Hanchate et al. 2015). Quantitative findings indicate the existence of ORs in mOSNs, iOSNs, tsOSNs, and NPC populations; however, only a minor proportion of NPCs exhibited OR expression **(Figure 2.3G-H)**. The combined results from olfactory sensory neuron trajectory inference and unbiased PAGA analysis robustly endorse the role of *Upf3b*, thereby implicating the nonsense mRNA decay pathway in olfactory sensory neuron differentiation.

2.3.4 RCT: a computational framework for cellular backtracking

To systematically assess the viability of the winner-takes-all or stochastic selection model for the OR gene selection, we initially sought to establish a connection among the cell types involved in OSN differentiation. We utilize the expression of ORs, their identities, and the transcriptional kinetics as guiding elements. This is accomplished through the development of an innovative, statistically-driven cellular lineage backtracking method, referred to as RCT **(Figure 2.5)**. RCT calculates a stepwise distance between a designated mOSN exhibiting a singular OR (e.g., ORx) and all other cellular types. The RCT's intrinsic property dictates that its direction opposes the cellular differentiation trajectory from mOSNs to NPCs. RCT encompasses a series of multi-step processes, wherein each step forges a connection with cells that express the corresponding OR (ORx) within the proximity of the UMAP embedding space **(Figure 2.5)**.

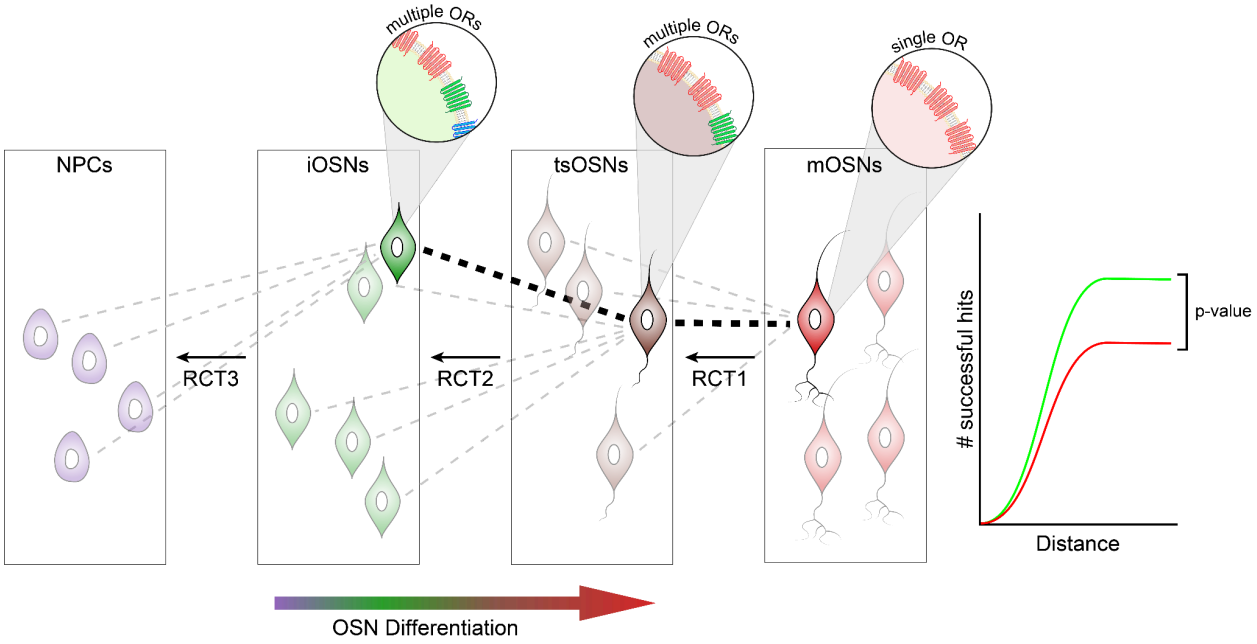


Figure 2.5 Reverse Cell Tracking (RCT) Framework for OSN Lineage Tracing

Scheme depicting the workflow of the Reverse Cell Tracking (RCT) computational framework. The scheme illustrates the sequential events of olfactory sensory neuron differentiation, starting from neuronal progenitor cells (NPCs) to mature olfactory sensory neurons (mOSNs). The Reverse Cell Tracking (RCT) algorithm computes Euclidean distances between mOSNs-tsOSNs, tsOSNs-iOSNs, and iOSNs-NPCs, represented by RCT1, RCT2, and RCT3, respectively. Arrowhead at the bottom represents the direction of cellular differentiation

To achieve comprehensive backtracking, a maximum of three RCT values are produced across various cell pairs, specifically mOSNx to tsOSNx (RCT1), tsOSNx to iOSNx (RCT2), and iOSNx to NPCx (RCT3). A fundamental characteristic of RCT is its recursive operation, which guarantees that all potential mOSNs within a specified dataset are mapped to their corresponding lineage predecessors. This creates a distance-based network linking each mOSN that expresses a specific OR to the nearest tsOSN, iOSN, and NPC that also express the same receptor.

2.3.5 RCT allows OSN backtracking

Upon acquiring the distance relationships, we subsequently evaluated our primary hypothesis, namely, whether the OR exhibiting the highest expression in the NPCs, tsOSNs, or iOSNs is predominantly chosen in the mOSNs (winner-takes-all), or if the selection occurs randomly, resulting in the selection of any OR among approximately one thousand potential options (expressed or non-expressed) in mice (stochastic selection). Initially, we re-validated the quantity of ORs expressed across all cell types and noted elevated OR frequencies in iOSNs compared to mOSNs in datasets reflecting the wild-type condition **(Figure 2.3G)**, corroborating recent findings in both mice and humans (L. Tan, Li, and Sunney

Xie 2015; Hanchate et al. 2015; Durante et al. 2020). Given that the cell cycle phase may impact the subsequent RCT analysis, we assessed the relative abundance of cells in the G2M and S phases and noted no significant differences. Next, we implemented RCT utilizing parameters that sequentially select each mOSN from the input data, associate the expressed OR (e.g., ORx in mOSNx) with a specific tsOSN exhibiting the minimal distance to the queried mOSN, and express the same receptor (ORx) with the highest expression relative to other ORs (e.g., ORx in tsOSNx). This distance is designated as RCT1 **(Figure 2.5)**. After pairing all potential mOSNs with their corresponding tsOSNs according to the aforementioned criteria, the RCT framework subsequently propagates and establishes the next relationship between the chosen tsOSN and iOSN, and further between the selected iOSN and NPC, employing the same premise of maximal expression and minimal distance. These distances are designated as RCT2 and RCT3, respectively **(Figure 2.5)**. This recursive process persists until all potential mOSNs are systematically mapped to their preceding cell types along the reversed differentiation trajectory.

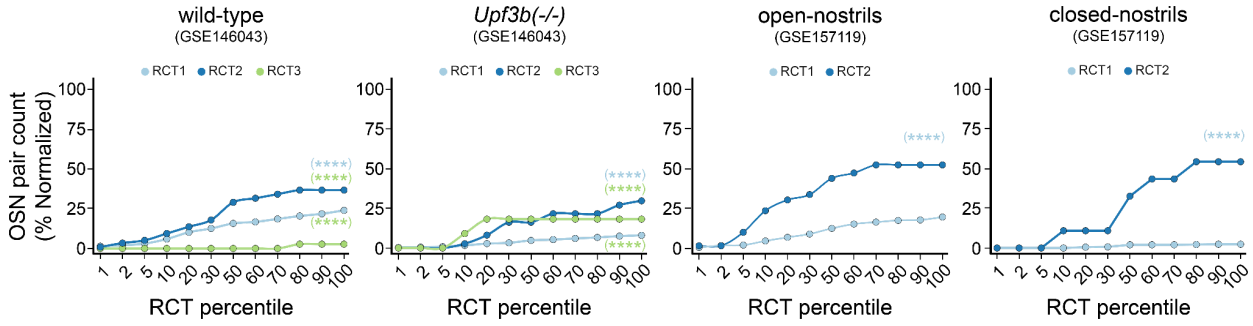


Figure 2.6 Normalized Success Rates for RCT1, RCT2, and RCT3 Across Conditions

Line graphs depicting the percentage of normalized hits (successful pairs) for the RCT1 (mOSN-tsOSNs), RCT2 (tsOSNs-iOSNs), and RCT3 (iOSNs-NPCs) in the indicated conditions. ANCOVA was performed on computed regression lines to compute the P-values. Asterisk represent significant p-value (< 0.05) and ns represent non-significant differences (≥ 0.05).

Initially, we conducted RCT measurements under four conditions (GSE146043 and GSE157119). In the GSE146043 dataset, RCT traced mOSNs back to NPCs (RCT1 + RCT2 + RCT3), while in the GSE157119 dataset, RCT traced back to iOSNs (RCT1 + RCT2) **(Figure 2.6)**. This may be attributable to the limited presence of OR-expressing NPCs in the GSE157119 dataset. Significantly, RCT facilitates the precise quantification of successful cellular pairs **(Figure 2.5)**. Of note, successful cellular pairs denote cell-cell connections that meet the criteria of maximal expression and minimal distance. We employed it as a metric to quantitatively assess the quantity of favorable cell-cell connections. This method is beneficial compared to direct distribution comparisons as it enables the concurrent estimation of successful pairs at various distance cutoffs; concurrently, this information is entirely obscured in comprehensive distance distribution comparisons. The distance matrix computed by RCT is contingent

upon the OR distribution and cell types; under certain conditions, cell backtracking occurs only up to RCT1/2, whereas in the comparative dataset, it extends to RCT3.

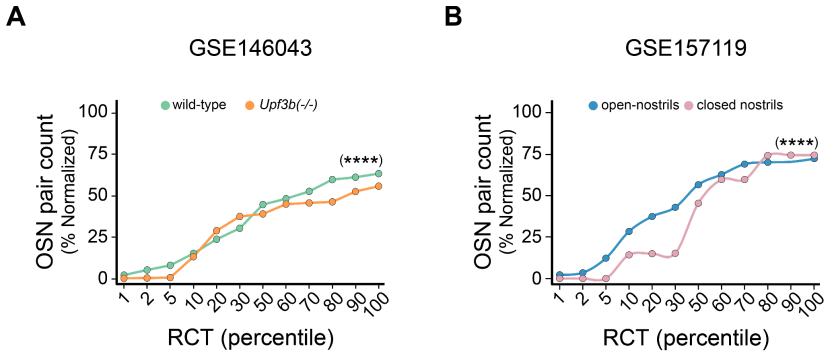


Figure 2.7 RCT Reveals Effect of Upf3b Loss and Sensory Input

(A) Line graphs depicting the percentage of normalized hits (successful pairs) in the wild-type and Upf3b knockout (Upf3b(-/-)) datasets. ANCOVA was performed on computed regression lines to compute the P-values. Asterisk and ns represent significant p-value (< 0.05) and non-significant differences (≥ 0.05), respectively. (B) Line graphs depicting the percentage of normalized hits (successful pairs) in the open-nostril and closed-nostril datasets. ANCOVA was performed on computed regression lines to compute the P-values. Asterisk and ns represent significant p-value (< 0.05) and non-significant differences (≥ 0.05), respectively.

To eliminate any distance-related bias, we calculated and compared the number of successful pairs at various distance cutoffs (categorized by percentiles) and noted significant differences between the wild-type and *Upf3b* knockout cells **(Figure 2.7A)**. Notably, *Upf3b* has previously been hypothesized to affect receptor selection (K. Tan et al. 2020). An analogous analysis was conducted on an independent dataset (GSE157119) comprising two biologically distinct conditions: cells collected and profiled from olfactory-deprived (closed nostrils) and olfactory-stimulated (open nostrils) mice. The intercomparison of successful pairs at various distance cutoffs demonstrated significant disparities between open and closed nostrils, indicating deficiencies in cellular differentiation under closed nostril conditions **(Figure 2.7B).** These findings align with prior reports that associate closed nostrils with compromised neurogenesis in the olfactory epithelium (van der Linden et al. 2020). Consistent with other reports, we observed diminished OR expression in a limited subset of NPCs (Hanchate et al. 2015; Fletcher et al. 2017) **(Figure 2.3H)**. This is primarily attributable to the restricted quantity of OR-expressing NPCs, which constrains the establishment of RCT-mediated connections between iOSNs and NPCs (designated as RCT3). This suggests that the selection of OR genes may transpire in the iOSNs, as a greater proportion of these cells expresses multiple ORs, and the quantity of successful pairs was elevated in the RCT2 (tsOSNx-iOSNx) across all datasets **(Figure 2.6)**.

2.3.6 RCT revealed selection bias for the highest-expressed OR

As an independent control, we conducted cumulative RCT-based mapping of mOSNs, wherein we selected the second (designated as Rank 2) or third highest expressed OR (designated as Rank 3) in predecessor cell types (tsOSN or iOSN), rather than the highest expressed OR (designated as Rank 1). We calculated and compared the quantity of successful pairs across all ranks under all tested conditions. In every instance, the quantity of successful pairs peaked at Rank 1, unlike Ranks 2 and 3, even at reduced distance cutoffs **(Figure 2.8A)**. These findings corroborate the hypothesis that the selection of OR genes is affected by OR expression during the iOSNs/tsOSNs phases under wild-type conditions. We subsequently questioned whether the observed preference for Rank 1 could be attributed to the relative representation of OR in the OSNs. To evaluate this, we analyzed the relative abundance of ORs that established successful connections across all three Ranks (1, 2, and 3) in wild-type datasets.

Our findings indicated comparable normalized OR frequencies across OSNs, implying an absence of discernible bias related to the chosen OR representation within each dataset. If the OR selection is stochastic, any random OR in mOSN, aside from the realistically expressed one, could yield a similar pattern across various distance cutoffs. To evaluate this, we compared the total number of successful pairs at specified distance thresholds between the experimental conditions and their corresponding randomized controls (50 controls per condition). In all instances, comparative analysis demonstrated a substantial percentage increase in successful pairs under favorable conditions (highest expression and minimal distance) relative to randomized controls (non-highest expression and minimal distance). The Chi-square test revealed a significant difference at each individual percentile cutoff, suggesting that the 'winners-takes-all' approach is the favored method of OR gene selection (K. Tan et al. 2020) **(Figure 2.8B)**. To substantiate the transcriptomic signature prevalent in the dominant ORs of the iOSNs in wild-type mice, which may affect the RCT outcome, we extracted their upstream regions (500 bps from the TSS) and conducted differential enrichment analysis utilizing MEME-ChIP (Ma, Noble, and Bailey 2014). We utilized the upstream sequences of the non-dominant ORs expressed in the iOSNs as a control group. Our differential enrichment analysis identified a group of Transcription Factors (NANOG, MEF2C, MEF2A, MEF2D) that the dominant ORs specifically share in the iOSNs state **(Figure 2.8C)**, which we hypothesize to be the principal factors responsible for conferring dominant expression.

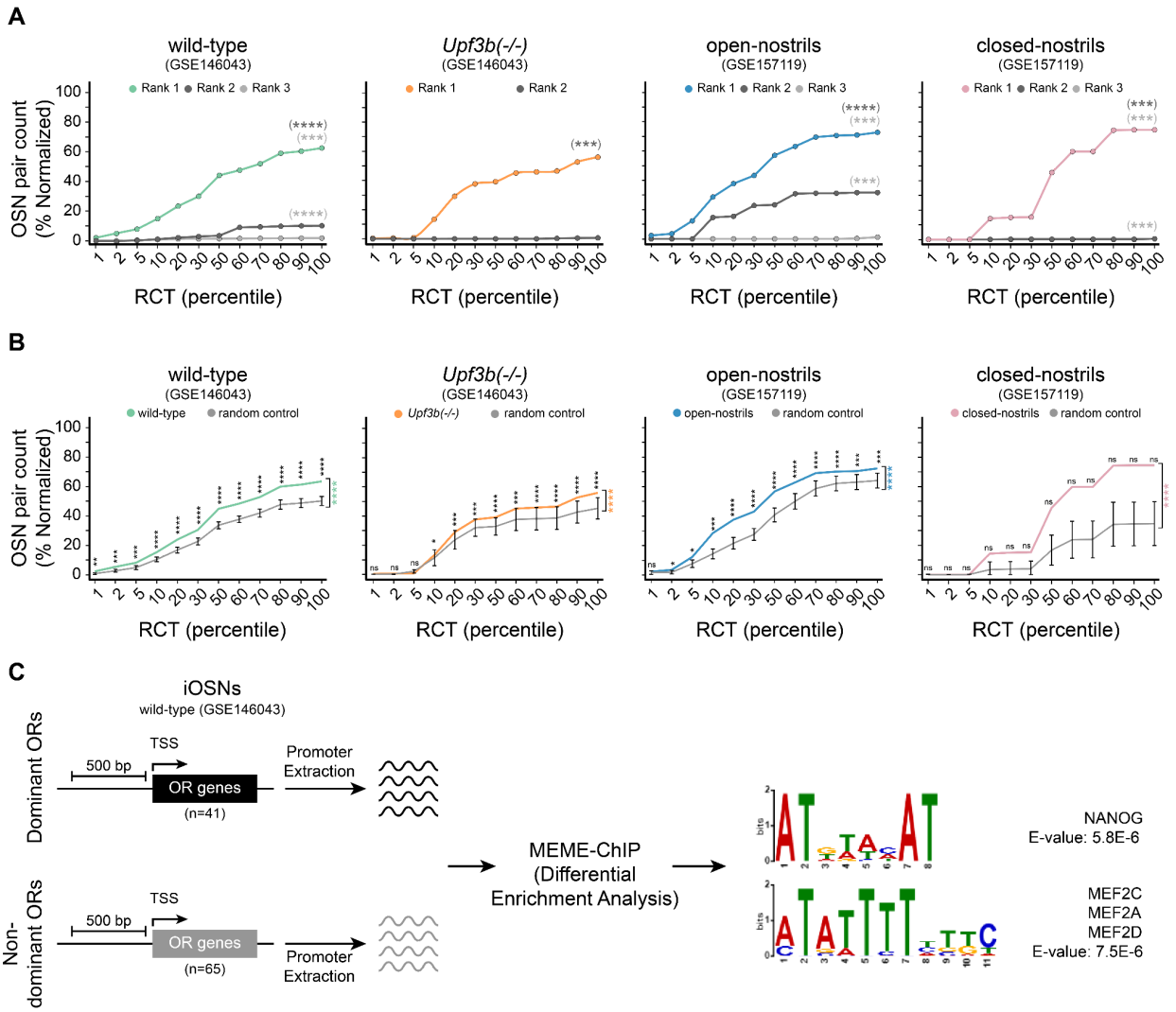


Figure 2.8 Rank-Based Success Pair Efficiency and Upstream Motif Enrichment

(A) Line graphs depicting the percentage of normalized hits (successful pairs) for the Rank 1, Rank 2, and Rank 3 conditions. (B) Line graphs depicting the percentage of normalized hits in the indicated conditions, compared with 50 randomized controls. Error bars represent standard deviations from the mean for the randomized control conditions. The Chi-Square test was performed for the test of significance between the enrichment pattern and random control at each percentile distance cutoff. In panels A and B, ANCOVA was performed on computed regression lines to compute the P-values. Asterisk and ns represent significant p-value (< 0.05) and non-significant differences (≥ 0.05), respectively. (C) The schematic representation on the left depicts the overall workflow used to extract the upstream nucleotide sequences of the dominant and non-dominant ORs in the iOSNs of the wild-type mouse (GSE146043). The right part depicts the key motifs enriched in the upstream region of the dominant ORs identified using Differential Enrichment analysis using MEME-ChIP.

Our comprehensive analysis of cellular trajectories, combined with a robust statistical cellular backtracking framework, strongly indicates that the most highly expressed ORs in the iOSN or tsOSN

stages are more likely to be selected in mOSNs during differentiation, thereby endorsing a winner-takes-all model for singular OR expression in mOSNs.

2.3.7 The Winner-Takes-All model is independent of OR zonal restriction or expression levels

The olfactory epithelium of mice expresses over a thousand functional ORs that are spatially organized into continuous and overlapping regions known as zones (Ruiz Tejada Segura et al. 2021; L. Tan and Xie 2018). Besides spatial segregation, recent RNA-seq profiling of the mouse olfactory mucosa demonstrated that not all ORs are expressed uniformly (Saraiva et al. 2019). We subsequently questioned whether the identified selection bias for the most highly expressed OR could be affected by the relative expression levels of the ORs within the olfactory mucosa or the specific regions of their expression. To resolve this, we initially collected the RCT output of the wild-type mice (GSE146043), which included distance data regarding the expressed OR in the mOSNs compared to the tsOSNs. Subsequently, to assess the potential impact of ORs relatively enriched in the olfactory mucosa on RCT, we initially categorized the ORs into high-expressed (H.E. ORs) and low-expressed (L.E. ORs) based on their expression levels within the bulk tissue (Saraiva et al. 2019).

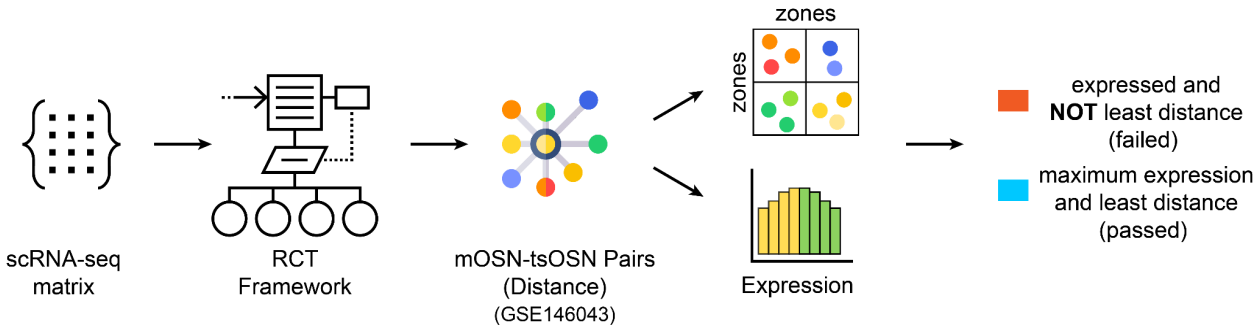


Figure 2.9 Workflow for Validation of RCT Results

Schematic representation depicting the entire workflow used to evaluate the performance of RCT on odorant receptors expressed at different levels (highly and low expressed) within the mouse olfactory mucosa. OSN mapping using OR expression is termed as passed only in case the odorant receptor is highest expressed within the immature state and possesses the least distance on the UMAP embeddings from RNA Velocity Workflow.

We then quantified and statistically analyzed the quantity of olfactory receptors (in mOSNs) that were closest to the transition state olfactory sensory neurons exhibiting the highest expression of the same olfactory receptor. ORs that meet this criterion are classified as "passed" for RCT mapping; conversely, those exhibiting the same OR with the highest expression but not minimally distant from tsOSN were designated as "failed" **(Figure 2.9)**. An analogous analysis was conducted on segregated ORs according to their zonal expression. Initially, we examined whether the ORs expressed in the iOSNs within the

wild-type background exhibit zonal limitations and found that the majority of iOSNs express multiple ORs associated with a single zone **(Figure 2.10A-B)**.

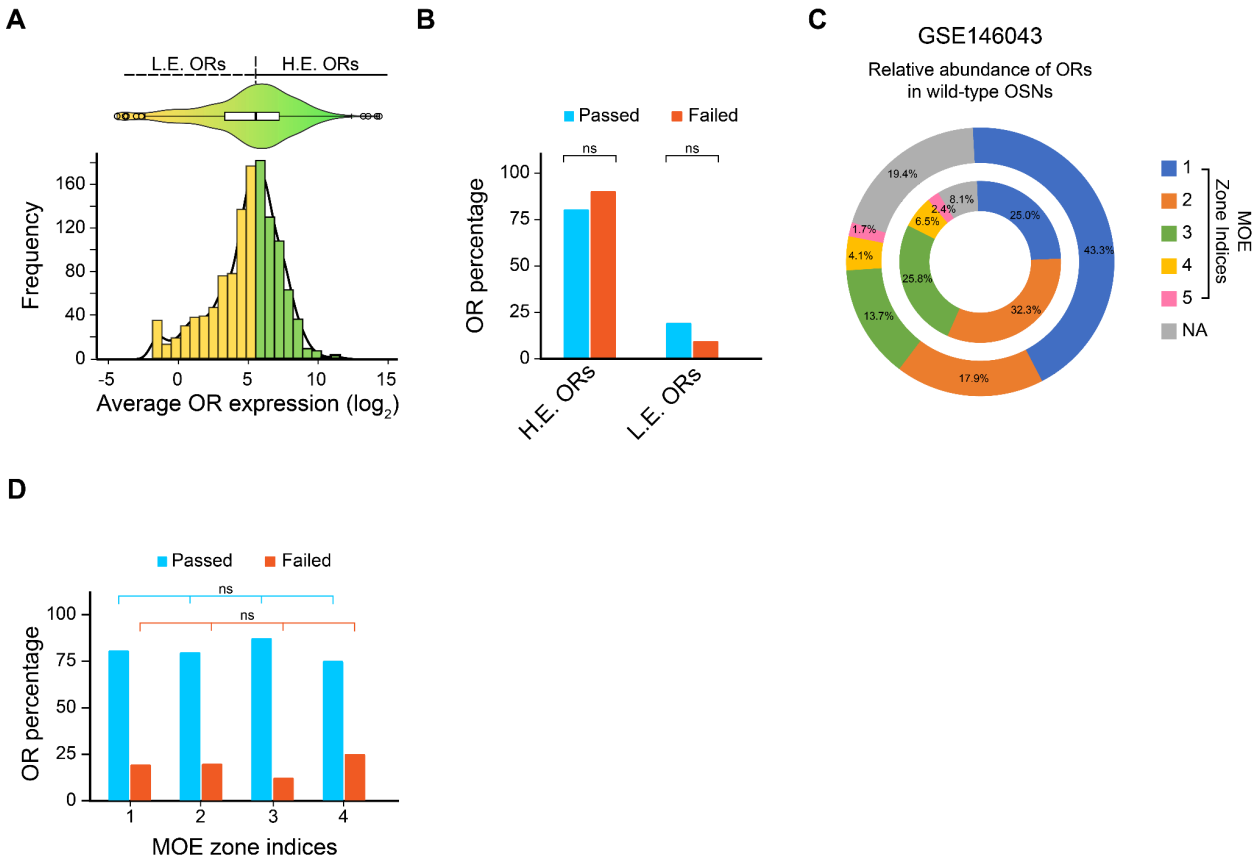


Figure 2.10 RCT results are independent from OR Expression Level and Spatial Proximity

(A) Histogram depicting the distribution of all functional odorant receptors expressed in the mouse olfactory mucosa. Of note, odorant receptors are segregated into two groups, i.e., L.E. (low expressed) and H.E. (highly expressed) groups, based on their overall expression values within the olfactory mucosa (bulk RNA-Sequencing). (B) Bar graph depicting the percentage of ORs that qualifies (passed) or disqualifies (failed) for the RCT mapping criteria, i.e., maximum expression in the transition state olfactory sensory neuron and possesses least distance to mature olfactory sensory neurons. (C) The outer ring of the donut plot depicts the percentage of odorant receptors in each zone (1 to 5) detected in single-cell RNA profiling of wild-type mice (GSE146043), where NA denotes the percentage of odorant receptors whose zonal information was unavailable. The inner ring depicts the percentage of odorant receptors in each zone (1 to 5) that were either qualified or disqualified based on the RCT criterion. (D) Bar graph depicting the percentage of ORs that qualify (passed) or disqualify (failed) for the RCT mapping criteria. The Chi-Square test was used to compute the P-values. ns represents statistical non-significance. MOE represents Main Olfactory Epithelium.

Subsequently, we categorized the ORs expressed in mOSNs within the scRNA-seq dataset (GSE146043) into specific zones according to the literature. It is important to mention that we exclude zone #5 from the

downstream analysis because it contains only a limited number of ORs, which are insufficient for drawing any statistical conclusions **(Figure 2.10C)**. The comparison of the relative quantities of passed and failed OR pairs among various zones showed no significant differences in the pairwise analysis, indicating that the winner-takes-all strategy is applicable to all ORs expressed across different zones **(Figure 2.10D)**.

2.4 Discussion

In the mouse olfactory system, each mature olfactory sensory neuron expresses exclusively one functional olfactory receptor from a repertoire exceeding a thousand different possibilities (Buck and Axel 1991). Efforts are underway to clarify the precise molecular mechanism and the foundational model that may elucidate the one-neuron-one-receptor principle of olfaction. Early research collectively suggested the involvement of epigenetics in determining olfactory receptor choice in mOSNs (Monahan et al. 2017; Bashkirova and Lomvardas 2019; Lomvardas et al. 2006). Recent advancements in scRNA-seq profiling of various cell types in mouse and human olfactory epithelia have demonstrated the expression of multiple olfactory receptors in immature olfactory sensory neurons (Hanchate et al. 2015; L. Tan, Li, and Sunney Xie 2015; Saraiva, Ibarra-Soria, et al. 2015; Durante et al. 2020). These results, however, contradict the previously accepted silence-all-and-activate-one model. Additionally, these findings suggest the existence of a highly intricate epigenetic mechanism that regulates the one-neuron-one-receptor principle.

In the current study, we seek to examine a fundamental question originally posed by Hanchate and colleagues (Hanchate et al. 2015), specifically whether OR selection in the mature olfactory sensory neuron (mOSN) state is determined by the expression levels of ORs in preceding states (winner-takes-all model for the highest expressed OR) or if it occurs randomly among all potential options, both expressed and non-expressed ORs (stochastic model; any OR among over 1000 options in rodents). To resolve this, we created RCT, an innovative computational framework that utilizes RNA Velocity-based trajectory data and UMAP embeddings. RCT facilitates the connection of ancestral and descendant cellular states throughout differentiation trajectories. Our examination of two datasets (pertaining to four conditions) demonstrated that OR selection is preferentially aligned with the most highly expressed OR. This "winner-takes-all" hypothesis suggests that the OR with the highest early expression gains a competitive advantage, ultimately dominating in mature OSNs, potentially due to epigenomic remodeling that reveals regulatory regions to pro-transcription factors. Nonetheless, one cannot dismiss the potential that the distinct subpopulations of mOSNs may adopt varying selection models. The primary limitation of RCT is that the current conclusions are derived from the analysis of a restricted, though randomly selected, olfactory cell type; consequently, it may not fully encompass the cellular heterogeneity in its entirety.

Secondly, not all sensory neurons (iOSNs/tsOSNs/NPCs) in the adult olfactory epithelium express multiple ORs, indicating that a singular model cannot entirely elucidate the OR selection process. Third, in our analysis, we can trace back only ~60% of the OSNs, which we hypothesize may be attributed to the absence of tsOSNs/iOSNs that are transcriptionally analogous to specific mOSNs regarded as their precursors by RCT, leading to unfavorable backtracking, or because certain OSNs do not meet the fundamental criteria employed for calculating RCT, namely, highest expression and minimal distance. The potential for an alternative yet parallel model is apparent from observations of *Upf3b* knockout or wild-type mice with closed nostrils, where significant impairment in olfactory sensory neuron differentiation has been noted, particularly during the transition from immature olfactory sensory neurons to mature olfactory sensory neurons (K. Tan et al. 2020; van der Linden et al. 2020). If a singular mechanism regulates OR choice, then in the *Upf3b* knockout scenario, the mOSN population would not be detectable. The existence of mOSNs under these conditions indicates either a partial impairment or modulation of the OR selection mechanism, or the activation of an alternative mechanism. Given that the refinement mechanism for the elimination of multigenic OR expression in mOSNs is established (Abdus-Saboor et al. 2016), one cannot dismiss the possibility of a comparable mechanism in the precursor OSNs. Furthermore, it is conceivable that tsOSNs or iOSNs expressing multiple receptors may not achieve maturity and instead undergo apoptosis; however, our Gene Set Variation Analysis indicated no selective enrichment for apoptotic markers, thereby negating this possibility. Lastly, and most critically, our analysis did not reveal statistical significance in certain conditions of the RCT at the minimal distance cutoffs. This partial traceability and variable statistical significance suggest that alternative or complementary mechanisms might operate alongside the winner-takes-all model. Bridging computational predictions with experimental validation, such as selective manipulation of OR genes or transcription factors and in vivo reporter tracking, will be critical to confirm and refine the winner-takes-all hypothesis.

A potential explanation for the expression of multiple ORs in iOSNs, as initially proposed by Tan et al. (L. Tan, Li, and Sunney Xie 2015), is the significant alterations in the epigenome during olfactory sensory neuron differentiation, which create a permissive environment for the expression of multiple ORs before ultimately restricting expression to a single OR upon maturation into mOSNs. This epigenomic remodeling may confer a transcriptional advantage to one OR over others by revealing its regulatory regions to pro-transcription factors. This computational study introduces a novel method that theoretically connects various OR differentiation stages while maintaining the identity of the OR, enabling the assessment of expression bias in OR selection in mOSNs.

Chapter 3 - Simplifying Machine Learning Integrated LBDD Techniques for GPCRs

3.1 Introduction

An unparalleled increase in the accessibility of ligand-target datasets, combined with advancements in high-throughput methodologies, has provided enhanced opportunities for extensive exploration of chemical space. There is an urgent necessity to swiftly upgrade the current tools or methodologies to extract meaningful information from these resource-intensive datasets. Recently, Machine Learning (ML) has facilitated numerous discoveries, particularly in the realm of biomedical sciences (Lavecchia 2015).

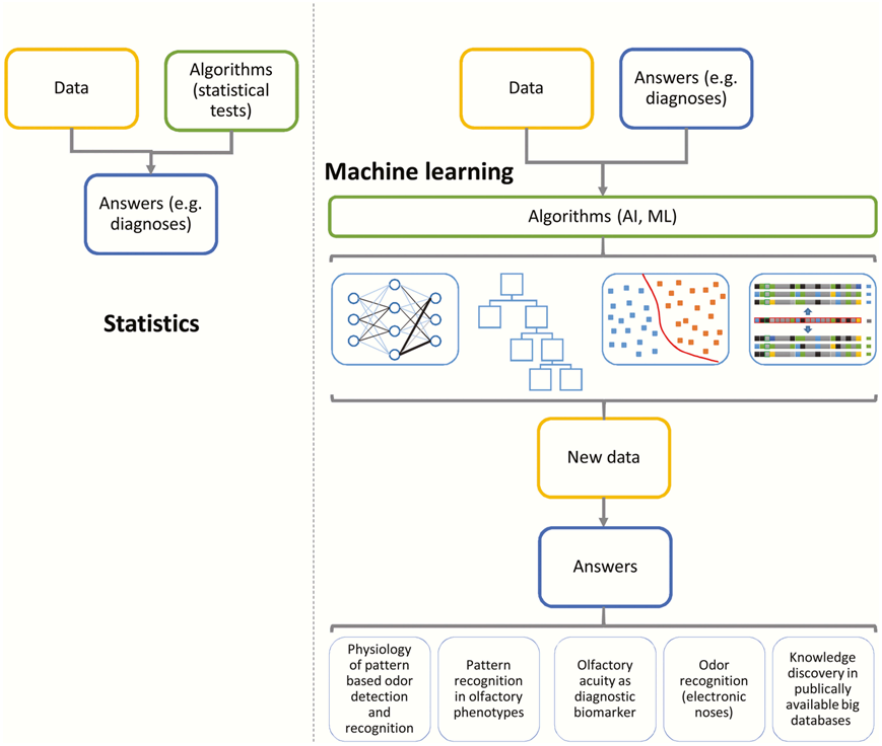


Figure 3.0 Machine Learning in Human Olfactory Research *(Lötsch, Kringel, and Hummel 2019)*
Overview about approaches to data processing, pursed either by statistics (left part) or by machine learning (right part).

ML-based methods are predominantly employed in critical phases including lead identification (Vamathevan et al. 2019), lead validation, biomarker discovery (Leclercq et al. 2019), pharmacodynamics, and pharmacokinetics (Poynton et al. 2009). Recently, these methodologies were applied in the domain of chemosensation, wherein machine learning models were developed to discern novel ligands for chemosensory receptors (olfactory and taste receptors) (Lötsch, Kringel, and Hummel 2019; Sanchez-Lengeling et al. 2019; Nozaki and Nakamoto 2018; Caballero-Vidal et al. 2020; Caroline Bushdid et al. 2018; C. Bushdid et al. 2018). For example, machine learning-based prediction tools like

BitterPredict (Dagan-Wiener et al. 2017) and BitterX (Huang et al. 2016)have facilitated the discovery of new bitter taste compounds, underscoring the significance of machine learning in elucidating the chemical landscape of relevant molecules. Similarly, a variety of new agonists has recently been identified for several human and rodent olfactory receptors (OR51E1, OR1A1, OR2W1, and MOR 256-3) (C. Bushdid et al. 2018). Notably, a recent report by Bushdid et al. experimentally validated predictive ligands using heterologous systems, achieving a remarkable hit rate of 39% to 50% (C. Bushdid et al. 2018), indicating the application of machine learning methods in elucidating the chemical space related to olfactory receptors. Likewise, Google, the technology behemoth, and others have recently explored the capacity of artificial intelligence to predict the correlation between a molecule's structure and its scent (Keller et al. 2017; Sanchez-Lengeling et al. 2019). In line with this, recently Jubeen et al. also applied the machine learning methods to identify potential agonists for one of the broadly tuned human olfactory receptors, OR1G1 (Jabeen and Ranganathan 2019). Comparable methodologies utilizing ML-based frameworks have been documented in the discovery of novel, yet behaviorally active antagonists for the insect odorant receptor-co-receptor (Orco) (Jabeen and Ranganathan 2019). Collectively, these factors underscore the critical significance of ML-based tools in chemoinformatics, thereby necessitating the urgent development and implementation of ML methodologies to facilitate and expedite the identification of novel agonists or antagonists. Given that most chemosensory receptors are GPCRs (Gaillard, Rouquier, and Giorgi 2004; Spehr and Munger 2009; Ahuja et al. 2018; Sharma et al. 2016; Saraiva, Ahuja, et al. 2015; Hussain et al. 2013), key regulators of vital cellular processes such as proliferation, migration, and immune responses (Ferrer et al. 2016b), their limited availability in atomic-resolution structures poses a significant challenge. Another significant challenge is the existence of "orphan" GPCRs, which are receptors for which the endogenous ligands remain unidentified (Alexander et al. 2019, 2023). The orphan GPCRs constitute a substantial segment of the GPCR superfamily (Davenport et al. 2013), and their deorphanization, or ligand identification, is essential for comprehending their physiological functions and therapeutic possibilities (Kroeze et al. 2015). Significantly, considerable efforts have been dedicated to creating a highly user-friendly interface for alternative methodologies (Banegas-Luna, Cerón-Carrasco, and Pérez-Sánchez 2018; Cosconati et al. 2010). Nonetheless, comparable initiatives are predominantly absent for ML-based methodologies, particularly within the realm of chemosensory research. These factors collectively underscore the pressing need for the creation of an intuitive, platform-agnostic, open-source computational framework driven by machine learning, which could enable researchers with minimal computational expertise.

To address this gap, we introduce Machine-OlF-Action, a user-friendly, open-source computational framework designed to empower researchers with minimal programming expertise. We hypothesized that

the integration of automated machine learning workflows into LBDD for GPCRs can streamline and democratize the identification of biologically relevant ligands. Specifically, a transparent and user-friendly computational platform, incorporating built-in descriptors, optimized models, and interpretability tools, will enable reliable ligand prediction even for researchers with limited computational expertise, thereby accelerating the pace and accessibility of GPCR-targeted drug discovery. MOA utilizes user-supplied SMILES (Simplified Molecular-Input Line-Entry System) strings of chemical compounds, along with their known activation statuses, to build robust classification models. By simplifying complex machine learning workflows into an accessible platform, MOA enables even researchers without a deep computational background to uncover meaningful GPCR-ligand relationships and advance the field of chemosensory biology. We tested the utility of MOA in identifying novel, albeit endogenous agonists for the human olfactory receptor OR1A1, recently identified to be predominantly expressed in multiple tumor types (Kalra et al. 2020). We further evaluated the utility of MOA on the olfactory receptor MOR174-9, as well as on olfactory receptor datasets curated by Bushdid et al. (2018) and Jabeen and Ranganathan (2019). As a LBDD framework, MOA is target-independent and emphasizes the bioactivity associated with the function of interest. This enables activity-based ligand screening across libraries without requiring detailed mechanistic knowledge of the underlying biological process. We demonstrated this capability of Machine-Olf-Action through the discovery of novel natural repellents. Our research addresses multiple deficiencies in traditional chemoinformatics methodologies and introduces a flexible, Machine Learning-based computational framework (MOA) designed to develop models for predicting novel agonists or antagonists by leveraging chemical characteristics. Significantly, our workflow incorporates model interpretation functionality, thereby emphasizing the critical molecular descriptors pertinent to the selection of molecules of interest. In summary, our findings reveal several potential natural repellent molecules that may possess a repellency index comparable to the standard repellent, DEET.

3.2 Materials and Methods

3.2.1 Data collection

Literature possessing information about the known repellents were manually curated from PUBMED using search string ("mosquito" OR "Culex" OR "Anopheles") AND ("behaviour" OR "behavior" OR "response" OR "repel" OR "attract") AND ("olfactory" OR "odour" OR "odor" OR "odorant" OR "olfaction" OR "smell" OR "ligand"). In total, 443 research articles were obtained. We manually gathered information about the positive dataset (repellants) and the negative dataset (behaviorally inactive or attractants). In total, we obtained 363 unique entries, out of which 228 molecules were labeled positive,

and the remaining 135 molecules were labeled negative. A similar approach was employed to collect data regarding the agonist or non-agonist for OR1A1 and MOR174-9 from PubMed. We acquired 355 unique entries for OR1A1, of which 68 molecules were classified as positive and 287 as negative. A comparable methodology was employed to collect data regarding the identified agonist or non-agonist for the mouse olfactory receptor MOR174-9. We acquired a total of 198 distinct entries, comprising 29 molecules in the positive dataset and 169 in the negative dataset. We have exclusively chosen entries from the literature that exhibited no discrepancies between two independent studies, as well as data points that were statistically significant according to the original study.

3.2.2 Model building in Machine-OlF-Action

The complete back-end source code of MOA is composed in Python. The workflow comprises several substeps: feature generation from chemical SMILES, data preprocessing, feature reduction, model construction, cross-validation, and elucidation of model explainability. MOA necessitates a user-provided input file containing three data categories: compound names, compound SMILES, and their activation status in binary format ('0' for non-activating compounds and '1' for activating compounds). For optimal results, it is advisable to maintain a ratio of activating compounds to non-activating compounds of at least 15:85. MOA employs two prevalent feature generation techniques, namely PaDEL (Yap 2011) and Mordred (Moriwaki et al. 2018), which utilize compound SMILES from the user-provided input file to produce numeric features (molecular descriptors). PaDEL produces a total of 1875 molecular descriptors (1444 one-dimensional, two-dimensional descriptors, and 431 three-dimensional descriptors), while Mordred, an alternative feature generation tool, generates over 1800 two-dimensional and three-dimensional descriptors. It is important to mention that all our analyses utilized PaDEL (Yap 2011). After feature generation, data undergoes preprocessing prior to being input into machine learning algorithms to ensure error-free training of the predictive model. Preprocessing encompasses: addressing missing values through mean imputation, eliminating features with minimal variance, examining feature correlation, normalizing data, and rectifying class imbalance via SMOTE (Blagus and Lusa 2013). Subsequent to data preprocessing, Boruta is employed for feature selection (Blagus and Lusa 2013; Kursa and Rudnicki 2010). Boruta conducts a top-down search for pertinent features by comparing the original attribute's significance with the attainable significance derived from random estimations using their permuted counterparts, systematically discarding irrelevant features. Boruta is set to utilize the standard RandomForestClassifier with 100 estimators (100 trees in the ensemble). We employed Principal Component Analysis (PCA) for feature extraction. According to our recommended settings, principal components that account for a minimum of 98% of the variance are retained for subsequent analyses. Machine-OlF-Action offers an array of classification models: Support Vector Machine (SVM), Extra Tree

(ET), Logistic Regression (LR), Gaussian Naive Bayes (GNB), Gradient Boosting Machine (GBM), Random Forest (RF), and Multi-Layer Perceptron (MLP). Data validation can be executed utilizing 3-Fold Cross-Validation, 5-Fold Cross-Validation, and Leave-One-Out Cross-Validation (LOOCV).

3.2.3 Inbuilt large library screening in Machine-OlF-Action

To create the test dataset from external databases, Machine-OlF-Action employs various similarity metrics derived from the RDKit package (http://www.rdkit.org). Methodologically, fingerprints for a specific chemical compound are initially generated utilizing the RDKit Fingerprint package. The fingerprints are subsequently compared to those of all compounds in the user-selected databases utilizing various similarity metrics such as “Tanimoto,” “Dice,” and “Cosine,” at varying threshold levels. The similarity calculation was conducted utilizing the RDKit package. The utilized databases include HMDB (Human Metabolome Database) (Wishart et al. 2017), ChEBI (Chemical Entities of Biological Interest) (Hastings et al. 2013), PubChem (Q. Li et al. 2010), FooDB (Harrington et al. 2019), and IMPPAT (Indian Medicinal Plants, Phytochemistry And Therapeutics) (Mohanraj et al. 2018) This step is intended to retrieve analogous chemical compounds. We selected fingerprint-based distance calculation to reduce computational intensity; however, the ultimate prediction of the molecules within each class does not rely on the distance matrix but rather considers the model generated by the user in the initial step. Upon acquiring the putative agonists (and non-agonists) from the user-designated databases via the distance matrix, Machine-OlF-Action maps these compounds onto the models (trained and validated in the initial phase) and assigns statistical significance to their classification as agonists or non-agonists. In the case of mosquito repellants, we also used the Ethnobotanical database (https://phytochem.nal.usda.gov/phytochem/).

3.2.4 Projecting molecules into the model interpretability space

The locally interpretable model-agnostic explanation (LIME) operates by executing a local linear approximation of the model's behavior. Although the model may exhibit global complexity, it is simpler to approximate it in the vicinity of a specific instance (Ribeiro, Singh, and Guestrin 2016). We limited the LIME analysis to the top 20 features identified using sklearn.feature_selection.selectKBest on the training dataset. The sklearn.feature_selection.selectKBest autonomously selects the optimal model by evaluating various classification algorithms while implementing appropriate hyperparameter tuning.

The Machine-Olf-Action pipeline utilizes the Submodular Pick (SP-LIME) functionality of the LIME Python package. SP-LIME provides feature importance scores for each data point, reflecting their contribution to binary decision-making in a bi-class classification context. We employed the feature importance scores derived from LIME to project the data points into a theoretical interpretability space utilizing PCA. This provided us with a two-dimensional representation of the data points, delineated by

principal components one and two. We hypothesized that such a representation would enhance our evaluation of data-point (molecule) similarities based on the alignment of the most critical chemical descriptors.

3.2.5 Single-Cell RNA sequencing data Analysis

The gene expression matrix at the single-cell level from the RNA-sequencing dataset was obtained from CancerSea (Experiment number 66) (H. Yuan et al. 2018). We conducted the majority of our analyses, including cell and gene filtering, clustering, and differential expression analysis, utilizing the widely recognized Seurat software suite (Satija et al. 2015) and the subsequent built-in functions. NormalizeData(), FindVariableFeatures(), ScaleData(), RunPCA(), DimPlot(), FindNeighbors(), and FindClusters() were employed for the standard procedures in single-cell expression data analysis.

3.2.6 Structure-based analysis of OR1A1

The OR1A1 protein sequence was obtained from NCBI and submitted to the GPCR-I-TASSER webserver (https://zhanglab.ccmb.med.umich.edu/GPCR-I-TASSER/) for three-dimensional model prediction(J. Zhang et al. 2015). The structural flexibility and subsequent selection of potential 3D conformations of the receptor were evaluated using the CABS-Flex webserver (http://biocomp.chem.uw.edu.pl/CABSflex2/) (Kuriata et al. 2018). The PPM server (Lomize et al. 2012) was utilized for aligning the lipid membrane. CHARMM-GUI (Jo et al. 2008) was employed for the insertion of the model into the lipid bilayer. GROMACS (Abraham et al. 2015) was employed for the Molecular Dynamics simulations in the lipid environment. The most stable 3D conformation of OR1A1 was chosen for subsequent docking utilizing AutoDock version 4.2 (Morris et al. 2009).

3.2.7 Ligand preparation

SMILES notations of the compounds, derived from the MOA with cutoff prediction probability, were obtained from PubChem utilizing the PubChemPy 1.0.4 Python package (https://pypi.python.org/pypi/PubChemPy/1.0.). The SMILES representations of the target molecules were initially transformed into 3D structures (Mol2 format) utilizing OpenBabel 2.3.2 (O'Boyle et al. 2011), and subsequently converted to PDBQT format employing the MGLTools ligand preparation script prepare_ligand4.py (http://mgltools.scripps.edu/).

3.2.8 AutoDock-based docking

AutoDock was employed to evaluate the binding affinity of both previously identified and predicted compounds to the receptor. The stable OR1A1 structure and the ligands underwent molecular docking, with the grid box encompassing its putative binding site. Scatter plots and correlation analyses between MOA probability scores and AutoDock binding energies (kcal/mol) were produced using the ggscatter()

function in R v3.6.3 (https://www.R-project.org/). LigPlot of the top ligand-receptor complexes was produced utilizing LigPlot+ v.2.1 (Roman A. Laskowski and Swindells 2011).

3.2.9 GUI of Machine-OlF-Action

The Graphical User Interface (GUI) of the Machine-OlF-Action was developed utilizing HTML, CSS, JavaScript, Bootstrap, and the Python-based web micro-framework Flask to present the entire machine learning pipeline code as a web interface. The following Python-based open-source machine learning packages were utilized: RDKit, scikit-learn, Matplotlib, Seaborn, BorutaPy, PadelPy, Mordred, and LIME.

3.2.10 Data and Software Availability

MOA is a cross-platform cheminformatics application compatible with Windows, Mac, and Linux operating systems. The software can be accessed at (https://ahuja-lab.in/). Source code, user manual, step-by-step guide, and support are accessible on GitHub (https://github.com/the-ahuja-lab/Machine-OlF-Action). Functionality of the latest version of our software can be quickly assessed using the link (https://youtu.be/bvvPT4rgRPA).

3.3 Results

3.3.1 Machine-Olf-Action: User-Friendly ML Framework for identification of novel ligands

A persistent challenge in investigating complex biological systems is to elucidate the information contained within intricate biochemical entities that elicit specific behavioral responses. To effectively tackle the aforementioned tasks, substantial expertise in chemistry, biology, and data analytics is required. To facilitate this challenging task, we have created a machine learning-based, end-to-end bioinformatics framework. The essential steps comprise the option to upload an input file containing details about the chemicals, including their names, SMILES, and activation status **(Figure 3.1).** After the input file is uploaded, Machine-Olf-Action calculates various molecular descriptors of the user-supplied chemicals utilizing advanced, open-source feature generation tools such as PaDEL (Yap 2011) or Mordred (Moriwaki et al. 2018) **(Figure 3.2A)**. To address the challenges associated with absent molecular descriptor entries, the elimination of non-contributing descriptors (zero or low variance) and class imbalance issues, Machine-Olf-Action is equipped with diverse algorithms that selectively and accurately tackle these data-related problems **(Figure 3.2A)**, thereby ensuring the flawless training of the predictive model. For the preselection of features capable of reliably distinguishing the classes, we employed Boruta, while principal component analysis was utilized for efficient feature extraction.

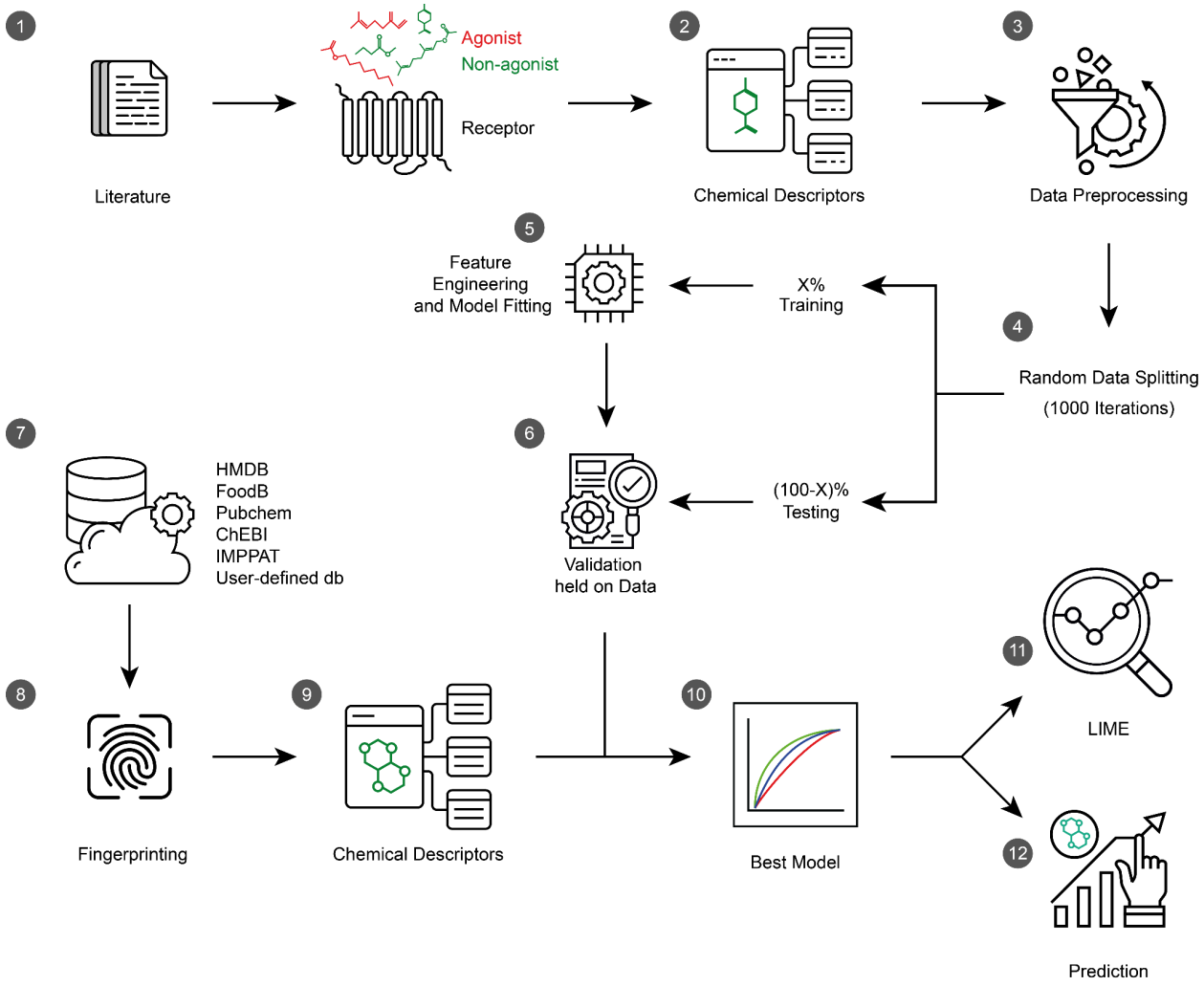
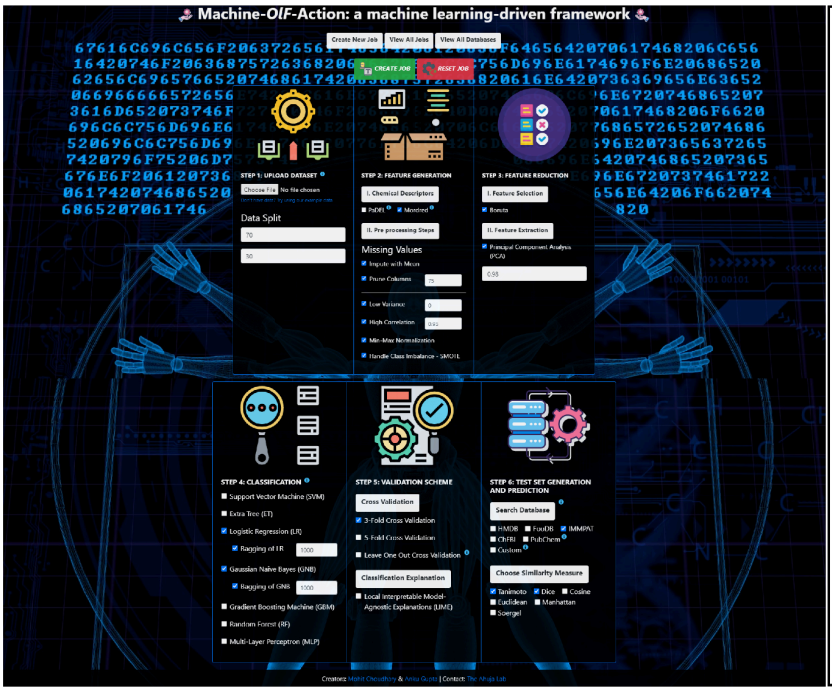


Figure 3.1 Machine-Olf-Action: A unified framework for chemoinformatics.

Schematic representation of the key steps embedded in the Machine-OlF-Action framework. The order of the steps is indicated by the numerical captions.

Machine-Olf-Action offers users the flexibility to select from a variety of distinct classification algorithms, including SVM, ET, LR, GNB, GBM, RF, and MLP. MOA is also equipped with Grid Search from scikit-learn, which systematically generates candidates from a grid of parameter values defined by the param_grid parameter to identify the most optimal parameters **(Figure 3.2A-B)**. Model validation may be conducted utilizing 3/5-fold cross-validation or LOOCV, while local explainability of the constructed model is achieved through LIME (see the methods section) **(Figure 3.2B)**.

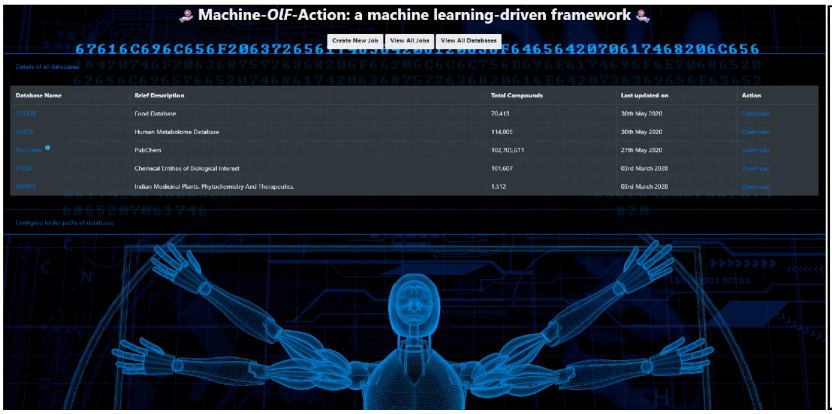
A



Keys Steps

- upload function:
  To select user-defined dataset
- Feature Generation
  (chemical descriptors)
  Using PaDel or Mordred
- Data Preprocessing Steps
  Prune Columns
  impute with Mean
  Variance removal
  Correlation check
  Normalization
  SMOTE
- Feature Selection
  Boruta
- Feature Selection
  Principle Component Analysis

B



Keys Steps

- Classification Methods
  Support Vector Machine
  Extra Tree Gaussian
  Naive Bayes
  Gradient Boosting Machine
  Random Forest
  Multi-Layer Perception
- Validation Scheme
  3-fold cross-validation
  5-fold cross-validation
  LOOCV
- Test Set Generation

Figure 3.2 Graphical User Interface for Job Submission and Result Visualization

(A) The graphical user interface of the Machine-Olf-Action job submission page (left), indicating the key features in a stepwise manner (right). (B) The graphical user interface of the Machine-Olf-Action view job page (left), indicating rest of the key features in a stepwise manner (right).

Ultimately, users are supplied with pre-embedded chemical databases for the creation of the test set, along with the option to select their preferred database. In summary, we created a comprehensive software platform for model generation and predictions related to chemical compounds (SI Video 1). We named our computational workflow “Machine-OlF-Action” because of its extensive applicability in identifying novel odorants for olfactory receptors.

3.3.2 Case study: MOA identifies novel agonists for the human olfactory receptor, OR1A1

To leverage the capabilities of the MOA, we evaluated its effectiveness in discovering novel, yet endogenous agonists for the human olfactory receptor OR1A1.

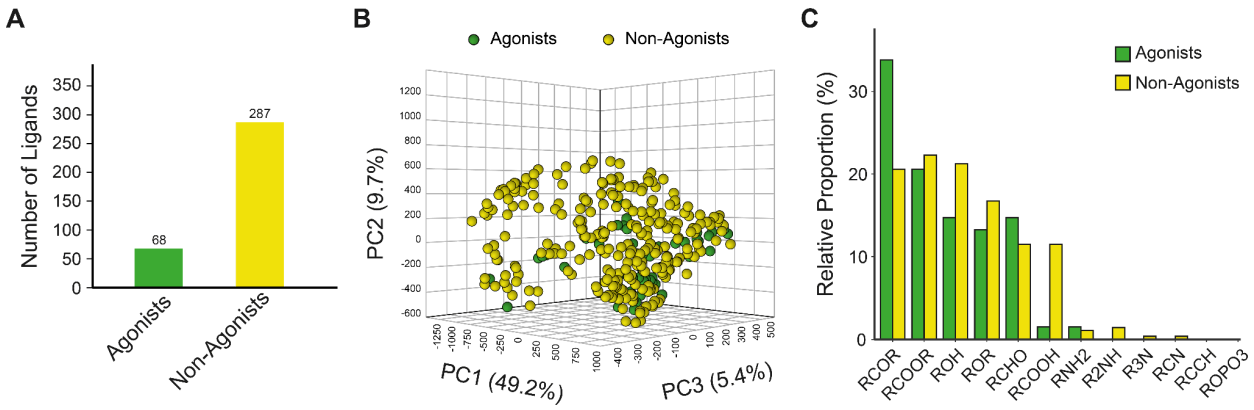


Figure 3.3 Chemical Composition and Distribution in the OR1A1 Dataset

(A) Bar graphs depicting the number of agonists and non-agonists in the OR1A1 dataset. (B) PCA of the known agonists and non-agonists of OR1A1. PCA was performed using the PaDEL-descriptors of these chemicals. (C) Bar graph representing the relative abundance of the 12 prominent functional groups enriched in the known agonists/non-agonists of OR1A1, collectively describing the chemical compositions of the input dataset.

We selected OR1A1 primarily because of its significant expression across various tumor types, and secondarily due to the greater availability of agonists and non-agonists **(Figure 3.3A)**, which provides sufficient data points for training the machine learning models utilizing the MOA framework (Kalra et al. 2020). We propose that metabolic intermediates may serve as potential ligands for the OR1A1 receptor and could be implicated in tumor pathology. To accomplish this, we initially conducted clustering and substructure analysis on the input agonists and non-agonists, noting similarities at the substructure level between the two groups **(Figure 3.3B-C)**. Upon acquiring confidence in the curated dataset, we subsequently developed prediction models utilizing MOA and screened for potential agonists through the HMDB database. In summary, we utilized PaDEL for feature generation, which identified 1,875 molecular descriptors for both OR1A1 agonists and non-agonists. Three high-performance classification models were developed, and their efficacy was assessed using multiple metrics, including accuracy, precision, recall, F1 score, Matthews Correlation Coefficient (MCC), specificity, sensitivity, Cohen's Kappa, AUC-ROC, and Average Precision (AP) **(Figure 3.4A-C)**.

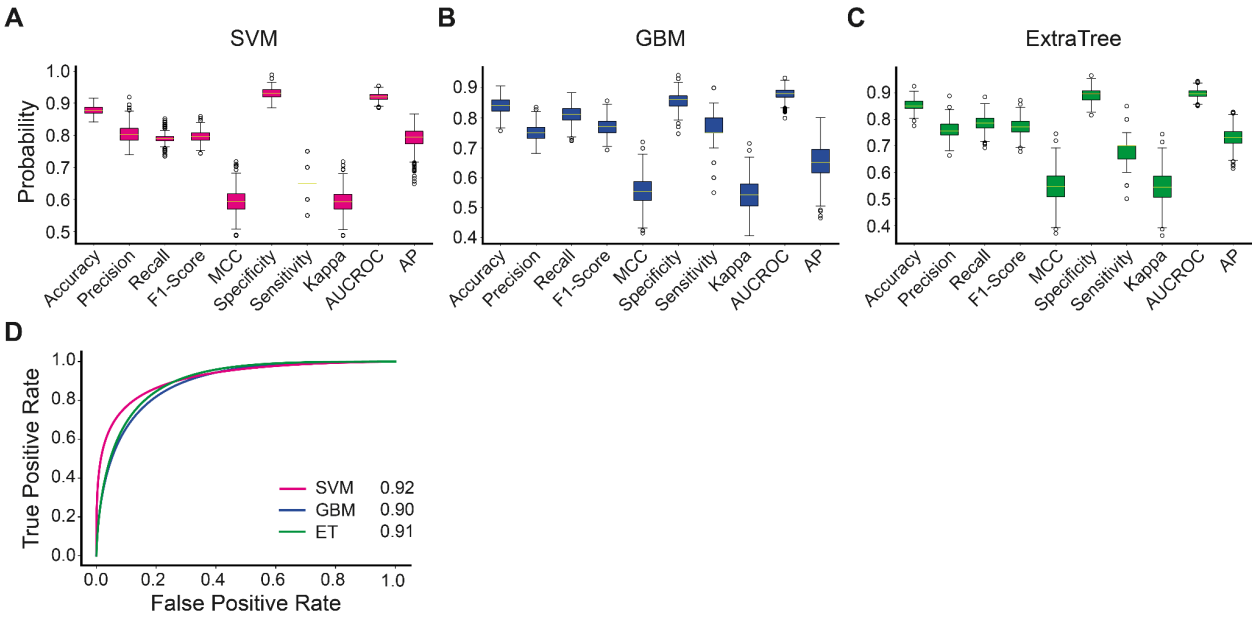


Figure 3.4 Performance Evaluation of ML Models for OR1A1 Ligand Classification

(A-C) Box plots representing the distributions of 1000 random iterations (70-30 split) of the key parameters collectively describing the potency of the indicated models. The parameters include model accuracy, precision, recall, F1 Score, MCC, model specificity, sensitivity, Cohen's Kappa, AUCROC, and AP. (D) AUC (Area under the curve) plots representing the performance of indicated models in classifying the agonists and non-agonists of OR1A1 using their chemical descriptors (PaDEL generated). Models are generated using three widely used classification algorithms namely GBM, SVM, and ET.

We additionally evaluated the model performances employing cross-validation techniques, including nested CV, 5-fold CV, and LOOCV. The most effective model was the SVM **(Figure 3.4D)**. Subsequently, we mapped the refined metabolites from HMDB (Tanimoto coefficient threshold) onto the SVM model of OR1A1, resulting in a list of putative agonists and non-agonists **(Figure 3.5A-C)**. Subsequently, we evaluated the efficacy of the highest predicted agonists and non-agonists through an orthogonal computational method, i.e. structure-based screening.

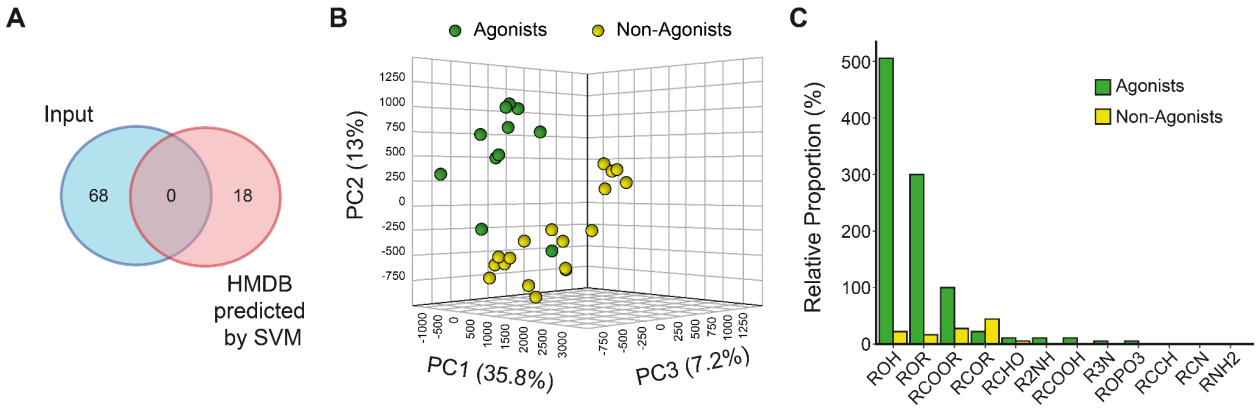


Figure 3.5 Distribution and Functional Profiling of HMDB-Predicted OR1A1 Ligands

(A) Venn diagram depicting the number of overlaps between the predicted agonists of OR1A1 from the HMDB dataset with the input dataset. (B) PCA depicting the top predicted agonists (prediction value >0.80) and non-agonists (prediction value < 0.20) by the SVM model of OR1A1 from the HMDB database. (C) Bar graph representing the relative abundance of the 12 prominent functional groups enriched in the top predicted agonists/non-agonists of OR1A1, collectively describing the chemical compositions of predicted agonists/non-agonists.

3.3.3 Machine-Olf-Action supported by conventional structure-based screening method

Traditional approaches to virtual screening and ligand-target interactions for identifying potential agonists typically entail virtual screening processes succeeded by molecular dynamics simulations, which are frequently assessed based on binding energies. In MOA, we assessed the potential ligands according to the prediction probabilities derived from machine learning models trained on verified data regarding known agonists and non-agonists. To facilitate a direct comparison of the predictive efficacy of Machine-Olf-Action and structure-based virtual screening, we initially constructed the structure of the OR1A1 protein **(Figure 3.6A)**. Subsequently, we enhanced the structure and assessed its quality utilizing the Ramachandran plot and the Discrete Optimized Protein Energy (DOPE) score (M.-Y. Shen and Sali 2006). We conducted a series of refinement analyses to achieve the most stable conformation of the OR1A1 protein **(Figure 3.6B)**. Subsequently, we incorporated this structure into a bilayer lipid membrane and conducted Molecular Dynamics simulations to further optimize the protein structure within the native membrane environment **(Figure 3.6C-D)**. Ultimately, we conducted molecular docking utilizing AutoDock with the ten highest-ranked predicted agonists and non-agonists from the HMDB database **(Figure 3.6A)**.

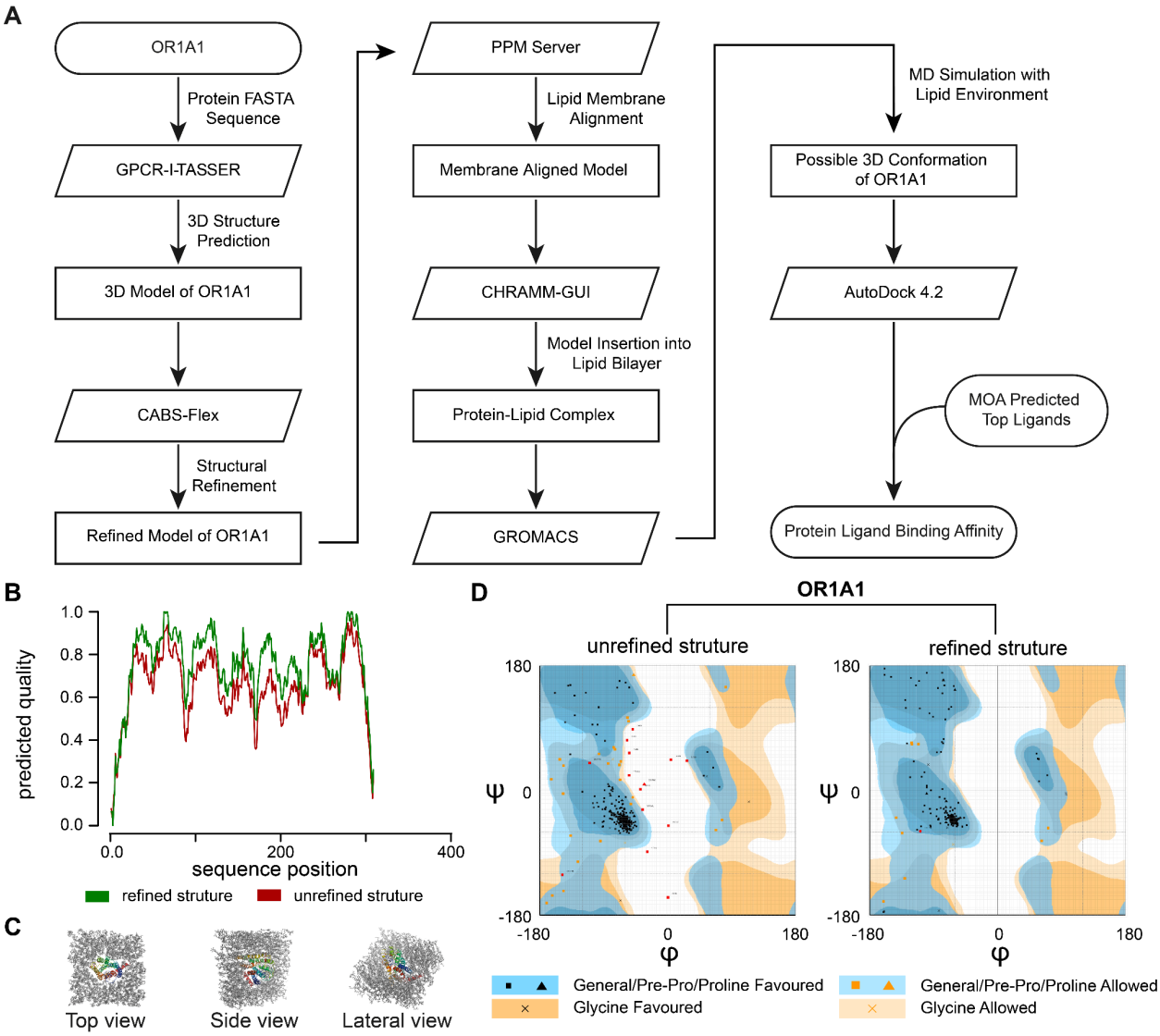


Figure 3.6 Computational Modeling and Quality Assessment of OR1A1 Structure

(A) Schematic representation depicting the key steps employed for the molecular modeling of OR1A1 protein structure, and its docking analysis. (B) Line plots depicting the Discrete Optimized Protein Energy (DOPE) scores of the modeled protein structure of the human olfactory receptor OR1A1, before (red) and after refinement (green). Y-axis represents the DOPE score, and the x-axis represents the alignment positions of the amino acid residues. (C) OR1A1 structure embedded in a bilayer membrane lipid. (D) Ramachandran plots depicting the location of the individual amino acids of the modeled OR1A1 protein structures (unrefined: on left and refined: on the right) in the favored, allowed, and outlier regions.

The comparative analysis of these two orthogonal techniques (prediction probabilities versus binding energies) demonstrated a significant negative correlation, indicating the efficacy of MOA in identifying novel agonists and non-agonists **(Figure 3.7A)**.

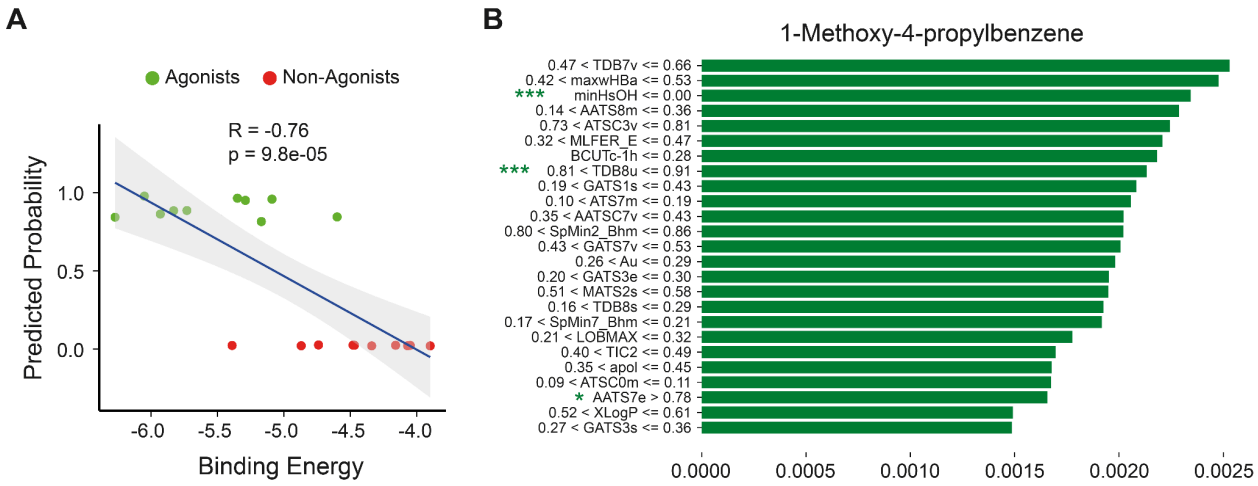


Figure 3.7 Molecular Descriptors and Binding Profiles of OR1A1 Agonist Candidates

*(A) Scatter plots depicting the relationship between predicted OR1A1 agonists (indicated in red) and non-agonist (indicated in green) from the SVM model of OR1A1 and binding energies values obtained using AutoDock. (B) Horizontal bar plot depicting the quantitative ranges of the key descriptors employed by the SVM model of OR1A1 to classify 1-Methoxy-4-propylbenzene as a potential agonist. The asterisk represents the prominent molecular descriptors shared among the top 10 predicted agonists of OR1A1. Notably, features shared by the top 10 molecules are marked as follows: '***' indicates features common to all 10, '**' for those present in 9 out of 10, and '*' for features found in 8 out of 10 molecules.*

Finally, the integrated model interpretability feature elucidates various insights into the model's behavior and uncovers essential chemical characteristics common among the top 10 predicted agonists of the human olfactory receptor, OR1A1 **(Figure 3.7B; Table 3.1)**.

Top Common Features	Description
minHsOH	Minimum atom-type H E-State: -OH
TDB8u	3D topological distance based autocorrelation lag 8 / unweighted
AATS7e	Average Broto-Moreau autocorrelation lag 7 / weighted by Sanderson electronegativities

Table 3.1 OR1A1 agonists prevalent molecular descriptors

Table describing the most prevalent molecular descriptors shared among the top 10 predicted OR1A1 agonists (SVM model) using LIME.

Collectively, our findings unequivocally endorse the idea that machine learning-based identification of novel ligands is relatively more efficient, rapid, and precise, and demonstrate that the mechanism of action could serve as a supplementary tool to structure-based screening methodologies.

3.3.4 Case study: MOA identifies novel agonists for the mouse olfactory receptor, MOR174-9

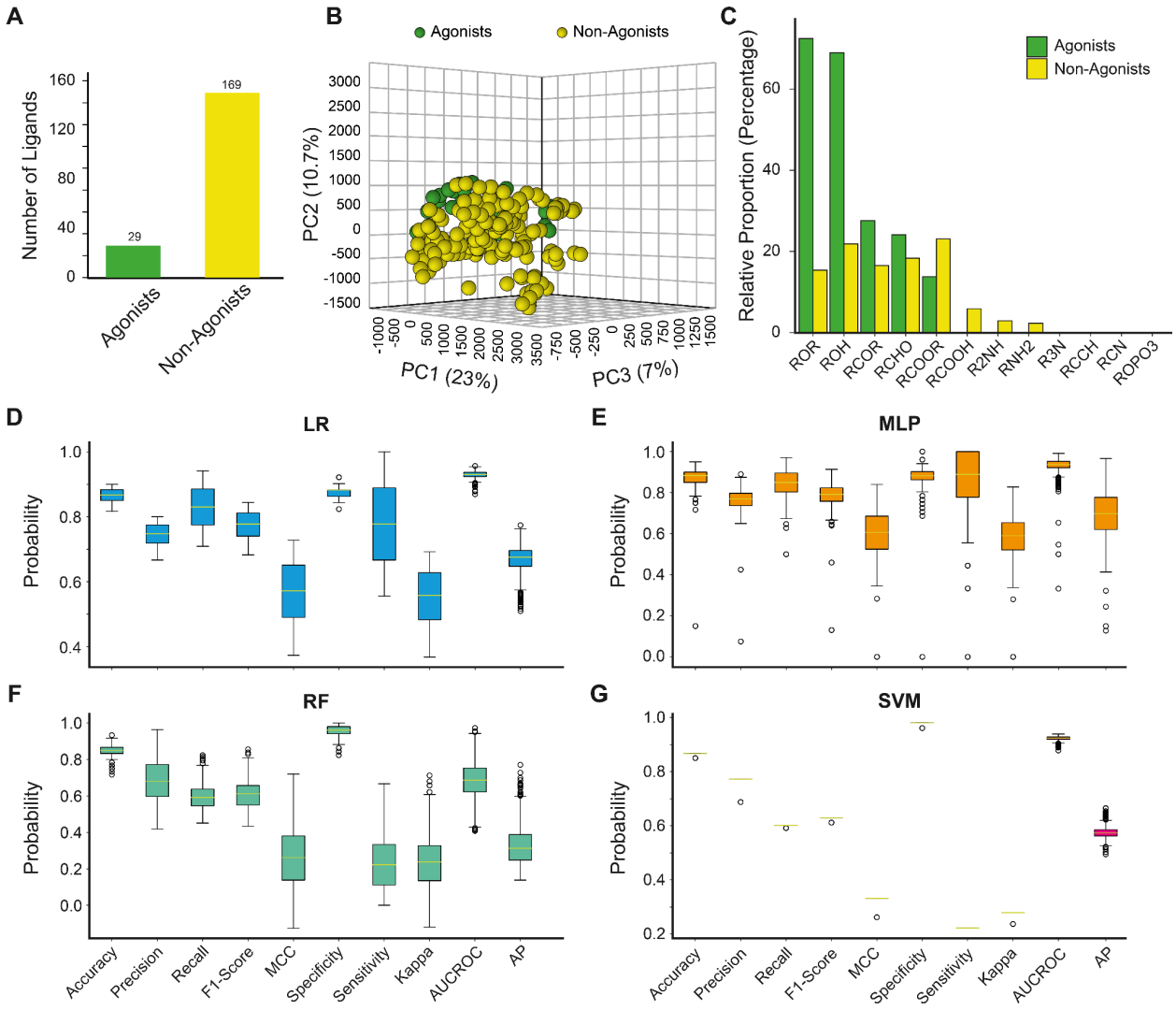


Figure 3.8 Dataset Characterization and Model Evaluation for MOR174-9 Ligand Prediction

(A) Bar graphs depicting the number of agonists and non-agonists in the MOR174-9 dataset. (B) PCA of the known agonists and non-agonists of mouse olfactory receptor MOR174-9. PCA was performed using the PaDEL-descriptors of the input chemicals. (C) Bar graph representing the relative abundance of the 12 prominent functional groups enriched in the known agonists/non-agonists of MOR174-9, collectively describing the chemical compositions of the input dataset. (D-G) Box plots representing the distributions of 1000 random iterations (70-30 split) of the key parameters collectively describing the potency of the indicated models. The parameters include model accuracy, precision, recall, F1 Score, MCC, model specificity, sensitivity, Cohen's Kappa, AUCROC, and AP.

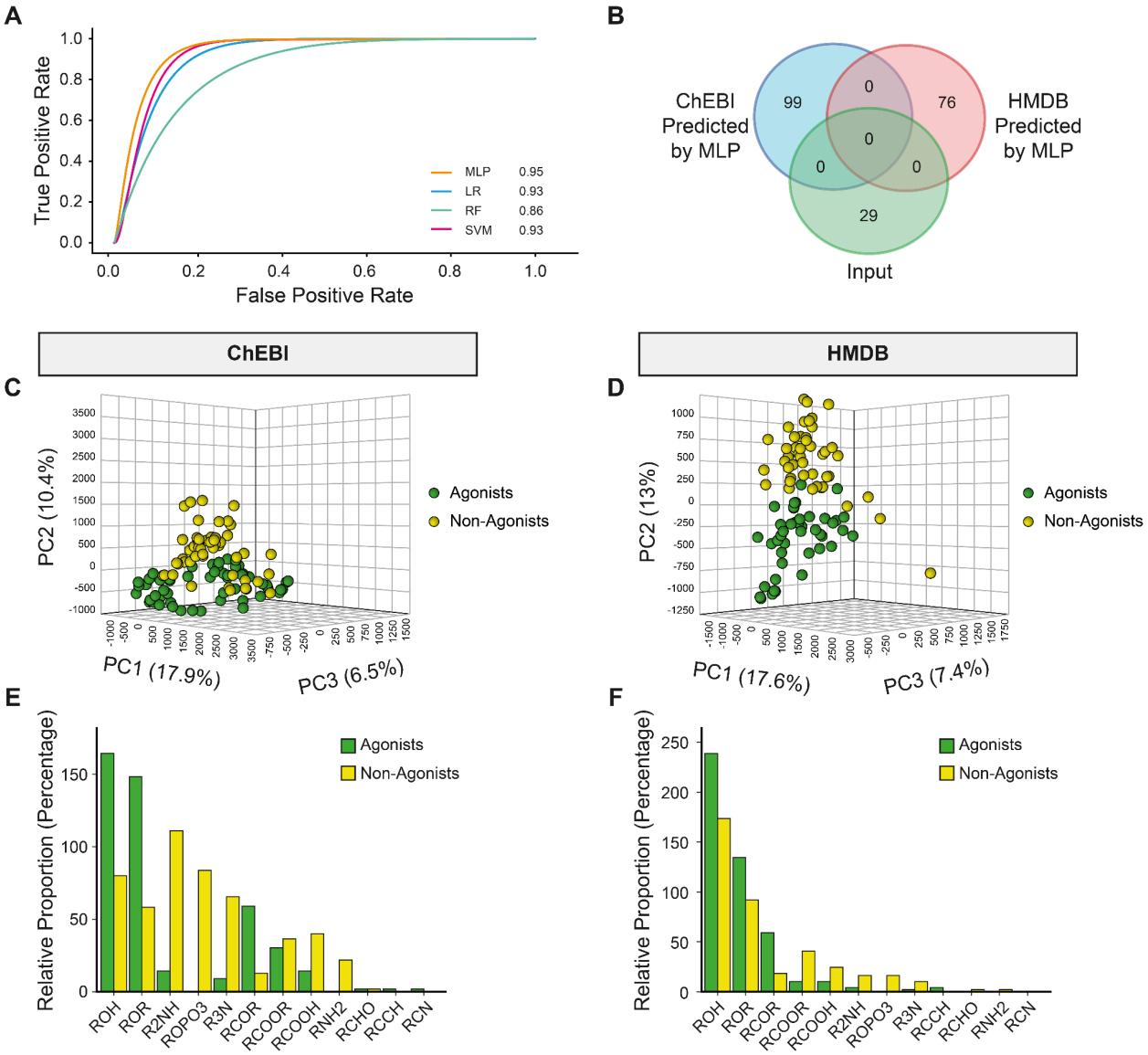


Figure 3.9 Predictive Modeling and Chemical Profiling of MOR174-9 Agonists

(A) AUC plots representing the performance of indicated models in classifying the agonists and non-agonists of MOR174-9 using their chemical descriptors (PaDEL generated). Models are generated using three widely used classification algorithms namely GBM, RF, LR, and MLP. (B) Venn diagram depicting the number of overlaps between the predicted agonists of MOR174-9 from HMDB and ChEBI datasets with input data. (C-D) PCA depicting the top predicted agonists (prediction value >0.80) and non-agonists (prediction value < 0.20) by best-performing model MLP of MOR174-9 from the indicated databases. (E-F) Bar graphs representing the relative abundance of the 12 prominent functional groups enriched in the top predicted agonists/non-agonists of MOR174-9, collectively describing the chemical compositions of predicted agonists/non-agonists from the indicated databases. Of note, functional group estimation was performed using ChemmineR, a Bioconductor package.

As with OR1A1, we utilized PaDEL for feature generation. To achieve a comprehensive understanding of the identified agonists and non-agonists, we conducted clustering and substructure analysis, revealing similarities at the substructure level between the two groups **(Figure 3.8B-C)**. Utilizing MOA, we developed four high-performance classification models employing the classifiers: MLP, SVM, LR, and RF **(Figure 3.8D-G)**. Model performance was evaluated through nested cross-validation, 5-fold cross-validation, and leave-one-out cross-validation. A meticulous evaluation of the performance metrics identified MLP as the superior model **(Figure 3.9A)**. The prediction of potential agonists on the MLP model was conducted utilizing the ChEBI and HMDB databases, adhering to a Tanimoto coefficient threshold. This analysis yielded several structurally significant molecules that may function as potential agonists of MOR174-9 **(Figure 3.9B-F)**.

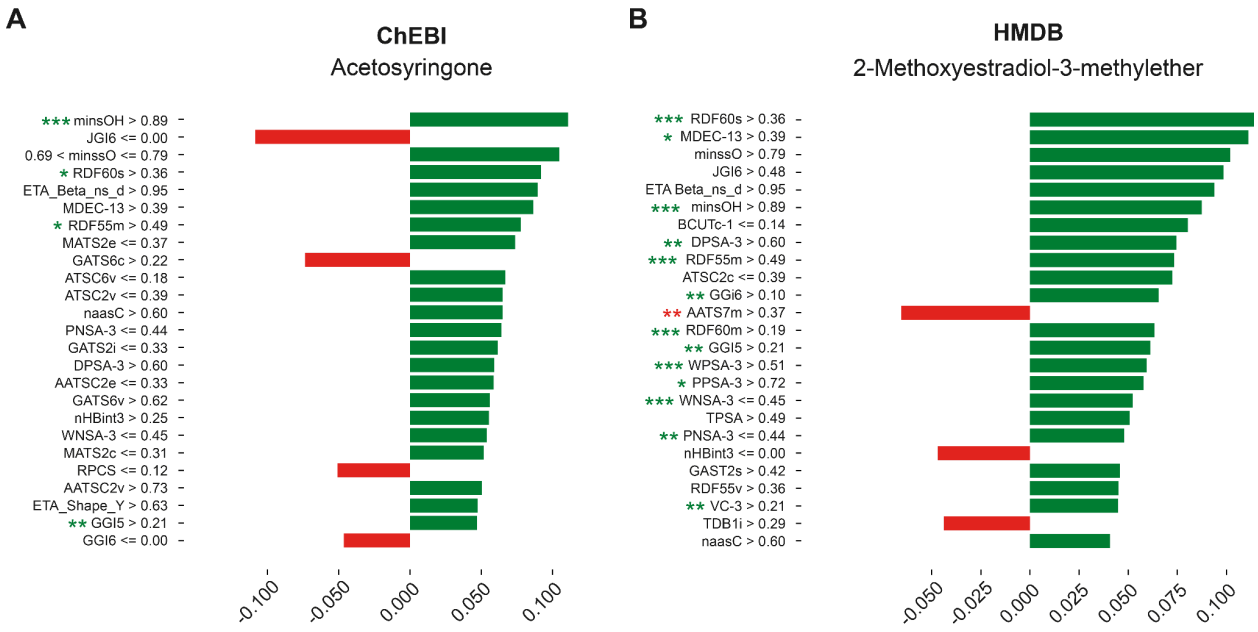


Figure 3.10 Descriptor Profiles of Top Predicted MOR174-9 Agonists

*(A-B) Horizontal bar plots depicting the quantitative ranges of the key descriptors employed by the MLP model of MOR174-9 to classify agonists and non-agonists among the screened molecules. The asterisk represents the prominent molecular descriptors shared among the top 10 predicted agonists of MOR174-9. Notably, features shared by the top 10 molecules are marked as follows: '***' indicates features common to all 10, '**' for those present in 9 out of 10, and '*' for features found in 8 out of 10 molecules.*

Prominent agonists include Acetosyringone (ChEBI) **(Figure 3.10A)** and 2-Methoxyestradiol-3-methyl ether (HMDB) **(Figure 3.10B)**. The LIME-based explainability analysis identified the most significant chemical features common to the highest predicted agonists from the ChEBI and HMDB databases **(Table 3.2)**. The features comprise minsOH, RDF60s, RDF55m, WPSA-3, and WNSA-3 **(Table 3.2)**.

Top Common Features	Description
minsOH	Minimum atom-type E-State: -OH
RDF60s	Radial distribution function - 060 / weighted by relative I-state
RDF55m	Radial distribution function - 055 / weighted by relative mass
WPSA-3	PPSA-3 * total molecular surface area / 1000
WNSA-3	PNSA-3 * total molecular surface area / 1000

Table 3.2 MOR174-9 agonists prevalent molecular descriptors

Table describing the most prevalent molecular descriptors shared among the top 10 predicted MOR174-9 agonists (MLP model) using LIME.

Clustering and substructure analysis of the predicted agonists and non-agonists identified prevalent chemical and functional groups, notably alcohol and ether groups, within each category.

3.3.5 Case study: Machine-OlF-Action validated Bushdid et.al 2018 datasets.

In 2018, Bushdid et al. developed SVM-based predictive models to identify potential agonists for four olfactory receptors. We evaluated the applicability of MOA on these datasets. Given that the authors employed a singular classifier, namely SVM, to construct the model, we consequently selected the same classifier for our analysis **(Figure 3.11)**.

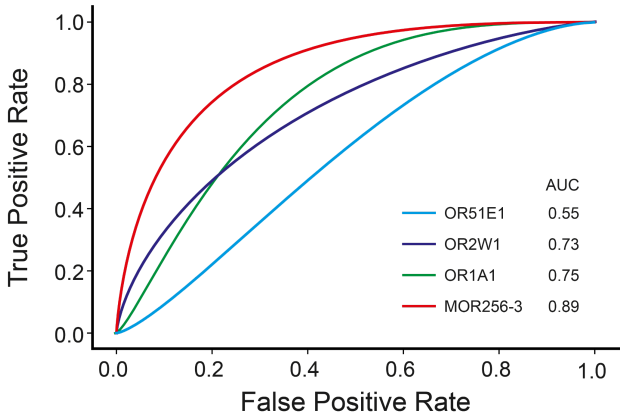


Figure 3.11 SVM-Based Classification Performance for Multiple Olfactory Receptors

AUC plots representing the performance of indicated models in classifying the agonists and non-agonists of four olfactory receptors: OR51E1, MOR256-3, OR2W1, and OR1A1. Models are generated using the SVM algorithm.

Notably, Bushdid et al. deliberately excluded certain non-agonists to address the class imbalance problem during model development, whereas MOA incorporates an inherent implementation of SMOTE to mitigate this issue. Secondly, the authors did not conduct an analysis of model performance parameters; instead, they employed the term “hit-rate,” which indicates the efficacy of the predictive chemicals in

accurately identifying true ligands as confirmed through experimental validation. Although MOA has effectively developed prediction models from these datasets, a comprehensive comparison is impractical due to the absence of one-to-one parameters.

3.3.6 Case study: Building prediction models to infer mosquito-repellency

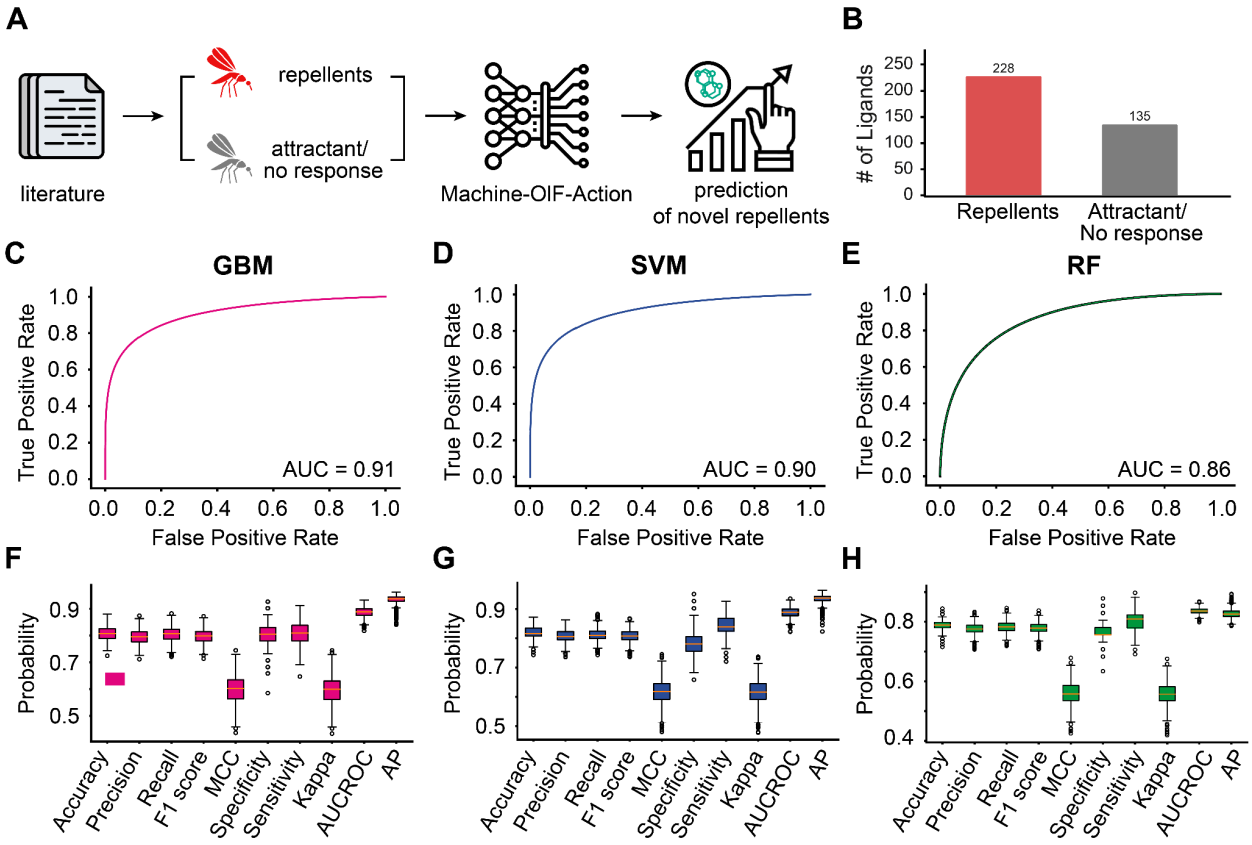


Figure 3.12 Machine-Olf-Action Framework for Repellent Screening

(A) Schematic representation indicating the key steps involved in the identification of natural mosquito repellants using Machine-Olf-Action workflow. (B) Bar graphs depicting the number of repellents and attractant/no-response in the dataset. (C-E) AUC plots representing the performance of indicated models in classifying repellants molecules using their chemical descriptors. Models are generated using three widely used classification algorithms namely GB, SVM, and RF. (F-H) Box plots representing the distributions of 1000 random iterations (70-30 split) of the key parameters collectively describing the potency of the indicated models. The parameters include model accuracy, precision, recall, F1 Score, MCC, model specificity, sensitivity, Cohen's Kappa, AUCROC, and AP.

The prediction of ligand-mediated biological responses solely based on the chemical characteristics of known ligands remains a significant challenge. Given that Machine-Olf-Action is a highly customizable computational framework that incorporates ligand features for subsequent model development and predictions, we hypothesized that it could also be utilized to construct models capable of predicting

biological activities **(Figure 3.12A)**. We employed Machine-Olf-Action to evaluate the repellent characteristics of natural phytochemicals and human metabolites in order to test our hypothesis. To accomplish this, we initially collected data on established repellents by systematically reviewing pertinent publications accessible on PubMed **(Figure 3.12B)**. Specifically, we collected data on the established experimentally-validated repellents and non-repellents (for further details, see Materials and Methods). Machine-Olf-Action developed seven models, three of which demonstrated satisfactory prediction accuracies. These encompass GBM, SVM, and Random Forest **(Figure 3.12B-G)**. To ensure a reliable assessment of the model's performance, we employed 3-fold, 5-fold, and leave-one-out cross-validation. Subsequently, we identified the most effective repellency prediction models and evaluated various phytochemicals and human metabolites for repellent-like characteristics.

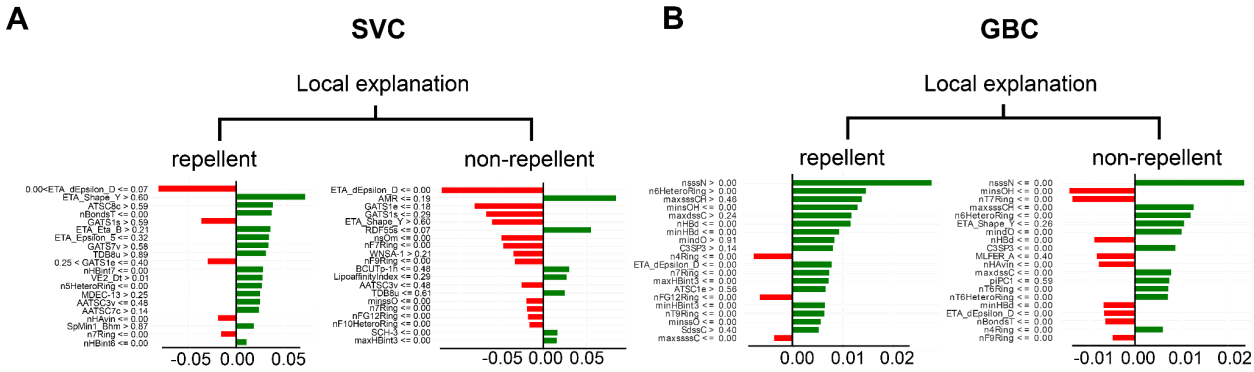


Figure 3.13 Descriptor Contribution in Predicting Mosquito Repellent Activity

(A-B) Horizontal bar plots depicting the quantitative ranges of the key descriptors employed by the indicated models to classify repellants and non-repellants among the screened molecules.

The cumulative outcomes from the three most effective models resulted in the identification of approximately 200 natural compounds exhibiting repellent characteristics. Subsequently, we meticulously reviewed the literature to identify previous reports supporting the repellent efficacy of several of our leading predictions. A substantial subset of the predicted compounds exhibiting mosquito repellent properties has been documented in the literature, indicating the effective extraction, interpretation, application, and understanding of the repellant-specific chemical space by the predictive models. Subsequently, we employed the LIME method to elucidate the model predictions and to ascertain the principal chemical descriptors that confer repellant-like characteristics to the predicted molecules **(Figure 3.13A-B)**. We identified the top 20 features (chemical descriptors) utilizing LIME and depicted all predicted chemicals, alongside DEET (the benchmark for repellents), in a 2D space via PCA **(Figure 3.14A-B)**.

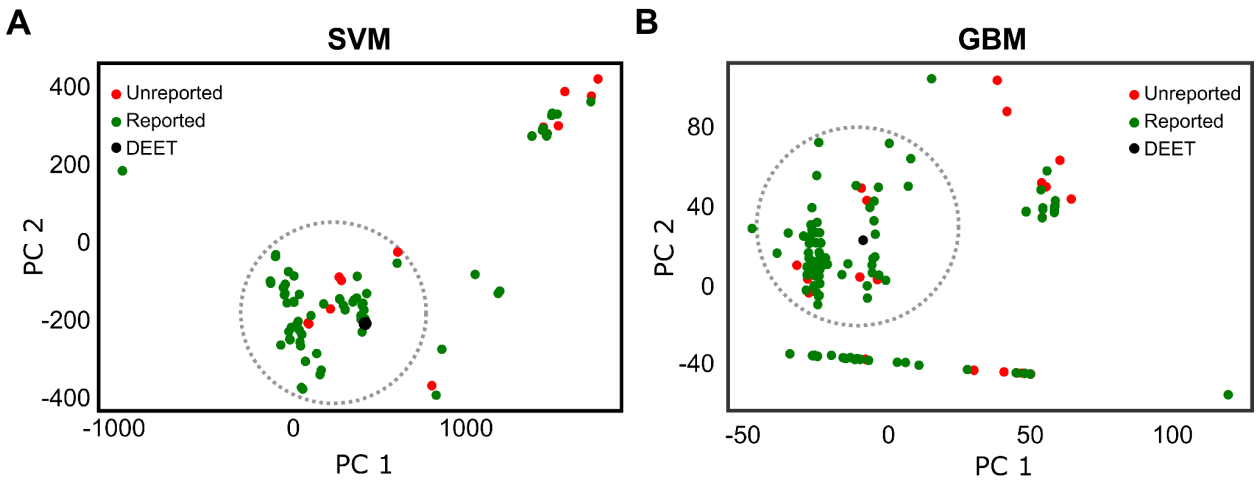


Figure 3.14 LIME-Guided PCA Segregation of Predicted Phytochemicals against DEET

(A-B) Principal Components Analysis segregating the predicted phytochemicals based on the top 20 chemical descriptors identified using LIME.

We systematically gathered the probabilistic scores for each sample and created a 2D map to analyze the chemical clusters within the interpretability space. Subsequently, we acquired the union of the features selected for each label, and utilizing these features, we projected other repellents onto a 2D plane to observe the clusterization of the predicted repellents with DEET. Our findings clearly show that most of the green-colored Machine-Olf-Action predicted repellents do, in fact, share a chemical space with DEET **(Figure 3.14A-B)**. This analysis also identified numerous untested phytochemicals/metabolites exhibiting repellant-like properties **(Figure 3.14A-B)**. In summary, we employed a comprehensive approach that integrates machine learning-based predictive models for the discovery of new natural repellents. Significantly, our research resulted in the discovery of several chemicals exhibiting repellant-like characteristics that occupy the same essential chemical space as DEET.

3.4 Discussion

Given the continually expanding volume of experimental data, it is essential to develop an intuitive, yet data-centric computational framework to enhance the comprehension of biological systems. Machine-learning models can be employed to identify novel ligands in ligand-receptor interactions by analyzing the chemical space of experimentally validated agonists and non-agonists (Haghighatlari et al. 2020). Despite the widespread adoption of such approaches across various scientific disciplines, the utility of these methods remains constrained due to the scarcity of highly user-friendly, flexible, customizable, open-source computational platforms. Furthermore, if accessible, these tools are frequently proprietary or require specialized technical skills.

A contributing factor in structure-activity relationship studies is the extraction of high-quality chemical descriptors. Numerous software applications exist that can calculate such 1D, 2D, 3D, or 4D descriptors (Moriwaki et al. 2018; Yap 2011; Tetko et al. 2005; Pirhadi, Sunseri, and Koes 2016; O'Boyle et al. 2011). Two of the predominant and extensively utilized methods are PaDEL (Yap 2011) or Mordred (Moriwaki et al. 2018), which we employed in the execution of the Machine-Olf-Action workflow. Mordred has been extensively employed in GPCR research for the efficient differentiation of agonist and non-agonist compounds (Jastrzębski et al. 2019). However, the selection of the most appropriate classification algorithm is not straightforward. SVM is the most prevalent classification technique utilized in drug or ligand discovery processes (Jabeen and Ranganathan 2019; He et al. 2016). For instance, it has been employed in the ligand prediction of several GPCRs, including cannabinoid receptors, adenosine receptors, and olfactory receptors OR51E1, OR1A1, OR2W1, and MOR256-3 (C. Bushdid et al. 2018). Besides SVM, classifiers like Random Forest and Naive Bayes (NB) have also been employed in GPCR research. A recent report by Jang et al. (2016) illustrated the application of structure-based and ligand-based virtual screening methodologies, integrating a machine learning technique to identify novel inhibitors for Metabotropic glutamate receptor 1 (mGluR1). In Machine-Olf-Action, we have implemented seven distinct ML classifiers, which we utilized to construct prediction models to hunt for target-specific agonists. A crucial phase in developing high-performance predictive models is data collection and preprocessing. MOA possesses integrated functionalities to address challenges related to input data and its processing, including class imbalance, missing values, and feature pruning issues. Significantly, for each iteration, it conducts a grid search for hyperparameter optimization. The workflow automates the robust estimation of accuracy for model performance evaluation through various cross-validation techniques. Crucially, MOA introduces an innovative method for elucidating the interpretation behind the individual predictions of black-box machine learning models (Carvalho, Pereira, and Cardoso 2019). It offers a list and quantitative elucidations of the principal descriptors considered by the models for the classification of individual chemicals. We assert that the application of this method in ML-driven chemoinformatics will yield profound, descriptor-level insights into the chemical landscape of the compounds being studied.

We evaluated the efficacy of our workflow in identifying novel agonists for the human olfactory receptor OR1A1 and the mouse olfactory receptor MOR174-9, as well as on curated olfactory receptor datasets from Bushdid et al. (2018) and Jabeen and Ranganathan (2019). Additionally, the workflow was applied to discover natural phytochemicals and metabolites with potential mosquito repellent properties. One of the primary limitations of ML-based ligand discovery is the lack of interpretability and repeatability of

ML-generated results, which often constrains their applications (Vamathevan et al. 2019). Machine-Olf-Action has addressed some of these fundamental limitations by enhancing the mathematical models and employing a robust statistical design. Consequently, the proposed tool is anticipated to serve as the foundation for innovative drug or ligand discovery pipelines, effecting a comprehensive paradigm shift that renders traditional, time-intensive methods obsolete. The Machine Learning-driven end-to-end computational framework (MOA) facilitates the rapid screening of numerous compounds. In summary, MOA endeavors to engage non-expert users, for whom the complexities of machine learning research may be intimidating, and to disseminate best practices and guidelines.

Chapter 4 - Addressing The Challenge of Data Scarcity In GPCR Research For Targeted Drug Discovery

4.1 Introduction

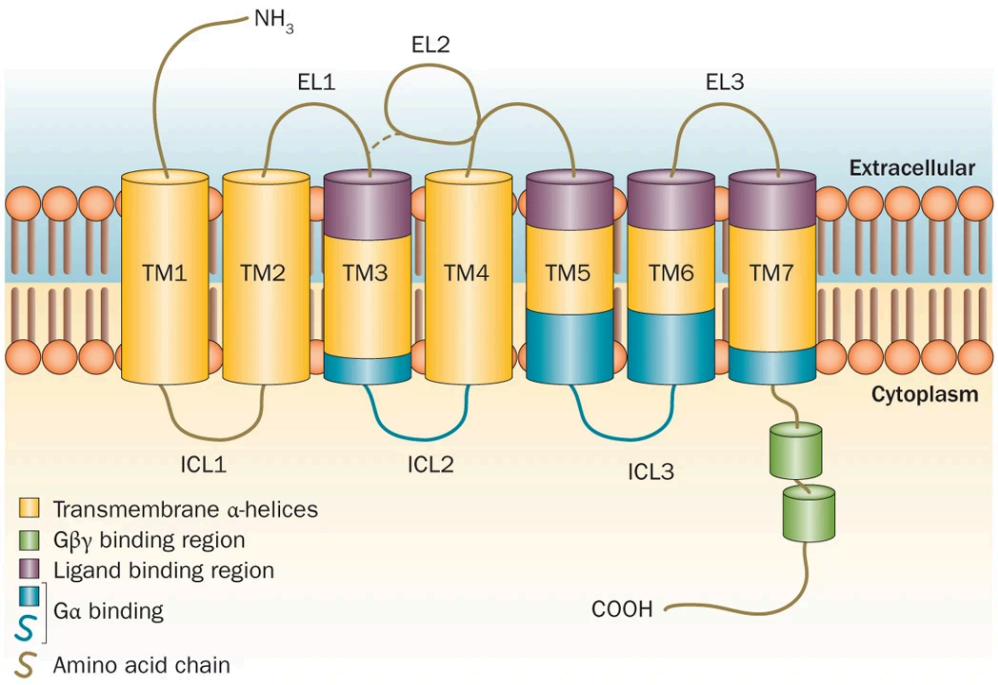


Figure 4.0 Schematic diagram of GPCR structure *(Neumann, Khawaja, and Müller-Ladner 2014)*
The different binding regions are shown in different colours as indicated. The EL2 domain forms a lid-like structure on top of the TM3 domain. Abbreviations: EL, extracellular loop; GPCR, G protein-coupled receptor; ICL, intracellular loop; TM, transmembrane.

Conventional therapeutic strategies aimed at modulating GPCR activity focus on the orthosteric or ligand binding sites (Wold and Zhou 2018; Bourque et al. 2022). Recent advancements in the structural and functional understanding of GPCRs have resulted in the identification of novel, albeit spatially distinct, allosteric sites that may modulate receptor activity responses (Bourque et al. 2022; Rees, Morrow, and Kenakin 2002; Leach, Sexton, and Christopoulos 2007). Allosteric sites can be utilized with chemical probes to modify GPCR-mediated physiological responses, either positively or negatively. Although numerous exogenous synthetic compounds with allosteric properties are documented in the literature, there is still less to no information regarding intracellular allosteric small molecule modulators, their regulatory mechanisms, and their physiological significance (van der Westhuizen et al. 2015; Stornaiuolo

et al. 2015; Doller 2017; Reyes-Alcaraz et al. 2020; P. Zhang, Covic, and Kuliopulos 2015; O'Callaghan, Kuliopulos, and Covic 2012). The practical design or identification of highly specific and high-affinity endogenous allosteric compounds requires efficient methods utilizing structure-based drug design, artificial intelligence (AI), and biochemical assays, each with distinct limitations. Classical computational SAR (structure-activity relationships) methodologies require a significant quantity of experimentally validated compounds for the specified cavity for predictive modeling, which, regrettably, is the limiting factor, particularly in the context of GPCRs (Basith et al. 2018; A. Gupta et al. 2021). Experimental methodologies, including Förster resonance energy transfer (FRET), bioluminescence resonance energy transfer (BRET), and single-molecule detection fluorescence, can be utilized to confirm potential allosteric compounds; however, their application in the search for novel intracellular allosteric modulators through high-throughput screening presents significant technical challenges (Zhou et al. 2021; Hoffmann and Bünemann 2010; Jaeger, Pfleger, and Eidne, n.d.). The identification and characterization of novel endogenous GPCR intracellular allosteric modulators present numerous challenges, including insufficient topology information for various GPCRs (Congreve et al. 2020), the necessary detection methods, extensive screening to assess their modulatory functions, and the vast chemical space involved, among others (Topiol 2018a; Yang et al. 2021). This requires the adoption of a hybrid computational strategy that harnesses predictive structural methodologies for the impartial identification of allosteric sites (cavities), facilitates *de novo* ligand synthesis based on cavity topological characteristics, and ultimately enhances the screening process through the integration of statistical techniques and advanced deep learning methods. Although numerous computational tools have been created for *de novo* drug design, they are predominantly impractical for identifying endogenous intracellular allosteric modulators of GPCRs due to computational limitations and the substantial requirement for specialized knowledge and technical expertise (Spiegel and Durrant 2020; Sicho et al. 2021; Y. Li, Pei, and Lai 2021; Böhm 1992; Nishibata and Itai 1991; Pearlman and Murcko 1993; Y. Yuan, Pei, and Lai 2011). Structure-based, cavity-dependent *de novo* drug design strategies have garnered interest for their capacity to evaluate the molecular conformation, physical and chemical characteristics of the druggable cavity, while considering both topological limitations and non-covalent interactions to identify high-affinity and specific binding ligands, thus enhancing sensitivity in drug screening. The implementation of deep learning, especially generative models that simultaneously utilize the chemical, topological, and physical properties of target cavities, significantly improves screening efficiencies. Despite its significant potential, there is a scarcity of computational tools that provide comprehensive solutions for cavity-based drug design.

To address these limitations, we created Gcoupler, which utilizes computational structural biology, statistical techniques, and deep learning methodologies, providing a novel, cost-effective, and highly

efficient means to systematically elucidate the endogenous intracellular allosteric modulators of GPCRs. We hypothesized that combining cavity detection methods and generative AI can overcome data scarcity in GPCR research. This approach, when combined with deep learning algorithms, is expected to accurately predict receptor-ligand interactions while successfully identifying new, druggable sites on receptor surface. Gcoupler is available as a Python package and a standalone Docker image, ensuring cross-platform compatibility and a seamless user experience. We illustrated the utility and relevance of Gcoupler in discovering novel endogenous modulators of GPCRs by utilizing the α-pheromone (α-factor)-induced mating or programmed cell death pathway of *S. cerevisiae*. Ste2p (GPCR) serves as the central protein in the yeast mating response, wherein the binding of pheromone to Ste2p catalyzes the exchange of GDP for GTP in a heterotrimeric G protein (Gpa1p), thereby initiating a MAPK signaling cascade that culminates in the expression of mating-related genes (Nakayama et al. 1988; Alvaro and Thorner 2016; Blumer, Reneke, and Thorner 1988). It has been previously documented that exposure to low doses of mating pheromones, either during unsuccessful mating or in the absence of an appropriate partner, induces programmed cell death (PCD) in haploid cells (Büttner et al. 2006; Severin and Hyman 2002; N.-N. Zhang et al. 2006; Carmona-Gutierrez et al. 2010). Pheromone-induced programmed cell death (PCD) mechanistically activates the *Ste2/Ste3*-mediated MAP kinase signaling pathway, resulting in elevated intracellular Ca2+ levels and enhanced mitochondrial activity, culminating in the release of cytochrome c (Büttner et al. 2006; Severin and Hyman 2002; N.-N. Zhang et al. 2006; Carmona-Gutierrez et al. 2010). Although pheromone-induced PCD promotes the diploid state, which is relatively more beneficial than the haploid state, it can be contended that evolution has exclusively favored one state over the other without providing an inherent defense mechanism to counteract unintentional or deliberate induction of pheromone-induced PCD. It is well established that the significantly elevated, non-physiological levels of α-factor induce PCD in fewer than half of the MATa population (N.-N. Zhang et al. 2006). Furthermore, an equivalent concentration of α-factor induces varying PCD kinetics among different laboratory strains; for instance, BY4741 exhibits greater resistance compared to the W303 strain (Sokolov et al. 2020). Previous research employing genetic screening techniques identified Yeast Suicidal Protein 1 (LAM1) and Yeast Suicidal Protein 2 (LAM2) as positive regulators of α-factor-induced PCD (Sokolov et al. 2020). The precise mechanisms underlying resistive or sensitive responses to pheromone-induced programmed cell death remain unclear; however, evidence indicates the existence of an unrecognized innate rescue mechanism that confers resistance to MATa in response to α-factor exposure. We hypothesized that a subset of pheromone-resistant cells may modulate the Ste2p-mediated programmed cell death signaling through endogenous intracellular metabolites by functioning at the Ste2-Gα binding interface. Recent single-cell RNA sequencing of yeast cells has elucidated the role of central metabolism in enhancing cellular fitness in response to adverse environmental challenges.

(Nadal-Ribelles et al. 2019). Utilizing Gcoupler, we discerned a subset of intracellular metabolites that may bind to Ste2p (GPCR) at the Gpa1 (Gα) binding interface and inhibit downstream signaling. Our computational results demonstrate that hydrophobic ligands, including sterols, augment the binding of Gpa1p to Ste2p and may trigger a unified response that potentially obstructs downstream signaling. Experimental evidence supports these findings, demonstrating that elevated intracellular levels of these metabolites reduce pheromone-induced PCD. To assess the evolutionary conservation and potential clinical applicability of this mechanism, we examined these metabolites in human and rat isoproterenol-induced, GPCR-mediated cardiac hypertrophy model systems, noting a diminished response in cardiomyocytes pretreated with metabolites that modulate the GPCR-Gα-protein interface.

This study introduces Gcoupler,a computational framework available as a Python package and Docker image, tthat uses *de novo* cavity identification on the GPCR surface, followed by *in silico* ligand synthesis. This is coupled with robust statistical methods to categorize synthesized compounds as high-affinity or low-affinity binders. Gcoupler also integrates deep learning (Graph Neural Networks) algorithms for predictive modeling and bioactivity-based ligand prioritization, enabling large-scale screening of user-specified compounds. Unlike other machine learning or deep learning approaches, Gcoupler offers a novel method that doesn't rely on cavity-specific experimentally validated compounds for model training. It also provides an efficient way to screen user-defined compounds in a seamless manner. Moreover, Gcoupler's flexible design makes it applicable not only to GPCRs but also to non-GPCR proteins. Furthermore, by utilizing Gcoupler, we substantiated the role of intracellular allosteric modulators (metabolites) and demonstrated their functional significance in regulating α-factor induced programmed cell death in yeast and isoproterenol-mediated hypertrophy models.

4.2 Materials and Methods

4.2.1 Backend code for the Gcoupler

The back-end code for the Gcoupler is written entirely in Python (3.8). The Gcoupler workflow consists of three distinct modules, namely, Synthesizer, Authenticator, and Generator. The Synthesizer module leverages the LigBuilder V3.0 (Y. Yuan, Pei, and Lai 2020) to identify the putative cavities on the protein surface with the potential to be an active or an allosteric site and perform *de novo in-silico* drug synthesis within the target cavity by following a hybrid approach of GROW and LINK method using Genetic Algorithm. Notably, among all the predicted cavities, the Synthesizer module pinpoints one cavity as the target for the subsequent *in-silico* ligand synthesis process. The decision is autonomous in nature and is taken based on user-supplied residue-position pairs of interest. The Synthesizer module employs the CAVITY (a structure-based protein binding site detection program) function of LigBuilder V3 for cavity

detection (Y. Yuan, Pei, and Lai 2011). It creates 3D grid points at a cushion of 0.5 Å to contain the whole protein and uses a Tripos force field (Clark, Cramer, and Van Opdenbosch 1989) to categorize these points into three groups: 'occupied grid points,' which contain protein atoms or lie within the range of solvent-accessible radii; 'vacant grid points,' located outside the protein atoms; and 'surface grid points,' positioned between the occupied and vacant grid points. Following this classification, a geometric approach is employed to identify potential sites accessible to ligands, also influenced by factors such as depth and volume. Finally, both geometric structural information and physical chemistry properties are utilized to pinpoint locations for ligand binding sites and to quantitatively assess the feasibility of drug binding within each potential site. Gcoupler identifies multiple potentially orthosteric and allosteric cavities across GPCR topological surfaces and offers users to select one of them (per run). Users are provided with three options to choose cavities in Gcoupler: (1) one of the identified cavities from the Gcoupler, (2) providing key amino acids to receive cavity suggestions, or utilizing third-party tools for cavity identification, and providing amino acid information. Synthesizer also performs *in-silico de novo* synthesis of the active compounds within the target cavity, complementing the pharmacophores in the cartesian space based on the target cavity. In summary, Synthesizer, the first module, generates and outputs the SMILES and PDBQT files of the synthesized ligands along with cavity information in the form of grid coordinates, which is required as input by the next module.

Authenticator, the next module of Gcoupler takes the PDBQT files and the cavity grid information generated using Synthesizer as input and further allows the segregation of these synthetic compounds based on their actual interaction at the molecular level. The Authenticator module leverages AutoDock Vina (1.2.3) python library for virtual screening (Trott and Olson 2010). The binding energies produced by the Authenticator module are further used to segregate the synthetic compounds into High-Affinity Binders (HABs) and Low-Affinity Binders (LABs) based on the distribution while preserving the class balance for the downstream deep learning steps. Of note, the default value of binding energy for the distribution split is set to -7 kcal/mol; however, the module provides enough flexibility to the users to try and test different cutoffs (loose to stringent) and can visualize the change in distribution and statistically compare the resulting distributions within the Gcoupler workflow.

The Authenticator uses the Kolmogorov-Smirnov test (Berger and Zhou 2014), Anderson-Darling test (Engmann and Cousineau 2011), and Epps-Singleton test (Goerg and Kaiser 2009) for hypothesis testing for the distribution comparison.

(1) Two Sample Kolmogorov-Smirnov test:

$$D_{m,n} = max_x|F(x) - G(x)|,$$

where,

$F(x)$ = observed cumulative distribution function of the first sample of size m

$G(x)$ = observed cumulative distribution function of the second sample of size n

(2) Anderson-Darling test:

$$ADK = \frac{n-1}{n^2(k-1)} \sum_{i=1}^{k} \left[\frac{1}{n_i} \sum_{j=1}^{L} h_j \frac{(nF_{ij} - n_i H_j)^2}{H_j(n-H_j) - \frac{nh_j}{4}} \right],$$

where,

h_j = the number of values in the combined samples equal to Z_j

H_j = the number of values in the combined samples less than Z_j plus one half the number of values in the combined samples equal to Z_j

F_{ij} = the number of values in the i-th group which are less than Z_j plus one half the number of values in this group which are equal to Z_j

k = the number of samples (groups)

n_i = the number of observations in group i

x_{ij} = the j^{th} observation in the i^{th} group, and

$Z_j = Z_1, Z_2 ..., Z_L$ are the distinct values in the combined data set ordered from smallest to largest (L is less than n if there are tied observations)

(3) Epps-Singleton test:

Epps-Singleton test was used to compare the binding energies distributions of HAB and LAB. It returns a p-value that signifies the probability of falsely rejecting the null hypothesis (H_o) that both HAB and LAB are taken from the same population.

The Authenticator module supports the graphical plots, such as overlapping density plots and Empirical Cumulative Distribution Function (ECDF), to visualize the distributions. In case either the Authenticator module fails to recognize the default cutoff as an optimal threshold for the distribution split or the user decides that the resulting distributions fail to obtain the statistical inference under any threshold of their choice, the module is also incorporated with the other options for providing negative datasets, such as

decoys (Gcoupler inbuilt function using Chem module of RDKit Python package (Landrum, n.d.)), and user-supplied negative datasets.

The next module of Gcoupler, known as Generator, leverages the DeepChem (2.6.1) (Ramsundar et al. 2019) python library to build the Graph-Neural Networks supported classification models. The Generator takes the segregated HAB and LAB/decoys from the Authenticator module and uses these synthetic compounds to build prediction models. By default, the Generator module supports four different graph-based models, i.e., Graph Convolution Model (GCM), Graph Convolution Network (GCN) (Kipf and Welling 2016), Attentive FP (AFP) (Xiong et al. 2020), and Graph Attention Network (GAT) (Veličković et al. 2018), to generate classification models. Before the model-building step, the Generator module evaluates the class imbalance in the synthetic dataset and implements the upsampling techniques to counter it, if required. The Generator, by default, tests the pre-processed synthetic data on all four deep learning-based models under the default hyperparameters and returns models' performance parameter list, allowing users to select the best-performing model for the downstream refinement process. Users can either manually set up the hyperparameter tuning grid parameters or opt for Generator recommended grid range along with the fold count for k-fold cross-validation to tune the selected model around the best hyperparameters and also evaluate the stability of the said optimized parameters. In the final step, Generator builds the resultant model with the best-performing hyperparameters by utilizing the complete synthetic compound library (HAB + LAB/decoys) for training, allowing large-scale screening of the user query compounds by using SMILES information.

The last module of Gcoupler, the BioRanker module, performs post-prediction analysis for functional activity-based compound screening. Positively predicted compounds are selected using a stringent probability threshold or adaptive methods such as G-means and Youden’s J statistic, which optimize sensitivity and specificity. The selected compounds are projected into biological activity spaces (Chemistry, Targets, Networks, Cells, Clinics) by comparing their biological activity descriptor vectors with those of HABs using cosine similarity (Bertoni et al. 2021). A modified PageRank algorithm ranks compounds based on activity-specific scores, with support for multi-activity ranking to refine results based on user-defined biological properties, ensuring precise and context-relevant compound prioritization.

4.2.2 Runtime Analysis

To showcase the speed and efficiency of our *de novo* coupled Deep Learning approach against conventional docking, we performed a comparative study involving analysis of the same query compound dataset against a single target of interest through both approaches. A system with 125 GB RAM, 16 Threads, and 8 CPU cores with no upgradation to the software or hardware within the duration of this analysis was used as the platform for this comparison. To compare the time complexity between Gcoupler and classical docking procedures using AutoDock (Morris et al. 2009), we selected human alpha-1A adrenergic receptors (ADRA1A) as our GPCR of interest. Since the experimentally elucidated structure of ADRA1A is unknown, we used AlphaFold (Jumper et al. 2021) predicted structure as our starting point. The ligand binding cavity at the extracellular site was determined using the LigBuilder V3.0 (Y. Yuan, Pei, and Lai 2020) cavity function. Moreover, we obtained the ADRA1A ligand information from the ChEMBL database (Version 31) corresponding to the accession ID: CHEMBL229 (Gaulton et al. 2012). We obtained a total of 3684 bioactive compounds against ADRA1A; however, to ensure uniformity, we only selected those ligands (#933) that were annotated as “single protein format experiments”. Files preparation for AutoDock-based docking includes conversion of ligand SMILES to MOL2 using OpenBabel (2.4.1) (O’Boyle et al. 2011), followed by conversion of MOL2 to PDBQT via MGLtools (1.5.6) (Dallakyan, n.d.), and receptor PDB to PDBQT using OpenBabel (2.4.1) (O’Boyle et al. 2011). Noteworthily, the grid parameter calculations were manually performed, and docking was performed using AutoDock (4.2.6).

The Gcoupler workflow was performed as per the aforementioned steps. To ensure uniformity of the target cavity between both approaches, the cavity information along with the grid coordinate obtained during the initial steps of the AutoDock approach (Morris et al. 2009) was reused by the Synthesizer and the Authenticator module for generating *de novo* ligands and their molecular interaction-based classification, respectively. The Authenticator module selected an optimal binding energy cutoff of -8 kcal/mol; however, we opted for decoys for the negative dataset generation due to severe class imbalance. Under default hyperparameters, AttentiveFP showed better overall performance scores among the four deep learning-based models tested. We used the AttentiveFP model with the following hyperparameters: {'num_layers': 2, 'num_timesteps': 2, 'graph_feat_size': 200, 'dropout': 0} after assessment with 10-fold cross-validation. The final model was generated using the complete HAB compound library as the positive and the respective decoys dataset as a negative class for the model learning. Post-model generation, all 933 ligands were screened for their binding prediction probabilities.

4.2.3 Gcoupler Benchmarking

Batch effect in the different runs of Gcoupler for a particular cavity was performed using the standard Gcoupler Docker image. Of note, intracellular cavity 4 (IC4) of the *Ste2* protein of yeast was used for benchmarking. A total of 100 molecules were *in-silico* synthesized by the Synthesizer module of Gcoupler in an iterative manner. Post-generation, atom pair fingerprints (ChemmineR; R package) were calculated for the synthesized molecules from each run, and the data was visualized using Principal Component Analysis (R package), and pairwise comparison using Tanimoto Similarity (ChemmineR, R package). Benchmarking of Gcoupler-generated models in classifying experimentally validated was performed using GPCRs taken from the DUD-E dataset, alongside the information about the active ligands and their randomly selected number-matched decoys (Mysinger et al. 2012). Benchmarking of Gcoupler in identifying experimentally elucidated allosteric sites and their modulates was performed on the PDB complexes taken from RCSB PDB.

4.2.4 Sequence-Structural-Functional level analysis

All the available, non-redundant sixty-six human GPCR-Gα complexes were downloaded from the RCSB PDB database (https://www.rcsb.org/). The sequence level information for each of these GPCRs was obtained from the UniProt database (https://www.uniprot.org). Multiple Sequence Alignment was performed to study the sequence level conservation, using the FASTA sequences of the GPCRs as input for the MUltiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm (Edgar 2004). Of note, MUSCLE uses the PAM substitution matrix and sum-of-pairs (SP) score for alignment scoring. The MUSCLE alignment output in CLW format was then subsequently used to compute amino acid level conservation using the WebLogo tool (https://weblogo.berkeley.edu/logo.cgi). To identify conserved motifs native to the GPCR-Gα interface, residues at the GPCR-Gα interface for each GPCR-Gα complex were retrieved using EMBL-EBI developed PDBePISA tool and then subsequently mapped on the WebLogos (Battle, n.d.). To test the structural level conservation, the pairwise similarity of the PDBePISA provided GPCR-Gα interfaces (cavity) of all 66 GPCRs was achieved using the normalized Root Mean Square Deviation (RMSD), calculated using the PyMOL software (https://pymol.org/2/). The RMSD values were normalized by dividing the cavity RMSD with their respective whole protein RMSD. Finally, to analyze the functional level conservation of the GPCR-Gα interfaces, we used the Gcoupler's Synthesizer module to compute 50 synthetic ligands for each GPCR-Gα interface (cavity) and calculated their physicochemical properties (descriptors) using Mordred (Moriwaki et al. 2018). For each GPCR-Gα-protein interface, we computed a single vector representing the aggregated physicochemical property of all the synthesized compounds for the given GPCR cavity and measured their similarities

using the Cosine similarity. Of note, cosine similarity is a measure of similarity between two non-zero vectors defined in an inner product space.

$$cosine\ similarity\ = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

where, A_i and B_i are components of vectors A and B of size n respectively.

4.2.5 Molecular Dynamics Simulation

The Molecular Dynamics simulation was performed using either GROMACS (Abraham et al. 2015) or AMBER (Case et al. 2005) software suite. For Ste2p, the experimentally elucidated structure with PDB ID: 7ad3 was downloaded from the RCSB PDB database (Velazhahan et al. 2021). Since the available structure was provided for the dimer, PDB preprocessing steps were used to isolate a single chain (chain B) that refers to the GPCR. Embedding in the lipid bilayer representing yeast plasma membrane composition was performed using the CHARM-GUI (Jo et al. 2008), followed by the simulations using GROMACS (Abraham et al. 2015). Notably, we performed six rounds of equilibration with the membrane system and nine rounds of production runs of 5 ns each. We finally selected the last frame of the simulation step from the production run for further analysis. Protein-ligand complexes were prepared for the molecular dynamics simulations using AMBER software (Case et al. 2005), as discussed above. Notably, we performed all analyses using the Ste2p monomers. The grid box center and several grids were chosen appropriately to restrict the conformational sampling within the specific binding site. By default, the software estimates docking energies for 2,50,000 conformations of the ligands within the binding sites and ranks them to find the top 10 binding modes. The best binding modes for the ligands in these two sites were analyzed. The subsequent molecular dynamics simulations were performed for the ligands in these two binding sites, and the input configuration is based on the aforementioned molecular docking study results. The charges for the ligands were computed using density functional theory calculations at the B3LYP/6-31+G* level of theory. We used a general amber force field (GAFF) to describe the molecular interactions (Sprenger, Jaeger, and Pfaendtner 2015). For protein, we employed FF19SB force-field (C. Tian et al. 2020), and for water, TIP3P force-field was used (H. Wang and Yang 2018). The Ste2p-ligand complexes were embedded in about 23,000 water molecules. The minimization run, simulation in the constant volume ensemble, and simulation in the isothermal-isobaric ensemble were carried out. The time step for integrating the equation of motion was 2 fs. Followed by the equilibration run, a time scale simulation for 100 ns (three replicates for IC4 and IC5) and 550 ns (one replicate for IC4 cavity only) time scale was carried out. All the Ste2p-ligand complex simulations were performed using

Amber20 software (Case et al. 2005). Various structural properties, such as RMSD and RMSF were computed and analyzed. The binding free energies using molecular mechanics-Generalizer Born surface area approach (MM/GBSA) were computed for 1000 configurations picked up from the last 10 ns trajectory. The binding free energies computed using the molecular mechanics-generalized Born surface area approach are the quantitative measure of the binding preference of ligands to a specific target binding site. The values are obtained from the time average of different contributions such as van der Waals, electrostatic, polar, and non-polar solvation energies. When compared to molecular docking approaches, these binding free energies are obtained as an average over multiple protein-ligand configurations from MD simulations, and so are more reliable and accurate. In addition, many of the molecular docking simulations do not account for the conformational flexibility, but in the MM/GBSA calculations, the effects of protein-conformational flexibility, segmental motion, and cavity breathing are also accounted for, making results more reliable compared to molecular docking based binding affinities.

4.2.6 Molecular Docking (AutoDock)

YMDB metabolites were downloaded from the YMDB database (Version 2.0) (Jewison et al. 2012). Of note, we took the *metabolites structure file (SDF format) provided on the YMDB database to extract 2D information of the metabolites, and to gain missing metadata, we intersected these metabolites with the YMDB full database and obtained a total of 1264 metabolites. When we finally converted these compounds into 3D structures for docking or into graphs for testing data for the DeepChem-based model, a fraction of these compounds failed to convert and we finally selected a unique set of 1198 metabolites.* The file preparation for the docking protocol involved converting the YMDB SMILES (1D) to MOL2 (3D) format. This was achieved using the OpenBabel software (2.4.1) (O'Boyle et al. 2011). Further, the compounds from the MOL2 format were converted into dockable PDBQT format by following the MGLtools (1.5.6) ligand preparation procedure (Dallakyan, n.d.). Of note, the sub-processing steps include detecting the ligand center, selecting single bonds as torsion/rotatable bonds, etc. As discussed above, the receptor PDBs were also converted into PDBQT format using the OpenBabel software (O'Boyle et al. 2011). AutoDock (4.2.6) (Morris et al. 2009) was used for molecular docking by implementing a Genetic algorithm with the following parameters (population size = 150, maximum number of evaluations = 2,50,00,000, and maximum number of generations = 27,000). Notably, ten complexes were generated per YMDB compound. The docking results, i.e., extracting the lowest binding energy from each docked protein-ligand complex, were analyzed using the MGLtools (1.5.6) (Dallakyan, n.d.). Notably, the ΔG values were calculated on the pre-simulated stable docked structure. In the case of blind docking, the grid was made across the complete Ste2p structure and was docked using the aforementioned parameters with the YMDB metabolites.

4.2.7 Functional Enrichment Analysis

Functional enrichment analysis of the differentially enriched metabolites or the metabolites of interest was performed using MetaboAnalyst (https://www.metaboanalyst.ca/) web server (Chong et al. 2018). In the case of the YMDB compound library, the union list of the lead compounds for each target cavity (IC4 & IC5) was used as input. In the case of the untargeted metabolomics data, the list of differentially enriched metabolites between the α-factor treated and untreated conditions were used as input. The hypergeometric test was used for hypothesis testing and for providing subsequent p-values for each resultant enrichment pathway.

4.2.8 Protein-Protein Docking

To determine the impact of metabolite binding on GPCR-Gα and GPCR-α-factor interaction, we performed Protein-Protein docking using the HADDOCK (High Ambiguity Driven protein-protein DOCKing) webserver (van Zundert et al. 2016). HADDOCK is an innovative information-guided, flexible docking methodology for biomolecular complex modeling, uniquely utilizing ambiguous interaction restraints (AIRs) derived from identified or predicted protein interfaces, along with explicit distance restraints (e.g., MS cross-links) and diverse experimental data like NMR residual dipolar couplings, pseudo contact shifts, and cryo-EM maps, setting it apart from ab-initio docking approaches. In the case of Ste2p-miniGpa1p or Ste2p–α-factor docking, the PDB chains of both proteins were extracted from the PDB complex (accession id: 7ad3) (Velazhahan et al. 2021). Ste2p in docked complex with the respective metabolites were simulated via the AMBER software suite (Case et al. 2005). Post simulation, the most stable conformation of protein-metabolite was taken, and the metabolite was manually removed to obtain a metabolite-influenced stable Ste2p structure. HADDOCK was subsequently used for the protein-protein docking between wild Ste2p or metabolite-influenced bound stable Ste2p against miniGpa1-protein or α-factor (extracted from PDB ID: 7ad3). The optimal representative complexes were obtained from the HADDOCK results (van Zundert et al. 2016) and were used as input in PRODIGY (PROtein binDIng enerGY prediction) (Xue et al. 2016) for K_d and ΔG calculations. PRODIGY is a web service that provides support for the prediction of binding affinity in biological complexes. Notably, the ΔG values were calculated on the pre-simulated stable docked structures.

4.2.9 Yeast strains

BY4741 *(MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*), and *ste2Δ* (*MATa ste2Δ::kanMX his3Δ1 leu2Δ0 ura3Δ0 met15Δ0*) strains of *Saccharomyces cerevisiae* were used in all the experiments. For the mating assay, BY4742 strain *(MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) was also used. All the knockouts were obtained from Yeast Deletion Collection (*MATa xxx::kanMX his3Δ1 leu2Δ0 ura3Δ0 met15Δ0*). Unless mentioned,

the yeast was grown at 30°C at 200 rpm in yeast extract–peptone–dextrose (YPD) (1% yeast extract, 2% peptone, and 2% dextrose). Agar (1.5%) was additionally added to YPD to prepare plates.

4.2.10 Pre-loading of Yeast cells with metabolite

Yeast cells were cultured in YPD medium at 30°C, 200 rpm for 16 hours in primary and secondary cultures. Equal cell densities (5 µL) from secondary cultures were inoculated into 96-well plates containing 145 µL YPD with metabolites Coenzyme Q6 (CoQ6, 900150O, Avanti® Polar Lipids), Zymosterol (ZST, 700068P, Avanti® Polar Lipids), and Lanosterol (LST, L5768, Sigma-Aldrich) at 0.1 µM, 1 µM, and 10 µM concentrations. Plates were incubated for 24 hours at 30°C, 200 rpm, with multiple biological replicates. Ethanol-treated wells served as solvent controls. For site-directed *STE2* mutants, the mutants were grown in YPD for primary and secondary cultures, but the metabolite pre-loading was performed in YPGR instead of YPD to induce *Ste2* expression. After pre-loading further assays were performed.

4.2.11 Metabolomics

Wildtype (BY4741) and *ste2Δ* yeast strains were inoculated in the primary culture, followed by secondary culture, in YPD medium at 30°C at 150 rpm for 16 hours each. 1.5 mL of the secondary cultures were aliquoted into a 96-well deep well plate to ensure an equal number of cells for both strains. α-factor (T6901, Sigma-Aldrich) was added to the wells to attain final concentrations of 10 µM, 20 µM, 30 µM, 40 µM and 50 µM, with eight biological replicates each. DMSO (28580, SRL) (volume equivalent to the 30 µM α-factor sample) was added to the solvent control. DMSO or α-factor was not added to untreated (WT) and *ste2Δ* conditions. The plate was covered with a breathe-easy membrane (Z380059, Sigma-Aldrich) and incubated at 30°C at 150 rpm for 4 hours. After incubation, an aliquot of 50 µL was used to perform Propidium Iodide (PI) fluorometric assay as described in the later section.

Post Propidium Iodide-based cell viability assay, four biological replicates were created by merging two biological replicates. These were transferred to 1.5 mL microcentrifuge tubes and pelleted down at 6000 rpm for 5 minutes at room temperature (RT). The cell pellet was treated with zymolyase enzyme (L2524, Sigma-Aldrich) at a final concentration of 40 units/mL in 1X PBS and incubated at 30°C for 1 hour. Following this, the cells were washed with 1X PBS, and the cell pellet was stored at -80°C before metabolomics.

Metabolomics was performed on 1290 Infinity HPLC coupled with 6545 QTOF (Agilent Technologies, Santa Clara, CA, USA) equipped with dual ESI Source. The internal controls used in metabolomics include Jasmonic acid, Gibberellic acid, zeatin, Tryptophan 15N, Estrone D4, Arginine 15N, Thiamine D4, 15N5-8-OH-2-dG, 13C12 dityrosine, and 8-PGF2-D4. For the sample preparation, the protein

precipitation method was used. Briefly, 10 µL of a cocktail containing commercial stable isotopes internal standards were added, followed by the addition of chilled 750 µL methanol: water (4:1) and probe sonicated. Then, 450µL of chilled chloroform was added and vortexed for 5 min, followed by adding 150 µL of chilled water, vortexed, and kept the sample at -20 ºC for 30 min. Next, samples were centrifuged at 5000 rpm at 4°C for 10 min to separate the methanol and chloroform layers. Both layers were collected and dried at 30 °C using a vacuum concentrator and reconstituted in 100 µL methanol: water (1:1). Quality Control (QC) samples were injected between every five samples to check the drift in the instrument and to evaluate the signal repeatability. Pearson's rank order showed that QCs (n = 6) were strongly correlated($r>0.9$) with each other, suggesting variations were insignificant.

Data analysis of the metabolomics data includes peak normalization. Furthermore, the normalized peak intensities of all the samples' replicates were merged, and the metabolites with a constant value or with more than 50% missing values were omitted. The missing values were imputed using the feature-based k nearest neighbors (kNN) via Metaboanalyst webserver. The data were filtered based on the interquartile range (IQR). To find the list of Differentially Enriched Metabolites (DEMs), $\log_2$ fold change ($\log_2$FC) values were calculated for each sample with respect to the solvent control, along with its statistical significance student's t-test. The metabolites with $|\log_2 FC \geq 1|$ and p-value < 0.05 were considered DEMs. Pathway Over Representation Analysis (ORA) was performed using MetaboAnalyst (Chong et al. 2018). Note ORA is a computational method used in metabolomics research to gain insights into the biological significance of a set of metabolites that exhibit changes under different experimental conditions or in response to a perturbation or those belonging to a particular group. This analysis is conducted using statistical tests such as hypergeometric tests or Fisher's exact tests to determine if the observed number of metabolites in a given pathway is significantly higher than what would be expected based on the background distribution of metabolites.

4.2.12 Genetic Screening

Fifty-three knockout strains from Yeast Deletion Collection, along with WT and *ste2Δ* strains, were grown in primary culture, followed by secondary culture, in YPD medium at 30°C at 150 rpm for 16 hours each. Ensuring an equal number of cells in all the strains, 5 µL of secondary culture was inoculated into 96-well plates containing 145 µL of YPD and grown for 16 hours at 30°C at 200 rpm with eight biological replicates. Afterward, the α-factor was added to half the replicates at a final concentration of 30 µM. An equal volume of DMSO was added to the other half to serve as solvent control. The plates were incubated for another 4 hours. After incubation, a 50 µL aliquot was mixed with 50 µL of staining solution (Propidium Iodide (PI; 11195, SRL) in 1x PBS, freshly prepared) to attain a final PI concentration of 5 µg/mL in 96-well black plates. The plates were incubated in the dark for 15 minutes.

Heat-killed (HK) cells were used as a positive control. The fluorescence was measured using Biotek Synergy HTX multi-mode reader at excitation and emission wavelengths of 530/25 nm and 590/25 nm, respectively. Another 50 µL aliquot from the original plates was used to measure OD_{600} after 1:2 dilution with 1X PBS. The fluorescence values were first normalized for data analysis with the blank normalized OD_{600} values of the respective well. These normalized fluorescence values were subjected to two more rounds of normalization, first with unstained cells followed by heat-killed cells. Following this, percentage fold change was calculated for the treated group with respect to the untreated group. The p-value was computed using a one-sample student's t-test.

4.2.14 Cardiomyocytes Hypertrophy Models

Human AC16 cardiomyocytes were cultured in DMEM-F12 (Thermo Scientific) with 12.5% fetal bovine serum (FBS) at 37°C and 5% CO2. Cells were seeded in a 24-well plate for size measurements, treated after 24 hours with metabolites (CoQ6, ZST, LST, FST, CoQ10) at 2.5 µM, and incubated overnight with 1% FBS. The medium was refreshed with fresh metabolites and isoproterenol (25 µM) for 48 hours. Cells were washed with PBS, fixed with 4% paraformaldehyde, and stained with wheat germ agglutinin (Thermo Scientific) and DAPI. Images were captured using a Leica DMI 6000 B microscope at 20X magnification, and cell area was measured using ImageJ. Neonatal rat cardiomyocytes were isolated from 1-3-day-old SD rat pups using Collagenase Type II. After heart explantation and digestion, the cells were centrifuged and pre-plated for 90 minutes to remove fibroblasts. The cardiomyocytes were seeded in a gelatin-coated 24-well plate, incubated overnight with 2.5 µM metabolites and 1% FBS, and then treated with metabolites (2.5 µM) and isoproterenol (10 µM) for 72 hours. Cells were fixed, stained with alpha-sarcomeric actinin and DAPI, and images were captured using a Leica DMI 6000 B at 20X magnification. Cell area was quantified using ImageJ. Additional details about methodology is available in **Supplementary Information**.

4.2.15 Statistical Analysis

Statistical analyses were performed using Past4 software or R-Programming. The Mann-Whitney U test was applied to compare medians between two distributions (non-parametric), while Student's t-test was used for pairwise comparisons of means. P-value correction was performed using the Bonferroni method when necessary. A significance threshold of 0.05 was set, with *, **, ***, and **** indicating p-values <0.05, <0.01, <0.001, and <0.0001, respectively.

4.2.16 Data Availability

The processed untargeted metabolomics data is provided as Supplementary Information. The raw RNA sequencing files are available at ArrayExpress under accession *E-MTAB-12992*.

4.2.17 Code and Software Availability

A Python package for Gcoupler is provided via pip https://test.pypi.org/project/Gcoupler/. A docker container pre-compiled with Gcoupler and all of its dependencies can be found at https://hub.docker.com/r/sanjayk741/gcoupler. The source code of Gcoupler is available on the project GitHub page: https://github.com/the-ahuja-lab/Gcoupler and also at Zenodo with DOI: 10.5281/zenodo.7835335, whereas the raw sequencing files can be accessed using DOI: 10.5281/zenodo.7834294.

4.3 Results

4.3.1 Gcoupler Architecture Overview

The customized creation of new target molecules, utilizing the topological, chemical, and physical properties of the target protein cavity, requires sophisticated neural networks to assimilate and represent this multidimensional data. Previous efforts employing deep learning methodologies, including Bicyclic Generative Adversarial Networks (GAN) (Skalic et al. 2019) and traditional Recurrent Neural Networks (RNN) (Xu, Ran, and Chen 2021), have demonstrated the efficacy of this strategy in pinpointing target hit molecules; nonetheless, as of now, comprehensive and versatile standalone solutions remain scarce, particularly for GPCRs.Identifying effective compounds for GPCRs is a formidable challenge, primarily due to the diversity among various GPCR classes and the complexities of the potential ligand chemical space (Hughes et al. 2011; Topiol 2018a). Presently, the standard methodology for identifying GPCR ligands, encompassing both orthosteric and allosteric sites, entails computationally intensive *in silico* techniques (such as docking) or experimental strategies (such as protein structure-based methods).These methodologies are technically demanding and require augmented resources. Consequently, to facilitate and expedite the generative design of ligands for specific cavities on the topological surface of GPCRs, we created Gcoupler and made it available to the community as a Python package and a Docker image. Gcoupler employs a comprehensive methodology that incorporates structure-based, cavity-dependent *de novo* ligand design, advanced statistical techniques, potent Graph Neural Networks, and bioactivity-based functional analysis. Gcoupler comprises four interlinked modules: Synthesizer, Authenticator, Generator, and BioRanker, collectively providing an intuitive workflow for ligand design, screening, and prioritization. The Synthesizer, the initial module of Gcoupler, accepts a protein structure in Protein Data Bank (PDB) format and detects potential cavities on the protein surface. The generation of cavity-dependent molecules is primarily influenced by the cavity's chemical composition and geometric constraints; thus, it is essential to choose the cavity for subsequent steps based on its chemical

characteristics (hydrophobicity/hydrophilicity) and functional significance (proximity to the active site or residue composition), among other factors.

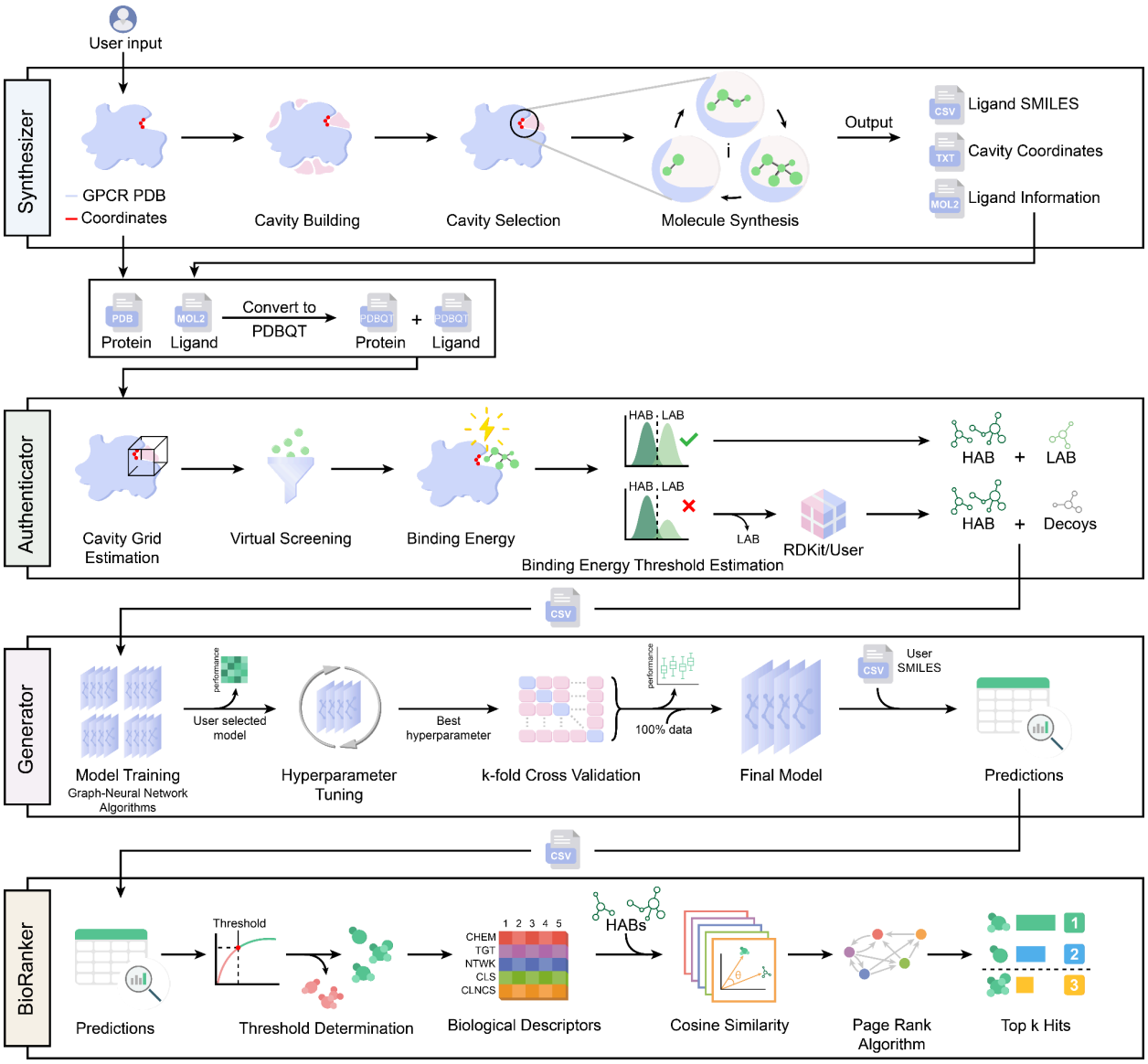


Figure 4.1 Gcoupler computational package framework

Schematic workflow depicting different modules of the Gcoupler package. Of note, Gcoupler possesses four major modules, i.e., Synthesizer, Authenticator, Generator and BioRanker.

Considering these factors, Gcoupler provides users the flexibility to choose between its predicted cavities based on user-supplied critical residues or by utilizing user-supplied cavity information (amino acids) through third-party software (e.g., Pocketome) (Hedderich et al. 2022). To improve user experience, Gcoupler calculates and presents all detected cavities alongside their druggability scores utilizing LigBuilder's V3 (Y. Yuan, Pei, and Lai 2020) cavity module. In summary, these druggability scores

evaluate solvent accessibility, cavity exposure or concealment, and identified pharmacophores and cavities, which are subsequently ranked according to this score. Following cavity selection, the Synthesizer module facilitates structure-based, cavity-dependent *de novo* ligand design to produce potential ligands and provides the SMILES data of the synthesized compounds, three-dimensional cavity coordinates, and additional files necessary for the subsequent Gcoupler modules **(Figure 4.1)**. The chemical composition of the *in silico* synthesized ligands by the Synthesizer module is affected by the three-dimensional cavity topology and its pharmacophoric composition. The Synthesizer module of Gcoupler utilizes LigBuilder V3 (Y. Yuan, Pei, and Lai 2020), which employs a genetic algorithm (GA) for *in silico* ligand synthesis. The fragment library, consisting of 177 unique molecular fragments in Mol2 format, facilitates the selection of various seed structures and extensions that optimally align with the cavity pharmacophores across multiple iterative cycles. Upon confirmation of a seed structure for each run, Gcoupler utilizes a hybrid methodology combining the Growing and Linking modes of the LigBuilder build module, facilitating the incremental incorporation of small fragments into the seed structure within the binding pocket of the target GPCR to synthesize ligands. Gcoupler, by default, generates 500 unique molecules, although this number can be customized by the user.

The Synthesizer module of Gcoupler augments the functionalities of LigBuilder V3 (Y. Yuan, Pei, and Lai 2020) for efficient ligand design in GPCRs; however, it is unable to screen user-defined chemical libraries and may occasionally suggest molecules that pose synthesis challenges through forward or retrosynthesis methods. Moreover, the synthetic compounds produced through this method frequently exhibit interaction potential across a range of binding affinities, necessitating further post-processing to isolate high-affinity binders from low-affinity binders. To mitigate this limitation, we created the second continuous module of Gcoupler, designated as Authenticator. This module processes output files from the Synthesizer module, performs downstream validation steps, and prepares results for the development of Deep Learning-based classification models. The Authenticator takes the input of protein 3D structures in PDB format, cavity coordinates, and all *in silico*-generated molecules from the synthesizer module. The Authenticator employs this information to categorize the synthesized compounds into HAB and LAB using a structure-based virtual screening method (AutoDock Vina) (Trott and Olson 2010) and statistically-supported hypothesis testing for distribution comparisons **(Figure 4.1)**. The Authenticator module provides the free binding energies of all generated compounds, which are subsequently categorized into HAB and LAB by the statistical submodule, while maintaining the optimal binding energy threshold and class balance. The Authenticator can utilize the Empirical Cumulative Distribution Function (ECDF) for comparing binding energy distributions of HABs and LABs, and it conducts the Kolmogorov–Smirnov test (Berger and Zhou 2014), Epps-Singleton test (Goerg and Kaiser 2009), and

Anderson-Darling test (Engmann and Cousineau 2011) for hypothesis testing. These tests are essential for assessing distributional disparities within datasets, facilitating the identification of unique patterns or outliers in binding energy distributions. The Kolmogorov–Smirnov test is adept at detecting variations in distribution shape and location, making it appropriate for comparing cumulative distribution functions. The Epps-Singleton test demonstrates resilience in detecting disparities in energy distributions, especially for small sample sizes and non-normal distributions. The Anderson-Darling Test improves the precision of classifying chemicals according to binding energies by evaluating the conformity of the distribution to theoretical models. These statistical tests yield significant insights into the distributional properties of binding energies, enabling informed decision-making in chemical segregation processes.

The Generator module, the fourth component of the Gcoupler, utilizes advanced GNN algorithms to develop classification models. It employs four distinct deep learning models: GCM, GCN, AFP, and GAT. These GNN algorithms are designed to extract features from the graph structure of the compounds produced by the Synthesizer and utilize them for the classification task by employing Authenticator-informed class information. Gcoupler offers pre-configured hyperparameter tuning to guarantee sufficient training, which is crucial for enhancing model performance. Upon selecting the optimal parameters and classification algorithm, Gcoupler further mitigates overfitting and enhances the accuracy of model performance estimation via k-fold cross-validation. Gcoupler utilizes three-fold cross-validation by default; however, users have the option to modify this parameter.

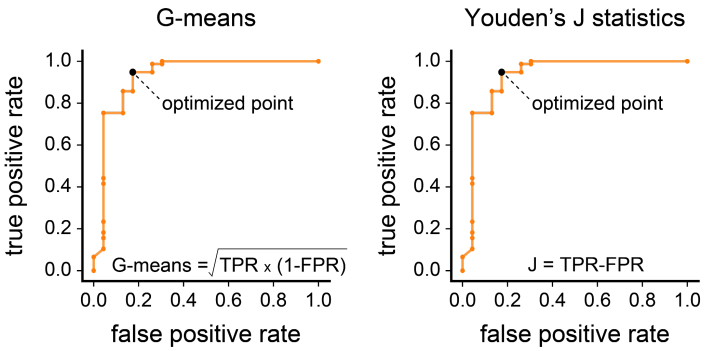


Figure 4.2 ROC Analysis Highlighting Optimal Probability Thresholds calculation

AUC-ROC curves depicting the optimal probability cutoff recommended by G-means and Youden J index algorithms, respectively. Notably, the optimal threshold is indicated by the black dot on the ROC curve.

Ultimately, BioRanker, the final module, prioritizes ligands utilizing statistical and bioactivity-based methodologies. The initial level ranking provided by BioRanker utilizes a statistical tool that incorporates two separate algorithms, G-means and Youden's J statistics, to aid users in determining the ideal probability threshold, thus enhancing the selection of high-confidence hit compounds **(Figure 4.2)**.

Furthermore, bioactivity embeddings generated through Signaturizer (Bertoni et al. 2021) facilitate multi-activity-based ranking employing a modified PageRank algorithm. Collectively , the four modules of Gcoupler constitute an interconnected system that offers a streamlined yet highly adaptable computational workflow to (a) design novel drugs by utilizing cavity-specific physical, chemical, and geometric attributes and (b) concurrently screen extensive chemical libraries against the designated cavity.

4.3.2 Gcoupler: open-source, feature-rich workflow for Drug Design

Currently, only a limited number of methods utilize generative AI models for cavity or pocket-based drug design.

Tools	Drug design Type	Open Source	Code Availability	End-to-End Solution	Command line support
Gcoupler	SBDD+LBDD	Yes	Yes	Yes	Yes
Pocket Crafter	SBDD	Partially	Partially	No	Yes
DeepLigBuilder	SBDD	--	No	--	--
AutoDesigner	LBDD	No	No	--	Yes

Table 4.1 Comparative Summary of de novo Drug Discovery Frameworks

Comparison of de novo drug design tools. SBDD refers to Structure-Based Drug Design, and LBDD refers to Ligand-Based Drug Design.

Gcoupler is an innovative, intuitive tool that amalgamates LBDD and Structure-Based Drug Design (SBDD) methodologies. In contrast to other platforms, Gcoupler is a completely open-source, end-to-end solution for drug design and extensive chemical library screening. PocketCrafter provides SBDD methodologies for *de novo* drug design, yet it varies in several critical respects (L. Shen et al. 2024). Not all modules of PocketCrafter are open-source, necessitating licensed software for the preparation of the protein pocket tertiary structure (specifically, the Molecular Operating Environment (MOE) QuickPrep module), which leads to only partial source code accessibility on GitHub. PocketCrafter lacks predictive model building modules and solely provides outputs of cavity-associated compounds, thereby constraining its applicability for extensive screenings **(Table 4.1)**. DeepLigBuilder, a recently introduced tool for SBDD, provides *de novo* drug design; however, its source code is not accessible for direct comparison **(Table 4.1)**. Moreover, akin to PocketCrafter, it is devoid of modules for constructing predictive models for extensive user-defined chemical screening. Finally, Schrodinger's pharmaceutical software suite utilizes a proprietary LBDD tool named AutoDesigner (Bos et al. 2022), which is not accessible to the public **(Table 4.1)**. The existing solutions are either partially or entirely closed source or are deficient in essential features found in Gcoupler.

Features	Gcoupler	SiteMap	Fpocket	CASTp	Pocket Query	DoGSite Scorer	Surfnet
Protein Structural input	✓	✓	✓	✓	✓	✓	✓
No additional input	✓	✓	✓	x	x	✓	✓
Grid based search	0.5 Å	1 Å	x	1.4 Å	x	✓	✓
All possible surface cavity	✓	✓	✓	✓	x	✓	✓
Cavity Residues	✓	✓	✓	✓	✓	✓	✓
Cavity coordinates	✓	x	x	x	x	x	x
Cavity druggability	✓	✓	x	x	✓	x	x
Cavity pharmacophore	✓	✓	x	x	✓	✓	x
Cavity additional properties	✓	✓	x	✓	✓	✓	x
Ligand complementation	✓	x	x	x	✓	x	x
Cavity specific ligand design	✓	x	x	x	✓	x	x
User defined # of ligands	✓	x	x	x	x	x	x

Table 4.2 Comparison of various receptor surface cavity detection tools

Consequently, we evaluated the individual modules of Gcoupler against alternative tools (Le Guilloux, Schmidtke, and Tuffery 2009; W. Tian et al. 2018; Koes and Camacho 2012; Volkamer et al. 2012; R. A. Laskowski 1995; Greener and Sternberg 2015; Huang et al. 2013; Song et al. 2017; Zha et al. 2023; Ngan et al. 2012; S. Wang et al. 2023), emphasizing the cavity detection module, as it is a pivotal step that directly impacts all subsequent modules **(Table 4.2; Table 4.3)**. A comparative analysis of the cavity detection module of Gcoupler with similar tools demonstrated superior precision in identifying cavity boundaries using higher resolution grids, along with distinctive features such as pharmacophore complementation and cavity-specific ligand design. It also includes options to synthesize a user-specified

number of compounds, crucial for downstream analysis by the subsequent modules **(Table 4.2; Table 4.3)**.

Features	Gcoupler	AlloPred	Allosite	Allosite Pro	Allo Reverse	FTMap	CavityPlus 2022
Protein Structural input	✓	✓	✓	✓	✓	✓	✓
No additional input	✓	x	✓	✓	x	✓	✓
Grid based search	0.5 Å	x	x	x	x	✓	x
All possible surface cavity	✓	x	x	x	✓	✓	✓
Cavity Residues	✓	✓	✓	✓	✓	✓	✓
Cavity coordinates	✓	x	x	✓	x	x	✓
Cavity druggability	✓	x	x	x	x	x	✓
Cavity pharmacophore	✓	x	x	x	x	x	✓
Cavity additional properties	✓	x	✓	✓	x	x	✓
Ligand complementation	✓	x	x	x	x	x	x
Cavity specific ligand design	✓	x	x	x	x	x	x
User defined # of ligands	✓	x	x	x	x	x	x

Table 4.3 Comparison of various protein allosteric cavity detection tools

Our analysis demonstrated that Gcoupler's cavity detection module surpasses competing tools in precision, distinctive features, and adaptability.

4.3.3 Benchmarking of Gcoupler Workflow for GPCR Ligand Classification

To assess the efficacy and reliability of Gcoupler, we conducted a series of stringent evaluations centered on its performance across various targets. Initially, we performed thorough performance evaluations of Gcoupler utilizing five different GPCRs (AA2AR, ADRB1, ADRB2, CXCR4, and DRD3), accompanied by their experimentally determined active ligands and corresponding number-matched decoys (Mysinger et al. 2012). The DUD-E datasets comprise five GPCRs, including data on their cavity coordinates, positive ligands, and decoys (https://dude.docking.org/subsets/gpcr). We utilized these five GPCRs as independent samples to assess various modules and sub-modules of Gcoupler. Initially, we assessed the efficacy of Synthesizer's cavity search algorithm in accurately identifying a specific orthosteric

ligand-binding site for a GPCR. Our findings indicate that Gcoupler accurately determined the coordinates of the orthosteric cavity, along with several additional allosteric sites for all five GPCRs, thereby validating the effectiveness of the employed *de novo* cavity identification algorithm **(Figure 4.3A)**. We subsequently inquired whether Gcoupler could also synthesize molecules analogous to the documented ligands for the respective orthosteric sites, based on the physical, chemical, and geometric characteristics of the cavities.

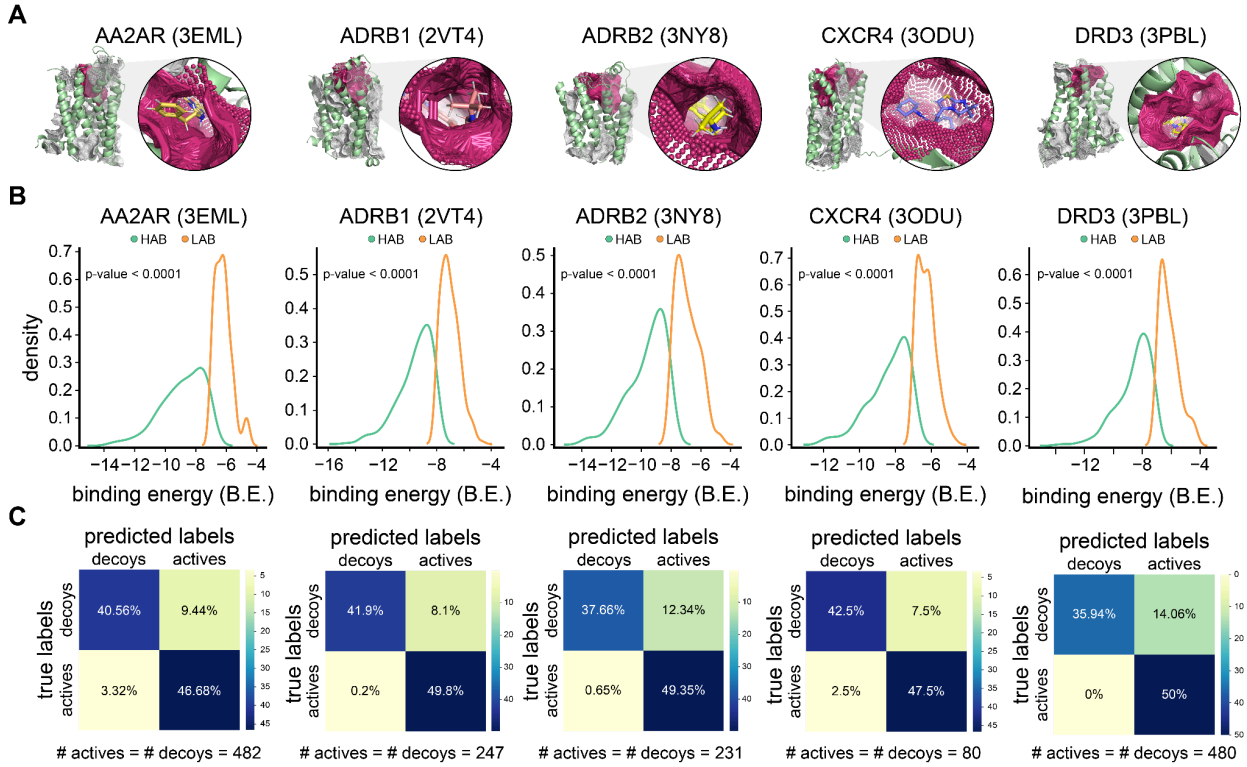


Figure 4.3 Gcoupler Benchmarking using experimentally validated orthosteric ligands of GPCRs

(A) Molecular representation depicting the indicated experimentally-validated orthosteric sites of the indicated GPCRs, with the zoom-in inlet on the right highlighting the ligand and the cavity topologies. (B) Overlapping density plots comparing the distributions of synthetic compounds predicted to target the indicated receptors into HAB and LAB using the Authenticator module of Gcoupler package. (C) Confusion matrices indicating the relative proportions of experimentally determined ligands and their respective decoys. Note, the number of active ligands and decoys is mentioned at the bottom.

To accomplish this, we conducted *de novo* ligand synthesis for the orthosteric cavities of each GPCR, resulting in an average of approximately 500 chemical compounds per GPCR. Subsequently, according to the Gcoupler default workflow, the Authenticator module performed a virtual screening of the newly synthesized compounds, categorizing them into HAB and LAB. Despite the Authenticator module's flexibility in determining an optimal threshold to differentiate between HAB and LAB, we opted for the

default cutoff of -7 kcal/mol for AA2AR, CXCR4, and DRD3. For ADRB1 and ADRB2, we established a threshold of -8 kcal/mol to reduce distribution overlap and prevent class imbalance, a crucial factor that could affect subsequent model generation using the Generator module **(Figure 4.3B)**. We utilized the integrated statistical test sub-module of Authenticator to statistically validate the distribution segregation of generated compounds according to the specified thresholds at a 95% confidence interval (p value < 0.0001) **(Figure 4.3B)**. Subsequently, the Generator module of the Gcoupler employs authenticator-classified synthetic molecules to construct graph-based models, consistently attaining high-performance metrics, with AUC-ROC values surpassing 0.95 in all instances **(Figure 4.3C)**.

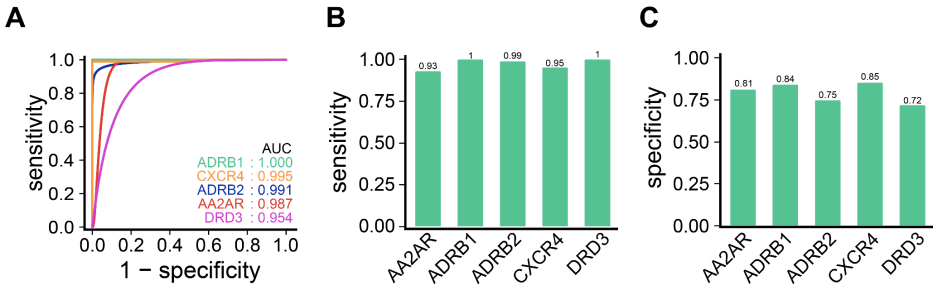


Figure 4.4 Predictive Performance of Gcoupler Models on GPCR Ligands

(A) AUC-ROC curves of the finally selected model for each of the indicated GPCRs. Note: Experimentally validated active ligands and decoys were used in the testing dataset. (B-C) Bar graphs depicting the sensitivities and specificities of the indicated GPCRs with experimentally validated active ligands and reported decoys.

Ultimately, the objective was to ascertain whether the generated models tailored to the orthosteric sites of each GPCR could effectively categorize the reported ligands and decoys into HAB and LAB classifications. To accomplish this, we mapped the identified cavity-specific ligands (orthosteric) and their decoys onto their corresponding models for each GPCR, noting satisfactory model efficacy in differentiating true positives from true negatives. This indicates Gcoupler's proficiency in accurately identifying cavity-specific HABs. Comparisons of the AUC-ROC values indicated elevated AUC values exceeding 0.95 in nearly all instances indicating strong model performance**(Figure 4.4A)**. We noted elevated sensitivities and specificities, highlighting the effectiveness of Gcoupler in differentiating between active ligands and decoys**(Figure 4.4B-C)**. These findings further substantiate that Gcoupler is a dependable and efficient instrument for detecting ligands at the orthosteric sites of GPCRs. Alongside assessing Gcoupler's efficacy for the orthosteric sites of GPCRs, we also confirmed its ability to detect allosteric sites and their associated ligands. Initially, we collected data regarding the experimentally confirmed GPCR-ligand complexes obtained from the PDB databank. We selected three GPCR-ligand complexes (β2AR-Cmpd-15PA, CCR2-CCR2-RA-[R], and CCR9-Vercirnon), recently identified by Shen and associates (S. Shen et al. 2023) as the principal data. For this validation, we extracted the ligands from the PDB files and performed the standard Gcoupler workflow using default parameters. In a manner

akin to the identification of the orthosteric site, Gcoupler proficiently identified the allosteric ligand-binding sites for all three GPCRs **(Figure 4.5A)**.

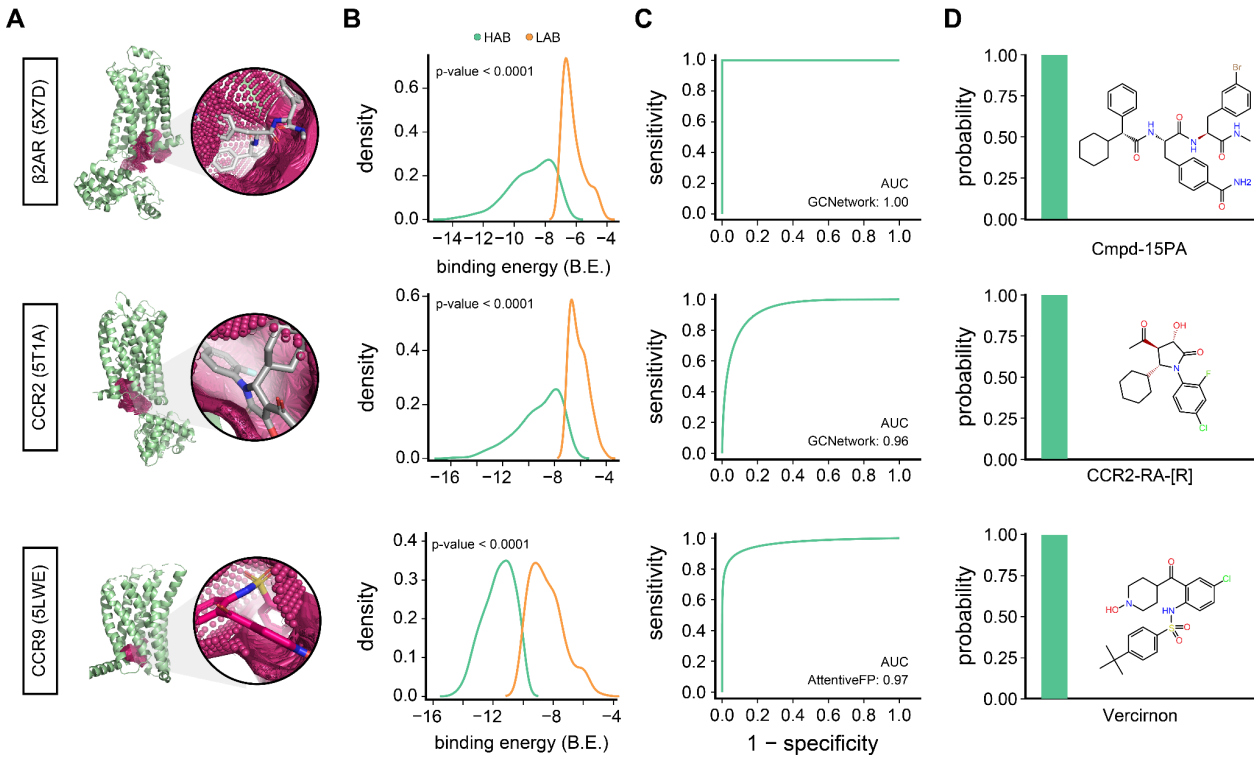


Figure 4.5 Gcoupler Benchmarking using experimentally validated allosteric ligands of GPCRs

(A) Molecular representation of the indicated GPCRs with highlighted experimentally validated intracellular allosteric cavity and ligand. (B) Overlapping density plots comparing the distributions of synthetic High-Affinity Binders (HAB) and Low-Affinity Binders (LAB) predicted to target the allosteric cavities of the indicated GPCRs using the Gcoupler package.. (C) The AUC-ROC curve indicates the model performance for classifying HABs and LABs. (D) Bar graphs indicating the prediction probabilities of the indicated ligands for each of the indicated experimentally validated ligands.

To further validate additional modules of Gcoupler, specifically the Authenticator and Generator, we categorized the synthetic compounds into HAB and LAB **(Figure 4.5B)** and assessed the performance metrics of the generated classification models for each GPCR utilizing the ground truth agonists/antagonists and non-agonists. The results exhibited exceptionally positive outcomes, with AUC-ROC values consistently exceeding 0.95 in all instances **(Figure 4.5C)**. Furthermore, the mapping of the experimentally validated ligands onto their corresponding models demonstrated Gcoupler's proficiency in accurately identifying hit compounds **(Figure 4.5D)**. This thorough assessment emphasizes Gcoupler's accuracy in detecting both orthosteric and allosteric binding sites, while also showcasing its potential as a multifaceted instrument for drug discovery across various GPCR targets.

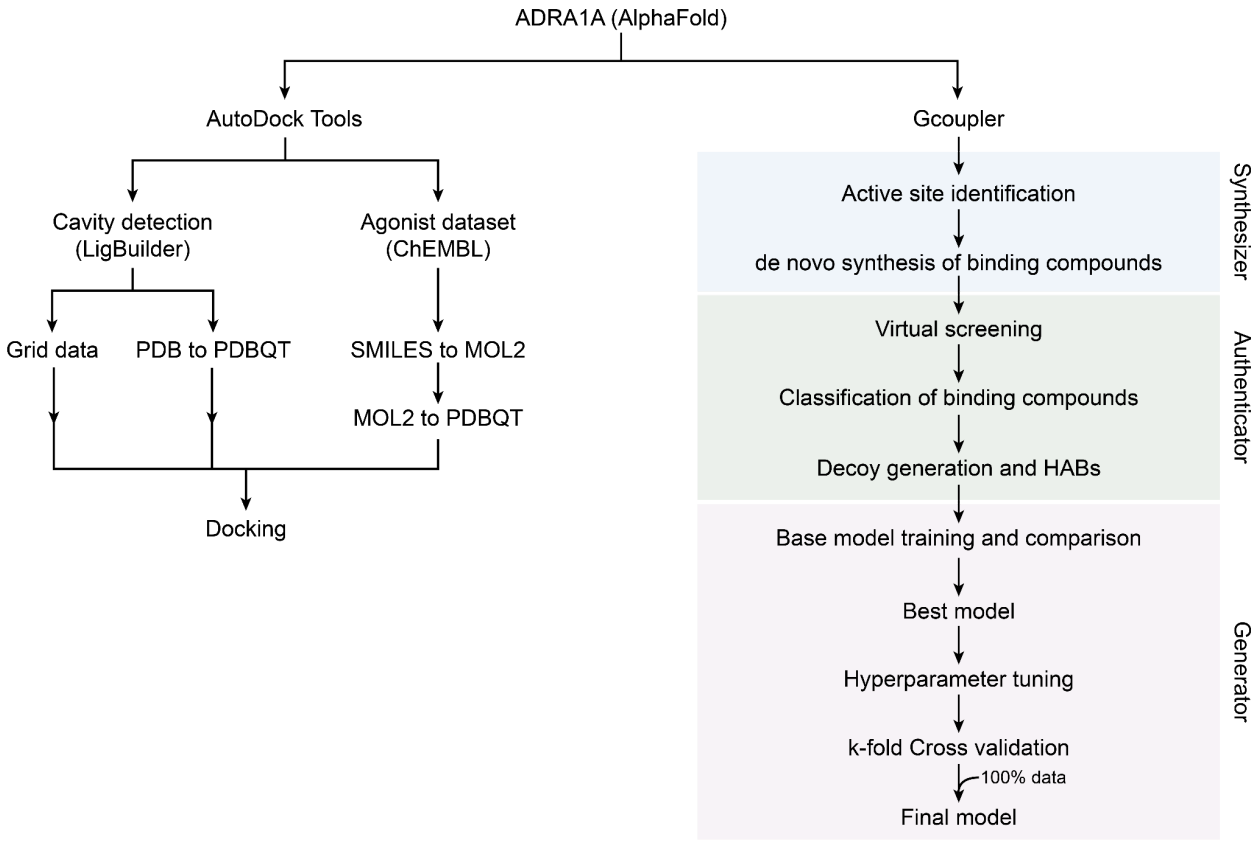


Figure 4.6 Comparative Workflow of Gcoupler and AutoDock for Ligand Evaluation

Flowchart depicting the entire workflow used to compute and compare the runtime using AutoDock and Gcoupler. Of note, the three major modules of Gcoupler, i.e., Synthesizer, Authenticator, and Generator, are indicated alongside their key processes.

Subsequently, to assess the efficacy of Gcoupler, we compared its runtime with that of the biophysics-based benchmark molecular docking method, AutoDock (Morris et al. 2009). To highlight the runtime efficiency, we initially employed the ChEMBL31 database (Gaulton et al. 2012) to pinpoint GPCRs with the greatest number of documented experimentally validated agonists. We chose the ADRA1A because it meets this criterion and comprises 993 agonists **(Figure 4.6; Figure 4.7A)**. Methodologically, we adhered to the standard procedures of AutoDock Tools for molecular docking, meticulously monitoring the execution time for each phase of the process until its conclusion **(Figure 4.6; Figure 4.7B)**. Simultaneously, we implemented the identical timestamp procedure for Gcoupler, encompassing its distinct module sub-functions **(Figure 4.6; Figure 4.7C)**. The comparative analysis of processing time between the two methodologies demonstrated that Gcoupler is 13.5 times more efficient than AutoDock **(Figure 4.7D)**.

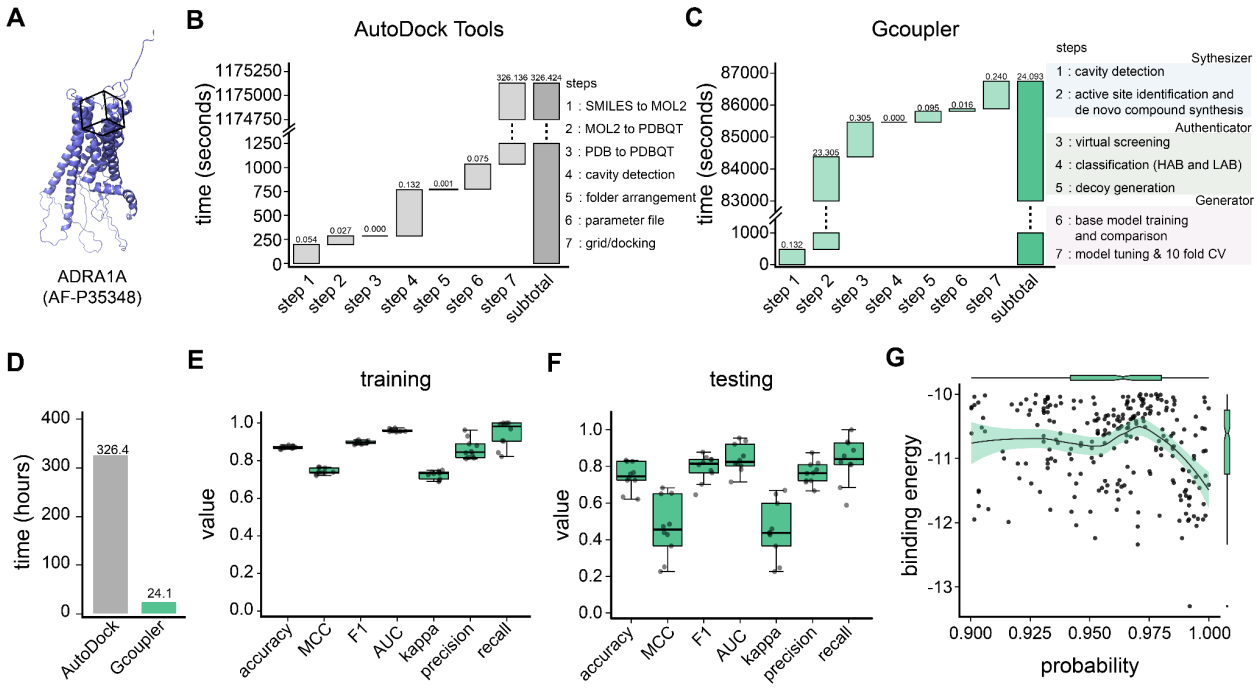


Figure 4.7 Evaluating Computational Efficiency: Gcoupler vs. AutoDock

(A) Diagram depicting the structure of ADRA1A predicted using Alphafold (ID: AF-P35348). (B) Waterfall chart depicting the timestamp information about the key steps involved in AutoDock. Of note, time consumed at the indicated steps is mentioned in seconds and hours along the y-axis and above bars, respectively. (C) Waterfall chart depicting the timestamp information about the key steps involved in Gcoupler. Of note, time consumed at the indicated steps is mentioned in seconds and hours along the y-axis and above bars, respectively. (D) Bar plot depicting the comparison between the total time consumed by the AutoDock and Gcoupler to predict the binding of experimentally validated ligands for ADRA1A (AF-P35348). (E) Boxplot depicting the performance metrics of the 10-fold cross-validation of the training data obtained using Gcoupler. (F) Boxplot depicting the performance metrics of the 10-fold cross-validation of the testing data obtained using Gcoupler. (G) Scatterplot depicting the relationship between AutoDock computed binding energies and Gcoupler computed prediction probabilities of the experimentally validated ligands of ADRA1A protein.

The notable decrease in runtime can be ascribed to the application of deep learning methodologies in the Generator module and possibly to the efficacy of AutoDock Vina (Liao et al. 2019; Gentile et al. 2020, 2022). Both methodologies produced similar predictions for the experimentally confirmed active compounds, distinguished by low binding energy and high predicted probability for all recognized agonists **(Figure 4.7E-G)**. The results indicate that Gcoupler provides an innovative and expedited method for ligand design, preserving predictive accuracy while significantly decreasing computational duration. This efficiency is especially beneficial for extensive screening and iterative drug design, underscoring Gcoupler's potential as a formidable asset in computational chemistry and drug discovery.

4.3.4 Gcoupler reveals endogenous, intracellular Ste2p allosteric modulators

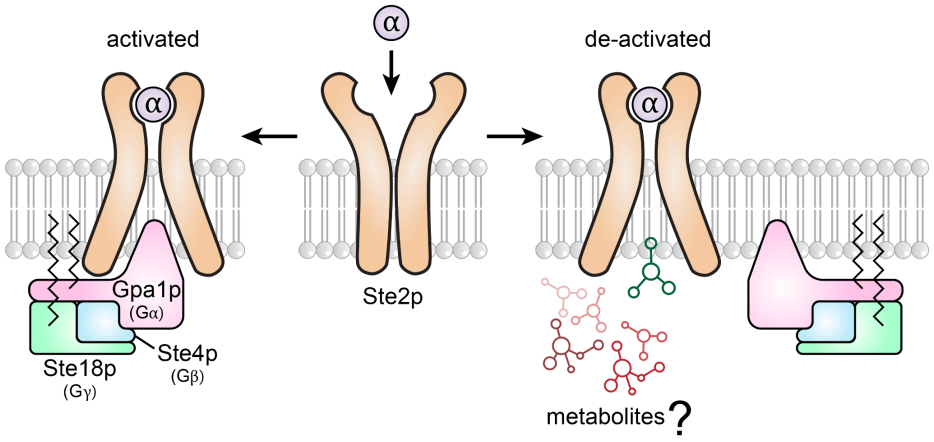


Figure 4.8 Proposed Model of Metabolite-Mediated Allosteric Modulation of GPCR signaling

Schematic diagram depicting the hypothesis that the intracellular yeast metabolites could allosterically modulate the GPCR-Gα-protein (Ste2p-Gpa1p) interface and, therefore, the programmed cell death pathway in yeast.

Subsequently, we employed Gcoupler to examine the α-factor-mediated programmed cell death/apoptosis in yeast, aiming to identify novel, previously uncharacterized intracellular allosteric modulating metabolites that confer natural resistance to PCD induction. The yeast mating pathway is among the most thoroughly investigated cellular pathways to date (N.-N. Zhang et al. 2006; Nakayama et al. 1988; Carmona-Gutierrez et al. 2010; Alvaro and Thorner 2016). To identify the endogenous metabolites that influence the GPCR-G-protein interface and elucidate their mechanisms of action, we employed the yeast mating receptor Ste2p, a pheromone-activated G protein-coupled receptor, as a model system. Our primary objective was to identify the intracellular metabolites that can interact with the Ste2p-Gpa1p interface and influence the signaling cascade **(Figure 4.8)**. We employed Gcoupler to identify these metabolites by screening them against the Yeast Metabolome Database (YMDB) (Jewison et al. 2012). We employed the cryo-EM structure of the *Ste2* protein (Velazhahan et al. 2021) and conducted a small-scale molecular dynamics simulation utilizing a yeast phospholipid composition-based lipid bilayer environment (Kaneko et al. 1976), subsequently applying the Gcoupler workflow to the simulated stable structure **(Figure 4.9A-C)**. The Synthesizer module of the Gcoupler package identified 17 potential surface cavities distributed across the extracellular, intracellular, and transmembrane regions of the *Ste2* protein **(Figure 4.9D-G)**. Meticulous examination of the structurally-supported Ste2p-Gpa1p interface identified two separate predicted cavities, designated as IC4 and IC5, which together encompass over 95% of the interface regions **(Figure 4.9H-I)**.

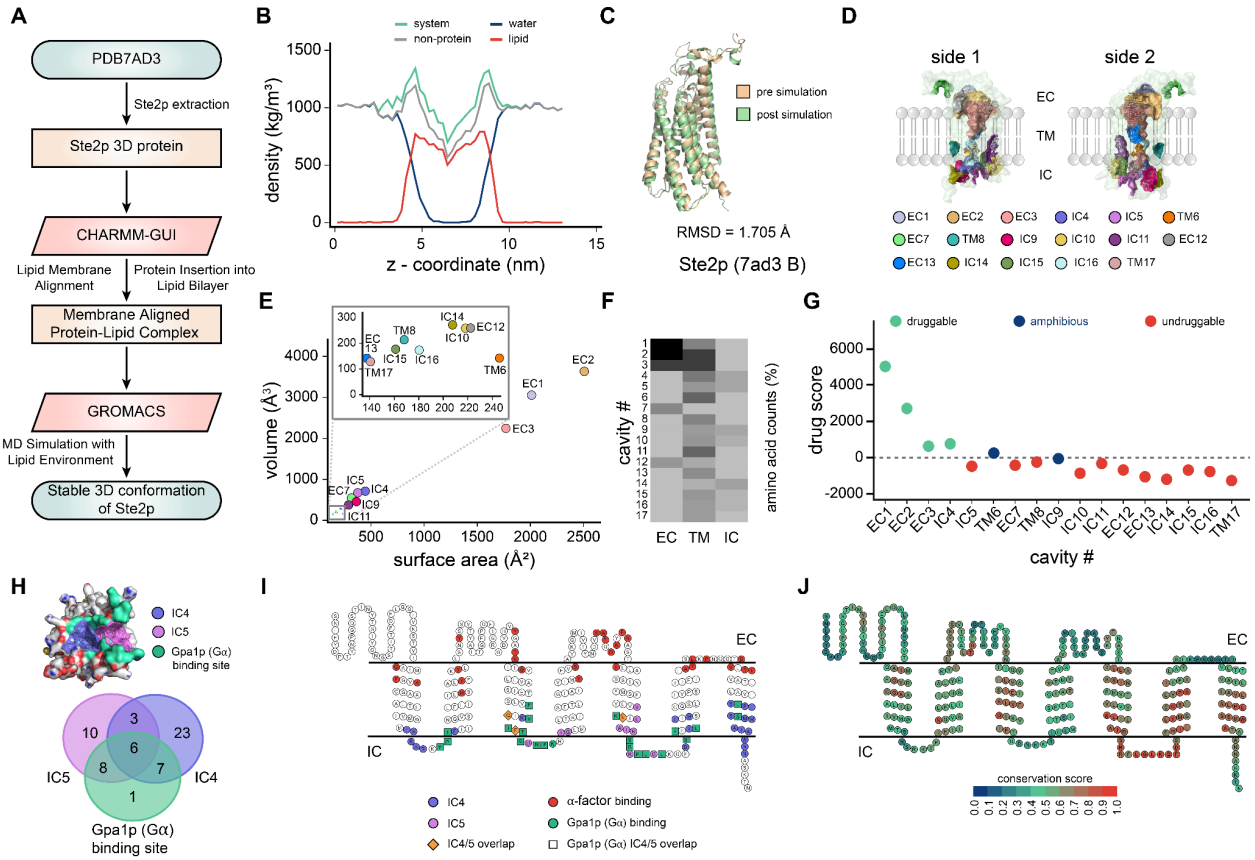


Figure 4.9 Ste2p Cavity Topology, and Interface Analysis Reveal Druggable Regions for Allosteric Modulation

(A) Flowchart depicting the key steps used for the molecular dynamics simulation of the Ste2p using CHARMM-GUI and GROMACS software. (B) Line plot depicting the changes in the density (kg/m³) along the z-axis of the three-dimensional Ste2p structure. (C) Overlapping Ste2p (GPCR) ribbon diagram depicting the structural similarity between the experimentally elucidated Ste2p structure before and after molecular dynamics simulation using GROMCS software. RMSD value is depicted at the bottom. (D) Schematic diagram depicting the topology of all the cavities identified using the Synthesizer module of the Gcoupler python package. Of note, the cavity nomenclature includes the cavity location, i.e., EC (extracellular), IC (intracellular), and TM (transmembrane), succeeded by numerical number. (E) Scatterplot depicting the relationship between the cavity volume and surface area in the indicated cavities of the Ste2 protein. (F) Heatmap depicting the percentage of amino acids of the indicated cavities residing in the EC, IC, and TM region of the Ste2 protein. (G) Scatterplot depicting the predicted drug score of the indicated cavities of the Ste2 protein. (H) Diagram depicting the three-dimensional view of the Ste2 protein, with highlighted Gα-protein binding site (Gpa1) and the Gcoupler intracellular cavities (IC4 and IC5). The bar plot at the bottom depicting the percentage overlap at the amino acid levels between the Gα-binding site and predicted IC4 and IC5. (I) Snake plot depicting the conservation of Ste2 protein at the amino acid levels. Of note, Ste2 proteins from 15-related yeast species were used for computing the conservation score. (J) Snake plot depicting the key amino acids of the Ste2 protein alongside their location within the indicated functional cavities.

We assessed the conservation of residue levels in the *Ste2* protein by utilizing *Ste2* protein sequences from 14 closely related yeast species. Residue level conservation scores indicated greater conservation in IC4 and IC5 **(Figure 4.9J)**.

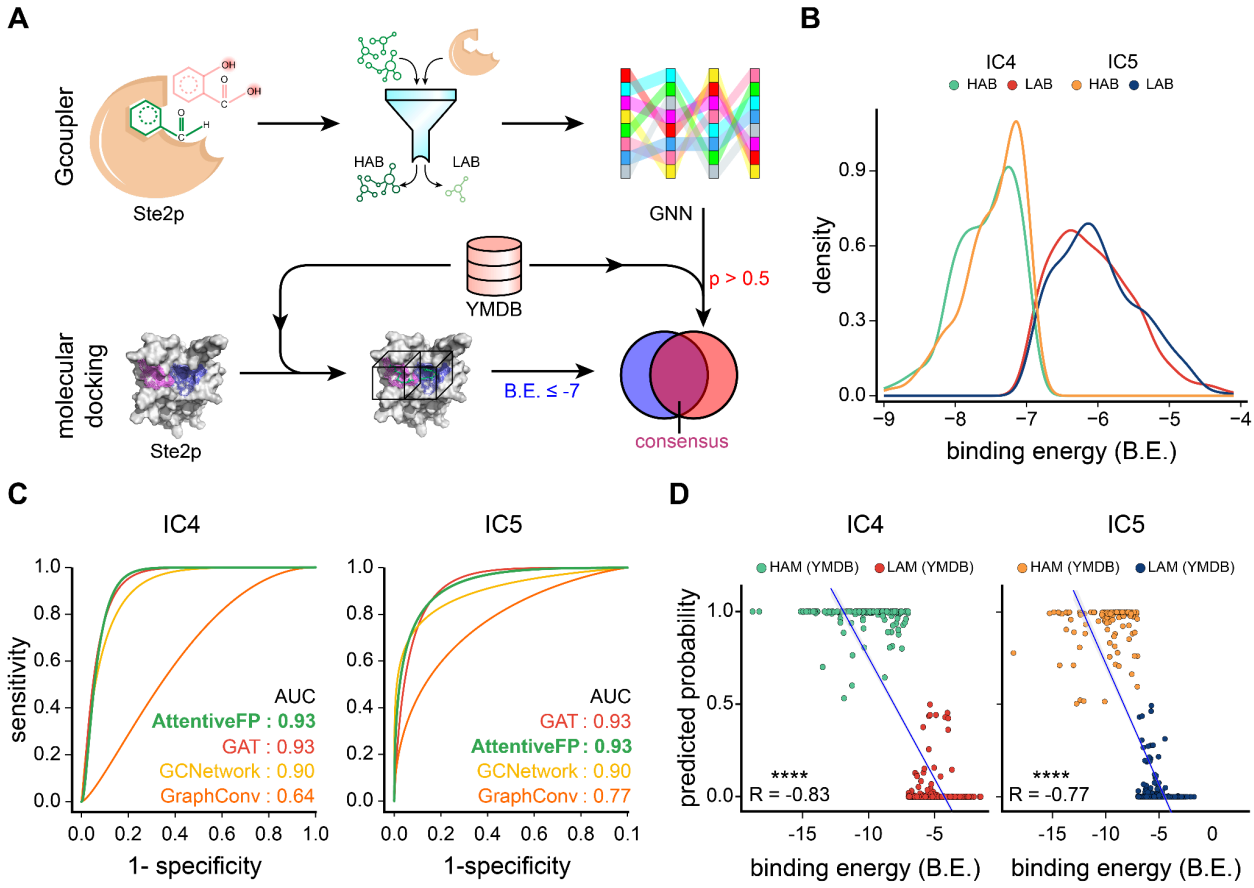


Figure 4.10 Integrative Prediction and Docking-Based Validation of Gcoupler-Predicted Ligands Targeting Ste2p *(A) Schematic representation of the overall workflow used to predict the endogenous intracellular allosteric modulators of Ste2 receptor using Gcoupler and molecular docking technique. Of note, Yeast Metabolome DataBase (YMDB) metabolites were used as query compounds. (B) Overlapping density plots depicting and comparing the distributions of synthetic compounds predicted to target the IC4 and IC5 of the Ste2 receptor using the Gcoupler package. Of note, the Authenticator module of Gcoupler segregated the synthesized compound for each cavity (IC4 or IC5) into HAB and LAB. (C) AUC (Area under the curve) plots representing the performance of the indicated models. Notably, the models were trained using the cavity-specific synthetic compounds generated using the Gcoupler package. (D) Scatter plots depicting the relationship (correlation) between the binding prediction probabilities using Gcoupler and binding free energies computed using molecular docking (AutoDock).*

We synthesized approximately 500 in-silico synthetic compounds for both IC4 and IC5 by utilizing the Synthesizer module of Gcoupler. Subsequently, we categorized the in-silico synthesized ligands for IC4 and IC5 into HABs and LABs utilizing the Authenticator module of Gcoupler. The estimated binding

energy threshold was established at -7 kcal/mol, a commonly recognized cutoff in virtual screening **(Figure 4.10A-B)**.

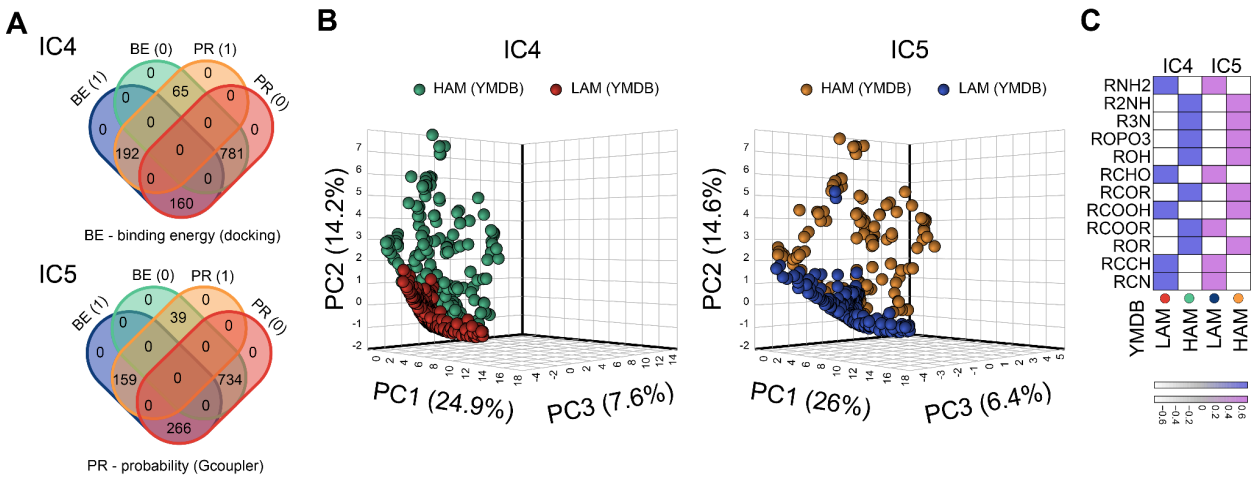


Figure 4.11 Chemical Diversity of Predicted Endogenous Allosteric Modulators of Ste2p

(A) Venn diagrams depicting the number of yeast metabolites from YMDB predicted to bind to the Ste2p-Gpa1p interface, predicted using Gcoupler and AutoDock. Of note, IC4 and IC5 represent intracellular cavities 4 and 5 of the Ste2 protein, respectively. (B) Principal Component Analysis depicting chemical heterogeneity of endogenous yeast metabolites predicted to bind to the intracellular cavities (IC4 and IC5) of the Ste2 protein. Notably, the compounds were segregated into High-Affinity Metabolites (HAM) and Low-Affinity Metabolites (LAM) by the Generator module of the Gcoupler. (C) Heatmap illustrating the relative enrichment of the indicated functional groups (RNH2: primary amine, R2NH: secondary amine, R3N: tertiary amine, ROPO3: monophosphate, ROH: alcohol, RCHO: aldehyde, RCOR: ketone, RCOOH: carboxylic acid, RCOOR: ester, ROR: ether, RCCH: terminal alkyne, RCN: nitrile) among the endogenous intracellular allosteric modulators (metabolites) of Ste2 receptor across IC4 and IC5.

An in-depth analysis of these classified synthetic compounds revealed a notable similarity between the HABs and LABs of the specified target cavities. The Generator module constructed classification models utilizing four distinct Graph Neural Network algorithms. The comparison of model performance metrics indicates that Attentive FP surpassed other algorithms for both cavities **(Figure 4.10C)**. We ultimately evaluated all known yeast metabolites, sourced from YMDB, using the optimal model (hyperparameter-tuned Attentive FP model) and identified metabolites likely to bind to the IC4 and IC5 regions of the Ste2p-Gpa1p interface, applying a binding probability threshold of > 0.5 **(Figure 4.10A)**. To further narrow down the lead metabolite list, we concurrently conducted standard molecular docking utilizing AutoDock with YMDB metabolites against IC4 and IC5 of Ste2p, ultimately selecting the consensus metabolites (binding energy < -7 kcal/mol and binding probability > 0.5) for subsequent analysis **(Figure 4.10A; Figure 4.11A)**. The consensus metabolites list was categorized into

High-Affinity Metabolites (HAM) and Low-Affinity Metabolites (LAM) according to the binding prediction probability (cutoff = 0.5).

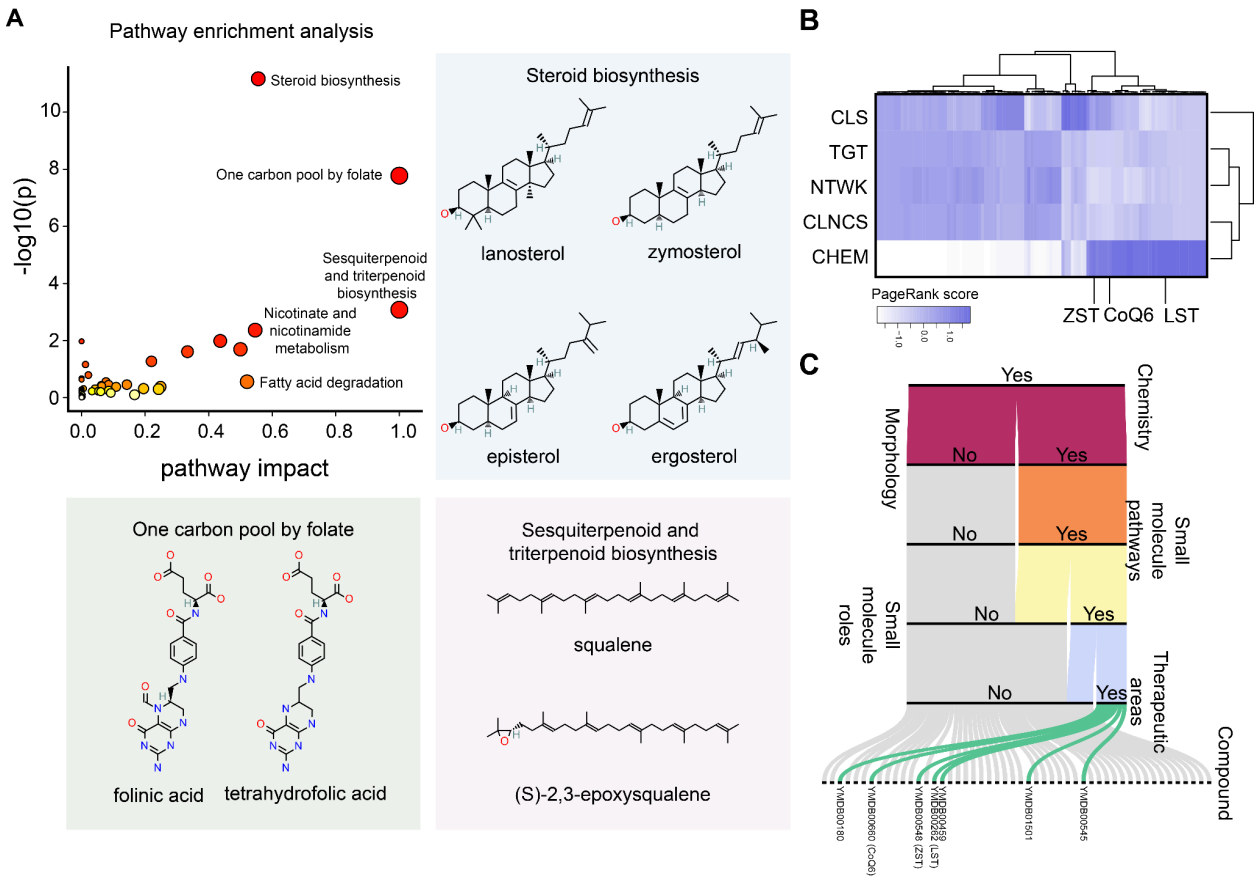


Figure 4.12 Activity-Space Screening of Predicted Allosteric Metabolites

(A) Scatterplot depicting the Pathway ORA results of the endogenous metabolites that were predicted to bind to the GPCR-Gα-protein (Ste2p-Gpa1p) interface using both Gcoupler and molecular docking. Structures of the top enriched endogenous metabolites are depicted on the right. (B) Heatmap depicting PageRank score of the selected metabolites. (C) Alluvial plot showing five level sub-activity spaces screening of the selected metabolites.

A comparative analysis of the binding prediction probabilities of Gcoupler and the binding energies from Autodock for HAB and LAB demonstrated a significant negative correlation, thereby corroborating the validity of our novel approach **(Figure 4.10D)**. Significantly, as anticipated, HAMs and LAMs exhibit unique atomic fingerprints, as demonstrated by the Principal Component Analysis **(Figure 4.11B)**. The functional group enrichment analysis of these metabolites indicates an abundance of R2NH, R3N, ROPO3, ROH, and ROR functional groups in the high-affinity binding metabolites for both cavities **(Figure 4.11C)**. To obtain pathway-level insights regarding the proposed endogenous intracellular allosteric modulators of Ste2p, we conducted Pathway-level ORA (Chong et al. 2018) and noted the

selective enrichment of metabolites associated with steroid, sesquiterpenoid, and triterpenoid biosynthesis, as well as the one-carbon pool by folate pathways **(Figure 4.12A)**. The BioRanker module identified sterols, specifically ZST, CoQ6, and LST, as prime candidates, demonstrating high prediction probabilities (>0.99) and structural resemblance to HABs **(Figure 4.12B-C)**.

4.3.5 Assessing the Specificity, Reproducibility, and Predictive Robustness of Gcoupler

To enhance confidence in the identified potential allosteric modulators through Gcoupler, we conducted comprehensive control analyses assessing the efficacy of the Synthesizer, Authenticator, and Generator modules, as well as executed blind docking of YMDB metabolites on Ste2p. To ascertain the cavity specificity of the Synthesizer module we conducted a rigorous assessment by comparing the chemical heterogeneity of 100 synthesized ligands, each derived from IC4 and IC5, with an extracellular cavity (EC1). Notably, EC1 does not contain any overlapping residues with IC4 or IC5 and exhibits unique pharmacophore characteristics **(Figure 4.13A)**. Subsequently, we calculated the atom pair fingerprints and illustrated the chemical heterogeneity in a low-dimensional space utilizing 2D and 3D PCA **(Figure 4.13B-C)**. The results indicate that the Synthesizer module of Gcoupler produced cavity-specific ligands by utilizing both the cavity topology (3D) and its composition (pharmacophore).

Additionally, we assessed the reproducibility of the Gcoupler executions. We chose the IC4 cavity of Ste2p and produced 100 in-silico compounds per iteration utilizing the Synthesizer module of the Gcoupler workflow. The visualization of chemical heterogeneity among compounds produced from various runs in a low-dimensional space, utilizing 2D/3D PCA and pairwise Tanimoto Similarity with atom pair fingerprints, indicates a heterogeneous and overlapping chemical composition among the synthesized ligands across all five runs **(Figure 4.13D-F)**. In case of the Authenticator module, we eliminated the class information (HAB and LAB labels) from the in-silico compounds created for the Ste2p's IC4 cavity and create a diverse pool of chemical compounds. Subsequently, we randomly partition the data into training and testing sets over five iterations and construct independent models.

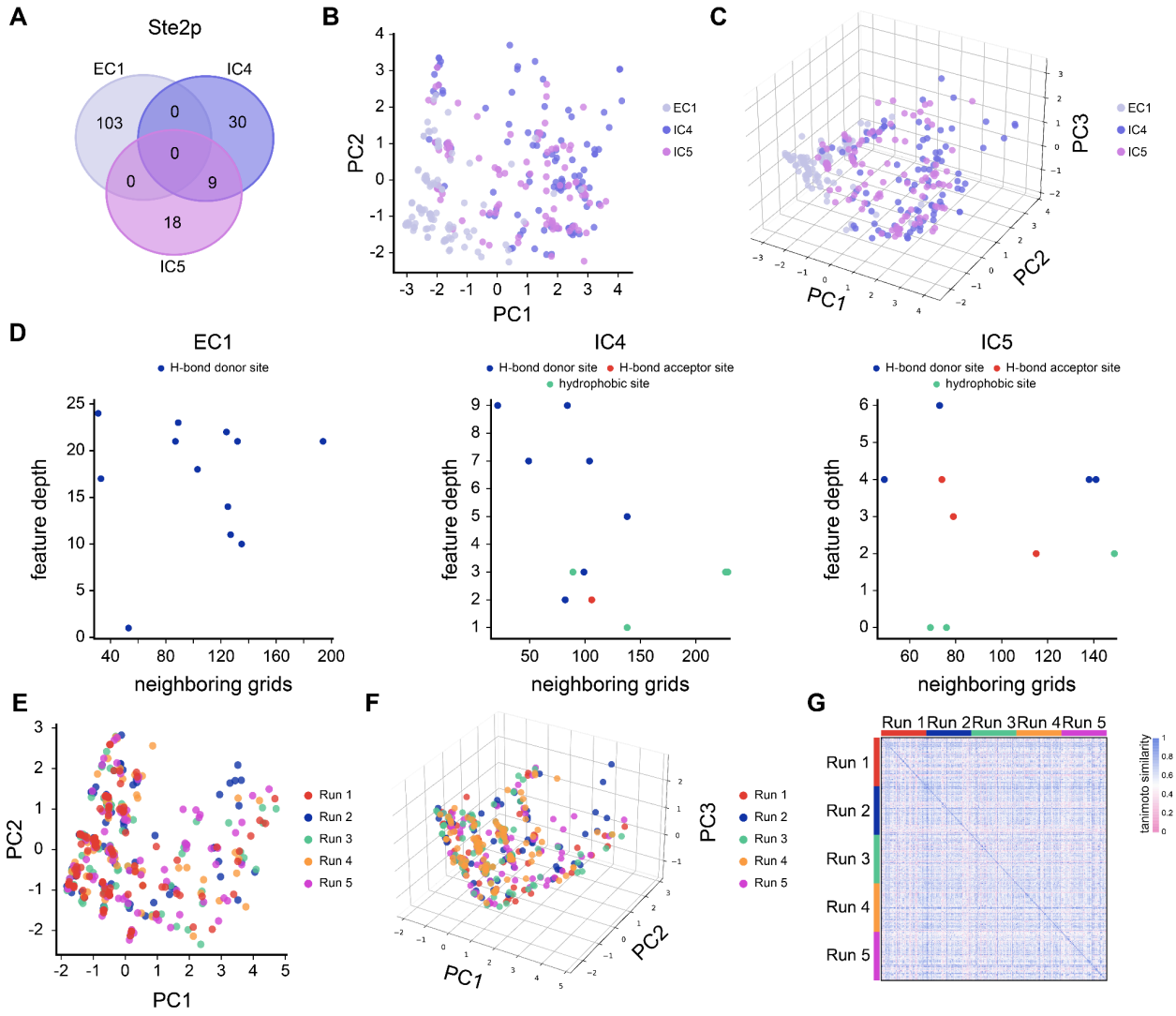


Figure 4.13 Cavity-Specific Ligand Diversity and Reproducibility Assessed by Gcoupler Synthesizer Module

(A) Venn diagram depicting the number of overlapping amino acids constituting the EC1, IC4, and IC5 of the Ste2p. (B) Two-dimensional and (C) three-dimensional Principal Component Analysis plots depicting the segregation of in silico synthesized ligands for the indicated cavities via Gcoupler. Of note, atom pair fingerprints were used as features for this analysis. (D) Scatter plot depicting the relationship between the depth of the pharmacophore features and the neighboring grids. (E) Two-dimensional and (F) three-dimensional Principal Component Analysis plots depicting the chemical properties of in silico synthesized ligands from five independent runs on the IC4 of the Ste2p using the Gcoupler Synthesizer module. Of note, atom pair fingerprints were used as features for this analysis. (G) Heatmap depicting the Tanimoto Similarities scores between the generated compounds for IC4 of Ste2p across five independent runs using Gcoupler.

Our findings indicate that, in contrast to the Authenticator-guided data splitting (HAB and LAB), random splitting yielded suboptimal model performance **(Figure 4.14A)**, thereby underscoring the robustness of

the Authenticator module. Furthermore, we assessed the influence of training data size on model performance. To accomplish this, we randomly selected 25%, 50%, 75%, and 100% of the in-silico synthesized compounds from the Ste2p IC4 cavity and constructed models utilizing default parameters. We assessed the model's performance by projecting metabolites from the YMDB database (Version 2.0). Our findings demonstrated a substantial enhancement in model performance correlating with the augmentation of data size **(Figure 4.14B-C)**.

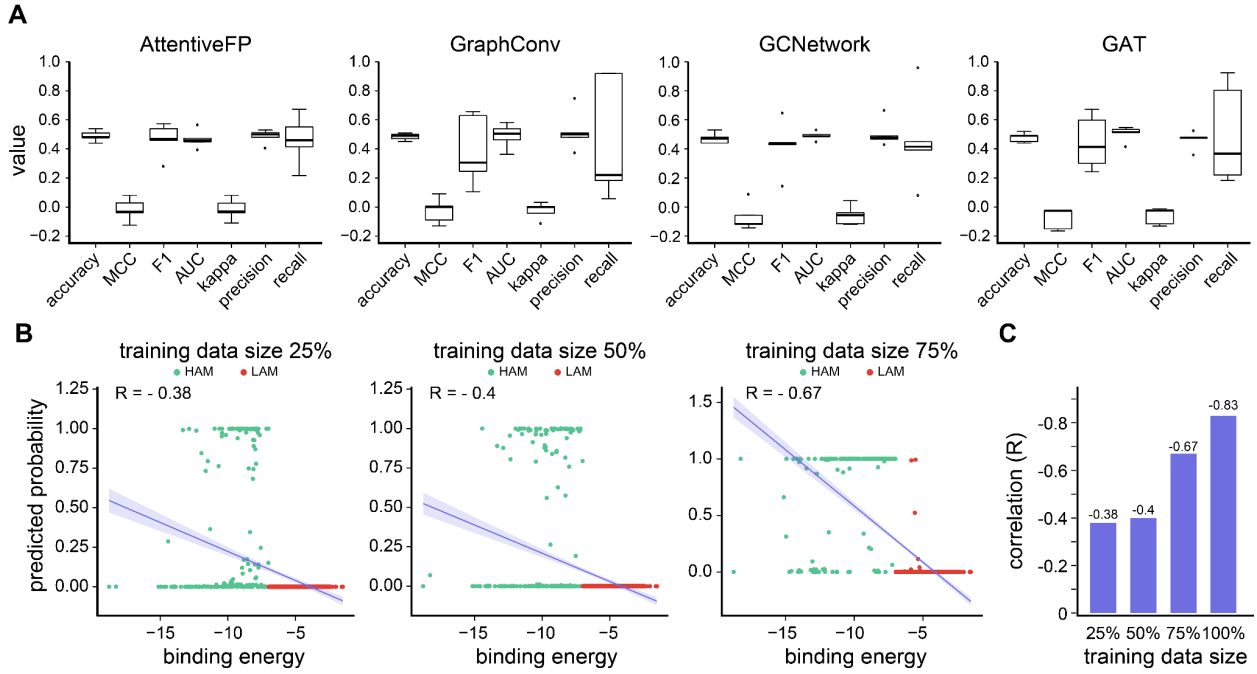


Figure 4.14 Impact of Training Data Size on Gcoupler Model Performance and Ligand Affinity Prediction

(A) Box plots depicting the performance parameters of the indicated models generated using Gcoupler against the IC4 cavity of Ste2p. Note the training and testing datasets were generated randomly (5 iterations) from the in silico synthesized ligands from Gcoupler. (B) Scatter plots depicting the segregation of HAM/LAM (indicated in green and red) identified using Gcoupler workflow with 100% training data. Of note, models trained on lesser training data size (25%, 50%, and 75% of HAB/LAB) severely failed to segregate HAM and LAM (along Y-axis). X-axis represents the binding affinity calculated using IC4-specific docking using AutoDock. (C) Bar plots depicting the correlation values obtained between the Gcoupler prediction probabilities and AutoDock computed binding energies in the indicated data size.

In the latter scenario, we conducted blind docking using AutoDock for Ste2p with YMDB metabolites and compared these findings with cavity-specific docking via AutoDock and Gcoupler predictions. As anticipated, unlike the cavity-specific Gcoupler and AutoDock, which demonstrated notable segregation of the HAM and LAM at a -7 kcal/mol Binding Energy (BE) threshold and a 0.5 Gcouplers' probability threshold, we did not observe any significant differences for the HAM **(Figure 4.15)**.

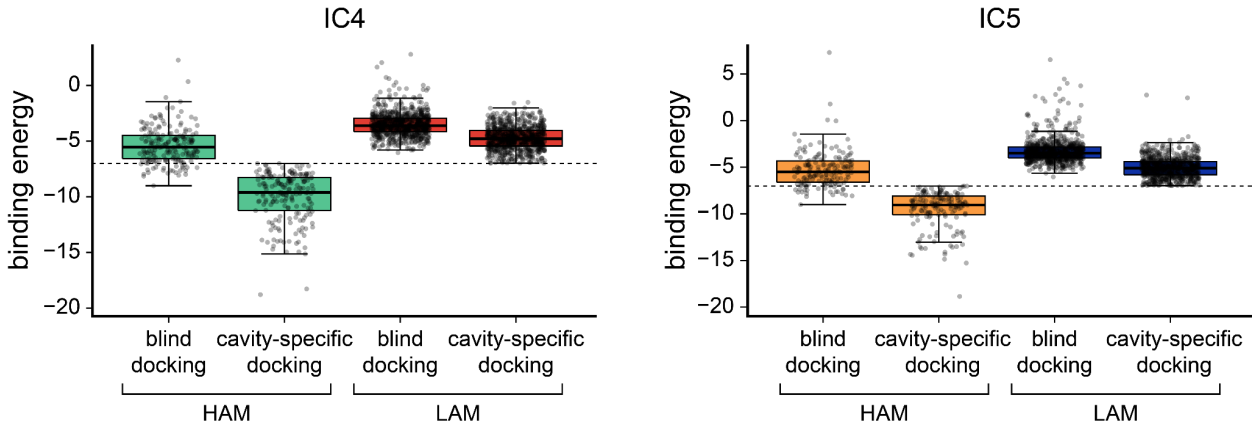


Figure 4.15 Validation of Predicted Affinities via Cavity-Specific and Global Docking Approaches

Box plots depicting the distributions for binding energies of the High-Affinity Metabolites and Low-Affinity Metabolites computed using cavity-specific docking and full docking via Autodock for IC4 and IC5 of the Ste2p.

Taken together, by utilizing Gcoupler, we identified cellular metabolites that may bind to the Ste2p-Gpa1p interface and influence pheromone-induced programmed cell death or mating response.

4.3.6 Uncovering Metabolic Pathways Influencing GPCR-Mediated PCD in Yeast

Gcoupler disclosed a collection of intracellular metabolites that may bind to the Ste2p-Gpa1p interface and modulate downstream signaling. We conducted a genetic screening of metabolic mutants to assess the influence of these metabolites on the *Ste2*-mediated signaling cascade. Initially, we mapped the anticipated allosteric-modulating metabolites to the biochemical pathway framework utilizing KEGG (Kanehisa and Goto 2000) and MetaCyc (Caspi et al. 2016), subsequently identifying the enzymes responsible for metabolite processing **(Figure 4.16A-B)**. These databases serve as extensive repositories of biological pathway information, including molecular interactions and metabolic networks, enabling sophisticated computational analysis. Of the 53 single metabolic mutants (+ *ste2Δ*) utilized for screening, only *Ste2* was documented in the KEGG pathway (ID: sce04011) for a modified mating response. Secondly, we conducted a comprehensive activity screening of individual metabolic mutants, positing that their deficiency results in the intracellular accumulation of specific metabolites, which can thus serve as a proxy to assess the interplay between metabolism (metabolites) and GPCR modulation **(Figure 4.16A)**. The activity of the signaling cascade was assessed through the α-factor-induced programmed cell death assay utilizing PI. Selected single metabolic mutants (MATa) were cultivated under optimal growth conditions until the late logarithmic phase (16 hours).

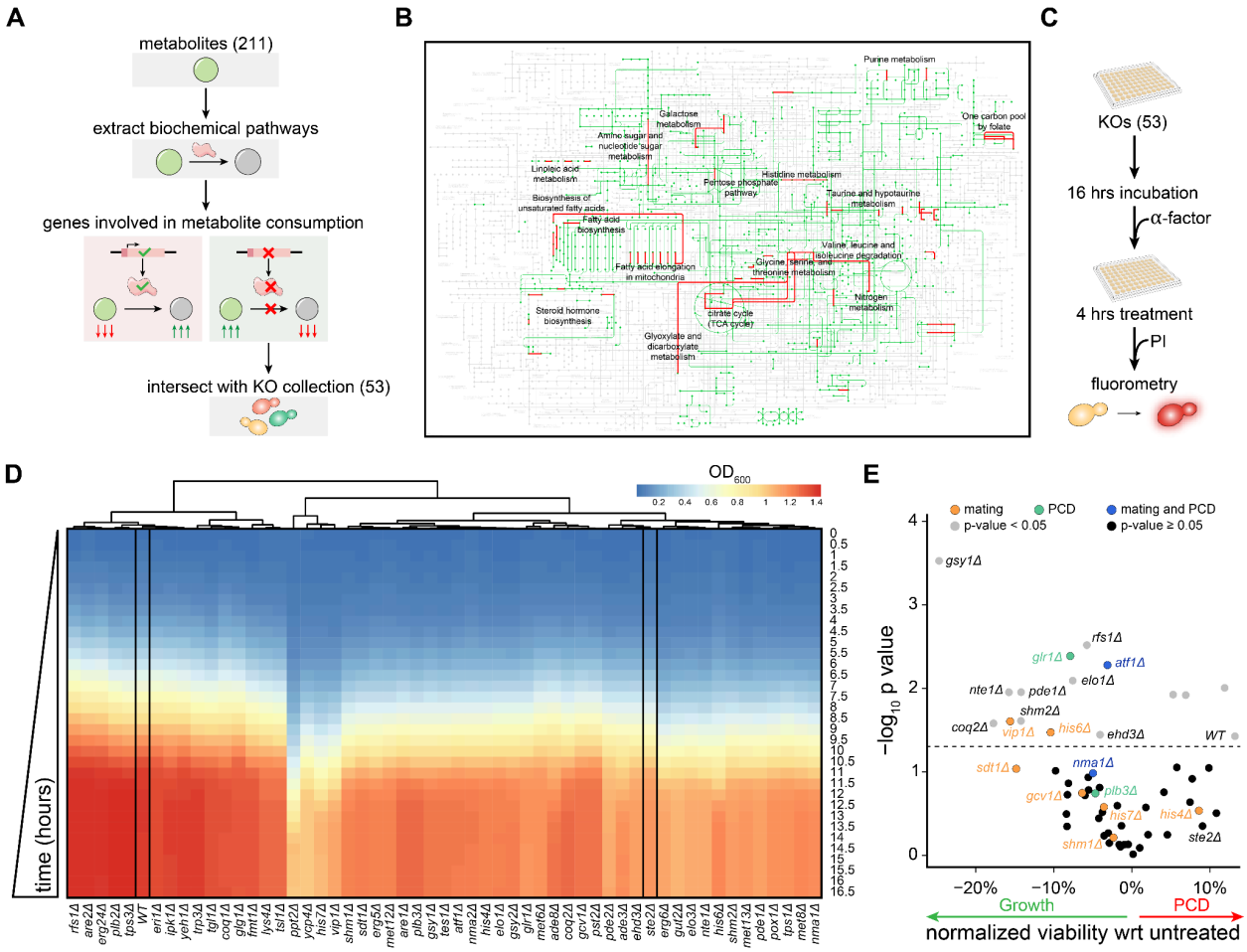


Figure 4.16 Validation of Gcoupler-Predicted Metabolic Modulators via Mutant Screening and Viability Assays

(A) Schematic diagram depicting the workflow opted to narrow down on the single metabolic gene mutants. (B) Screenshot of the central yeast metabolism from the Kyoto Encyclopedia of Genes and Genomes (KEGG) indicating the key metabolic pathways (highlighted in red) harboring metabolic intermediates predicted as high-affinity binders of GPCR-Gα-protein (Ste2p-Gpa1p) interface by Gcoupler. (C) Schematic diagram depicting the experimental design used to screen single metabolic gene mutants for α-factor-induced PCD. Cell viability was assessed using Propidium iodide-based fluorometric assay. (D) Heatmap depicting the growth curve profiles of the wild-type, single metabolic mutants, and ste2Δ under optimal growth conditions. (E) Scatter plot depicting the impact of α-factor stimuli on cellular viability, assessed using Propidium iodide-based fluorometric assay. The y-axis represents -log10(p-value) of the one-sample student's t-test between the normalized PI fluorescence of untreated and treated conditions. The x-axis represents the percentage inhibition or increase in cellular viability, estimated using a Propidium Iodide-based assay. The mutants reported to be involved in mating, PCD, or both are indicated in orange, green, and blue, respectively. The statistically non-significant mutants are indicated below the dashed line in black.

The intercomparison of growth profiles among these single metabolic mutants demonstrated diverse growth responses compared to the wild-type, with most mutants exhibiting a delayed growth response **(Figure 4.16D)**. Cells in the late log phase were subjected to α-factor treatment, and programmed cell death induction was assessed using the PI-based cell viability assay **(Figure 4.16C-D)**. Heat-killed cells served as the positive control (100% cell death). As anticipated, we noted a substantial elevation in PI fluorescence in the wild-type BY4741 yeast strains subjected to α-factor, signifying pheromone-induced programmed cell death, whereas no considerable mortality was detected in the *STE2* loss-of-function mutants. Notably, the majority of single metabolic mutants (94.4%) exhibited resistance to α-factor-induced cell death, while a subset demonstrated enhanced growth in the presence of α-factor, suggesting an interaction between central metabolism (or metabolites) and *Ste2* signaling. Our screening procedure identified several metabolic mutants exhibiting increased sensitivity to the α-factor **(Figure 4.16E)**. The results suggest that a substantial fraction of Gcoupler-predicted metabolites may directly or indirectly affect the *Ste2* signaling pathway, thereby establishing a connection between metabolism and *Ste2* signaling.

4.3.7 Metabolomic Evidence Linking Metabolites to α-Factor-Induced Cell Death

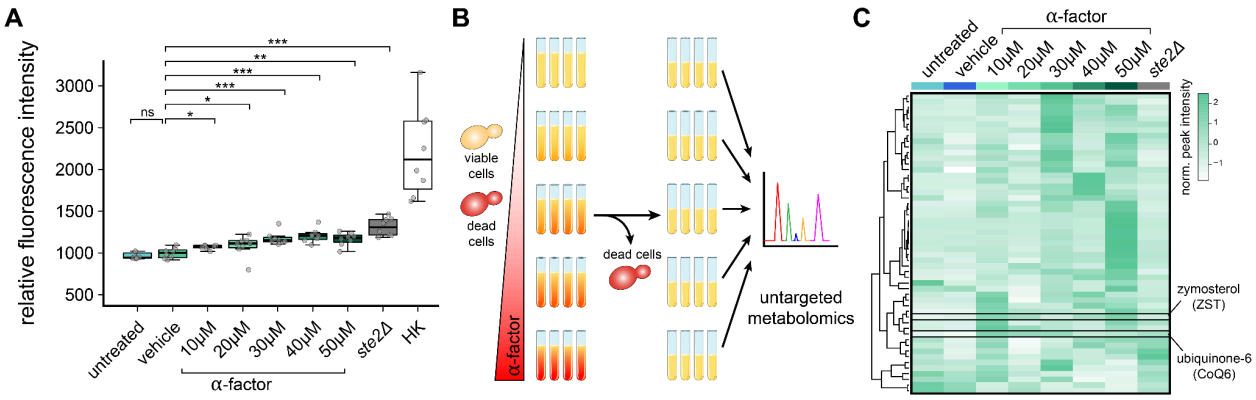


Figure 4.17 Untargeted Metabolomics Reveals Metabolic Shifts in Yeast Cells Surviving α-Factor-Induced PCD

(A) Heatmap depicting the relative enrichment of all the detected metabolites in the indicated condition. HK, untreated, and vehicle (DMSO-treated) were used as controls, whereas the increasing concentration of α-factor survived wild-type cells were used as test conditions. (B) Schematic diagram depicting the experimental workflow of the untargeted metabolomics experiment. Notably, the experiment involves the usage of increasing doses of α-factor to induce PCD and to enrich the survival cells selectively. (C) Heatmap depicting the relative enrichment/de-enrichment of differentially enriched metabolites in the indicated conditions. Of note, four biological replicates per condition were used in the untargeted metabolomics.

Subsequently, we employed a high-resolution metabolomics methodology to elucidate the metabolic pathways, or metabolomes, contributing to α-factor-induced programmed cell death resistance. To

accomplish this, we initially enriched the cells that withstood α-factor-induced programmed cell death at varying pheromone concentrations, referred to as survivors **(Figure 4.17A)**, and subsequently conducted high-resolution metabolomics to elucidate their metabolic profiles **(Figure 4.17B)**. We subjected yeast cells to escalating concentrations of α-factor to identify differentially enriched metabolites in the surviving population. Alongside the wild type, we analyzed the *Ste2* knockout. An impartial high-resolution metabolome analysis identified a distinct subset of metabolites that were differentially and significantly enriched in the surviving population **(Figure 4.17C)**.

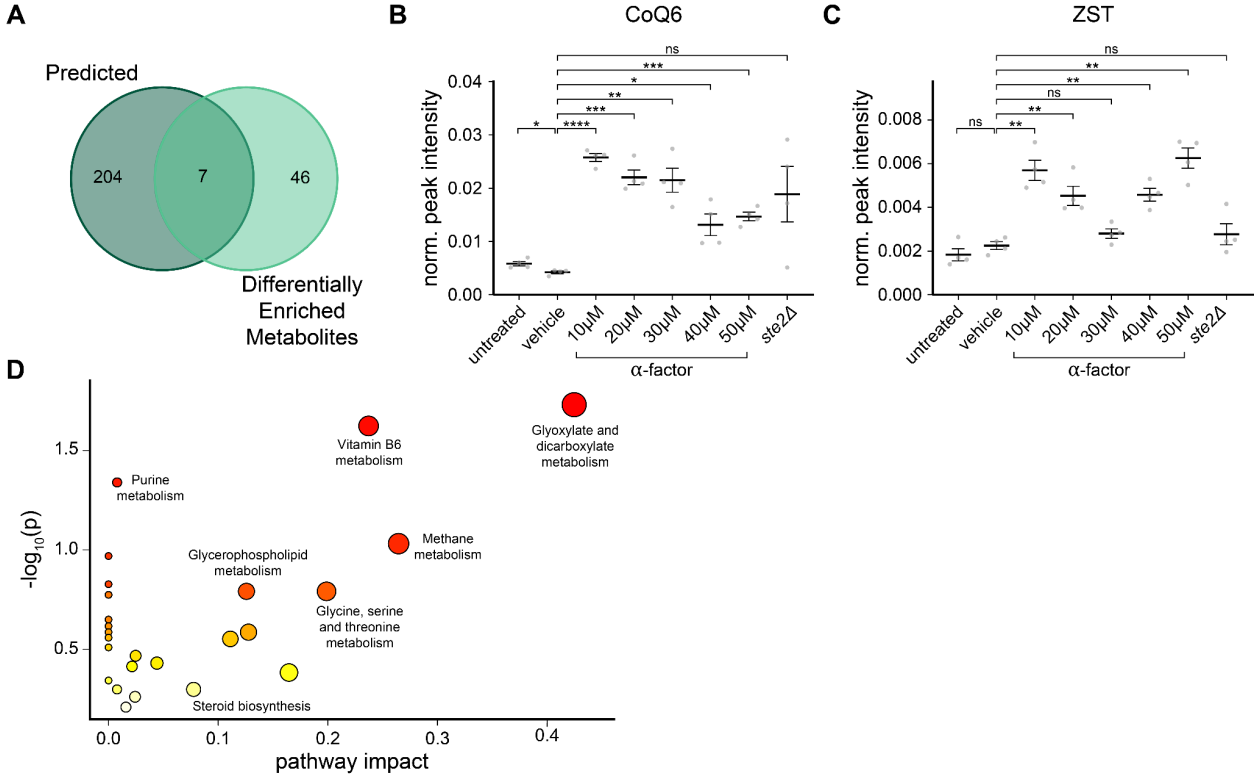


Figure 4.18 Experimental Confirmation and Pathway Enrichment of Predicted Intracellular Modulators of Ste2p *(A) Venn diagram depicting the overlap between the predicted endogenous intracellular allosteric modulators of Ste2p and DEMs identified using untargeted metabolomics. (B-C) Mean-whisker plot depicting the relative abundance of CoQ6 and ZST in the indicated conditions. Student's t-test was used to compute statistical significance. Asterisks indicate statistical significance, whereas ns represents non-significance. (D) Scatterplot depicting the Pathway ORA results of the differentially enriched metabolites (treated vs. vehicle control) identified using untargeted metabolomics.*

In a comparative analysis, we identified an intersection of seven metabolites between the Gcoupler prediction of Ste2p intracellular allosteric modulators and differentially enriched metabolites from the surviving cells. Notably, Ubiquinone-6 and Zymosterol exhibited significant enrichment in the surviving cells across all tested concentrations **(Figure 4.18A-C)**. Furthermore, ORA of the differentially enriched

metabolites indicates their participation in glyoxylate and dicarboxylate metabolism, purine metabolism, and vitamin B6 metabolism, among others **(Figure 4.18D)**.The findings from genetic screening of metabolic mutants and untargeted metabolomics of PCD-survived mutants suggest an interaction between central metabolism and *Ste2* signaling. These results identified two significant metabolites that may confer resistance to α-factor-induced PCD.

4.3.8 Evaluating Interface Stability of Gcoupler-Predicted Metabolites via MD Simulations

We subsequently examined whether the interaction of these metabolites at the Ste2p-Gpa1p interface is stable or transient.

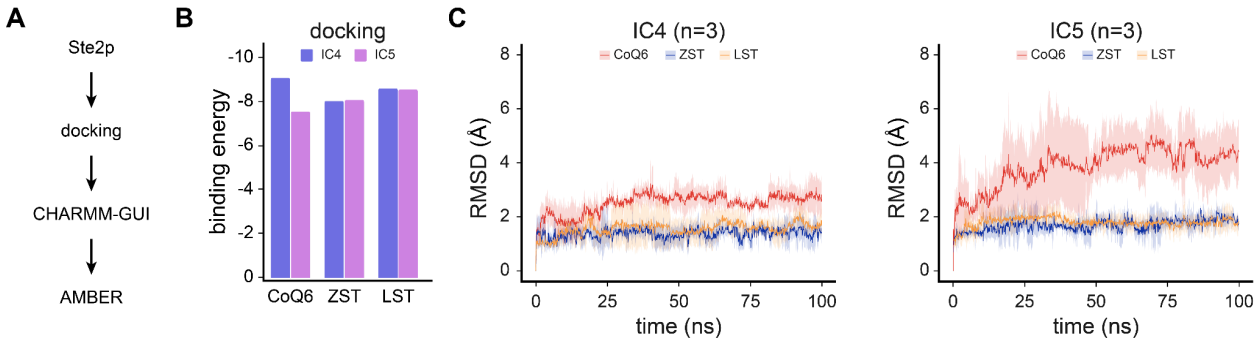


Figure 4.19 Molecular Dynamics and Binding Energy Analysis of Ste2p–Metabolite Complexes

(A) Workflow depicting the steps involved in the molecular dynamics simulation of the Ste2p-metabolite docked complex using the AMBER software suite. (B) Barplots depicting the binding energies obtained by the docking of Ste2p and indicated metabolites across IC4 and IC5. (C) Line plot depicting the RMSD changes over simulation timeframes from the three independent replicates of the indicated conditions. The spread of the data is indicated as Standard Deviation (SD). Notably, RMSD is provided in Angstroms (Å), whereas the simulation time is in nanoseconds (ns).

To accomplish this, we conducted three independent replicates of molecular dynamics (MD) simulations (100 ns each) of the Ste2p-metabolite complex for both cavities **(Figure 4.19A-B)**. MD simulation results indicate that the interactions between the metabolites and Ste2p at the Ste2p-Gpa1p interface are thermodynamically stable in nearly all instances across both cavities (IC4 and IC5), with the exception of CoQ6, which exhibited a fluctuating RMSD throughout the simulation period **(Figure 4.19C)**. Significantly, variable RMSD is detected solely in the instance of IC5, whereas it remains within the acceptable range for IC4 **(Figure 4.19C; Table 4.4; Table 4.5)**.

IC4	CoQ6			ZST			LST		
	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
ΔE_{psolv}	-62.61 ± 0.06	-64.31 ± 0.17	-60.08 ± 0.24	-42.71 ± 0.05	-45.58 ± 0.19	-45.06 ± 0.17	-46.96 ± 0.05	-45.89 ± 0.17	-29.91 ± 0.57
ΔE_{elec}	0 ± 0	0 ± 0	0 ± 0	-1.53 ± 0.02	-2.41 ± 0.07	-1.23 ± 0.07	-0.07 ± 0.02	-0.12 ± 0.06	-0.35 ± 0.06
ΔE_{psolv}	18.87 ± 0.02	19.96 ± 0.07	17.07 ± 0.07	9.25 ± 0.02	10.86 ± 0.09	9.2 ± 0.08	9.26 ± 0.03	9.49 ± 0.09	9.19 ± 0.19
ΔE_{npsolv}	-7.86 ± 0.01	-8 ± 0.02	-7.7 ± 0.03	-5.33 ± 0.01	-5.63 ± 0.02	-5.45 ± 0.02	-5.53 ± 0.01	-5.49 ± 0.02	-3.82 ± 0.07
ΔG (kcal/mol)	-51.59 ± 0.05	-52.36 ± 0.18	-50.71 ± 0.25	-40.32 ± 0.05	-42.76 ± 0.19	-42.54 ± 0.19	-43.31 ± 0.06	-42.02 ± 0.18	-24.89 ± 0.49

Table 4.4 Binding Energy Components of Metabolites in Ste2p IC4

Table indicating the Van der Waals (ΔEvdw), electrostatic (ΔEelec), polar solvation (ΔEpsolv), and non-polar solvation (ΔEnpsolv) energies along with net binding free-energy (ΔGbind) in kcal/mol of the three independent replicates of the indicated metabolites across intracellular cavities IC4 of the Ste2p. Note the values in the table are provided as mean ± standard errors.

IC5	CoQ6			ZST			LST		
	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
ΔE_{psolv}	-48.27 ± 0.09	-39.4 ± 0.35	-63.7 ± 0.17	-31.64 ± 0.05	-36.98 ± 0.13	-43.69 ± 0.15	-34.05 ± 0.06	-42.37 ± 0.18	-42.7 ± 0.52
ΔE_{elec}	-2.75 ± 0.04	-2.13 ± 0.2	-1.97 ± 0.12	-0.44 ± 0.02	-1.88 ± 0.09	-0.25 ± 0.07	0.45 ± 0.01	-1.33 ± 0.07	0.11 ± 0.05
ΔE_{psolv}	20.05 ± 0.06	15.23 ± 0.35	28.47 ± 0.15	8.61 ± 0.03	10.25 ± 0.09	7.89 ± 0.08	6.46 ± 0.02	11.11 ± 0.1	8.89 ± 0.13
ΔE_{npsolv}	-6.13 ± 0.01	-5.36 ± 0.03	-8.3 ± 0.02	-4 ± 0.01	-4.79 ± 0.02	-5.51 ± 0.02	-4.05 ± 0.01	-5.07 ± 0.02	-5.16 ± 0.06
ΔG (kcal/mol)	-37.1 ± 0.08	-31.65 ± 0.23	-45.5 ± 0.16	-27.47 ± 0.05	-33.4 ± 0.14	-41.57 ± 0.16	-31.19 ± 0.06	-37.66 ± 0.2	-38.86 ± 0.48

Table 4.5 Binding Energy Components of Metabolites in Ste2p IC5

Table indicating the Van der Waals (ΔEvdw), electrostatic (ΔEelec), polar solvation (ΔEpsolv), and non-polar solvation (ΔEnpsolv) energies along with net binding free-energy (ΔGbind) in kcal/mol of the three independent replicates of the indicated metabolites across intracellular cavities IC5 of the Ste2p. Note the values in the table are provided as mean ± standard errors.

We conducted an extended simulation of IC4 for 550 ns involving three Ste2p-metabolite complexes, noting stable complexes for ZST and LST, while observing fluctuations in the ligand RMSD for CoQ6 after ~150 ns **(Figure 4.20)**. To obtain additional insight into the contributing residues from the MD simulations, we conducted a residue-wise decomposition analysis that elucidates the energy contributions of various residues to the total binding free energies **(Figure 4.21)**. The results indicate that residues specific to IC4 and IC5 primarily influence the overall binding free energies. The binding free energies are derived as an average from ~500 configurations corresponding to the final 50 ns of the MD simulations. Collectively, the results demonstrate that the interaction between the metabolites and the Ste2p–Gα (Gpa1p) interface is not transient, suggesting their potential to stably modulate receptor activity and influence downstream signaling.

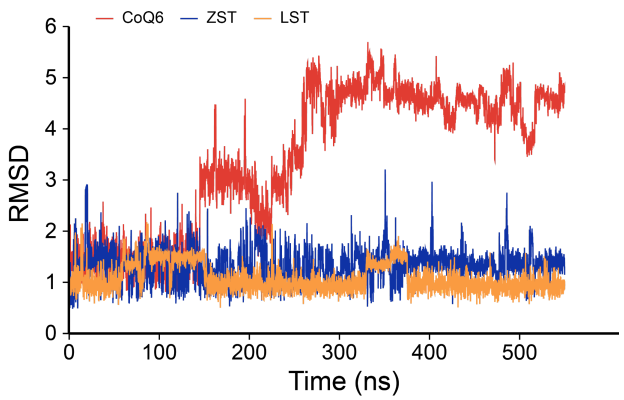


Figure 4.20 Long-Timescale MD Reveals RMSD Stability of Metabolites Bound to Ste2p Cavity

Line plot depicting the variation in the RMSD of the indicated metabolites bound to the intracellular cavity 4 (IC4) at the GPCR-Gα interface of the Ste2p, obtained using longer MD simulation runs. Notably, RMSD is provided in Angstroms (Å), whereas the simulation time is in ns.

Though we observed fluctuating RMSD values in the case of CoQ6, this may be attributed to its longer and more flexible molecular structure compared to ZST and LST.

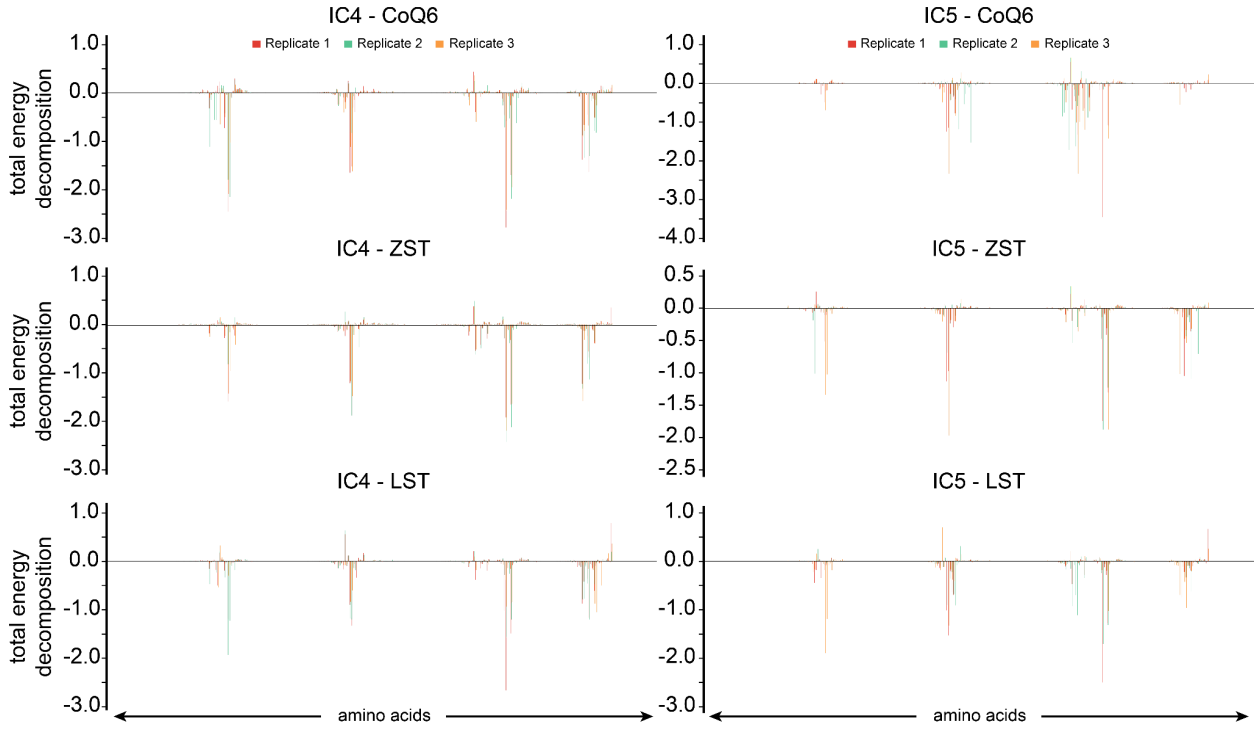


Figure 4.21 Energy Decomposition of Ste2p in IC4 and IC5 from MD Simulations

Line plots depicting the total energy decomposition of the individual amino acids of the Ste2p (three independent replicates) computed using MD simulations in IC4 (left) and IC5 (right).

4.3.9 Mutational Analysis for Key Residues Mediating Metabolite-Driven Ste2p Stabilization

To obtain a more profound understanding of the mechanism by which these metabolites inhibit Ste2p signaling, we initially examined their effects at the orthosteric site, specifically α-factor binding. We conducted protein-peptide docking of Ste2p and α-factor, revealing that metabolite binding at the Ste2p-Gpa1p interface enhances α-factor interaction, as indicated by the binding free energies ΔG (kcal/mol) **(Figure 4.22A-B)**.

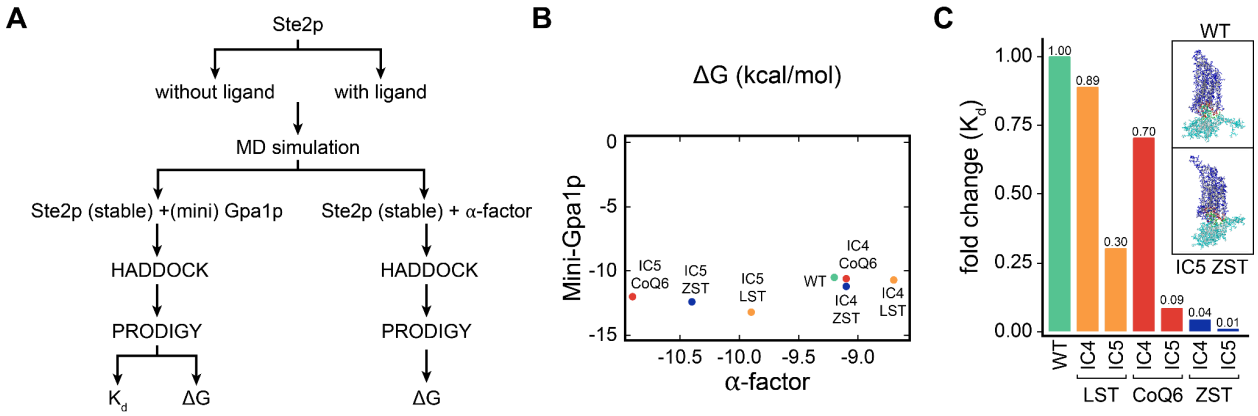


Figure 4.22 Effect of Metabolite Binding on Ste2p Interactions with miniG-Protein and α-Factor

(A) Workflow depicting the steps involved in Ste2p-miniG-protein and Ste2p-α-factor docking using HADDOCK and PRODIGY web servers. (B) Scatterplot depicting the net binding free-energy ΔG (kcal/mol) of the Ste2p (with and without indicated metabolites) with miniG-protein (y-axis) and α-factor (x-axis). (C) Barplots depicting the fold change of the dissociation constant (Kd) in the indicated conditions. Notably, fold change was computed with respect to the wild-type condition (Ste2p-miniG-protein). Inlets represent molecular representations of Ste2p-miniG-protein and the highlighted interface residues.

We examined the protein-protein interaction between Ste2p (GPCR) and miniGpa1-protein (Velazhahan et al. 2021) by selecting GPCR configurations with and without metabolite-induced alterations in cavity topologies (IC4 and IC5), respectively, derived from the previously mentioned Ste2p-metabolite complex simulations. We calculated the dissociation constant (Kd), binding affinity (ΔG), and the structural modifications in the overall topology of Ste2p **(Figure 4.22A)**. The computational analyses demonstrated that, unlike the metabolite-free Ste2p-Gpa1p interaction, designated as the WT condition, the Kd value is significantly lower in the presence of metabolites, indicating a synergistic response elicited by these metabolites **(Figure 4.22C)**. A significantly reduced Kd value further suggests and potentially elucidates that the binding of metabolites preferentially supports the interaction between Ste2p (GPCR) and miniGpa1, facilitating the formation of a stable complex that may affect the shielding of the effector-regulating domains of Gpa1p or its interaction with the Ste4p (Gβ)-Ste18p (Gγ) complex.

To enhance our comprehension, we subsequently inquired whether obstructing the metabolite binding at Ste2p could alter the interaction dynamics between Ste2p and Gpa1p. *In silico* site-directed mutagenesis of Ste2p identified critical metabolite-binding residues: S75 and R233 for CoQ6, L289 for ZST, and T155, V152, and I153 for LST **(Figure 4.23A)**.

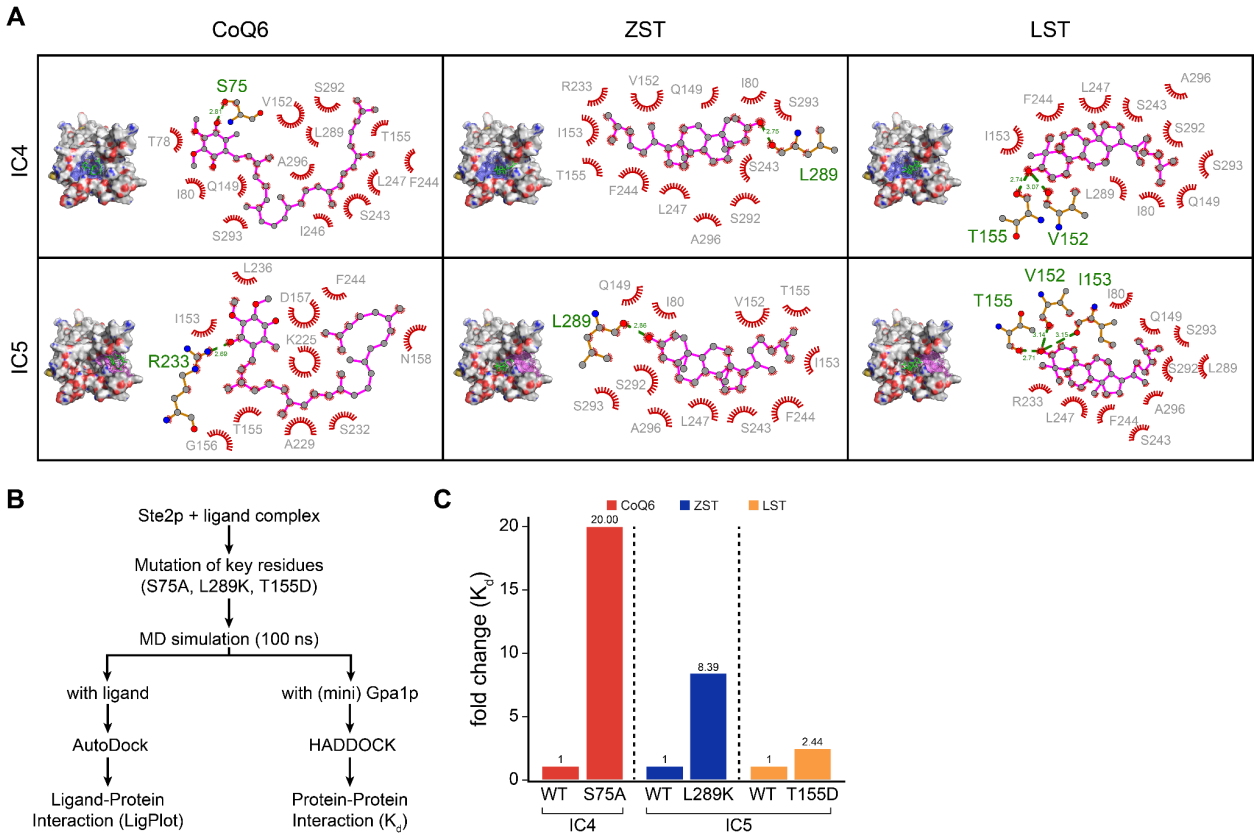


Figure 4.23 Functional Assessment of Key Residues Mediating Metabolite–Ste2p Interactions

(A) Ligplots depicting the interacting amino acid residues and atoms of the Ste2 protein and indicated metabolites for the intracellular cavities (IC4 and IC5). (B) Workflow depicting the steps involved in Ste2p-miniG-protein docking of the wild-type and site-directed Ste2p mutants. Notably, docking was performed using HADDOCK and PRODIGY web servers. (C) Barplots depicting the dissociation constant (Kd) fold change in Ste2p site-directed mutants and wild-type. Notably, fold change was computed with respect to the metabolite influenced wild-type condition (Ste2p-miniG-protein).

We selected the mutants S75A, L289K, and T155D, prioritizing robust hydrogen bonds within the optimal range of 2.7–3.3 Å. The mutations markedly elevated the dissociation constant (Kd) of the Ste2p-Gpa1p complex, signifying diminished interactions relative to wild-type Ste2p **(Figure 4.23B-C)**. This computational evidence corroborates the hypothesis that metabolite binding stabilizes the Ste2p-Gpa1p complex, thereby enhancing the rescue response to α-factor-induced PCD.

4.3.10 Gcoupler Unveiled Functional Conservation of GPCR-Gα interface

Human GPCRs expressed in yeast can activate yeast G-alpha proteins and vice versa, illustrating functional conservation despite approximately one billion years of evolutionary divergence. This phenomenon underscores the evolutionary conservation of the GPCR-Gα interface's architecture across species. The capacity of human GPCRs to interact efficiently with yeast G-alpha proteins highlights a significant conservation of the essential regions implicated in GPCR signaling pathways. This observation aligns with the extensively documented conservation of both sequence and structural elements across all six classes of GPCRs in mammals (45–47).

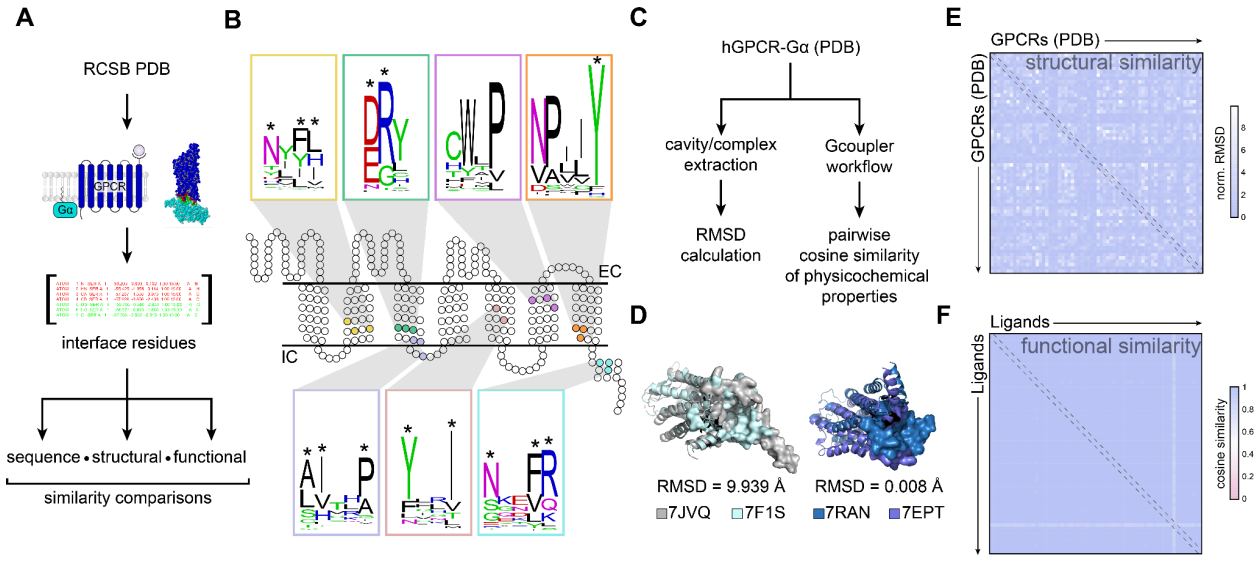


Figure 4.24 Conserved Architecture and Ligand Space of Human GPCR–Gα-Protein Binding Interfaces

(A) Schematic workflow depicting the major steps in determining the sequence, structural and functional conservation of the human GPCRs-Gα-protein interfaces. (B) Snakeplot depicting the standard human GPCR two-dimensional sequence level information. Conserved motifs of the GPCR-Gα-protein interfaces are depicted as WebLogo. Asterisks represent residues of conserved motifs present in the GPCRs-Gα-protein interfaces. Of note, the location of the motifs indicated in the exemplary GPCR snake plot is approximated. (C) Schematic workflow illustrating the steps involved in measuring and comparing the structural conservation of the GPCR-Gα-protein interfaces across human GPCRs. (D) Representative structures of the proteins depicting highly conserved (low RMSD) and highly divergent (high RMSD) GPCR-Gα-protein interfaces. PDB accession numbers are indicated at the bottom. (E) Heatmap depicting the RMSD values obtained by the pairwise comparison of all the GPCR-Gα-protein interfaces of the available human GPCRs from the protein databank. Of note, the RMSD of the Gα-protein cavity was normalized with the RMSDs of the respective whole proteins across all pairwise comparisons. (F) Heatmap depicting the pairwise cosine similarities between the in-silico synthesized ligands of the GPCR-Gα-protein interfaces of the available human GPCRs using Gcoupler.

We aimed to determine if this conservation applies at the functional level, particularly whether the small molecules that can modulate these sites across various GPCRs exhibit chemical similarities. To achieve a more profound comprehension of the GPCR-Gα protein interface from various perspectives i.e. sequence, structural, and functional basis, we initiated an extensive analysis. We sought to investigate the potential for direct small molecule interaction with the GPCR-Gα interface to influence downstream signaling pathways.Our inquiry commenced with the acquisition of all experimentally determined structures of human GPCR-Gα complexes from the PDB **(Figure 4.24A)**. This method enabled us to analyze the conserved characteristics of the interface and pinpoint potential binding sites for small molecule modulators **(Figure 4.24A)**. We examined these structures by isolating the residues that constitute the GPCR-Gα interaction interface. Subsequently, by examining the amino acid conservation at the GPCR-Gα-protein interface, we identified several conserved motifs, including the canonical DRY, CWxL, and NPxxY motifs (Calebiro et al. 2021) **(Figure 4.24B)**. To assess the topological similarity of the GPCR-Gα protein interface, we conducted a comprehensive structural analysis of various GPCR-Gα protein complexes. This analysis entailed identifying and extracting the cavities within each complex. By concentrating on these pivotal regions, we sought to evaluate the extent of structural conservation and quantify it using normalized RMSD values. Our findings demonstrated significant structural conservation at the GPCR-Gα interface. The normalized RMSD values, which quantify the average distance between atoms of superimposed proteins, demonstrated a significant level of similarity. The average RMSD value was determined to be 1.47 Å, whereas the median RMSD value was lower at 0.86 Å. The values indicate that the overall topology of the GPCR-Gα interface is highly conserved across various complexes, underscoring the resilience of this interaction site **(Figure 4.24C-E)**. To evaluate whether the conservation of topology and sequence influences the ligand profiles that may interact with this interface, we executed the Gcoupler workflow on all 66 GPCRs and synthesized approximately 50 distinct ligands for each GPCR **(Figure 4.24C)**. Subsequently, we calculated and compared the physicochemical properties of the synthesized ligands using Mordred (Moriwaki et al. 2018) and noted an increased cosine similarity, suggesting functional conservation of the ligand profiles capable of binding to these interfaces **(Figure 4.24C,F)**. In conclusion, we employed Gcoupler to methodically assess and analyze the ligand profiles of the GPCR-Gα-protein interface, revealing a greater extent of sequence, topological, and functional conservation.

4.3.11 Attenuating GPCR-mediated Hypertrophy Response via Allosteric Modulators

The yeast Gα is generally recognized to transduce both transgenically expressed human GPCRs in yeast and vice versa. This is likely attributable to the significant conservation of sequence and structure at the

GPCR-Gα interface. Our thorough computational analysis of all existing human GPCR-Gα complexes demonstrated a significant level of functional conservation **(Figure 4.24E,F).**

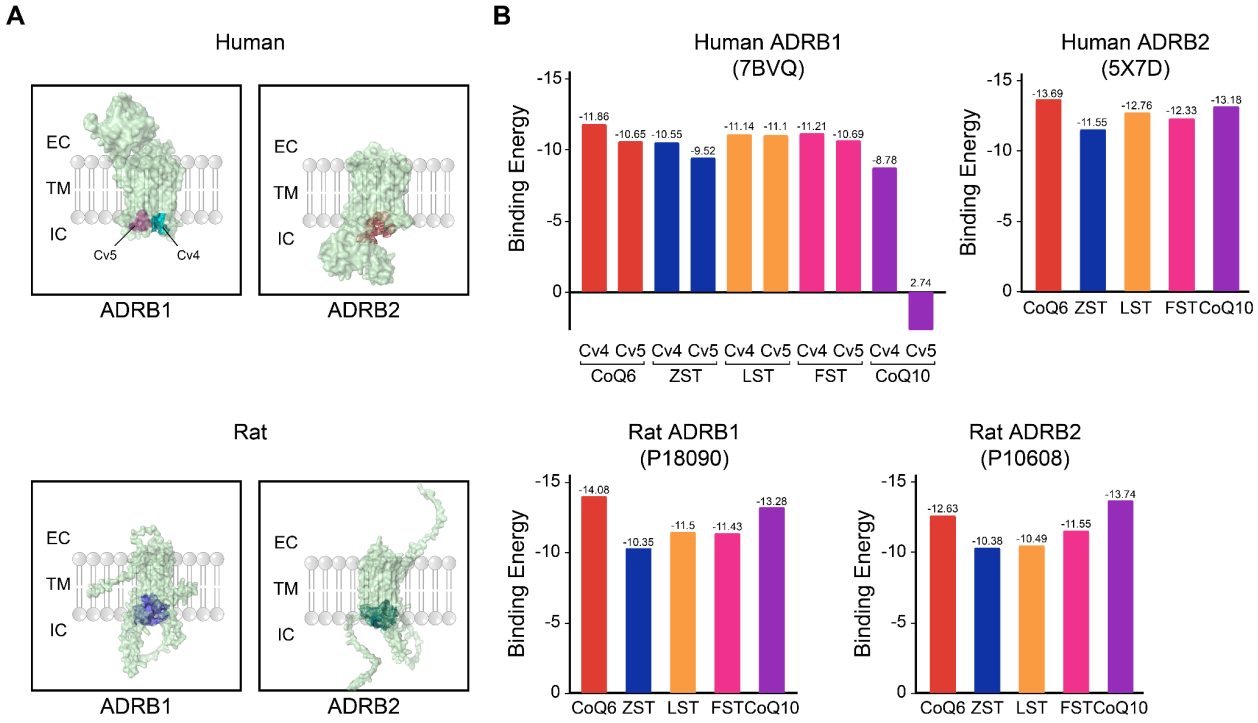


Figure 4.25 Evaluation of Metabolite Binding in Human and Rat β-Adrenergic Receptors

(A) Molecular representations depicting the topology of human and rat ADRB1 and ADRB2 receptors. GPCR-Gα interface cavities are color-coded at the intracellular sites. Of note, in the case of human ADRB1, the nomenclature of the two cavities detected at the interface includes 'Cv' (cavity) succeeded by the numerical number. (B) Barplots depicting the binding energies obtained by the docking of Human ADRB1 (7BVQ), Human ADRB2 (5X7D), Rat ADRB1 (P18090), and Rat ADRB2 (P10608) with the indicated metabolites at the GPCR-Gα interfaces.

Consequently, we inquired whether the recognized allosteric, intracellular metabolic modulators, including CoQ6, ZST, LST, FST, and CoQ10, could also influence human GPCR signaling.

								Consensus				
Residue	Location	MSA location	Ste2 *yeast*	ADRB1 *human*	ADRB2 *human*	Adrb1 *rat*	Adrb2 *rat*	Residue	Conservation Score	Property	Ste2 Match	Ste2 Property
V	152	181	V	S	S	S	S	S	80 %	polar	No	non-polar
I	153	182	I	L	L	L	L	L	80 %	non-polar	No	non-polar
T	155	184	T	T	T	T	T	T	100 %	polar	Yes	polar
R	233	263	R	R	R	R	R	R	100 %	basic	Yes	basic
L	289	327	L	-	L	-	L	L	60 %	non-polar	Yes	non-polar

Table 4.6 Ste2p Binding Residue Conservation in Human and Rat GPCRs

Table depicting the conservation of yeast Ste2p-metabolite (ZST, CoQ6, and LST) interacting residues in the adrenergic receptors from humans and rats.

We aimed to examine whether sterols could intrinsically and intracellularly modulate GPCR activity in higher vertebrates. To evaluate this hypothesis, we initially conducted a computational analysis on the human and rat beta1/2-adrenergic receptors. In summary, utilizing Gcoupler, we delineated the putative GPCR-Gα interface and conducted molecular docking with the specified metabolites **(Figure 4.25A-B)**. Docking results demonstrated a strong binding affinity of the selected metabolites at the GPCR-Gα interface of adrenergic receptors, analogous to Ste2p-metabolite interactions **(Figure 4.25A-B)**.

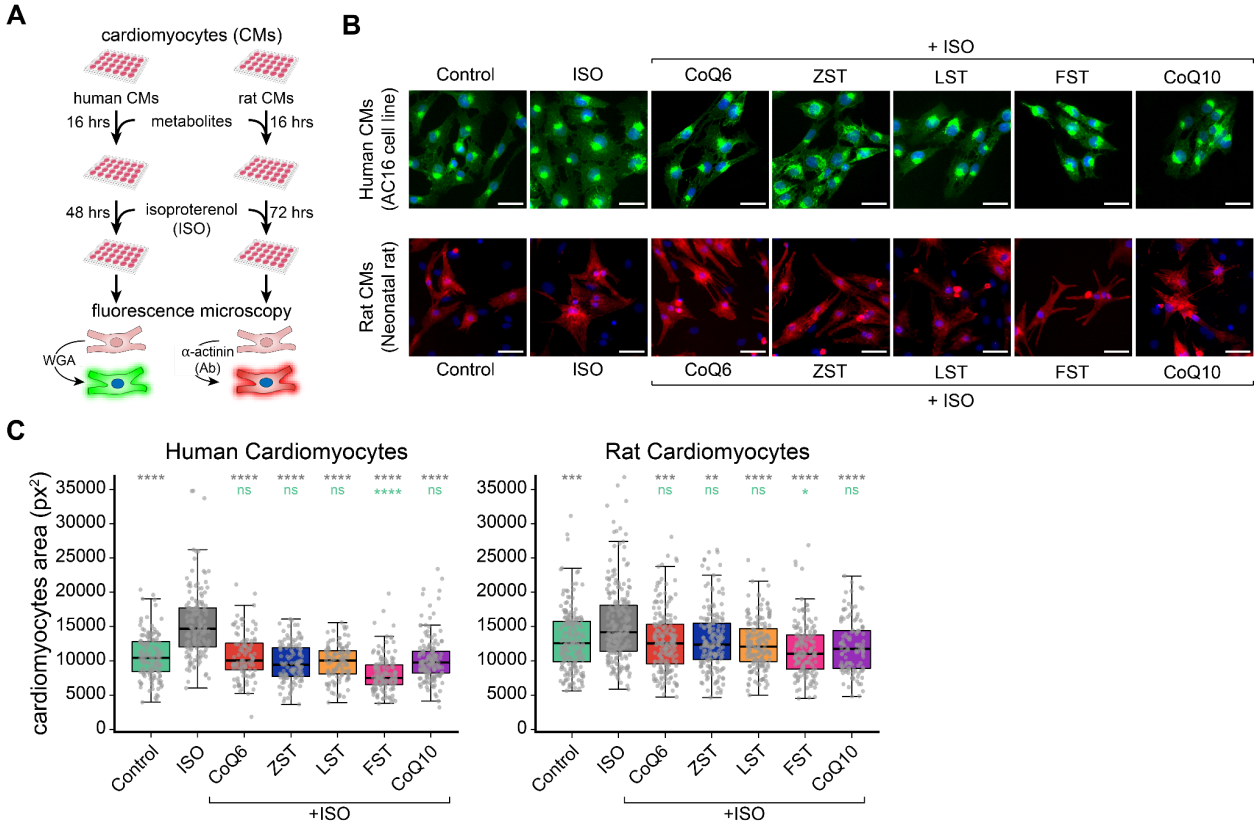


Figure 4.26 Experimental Assessment of Metabolite Impact on ISO-Induced Cardiomyocyte Hypertrophy

(A) Schematic representation of the experimental workflow followed to deduce the impact of indicated metabolites treatment on isoproterenol (ISO)-induced, GPCR-mediated hypertrophy response in human (AC16) and neonatal rat cardiomyocytes. Notably, in the case of AC16 cells, Wheat germ agglutinin (WGA) was used to stain the cardiomyocytes, whereas, for neonatal cardiomyocytes, alpha-sarcomeric actinin staining was used. (B) Micrographs depicting the human (above; green colored) and neonatal rat (below; red colored) cardiomyocytes in the indicated conditions. Scale 50 µm. (C) Box plot depicting the surface area of human (AC16) and neonatal rat cardiomyocytes in the indicated conditions. Statistical significance of indicated metabolites with untreated control and isoproterenol-treated conditions are indicated in green and grey text, respectively. Mann Whitney U test with Bonferroni corrected p-values was used to compute statistical significance.

To assess evolutionary conservation, we conducted a computational sequence conservation analysis involving Ste2p from yeast and adrenergic receptors from humans and rats. A meticulous examination of

the GPCR-Gα interface, especially the residues involved in the metabolites-GPCR interaction, demonstrated a significant level of conservation **(Table 4.6)**. To clarify the functional significance of this anticipated interaction, we employed isoproterenol-induced, adrenergic receptor-mediated cardiac hypertrophy models in cultured human cardiac cell line (AC16) and primary neonatal rat cardiomyocytes **(Figure 4.26A)**. Our findings indicate that both AC16 and neonatal cardiomyocytes preloaded with these metabolites mitigate the hypertrophic response induced by isoproterenol, as evidenced by the quantitative analysis of the surface area of individual cardiomyocytes **(Figure 4.26B-C)**. Collectively, these results affirm that the modulation of GPCR activity by endogenous intracellular metabolites is an evolutionarily conserved phenomenon.

4.4 Discussion

Comprehensive examination of orthosteric or allosteric sites, including the GPCR-Gα interface or adjacent areas, necessitates various resource-intensive computational tools, advanced technical expertise, and specialized knowledge (Basith et al. 2018; Congreve et al. 2020; Hedderich et al. 2022). Although conventional computational structural biology and empirical machine learning techniques can identify novel intracellular allosteric modulators of GPCRs, the scarcity of information regarding compounds that regulate new allosteric sites renders the application of these methods largely impractical (Chatzigoulas and Cournia 2021; Bartuzi, Kaczor, and Matosiuk 2018; Hou et al. 2021; Topiol 2018b). Comparable problems are recognized at the orthosteric sites, leading to the orphan designation of a substantial proportion of GPCRs. Given the recently documented high success rate of cavity-based drug design in identifying hit ligands (Zhavoronkov et al. 2019), existing solutions that provide an end-to-end workflow are either functionally constrained or encounter installation difficulties due to dependency errors.

We developed Gcoupler, a computational framework that utilizes *de novo* cavity identification on GPCR topological surfaces and subsequent *in silico* ligand synthesis, combined with robust statistical methods to categorize synthesized compounds into high-affinity and low-affinity binders. The deep learning algorithms (Graph Neural Networks) in Gcoupler enhance predictive modeling and facilitate large-scale screening of user-specified compounds. Ultimately, bioactivity-based ligand prioritization conducts post-prediction analysis for the classification of compounds based on functional activity. Unlike other ML/DL-based methodologies, Gcoupler presents an innovative solution that eliminates the necessity for cavity-specific experimentally validated compounds for model training and facilitates the large-scale screening of user-specified compounds efficiently and effortlessly. Furthermore, owing to its universal design, Gcoupler can theoretically be utilized for non-GPCR proteins as well. In contrast to other established site identification tools for GPCRs, such as Allosite (Huang et al. 2013), AllositePro (Song et

al. 2017), and AlloReverse (Zha et al. 2023), which predominantly utilize machine learning models or necessitate orthosteric ligand-bound structures as input, Gcoupler (LigBuilderV3) features a cavity detection capability that is not confined to allosteric sites. Instead, it identifies all potential cavity-like regions on the protein surface, subsequently classifying them as druggable, undruggable, or amphibious based on individual scoring and ligandability, thereby ensuring an unbiased and more precise approach towards the query protein. Moreover, Gcoupler accommodates user-defined regions and residues, enhancing its flexibility for users. Notably, a further justification for selecting LigBuilder V3 for cavity identification over comparable tools like Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) is that the former employs a hydrogen atom probe that traverses the protein surface grid at 0.5 Å for cavity detection, thereby offering superior precision in delineating cavity boundaries in both breadth and depth mapping, whereas the latter relies on clusters of alpha spheres.

Recent tools such as DeepDock (Liao et al. 2019) and DeepDocking (Gentile et al. 2022) integrate AI algorithms with conventional docking methodologies, exhibiting exceptional efficacy in the identification of lead molecules. Nonetheless, their methodology for identifying high-confidence hit compounds markedly diverges from that of Gcoupler. The innovative approach of concentrating on cavity-associated hit compounds distinguishes Gcoupler from other Deep Learning-based docking protocols. Conversely, Deep Learning-based docking protocols typically depend on extensive pre-docked structures from various proteins for model training or employ a minimal subset of protein-ligand complexes as input (usually 1%), while the majority of ligands (99%) are allocated for testing in an iterative framework. A significant problem with these protocols is their propensity to demonstrate bias towards the 1% of data utilized for model training, leading to considerable variability in results due to the restricted compounds employed for training. Furthermore, these methodologies are most efficacious for ultra-large-scale libraries, necessitating extensive pre-computation procedures to ascertain the optimal training data volume. These complexities exacerbate the overall workflow. Significantly, Gcoupler lacks these constraints and provides a comprehensive, end-to-end solution for the complete analysis. Thus far, only a limited number of methods utilize generative AI models for cavity or pocket-based drug design. Gcoupler is an open-source, comprehensive platform that integrates LBDD and SBDD for drug development and extensive screening. In contrast to Pocket Crafter (L. Shen et al. 2024), which necessitates proprietary tools (e.g., MOE QuickPrep) and is devoid of predictive model-building modules, Gcoupler provides extensive functionality. Likewise, DeepLigBuilder (Y. Li, Pei, and Lai 2021) and Schrodinger’s AutoDesigner are either proprietary or feature-restricted in comparison to Gcoupler. The comparative analysis underscores Gcoupler's distinctive advantages in precision, flexibility, and functionality (Supplementary Table 11). It is accessible to the community as a Docker image and a Python package.

We employed Gcoupler to elucidate the molecular foundation of innate resistance to α-factor-induced PCD in yeast (Büttner et al. 2006; N.-N. Zhang et al. 2006; Carmona-Gutierrez et al. 2010; Teng and Hardwick 2009). In contrast to the human genome, which encodes approximately 800 GPCRs, the yeast genome contains merely two GPCR systems responsible for pheromone and glucose detection (Horn et al. 2003). The pheromone-sensing pathway in yeast has been extensively studied in recent decades, resulting in the identification of feedback inhibition systems associated with this pathway; however, to our knowledge, most known regulatory mechanisms function downstream of the signaling cascade (Carmona-Gutierrez et al. 2010; Alvaro and Thorner 2016). This indicates that the rescue mechanism conferring innate resistance to α-factor-induced programmed cell death likely functions at an upstream level of the signaling cascade through a direct, physical interaction-based regulatory mechanism, as opposed to the comparatively slower transcriptome or epigenome-mediated feedback mechanisms. One example may involve the regulation at the receptor level through endogenous metabolites that can induce allosteric modulation of the *Ste2* protein intracellularly. This may be viable, as several biochemically diverse, endogenous intracellular allosteric modulators for various GPCRs have been documented in the literature (van der Westhuizen et al. 2015; Stornaiuolo et al. 2015; Doller 2017). Gcoupler facilitated the identification of a specific subset of cellular metabolites capable of physically interacting with the Ste2p-miniGpa1 interface and modulating the downstream signaling pathway (Velazhahan et al. 2021). Computational and experimental evidence identified specific metabolites that bind to the Ste2p-miniGpa1 interface, thereby modulating signaling. Site-directed mutagenesis (data not shown) validated the functional significance of these metabolite-interacting residues. The mutagenesis experiments identified several essential amino acid residues that coincide with IC4 and IC5, indicating their functional significance in Ste2p downstream signaling (Supplementary Table 12).

This work clearly demonstrates the interlink; however, it has certain limitations. Initially, although genetic screening for individual metabolic mutants suggests a connection, it remains uncertain whether the observed variations in α-factor mediated programmed cell death are due to metabolites or arise from the pleiotropic effects in these mutants. Secondly, the concentrations of the metabolites employed in the exogenous treatment experiments may differ from those found in natural conditions (data not shown). Third, the predictions regarding metabolites-protein interactions can be corroborated by mutagenesis experiments involving the IC4 and IC5 residues; however, these experiments complicate interpretation and present challenges in determining whether the hypothesized activity loss from mutations results from metabolites-induced binding enhancement or inhibition. Fourth, The RNA-sequencing results identified potential genes that may contribute to the observed innate resistance response to the α-factor; however, a series of biochemical assays must be performed on both loss (knock-out/down) and gain

(over-expression) of function mutants for these genes to confirm the findings. Ultimately, evaluating the specificity of the observed mechanism using an alternative yeast GPCR system (Gpr1p) is technically demanding due to the significant sequence, structural, and functional conservation of the GPCR-Gα interface. Another significant limitation of our study is the lack of direct binding assays to confirm the interaction between the metabolites and Ste2p. Although our findings from genetic interventions, molecular dynamics simulations, and docking studies strongly indicate that the metabolites engage with the Ste2p-Gpa1 interface, these results are still indirect. Direct binding confirmation via techniques such as surface plasmon resonance, isothermal titration calorimetry, or co-crystallization would yield conclusive evidence of this interaction. A significant limitation of our findings is the dependence on tools such as AutoDock and PRODIGY for initial binding affinity estimates, which do not possess the thermodynamic accuracy of more sophisticated methods. To tackle this, we utilized molecular dynamics simulations with MM/GBSA, integrating factors such as protein flexibility and solvation effects for enhanced ΔG calculations. This study did not encompass computationally intensive methods; however, we ensured that the reported ΔG values accounted for system conformational flexibility by deriving them from pre-simulated docked structures obtained through molecular dynamics simulations. Moreover, our findings indicate that the metabolite interacts with the Ste2p-Gpa1 interface and regulates receptor activity in response to pheromone stimulation, as corroborated by multiple assays (data partially presented). The exact sequence of interactions among Ste2p, the metabolite, and Gpa1 has not been investigated, as it necessitates sequential experiments beyond the scope of this study. Collectively, rectifying these limitations in subsequent research would substantially enhance our conclusions and yield more profound insights into the specific molecular mechanisms driving the observed phenotypic effects.

The functional characterization of sterol biosynthesis intermediate (*ergΔ*) mutants demonstrated a compromised mating response. Mechanistic insights indicated that the phenotypic defects are markedly heterogeneous, encompassing aberrant schmoo morphology, diminished sterol accumulation at the mating projection tip, compromised membrane fusion, and reduced expression levels of FUS1 (Bagnat and Simons 2002; Jin, McCaffery, and Grote 2008; Heese-Peck et al. 2002; Aguilar et al. 2010; Tiedje et al. 2007). This study reveals a novel molecular mechanism that elucidates a new role of cellular sterols and accounts for the diminished efficacy of mating-dependent yeast microbial cell factories or biosensors in mutants with modified sterol levels (Shaw et al. 2019; Ostrov et al. 2017). Significantly, besides the regulation of Ste2p by endogenous intracellular sterols, their dysregulation confers tolerance to mycotoxic compounds, ethanol, and elevated temperatures; however, the underlying mechanism remains largely obscure (Johnston, Moses, and Rosser 2020). Due to the diverse biological roles of sterols, their in vivo functional characterization, including the simultaneous measurement of interactions with cellular proteins

and resultant phenotypic alterations, presents significant technical challenges. This study employs computational and experimental methods to demonstrate that sterols, along with other endogenous metabolites, can interact with the conserved Gα binding sites of GPCRs and influence their activity. The hypertrophy experiments utilizing human and rat cells further substantiate the specificity of this evolutionarily conserved regulatory mechanism. In summary, this study not only presents a novel computational method that could enhance and deepen our comprehension of various unexplored allosteric sites of GPCRs but also reveals a new regulatory mechanism of GPCRs mediated by endogenous intracellular metabolites. The results from the computational and experimental analysis suggest that GPCRs can be intrinsically regulated through their Gα-binding interface by intracellular metabolites.

Conclusion

Chapter 5 - Conclusion and Summary

GPCRs occupy a central role in mediating signal transduction and intercellular communication. This study aims to advance the understanding of GPCRs and their roles through the development of innovative tools. By providing a deeper mechanistic understanding of GPCR expression, signaling specificity, ligand engagement, and modulation strategies, the research presented herein seeks to advance our fundamental understanding of GPCR biology. It is hypothesized that the integrative approach, bridging mechanistic modeling, simplified machine learning, and data efficient artificial intelligence, will provide both conceptual advances and practical tools in GPCR research. Ultimately, this work aims to facilitate the development of targeted therapeutic interventions that exploit the inherent biological versatility of GPCRs, thus contributing meaningfully to the broader goals of precision therapeutics

In Chapter 2, we focused more on a specific subset of GPCRs i.e. odorant receptors. More precisely, the mechanism known as the "one neuron, one receptor" rule of olfaction. Our computational analysis employing the RCT framework provides compelling evidence supporting a preferential, rather than purely stochastic, OR selection mechanism guided predominantly by the expression level of ORs in precursor neuronal states. This supports a refined "winner-takes-all" model, though the presence of multiple models across distinct olfactory sensory neuron subpopulations cannot be ruled out. Despite notable methodological limitations, including incomplete cellular trajectory reconstruction and partial coverage of olfactory neuron heterogeneity, our findings significantly challenge the classical "silence-all-and-activate-one" hypothesis. Instead, they align with emerging evidence suggesting a dynamic and complex epigenetic landscape that transiently permits multi-OR expression early in differentiation before consolidating receptor choice in mature neurons. Future research should leverage both computational advancements and experimental validation to further resolve these mechanisms and fully characterize the nuanced regulatory processes underlying olfactory receptor expression.

Further, In Chapter 3, we addressed the existing limitations in GPCR ligand-receptor interaction research by providing an intuitive, user-friendly, and highly customizable platform for the rapid identification and classification of potential ligands. Leveraging built-in descriptor extraction methods and an array of robust machine learning classifiers, MOA enables researchers, even those without extensive computational backgrounds, to employ state-of-the-art techniques for selective, effective, and rapid identification of biologically relevant molecules from vast chemical databases. By integrating automated data preprocessing, hyperparameter optimization, rigorous validation protocols, and innovative model interpretability strategies, this approach significantly enhances the transparency, reliability, and reproducibility of predictions. Through successful applications ranging from identifying novel agonists

for olfactory receptors to discovering natural mosquito repellents, MOA demonstrates its utility and versatility. Thus, it not only broadens the accessibility and applicability of machine learning in chemoinformatics but also empowers researchers to uncover valuable GPCR-ligand relationships from large libraries.

Continuing on this, in Chapter 4, we tackled the challenge of data scarcity in GPCR research by leveraging the Gcoupler toolkit, an AI platform to identify receptor surface interaction sites and predict high-affinity ligands, facilitating targeted drug discovery. By integrating cavity detection, generative AI, and deep learning-based ligand synthesis and prioritization, Gcoupler addresses critical challenges posed by data scarcity and the technical complexity inherent in GPCR-targeted drug discovery. Our computational analyses, complemented by targeted experimental validation, revealed an unprecedented molecular mechanism where intracellular metabolites modulate GPCR signaling through direct interaction with conserved Gα-binding sites. This approach provided meaningful mechanistic insights into GPCR regulation in both yeast and mammalian systems. Moreover, the identification of sterols as endogenous modulators of GPCR activity highlights a previously unexplored field of cellular regulation, with profound implications for understanding biological processes such as stress tolerance and cellular signaling. In this way, Gcoupler could help foster deeper biological insights even in data-scarce GPCR contexts, paving the way for more targeted and effective drug discovery.

Collectively, building on the multifaceted exploration of GPCRs, this research unifies computational innovation and biological experimentation to unravel the layered complexity of GPCR regulation. Through the development of specialized frameworks, RCT for decoding olfactory receptor expression dynamics, MOA for democratizing ligand prediction, and Gcoupler for cavity-driven allosteric modulation, this work illuminates previously inaccessible aspects of GPCR function and ligand interaction. This research not only enhances our understanding of GPCR-mediated cellular communication but also equips the scientific community with flexible, open-source tools to interrogate receptor function across diverse biological contexts. The integration of algorithmic rigor with biological relevance ensures that the insights gained are not only mechanistically sound but also translationally meaningful, paving the way for next-generation drug discovery, receptor deorphanization, and precision therapeutics in complex diseases.

Building on the integrative computational frameworks developed in this thesis, several promising directions emerge to further advance GPCR biology and drug discovery. First, refining and scaling the RCT framework to incorporate higher-resolution, multi-omics datasets could further clarify olfactory receptor selection mechanisms and extend its utility to track target gene behavior during cellular

differentiation. Second, enhancing the MOA platform by incorporating real-time experimental feedback and active learning strategies could enable iterative refinement and better bridge in silico predictions with biological validation. Third, further development of the Gcoupler toolkit, through the inclusion of alternative cavity detection algorithms and customizable features, will increase its utility beyond GPCRs to a wider range of novel target proteins. Ultimately, close collaboration with experimental biologists and clinicians will be essential to validate computational predictions, facilitate translation to in vivo systems, and support the rational design of next-generation therapeutics targeting GPCR-mediated pathways in cancer, neurological, and metabolic diseases.

References

References

Abdus-Saboor, Ishmail, Mohammed J. Al Nufal, Maha V. Agha, Marion Ruinart de Brimont, Alexander Fleischmann, and Benjamin M. Shykind. 2016. "An Expression Refinement Process Ensures Singular Odorant Receptor Gene Choice." *Current Biology: CB* 26 (8): 1083–90.

Abraham, Mark James, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. 2015. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers." *SoftwareX* 1-2 (September):19–25.

Aguilar, Pablo S., Maxwell G. Heiman, Tobias C. Walther, Alex Engel, Dominik Schwudke, Nathan Gushwa, Teymuras Kurzchalia, and Peter Walter. 2010. "Structure of Sterol Aliphatic Chains Affects Yeast Cell Shape and Cell Fusion during Mating." *Proceedings of the National Academy of Sciences* 107 (9): 4170–75.

Ahmad, Raise, Stefanie Wojciech, and Ralf Jockers. 2015. "Hunting for the Function of Orphan GPCRs - beyond the Search for the Endogenous Ligand: Function of Orphan GPCRs." *British Journal of Pharmacology* 172 (13): 3212–28.

Ahuja, Gaurav, Vera Reichel, Daniel Kowatschew, Adnan S. Syed, Aswani Kumar Kotagiri, Yuichiro Oka, Franco Weth, and Sigrun I. Korsching. 2018. "Overlapping but Distinct Topology for Zebrafish V2R-like Olfactory Receptors Reminiscent of Odorant Receptor Spatial Expression Zones." *BMC Genomics* 19 (1): 383.

Alberts, Bruce, Dennis Bray, Karen Hopkin, Alexander D. Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2015. *Essential Cell Biology*. 4th ed. Boca Raton, FL: CRC Press.

Alexander, Stephen P. H., Arthur Christopoulos, Anthony P. Davenport, Eamonn Kelly, Alistair A. Mathie, John A. Peters, Emma L. Veale, et al. 2023. "The Concise Guide to PHARMACOLOGY 2023/24: G Protein-Coupled Receptors." *British Journal of Pharmacology* 180 Suppl 2 (October):S23–144.

Alexander, Stephen P. H., Arthur Christopoulos, Anthony P. Davenport, Eamonn Kelly, Alistair Mathie, John A. Peters, Emma L. Veale, et al. 2019. "THE CONCISE GUIDE TO PHARMACOLOGY 2019/20: G Protein-Coupled Receptors." *British Journal of Pharmacology* 176 Suppl 1 (Suppl 1): S21–141.

Alvaro, Christopher G., and Jeremy Thorner. 2016. "Heterotrimeric G Protein-Coupled Receptor Signaling in Yeast Mating Pheromone Response." *The Journal of Biological Chemistry* 291 (15): 7788–95.

Armstrong, Jane F., Elena Faccenda, Simon D. Harding, Adam J. Pawson, Christopher Southan, Joanna L. Sharman, Brice Campo, et al. 2020. "The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: Extending Immunopharmacology Content and Introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY." *Nucleic Acids Research* 48 (D1): D1006–21.

Bagnat, Michel, and Kai Simons. 2002. "Cell Surface Polarization during Yeast Mating." *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14183–88.

Baltimore, D. 1970. "RNA-Dependent DNA Polymerase in Virions of RNA Tumour Viruses." *Nature* 226 (5252): 1209–11.

Banegas-Luna, Antonio-Jesús, José P. Cerón-Carrasco, and Horacio Pérez-Sánchez. 2018. "A Review of Ligand-Based Virtual Screening Web Tools and Screening Algorithms in Large Molecular Databases in the Age of Big Data." *Future Medicinal Chemistry* 10 (22): 2641–58.

Bar-Shavit, Rachel, Myriam Maoz, Arun Kancharla, Jeetendra Kumar Nag, Daniel Agranovich, Sorina Grisaru-Granovsky, and Beatrice Uziely. 2016. "G Protein-Coupled Receptors in Cancer." *International Journal of Molecular Sciences* 17 (8): 1320.

Bartuzi, Damian, Agnieszka A. Kaczor, and Dariusz Matosiuk. 2018. "Opportunities and Challenges in the Discovery of Allosteric Modulators of GPCRs." *Methods in Molecular Biology* 1705:297–319.

Bashkirova, Elizaveta, and Stavros Lomvardas. 2019. "Olfactory Receptor Genes Make the Case for Inter-Chromosomal Interactions." *Current Opinion in Genetics & Development* 55 (April):106–13.

Basith, Shaherin, Minghua Cui, Stephani J. Y. Macalino, Jongmi Park, Nina A. B. Clavio, Soosung Kang, and Sun Choi. 2018. “Exploring G Protein-Coupled Receptors (GPCRs) Ligand Space via Cheminformatics Approaches: Impact on Rational Drug Design.” *Frontiers in Pharmacology* 9 (March):128.

Battle, G. M. n.d. “PDBePISA: Identifying and Interpreting the Likely Biological Assemblies of a Protein Structure.” *Chemistry* .

Bergen, Volker, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. 2020. “Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling.” *Nature Biotechnology* 38 (12): 1408–14.

Berger, Vance W., and Yanyan Zhou. 2014. “Kolmogorov–Smirnov Test: Overview.” *Wiley StatsRef: Statistics Reference Online*. Wiley. https://doi.org/10.1002/9781118445112.stat06558.

Berg, J. M., J. L. Tymoczko, G. J. Gatto, and L. Stryer. 2015. “Chapter 7: Hemoglobin: Portrait of a Protein in Action.” *Biochemistry, 8th Edition. New York: WH Freeman and Co*, 191–214.

Bertoni, Martino, Miquel Duran-Frigola, Pau Badia-I-Mompel, Eduardo Pauls, Modesto Orozco-Ruiz, Oriol Guitart-Pla, Víctor Alcalde, et al. 2021. “Bioactivity Descriptors for Uncharacterized Chemical Compounds.” *Nature Communications* 12 (1): 3932.

Blagus, Rok, and Lara Lusa. 2013. “SMOTE for High-Dimensional Class-Imbalanced Data.” *BMC Bioinformatics* 14 (1): 106.

Blumer, K. J., J. E. Reneke, and J. Thorner. 1988. “The STE2 Gene Product Is the Ligand-Binding Component of the Alpha-Factor Receptor of Saccharomyces Cerevisiae.” *The Journal of Biological Chemistry* 263 (22): 10836–42.

Böhm, H. J. 1992. “The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors.” *Journal of Computer-Aided Molecular Design* 6 (1): 61–78.

Bos, Pieter H., Evelyne M. Houang, Fabio Ranalli, Abba E. Leffler, Nicholas A. Boyles, Volker A. Eyrich, Yuval Luria, et al. 2022. “AutoDesigner, a De Novo Design Algorithm for Rapidly Exploring Large Chemical Space for Lead Optimization: Application to the Design and Synthesis of D-Amino Acid Oxidase Inhibitors.” *Journal of Chemical Information and Modeling* 62 (8): 1905–15.

Bourque, Kyla, Juliana C. C. Dallagnol, Hassan Nassour, David Chatenet, Bruce G. Allen, and Terence E. Hébert. 2022. “Exploring the Use of Intracellular and Extracellular Allosteric Modulators to Understand GPCR Signaling.” *Allosteric Modulation of G Protein-Coupled Receptors*. https://doi.org/10.1016/b978-0-12-819771-4.00008-7.

Buck, L., and R. Axel. 1991. “A Novel Multigene Family May Encode Odorant Receptors: A Molecular Basis for Odor Recognition.” *Cell* 65 (1): 175–87.

Bushdid, Caroline, Claire A. de March, Hiroaki Matsunami, and Jérôme Golebiowski. 2018. “Numerical Models and In Vitro Assays to Study Odorant Receptors.” *Methods in Molecular Biology (Clifton, N.J.)* 1820:77–93.

Bushdid, C., C. A. de March, S. Fiorucci, H. Matsunami, and J. Golebiowski. 2018. “Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features.” *The Journal of Physical Chemistry Letters* 9 (9): 2235–40.

Büttner, Sabrina, Tobias Eisenberg, Eva Herker, Didac Carmona-Gutierrez, Guido Kroemer, and Frank Madeo. 2006. “Why Yeast Cells Can Undergo Apoptosis: Death in Times of Peace, Love, and War.” *The Journal of Cell Biology* 175 (4): 521–25.

Bystrova, M. F., and S. S. Kolesnikov. 2021. “The ‘One Neuron–One Receptor’ Rule in the Physiology and Genetics of Olfaction.” *Neuroscience and Behavioral Physiology* 51 (7): 1008–17.

Caballero-Vidal, Gabriela, Cédric Bouysset, H. Grunig, S. Fiorucci, N. Montagné, J. Golebiowski, and E. Jacquin‐Joly. 2020. “Machine Learning Decodes Chemical Features to Identify Novel Agonists of a Moth Odorant Receptor.” *Scientific Reports* 10 (February). https://doi.org/10.1038/s41598-020-58564-9.

Calebiro, Davide, Zsombor Koszegi, Yann Lanoiselée, Tamara Miljus, and Shannon O’Brien. 2021. “G Protein-Coupled Receptor-G Protein Interactions: A Single-Molecule Perspective.” *Physiological Reviews* 101 (3): 857–906.

Calebiro, Davide, Viacheslav O. Nikolaev, Luca Persani, and Martin J. Lohse. 2010. “Signaling by Internalized G-Protein-Coupled Receptors.” *Trends in Pharmacological Sciences* 31 (5): 221–28.
Carmona-Gutierrez, D., T. Eisenberg, S. Büttner, C. Meisinger, G. Kroemer, and F. Madeo. 2010. “Apoptosis in Yeast: Triggers, Pathways, Subroutines.” *Cell Death and Differentiation* 17 (5): 763–73.
Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. 2019. “Machine Learning Interpretability: A Survey on Methods and Metrics.” *Electronics* 8 (8): 832.
Case, David A., Thomas E. Cheatham 3rd, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. “The Amber Biomolecular Simulation Programs.” *Journal of Computational Chemistry* 26 (16): 1668–88.
Casiraghi, Marina, Marjorie Damian, Ewen Lescop, Jean-Louis Banères, and Laurent J. Catoire. 2018. “Illuminating the Energy Landscape of GPCRs: The Key Contribution of Solution-State NMR Associated with Escherichia Coli as an Expression Host.” *Biochemistry* 57 (16): 2297–2307.
Caspi, Ron, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, et al. 2016. “The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/genome Databases.” *Nucleic Acids Research* 44 (D1): D471–80.
Chatzigoulas, Alexios, and Zoe Cournia. 2021. “Rational Design of Allosteric Modulators: Challenges and Successes.” *Wiley Interdisciplinary Reviews. Computational Molecular Science* 11 (6). https://doi.org/10.1002/wcms.1529.
Chess, A., I. Simon, H. Cedar, and R. Axel. 1994. “Allelic Inactivation Regulates Olfactory Receptor Gene Expression.” *Cell* 78 (5): 823–34.
Chong, Jasmine, Othman Soufan, Carin Li, Iurie Caraus, Shuzhao Li, Guillaume Bourque, David S. Wishart, and Jianguo Xia. 2018. “MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis.” *Nucleic Acids Research* 46 (W1): W486–94.
Clark, Matthew, Richard D. Cramer, and Nicole Van Opdenbosch. 1989. “Validation of the General Purpose Tripos 5.2 Force Field.” *Journal of Computational Chemistry* 10 (8): 982–1012.
Congreve, Miles, Chris de Graaf, Nigel A. Swain, and Christopher G. Tate. 2020. “Impact of GPCR Structures on Drug Discovery.” *Cell* 181 (1): 81–91.
Cosconati, Sandro, Stefano Forli, Alex L. Perryman, Rodney Harris, David S. Goodsell, and Arthur J. Olson. 2010. “Virtual Screening with AutoDock: Theory and Practice.” *Expert Opinion on Drug Discovery* 5 (6): 597–607.
Crick, F. 1970. “Central Dogma of Molecular Biology.” *Nature* 227 (5258): 561–63.
Dagan-Wiener, Ayana, Ido Nissim, Natalie Ben Abu, Gigliola Borgonovo, Angela Bassoli, and Masha Y. Niv. 2017. “Bitter or Not? BitterPredict, a Tool for Predicting Taste from Chemical Structure.” *Scientific Reports* 7 (1): 12074.
Dallakyan, S. n.d. “MGLTools.” *Reference Source*.
Davenport, Anthony P., Stephen P. H. Alexander, Joanna L. Sharman, Adam J. Pawson, Helen E. Benson, Amy E. Monaghan, Wen Chiy Liew, et al. 2013. “International Union of Basic and Clinical Pharmacology. LXXXVIII. G Protein-Coupled Receptor List: Recommendations for New Pairings with Cognate Ligands.” *Pharmacological Reviews* 65 (3): 967–86.
Doller, Dario. 2017. “Endogenous Allosteric Modulators of G Protein-Coupled Receptors: Implications in Drug Design.” *2017 Medicinal Chemistry Reviews*. https://doi.org/10.29200/acsmedchemrev-v52.ch18.
Durante, Michael A., Stefan Kurtenbach, Zoukaa B. Sargi, J. William Harbour, Rhea Choi, Sarah Kurtenbach, Garrett M. Goss, Hiroaki Matsunami, and Bradley J. Goldstein. 2020. “Single-Cell Analysis of Olfactory Neurogenesis and Differentiation in Adult Humans.” *Nature Neuroscience* 23 (3): 323–26.
Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97.
Engmann, Sonja, and Denis Cousineau. 2011. “Comparing Distributions: The Two-Sample Anderson-Darling Test as an Alternative to the Kolmogorov-Smirnoff Test.” *Journal of Applied*

Quantitative Methods 6 (3). http://jaqm.ro/issues/volume-6,issue-3/pdfs/jaqm_vol6_issue3.pdf#page=5.

Ferrer, Isidro, Paula Garcia-Esparcia, Margarita Carmona, Eva Carro, Eleonora Aronica, Gabor G. Kovacs, Alice Grison, and Stefano Gustincich. 2016a. "Olfactory Receptors in Non-Chemosensory Organs: The Nervous System in Health and Disease." *Frontiers in Aging Neuroscience* 8 (July):163.

———. 2016b. "Olfactory Receptors in Non-Chemosensory Organs: The Nervous System in Health and Disease." *Frontiers in Aging Neuroscience* 8 (July):163.

Flegel, Caroline, Stavros Manteniotis, Sandra Osthold, Hanns Hatt, and Günter Gisselmann. 2013. "Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing." *PloS One* 8 (2): e55368.

Fletcher, Russell B., Diya Das, Levi Gadye, Kelly N. Street, Ariane Baudhuin, Allon Wagner, Michael B. Cole, et al. 2017. "Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution." *Cell Stem Cell* 20 (6): 817–30.e8.

Flock, Tilman, Alexander S. Hauser, Nadia Lund, David E. Gloriam, Santhanam Balaji, and M. Madan Babu. 2017. "Selectivity Determinants of GPCR-G-Protein Binding." *Nature* 545 (7654): 317–22.

Fredriksson, Robert, Malin C. Lagerström, Lars-Gustav Lundin, and Helgi B. Schiöth. 2003. "The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints." *Molecular Pharmacology* 63 (6): 1256–72.

Fredriksson, R., and H. B. Schiöth. 2006. *Ligand Design for G Protein-Coupled Receptors*. Edited by Didier Rognan. 1st ed. Methods and Principles in Medicinal Chemistry, v. 30. Weinheim, Germany: Wiley-VCH Verlag.

Gahbauer, Stefan, and Rainer A. Böckmann. 2016. "Membrane-Mediated Oligomerization of G Protein Coupled Receptors and Its Implications for GPCR Function." *Frontiers in Physiology* 7 (October):494.

Gaillard, I., S. Rouquier, and D. Giorgi. 2004. "Olfactory Receptors." *Cellular and Molecular Life Sciences: CMLS* 61 (4): 456–69.

Gaitonde, Supriya A., and Javier González-Maeso. 2017. "Contribution of Heteromerization to G Protein-Coupled Receptor Function." *Current Opinion in Pharmacology* 32 (February):23–31.

Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, et al. 2012. "ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery." *Nucleic Acids Research* 40 (Database issue): D1100–1107.

Gentile, Francesco, Vibudh Agrawal, Michael Hsing, Anh-Tien Ton, Fuqiang Ban, Ulf Norinder, Martin E. Gleave, and Artem Cherkasov. 2020. "Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery." *ACS Central Science* 6 (6): 939–49.

Gentile, Francesco, Jean Charle Yaacoub, James Gleave, Michael Fernandez, Anh-Tien Ton, Fuqiang Ban, Abraham Stern, and Artem Cherkasov. 2022. "Artificial Intelligence–enabled Virtual Screening of Ultra-Large Chemical Libraries with Deep Docking." *Nature Protocols* 17 (3): 672–97.

Goerg, Sebastian J., and Johannes Kaiser. 2009. "Nonparametric Testing of Distributions—the Epps–Singleton Two-Sample Test Using the Empirical Characteristic Function." *The Stata Journal* 9 (3): 454–65.

Greener, Joe G., and Michael J. E. Sternberg. 2015. "AlloPred: Prediction of Allosteric Pockets on Proteins Using Normal Mode Perturbation Analysis." *BMC Bioinformatics* 16 (October):335.

Gupta, Anku, Mohit Choudhary, Sanjay Kumar Mohanty, Aayushi Mittal, Krishan Gupta, Aditya Arya, Suvendu Kumar, et al. 2021. "Machine-OlF-Action: A Unified Framework for Developing and Interpreting Machine-Learning Models for Chemosensory Research." *Bioinformatics* , January. https://doi.org/10.1093/bioinformatics/btaa1104.

Gupta, Krishan, Sanjay Kumar Mohanty, Aayushi Mittal, Siddhant Kalra, Suvendu Kumar, Tripti Mishra, Jatin Ahuja, Debarka Sengupta, and Gaurav Ahuja. 2020. "The Cellular Basis of the Loss of Smell in 2019-nCoV-Infected Individuals." *Briefings in Bioinformatics*, August. https://doi.org/10.1093/bib/bbaa168.

Gurevich, Vsevolod V., and Eugenia V. Gurevich. 2018. "GPCRs and Signal Transducers: Interaction

Stoichiometry." *Trends in Pharmacological Sciences* 39 (7): 672–84.
Haghighatlari, Mojtaba, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava U. Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann. 2020. "ChemML : A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data." *Wiley Interdisciplinary Reviews. Computational Molecular Science* 10 (4). https://doi.org/10.1002/wcms.1458.
Hanchate, Naresh K., Kunio Kondoh, Zhonghua Lu, Donghui Kuang, Xiaolan Ye, Xiaojie Qiu, Lior Pachter, Cole Trapnell, and Linda B. Buck. 2015. "Single-Cell Transcriptomics Reveals Receptor Transformations during Olfactory Neurogenesis." *Science* 350 (6265): 1251–55.
Hanyaloglu, Aylin C., and Mark von Zastrow. 2008. "Regulation of GPCRs by Endocytic Membrane Trafficking and Its Potential Implications." *Annual Review of Pharmacology and Toxicology* 48 (1): 537–68.
Harrington, Richard A., Vyas Adhikari, M. Rayner, and P. Scarborough. 2019. "Nutrient Composition Databases in the Age of Big Data: foodDB, a Comprehensive, Real-Time Database Infrastructure." *BMJ Open* 9 (June). https://doi.org/10.1136/bmjopen-2018-026652.
Hartl, F. Ulrich. 2017. "Protein Misfolding Diseases." *Annual Review of Biochemistry* 86 (June):21–26.
Hastings, Janna, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, et al. 2013. "The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013." *Nucleic Acids Research* 41 (Database issue): D456–63.
Hauser, Alexander S., Misty M. Attwood, Mathias Rask-Andersen, Helgi B. Schiöth, and David E. Gloriam. 2017. "Trends in GPCR Drug Discovery: New Agents, Targets and Indications." *Nature Reviews. Drug Discovery* 16 (12): 829–42.
Hedderich, Janik B., Margherita Persechino, Katharina Becker, Franziska M. Heydenreich, Torben Gutermuth, Michel Bouvier, Moritz Bünemann, and Peter Kolb. 2022. "The Pocketome of G-Protein-Coupled Receptors Reveals Previously Untargeted Allosteric Sites." *Nature Communications* 13 (1): 2567.
Heese-Peck, Antje, Harald Pichler, Bettina Zanolari, Reika Watanabe, Günther Daum, and Howard Riezman. 2002. "Multiple Functions of Sterols in Yeast Endocytosis." *Molecular Biology of the Cell* 13 (8): 2664–80.
Heifetz, Alexander, Michelle Southey, Inaki Morao, Andrea Townsend-Nicholson, and Mike J. Bodkin. 2018. "Computational Methods Used in Hit-to-Lead and Lead Optimization Stages of Structure-Based Drug Discovery." *Methods in Molecular Biology (Clifton, N.J.)* 1705:375–94.
Heng, Boon Chin, Dominique Aubel, and Martin Fussenegger. 2013. "An Overview of the Diverse Roles of G-Protein Coupled Receptors (GPCRs) in the Pathophysiology of Various Human Diseases." *Biotechnology Advances* 31 (8): 1676–94.
He, Song-Bing, Ben Hu, Zheng-Kun Kuang, Dong Wang, and De-Xin Kong. 2016. "Predicting Subtype Selectivity for Adenosine Receptor Ligands with Three-Dimensional Biologically Relevant Spectrum (BRS-3D)." *Scientific Reports* 6 (November). https://doi.org/10.1038/srep36595.
Hilger, Daniel, Matthieu Masureel, and Brian K. Kobilka. 2018. "Structure and Dynamics of GPCR Signaling Complexes." *Nature Structural & Molecular Biology* 25 (1): 4–12.
Hoffmann, Carsten, and Moritz Bünemann. 2010. "Fluorescence and Resonance Energy Transfer Shine New Light on GPCR Function." *GPCR Molecular Pharmacology and Drug Targeting*. https://doi.org/10.1002/9780470627327.ch10.
Horn, Florence, Emmanuel Bettler, Laerte Oliveira, Fabien Campagne, Fred E. Cohen, and Gerrit Vriend. 2003. "GPCRDB Information System for G Protein-Coupled Receptors." *Nucleic Acids Research* 31 (1): 294–97.
Hou, Tianling, Yuemin Bian, Terence McGuire, and Xiang-Qun Xie. 2021. "Integrated Multi-Class Classification and Prediction of GPCR Allosteric Modulators by Machine Learning Intelligence." *Biomolecules* 11 (6). https://doi.org/10.3390/biom11060870.
Huang, Wenkang, Shaoyong Lu, Zhimin Huang, Xinyi Liu, Linkai Mou, Yu Luo, Yanlong Zhao, et al.

2013. “Allosite: A Method for Predicting Allosteric Sites.” *Bioinformatics* 29 (18): 2357–59.
Huang, Wenkang, Qiancheng Shen, Xubo Su, Mingfei Ji, Xinyi Liu, Yingyi Chen, Shaoyong Lu, Hanyi Zhuang, and Jian Zhang. 2016. “BitterX: A Tool for Understanding Bitter Taste in Humans.” *Scientific Reports* 6 (April). https://doi.org/10.1038/srep23450.
Hughes, J. P., S. Rees, S. B. Kalindjian, and K. L. Philpott. 2011. “Principles of Early Drug Discovery.” *British Journal of Pharmacology* 162 (6): 1239–49.
Hussain, Ashiq, Luis R. Saraiva, David M. Ferrero, Gaurav Ahuja, Venkatesh S. Krishna, Stephen D. Liberles, and Sigrun I. Korsching. 2013. “High-Affinity Olfactory Receptor for the Death-Associated Odor Cadaverine.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (48): 19579–84.
Inoue, Asuka, Francesco Raimondi, Francois Marie Ngako Kadji, Gurdeep Singh, Takayuki Kishi, Akiharu Uwamizu, Yuki Ono, et al. 2019. “Illuminating G-Protein-Coupling Selectivity of GPCRs.” *Cell* 177 (7): 1933–47.e25.
Insel, Paul A., Krishna Sriram, Shu Z. Wiley, Andrea Wilderman, Trishna Katakia, Thalia McCann, Hiroshi Yokouchi, et al. 2018. “GPCRomics: GPCR Expression in Cancer Cells and Tumors Identifies New, Potential Biomarkers and Therapeutic Targets.” *Frontiers in Pharmacology* 9 (May):431.
Irannejad, Roshanak, Jin C. Tomshine, Jon R. Tomshine, Michael Chevalier, Jacob P. Mahoney, Jan Steyaert, Søren G. F. Rasmussen, et al. 2013. “Conformational Biosensors Reveal GPCR Signalling from Endosomes.” *Nature* 495 (7442): 534–38.
Jabeen, Amara, and Shoba Ranganathan. 2019. “Applications of Machine Learning in GPCR Bioactive Ligand Discovery.” *Current Opinion in Structural Biology* 55 (April):66–76.
Jaeger, Werner C., Kevin D. G. Pfleger, and Karin A. Eidne. n.d. “Monitoring GPCR-Protein Complexes Using Bioluminescence Resonance Energy Transfer.” *G Protein-Coupled Receptors*. https://doi.org/10.1002/9780470749210.ch6.
Jastrzębski, Stanisław, Igor Sieradzki, Damian Leśniak, Jacek Tabor, Andrzej J. Bojarski, and Sabina Podlewska. 2019. “Three-Dimensional Descriptors for Aminergic GPCRs: Dependence on Docking Conformation and Crystal Structure.” *Molecular Diversity* 23 (3): 603–13.
Jensen, Ole N. 2006. “Interpreting the Protein Language Using Proteomics.” *Nature Reviews. Molecular Cell Biology* 7 (6): 391–403.
Jewison, Timothy, Craig Knox, Vanessa Neveu, Yannick Djoumbou, An Chi Guo, Jacqueline Lee, Philip Liu, et al. 2012. “YMDB: The Yeast Metabolome Database.” *Nucleic Acids Research* 40 (Database issue): D815–20.
Jin, Hui, J. Michael McCaffery, and Eric Grote. 2008. “Ergosterol Promotes Pheromone Signaling and Plasma Membrane Fusion in Mating Yeast.” *The Journal of Cell Biology* 180 (4): 813–26.
Johnston, Emily J., Tessa Moses, and Susan J. Rosser. 2020. “The Wide-Ranging Phenotypes of Ergosterol Biosynthesis Mutants, and Implications for Microbial Cell Factories.” *Yeast* 37 (1): 27–44.
Jo, Sunhwan, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. 2008. “CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM.” *Journal of Computational Chemistry* 29 (11): 1859–65.
Jovancevic, Nikolina, A. Dendorfer, M. Matzkies, M. Kovarova, J. C. Heckmann, M. Osterloh, M. Boehm, et al. 2017. “Medium-Chain Fatty Acids Modulate Myocardial Function via a Cardiac Odorant Receptor.” *Basic Research in Cardiology* 112 (2): 13.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.
Kalra, Siddhant, Aayushi Mittal, Krishan Gupta, Vrinda Singhal, Anku Gupta, Tripti Mishra, Srivatsava Naidu, Debarka Sengupta, and Gaurav Ahuja. 2020. “Analysis of Single-Cell Transcriptomes Links Enrichment of Olfactory Receptors with Cancer Cell Differentiation Status and Prognosis.” *Communications Biology* 3 (1): 506.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.
Kaneko, Hiroshi, Masako Hosohara, Masamichi Tanaka, and Toshihiro Itoh. 1976. “Lipid Composition of 30 Species of Yeast.” *Lipids* 11 (12): 837–44.
Katritch, Vsevolod, Vadim Cherezov, and Raymond C. Stevens. 2013. “Structure-Function of the G Protein-Coupled Receptor Superfamily.” *Annual Review of Pharmacology and Toxicology* 53 (1): 531–56.
Keller, Andreas, Richard C. Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D. Mainland, et al. 2017. “Predicting Human Olfactory Perception from Chemical Features of Odor Molecules.” *Science (New York, N.Y.)* 355 (6327): 820–26.
Khoury, George A., Richard C. Baliban, and Christodoulos A. Floudas. 2011. “Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database.” *Scientific Reports* 1 (1): 1–5.
Kipf, Thomas N., and Max Welling. 2016. “Semi-Supervised Classification with Graph Convolutional Networks.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1609.02907.
Kobilka, Brian. 2013. “The Structural Basis of G-Protein-Coupled Receptor Signaling (Nobel Lecture).” *Angewandte Chemie (International Ed. in English)* 52 (25): 6380–88.
Koes, David Ryan, and Carlos J. Camacho. 2012. “PocketQuery: Protein-Protein Interaction Inhibitor Starting Points from Protein-Protein Interaction Structure.” *Nucleic Acids Research* 40 (Web Server issue): W387–92.
Kooistra, Albert J., Stefan Mordalski, Gáspár Pándy-Szekeres, Mauricio Esguerra, Alibek Mamyrbekov, Christian Munk, György M. Keserű, and David E. Gloriam. 2021. “GPCRdb in 2021: Integrating GPCR Sequence, Structure and Function.” *Nucleic Acids Research* 49 (D1): D335–43.
Krishnan, Arunkumar, Markus Sällman Almén, Robert Fredriksson, and Helgi B. Schiöth. 2012. “The Origin of GPCRs: Identification of Mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in Fungi.” *PloS One* 7 (1): e29817.
Kroeze, Wesley K., Maria F. Sassano, Xi-Ping Huang, Katherine Lansu, John D. McCorvy, Patrick M. Giguère, Noah Sciaky, and Bryan L. Roth. 2015. “PRESTO-Tango as an Open-Source Resource for Interrogation of the Druggable Human GPCRome.” *Nature Structural & Molecular Biology* 22 (5): 362–69.
Kuriata, Aleksander, Aleksandra Maria Gierut, Tymoteusz Oleniecki, Maciej Pawel Ciemny, Andrzej Kolinski, Mateusz Kurcinski, and Sebastian Kmiecik. 2018. “CABS-Flex 2.0: A Web Server for Fast Simulations of Flexibility of Protein Structures.” *Nucleic Acids Research* 46 (W1): W338–43.
Kursa, Miron B., and Witold R. Rudnicki. 2010. “Feature Selection with theBorutaPackage.” *Journal of Statistical Software* 36 (11): 1–13.
La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. “RNA Velocity of Single Cells.” *Nature* 560 (7719): 494–98.
Landrum, Greg. n.d. “RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling.” Accessed August 11, 2023. http://www.rdkit.org/RDKit_Overview.pdf.
Lange, Marius, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, et al. 2022. “CellRank for Directed Single-Cell Fate Mapping.” *Nature Methods* 19 (2): 159–70.
Lappano, Rosamaria, and Marcello Maggiolini. 2011. “G Protein-Coupled Receptors: Novel Targets for Drug Discovery in Cancer.” *Nature Reviews. Drug Discovery* 10 (1): 47–60.
Laschet, Céline, Nadine Dupuis, and Julien Hanson. 2018. “The G Protein-Coupled Receptors Deorphanization Landscape.” *Biochemical Pharmacology* 153 (July):62–74.
Laskowski, R. A. 1995. “SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions.” *Journal of Molecular Graphics* 13 (5): 323–30, 307–8.
Laskowski, Roman A., and Mark B. Swindells. 2011. “LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery.” *Journal of Chemical Information and Modeling* 51 (10): 2778–86.
Latorraca, Naomi R., A. J. Venkatakrishnan, and Ron O. Dror. 2017. “GPCR Dynamics: Structures in

Motion." *Chemical Reviews* 117 (1): 139–55.
Lavecchia, Antonio. 2015. "Machine-Learning Approaches in Drug Discovery: Methods and Applications." *Drug Discovery Today* 20 (3): 318–31.
Leach, Katie, Patrick M. Sexton, and Arthur Christopoulos. 2007. "Allosteric GPCR Modulators: Taking Advantage of Permissive Receptor Pharmacology." *Trends in Pharmacological Sciences* 28 (8): 382–89.
Leclercq, Mickael, Benjamin Vittrant, Marie Laure Martin-Magniette, Marie Pier Scott Boyer, Olivier Perin, Alain Bergeron, Yves Fradet, and Arnaud Droit. 2019. "Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data." *Frontiers in Genetics* 10 (May):452.
Lee, Sung-Joon, Inge Depoortere, and Hanns Hatt. 2019. "Therapeutic Potential of Ectopic Olfactory and Taste Receptors." *Nature Reviews. Drug Discovery* 18 (2): 116–38.
Le Guilloux, Vincent, Peter Schmidtke, and Pierre Tuffery. 2009. "Fpocket: An Open Source Platform for Ligand Pocket Detection." *BMC Bioinformatics* 10 (June):168.
Levine, Michael, and Robert Tjian. 2003. "Transcription Regulation and Animal Diversity." *Nature* 424 (6945): 147–51.
Liao, Zhirui, Ronghui You, Xiaodi Huang, Xiaojun Yao, Tao Huang, and Shanfeng Zhu. 2019. "DeepDock: Enhancing Ligand-Protein Interaction Prediction by a Combination of Ligand and Structure Information." In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 311–17.
Linden, Carl J. van der, Pooja Gupta, Ashraful Islam Bhuiya, Kelci R. Riddick, Kawsar Hossain, and Stephen W. Santoro. 2020. "Olfactory Stimulation Regulates the Birth of Neurons That Express Specific Odorant Receptors." *Cell Reports* 33 (1): 108210.
Lionta, Evanthia, George Spyrou, Demetrios K. Vassilatis, and Zoe Cournia. 2014. "Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances." *Current Topics in Medicinal Chemistry* 14 (16): 1923–38.
Li, Qingliang, Tiejun Cheng, Yanli Wang, and Stephen H. Bryant. 2010. "PubChem as a Public Resource for Drug Discovery." *Drug Discovery Today* 15 (23-24): 1052–57.
Li, Yibo, Jianfeng Pei, and Luhua Lai. 2021. "Structure-Based de Novo Drug Design Using 3D Deep Generative Models." *Chemical Science* 12 (41): 13664–75.
Lomize, Mikhail A., Irina D. Pogozheva, Hyeon Joo, Henry I. Mosberg, and Andrei L. Lomize. 2012. "OPM Database and PPM Web Server: Resources for Positioning of Proteins in Membranes." *Nucleic Acids Research* 40 (Database issue): D370–76.
Lomvardas, Stavros, Gilad Barnea, David J. Pisapia, Monica Mendelsohn, Jennifer Kirkland, and Richard Axel. 2006. "Interchromosomal Interactions and Olfactory Receptor Choice." *Cell* 126 (2): 403–13.
Lötsch, Jörn, Dario Kringel, and Thomas Hummel. 2019. "Machine Learning in Human Olfactory Research." *Chemical Senses* 44 (1): 11–22.
Lykke-Andersen, Søren, and Torben Heick Jensen. 2015. "Nonsense-Mediated mRNA Decay: An Intricate Machinery That Shapes Transcriptomes." *Nature Reviews. Molecular Cell Biology* 16 (11): 665–77.
Lyons, David B., William E. Allen, Tracie Goh, Lulu Tsai, Gilad Barnea, and Stavros Lomvardas. 2013. "An Epigenetic Trap Stabilizes Singular Olfactory Receptor Expression." *Cell* 154 (2): 325–36.
Lyons, David B., Angeliki Magklara, Tracie Goh, Srihari C. Sampath, Anne Schaefer, Gunnar Schotta, and Stavros Lomvardas. 2014. "Heterochromatin-Mediated Gene Silencing Facilitates the Diversification of Olfactory Neurons." *Cell Reports* 9 (3): 884–92.
Lyu, Jiankun, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, et al. 2019. "Ultra-Large Library Docking for Discovering New Chemotypes." *Nature* 566 (7743): 224–29.
Malnic, Bettina, Paul A. Godfrey, and Linda B. Buck. 2004. "The Human Olfactory Receptor Gene Family." *Proceedings of the National Academy of Sciences of the United States of America* 101 (8): 2584–89.

Malnic, B., J. Hirono, T. Sato, and L. B. Buck. 1999. “Combinatorial Receptor Codes for Odors.” *Cell* 96 (5): 713–23.
Masjedi, Shirin, Laurence J. Zwiebel, and Todd D. Giorgio. 2019. “Olfactory Receptor Gene Abundance in Invasive Breast Carcinoma.” *Scientific Reports* 9 (1): 13736.
Maßberg, Désirée, Nikolina Jovancevic, Anne Offermann, Annika Simon, Aria Baniahmad, Sven Perner, Thanakorn Pungsrinont, et al. 2016. “The Activation of OR51E1 Causes Growth Suppression of Human Prostate Cancer Cells.” *Oncotarget* 7 (30): 48231–49.
Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. “Transcriptional Regulatory Elements in the Human Genome.” *Annual Review of Genomics and Human Genetics* 7 (1): 29–59.
Ma, Wenxiu, William S. Noble, and Timothy L. Bailey. 2014. “Motif-Based Analysis of Large Nucleotide Data Sets Using MEME-ChIP.” *Nature Protocols* 9 (6): 1428–50.
Mendoza, Alex de, Arnau Sebé-Pedrós, and Iñaki Ruiz-Trillo. 2014. “The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity.” *Genome Biology and Evolution* 6 (3): 606–19.
Michel, Martin C., and Steven J. Charlton. 2018. “Biased Agonism in Drug Discovery-Is It Too Soon to Choose a Path?” *Molecular Pharmacology* 93 (4): 259–65.
Mohanraj, Karthikeyan, Bagavathy Shanmugam Karthikeyan, R. P. Vivek-Ananth, R. P. Bharath Chand, S. R. Aparna, Pattulingam Mangalapandi, and Areejit Samal. 2018. “IMPPAT: A Curated Database of Indian Medicinal Plants, Phytochemistry And Therapeutics.” *Scientific Reports* 8 (1): 4329.
Mombaerts, Peter. 2004. “Odorant Receptor Gene Choice in Olfactory Sensory Neurons: The One Receptor-One Neuron Hypothesis Revisited.” *Current Opinion in Neurobiology* 14 (1): 31–36.
Monahan, Kevin, Ira Schieren, Jonah Cheung, Alice Mumbey-Wafula, Edwin S. Monuki, and Stavros Lomvardas. 2017. “Cooperative Interactions Enable Singular Olfactory Receptor Expression in Mouse Olfactory Neurons.” *eLife* 6 (September). https://doi.org/10.7554/eLife.28620.
Moran, D. T., J. C. Rowley 3rd, and B. W. Jafek. 1982. “Electron Microscopy of Human Olfactory Epithelium Reveals a New Cell Type: The Microvillar Cell.” *Brain Research* 253 (1-2): 39–46.
Moriwaki, Hirotomo, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. 2018. “Mordred: A Molecular Descriptor Calculator.” *Journal of Cheminformatics* 10 (1): 4.
Morris, Garrett M., Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. 2009. “AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility.” *Journal of Computational Chemistry* 30 (16): 2785–91.
Munk, Christian, Eshita Mutt, Vignir Isberg, Louise F. Nikolajsen, Janne M. Bibbe, Tilman Flock, Michael A. Hanson, Raymond C. Stevens, Xavier Deupi, and David E. Gloriam. 2019. “An Online Resource for GPCR Structure Determination and Analysis.” *Nature Methods* 16 (2): 151–62.
Mysinger, Michael M., Michael Carchia, John J. Irwin, and Brian K. Shoichet. 2012. “Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.” *Journal of Medicinal Chemistry* 55 (14): 6582–94.
Nadal-Ribelles, Mariona, Saiful Islam, Wu Wei, Pablo Latorre, Michelle Nguyen, Eulàlia de Nadal, Francesc Posas, and Lars M. Steinmetz. 2019. “Sensitive High-Throughput Single-Cell RNA-Seq Reveals within-Clonal Transcript Correlations in Yeast Populations.” *Nature Microbiology* 4 (4): 683–92.
Nakayama, N., Y. Kaziro, K. Arai, and K. Matsumoto. 1988. “Role of STE Genes in the Mating Factor Signaling Pathway Mediated by GPA1 in Saccharomyces Cerevisiae.” *Molecular and Cellular Biology* 8 (9): 3777–83.
Neumann, Elena, Kiran Khawaja, and Ulf Müller-Ladner. 2014. “G Protein-Coupled Receptors in Rheumatology.” *Nature Reviews. Rheumatology* 10 (7): 429–36.
Ngan, Chi Ho, Tanggis Bohnuud, Scott E. Mottarella, Dmitri Beglov, Elizabeth A. Villar, David R. Hall, Dima Kozakov, and Sandor Vajda. 2012. “FTMAP: Extended Protein Mapping with User-Selected Probe Molecules.” *Nucleic Acids Research* 40 (Web Server issue): W271–75.
Nieto Gutierrez, Ainhoa, and Patricia H. McDonald. 2018. “GPCRs: Emerging Anti-Cancer Drug Targets.” *Cellular Signalling* 41 (January):65–74.

Niimura, Yoshihito, Atsushi Matsui, and Kazushige Touhara. 2014. “Extreme Expansion of the Olfactory Receptor Gene Repertoire in African Elephants and Evolutionary Dynamics of Orthologous Gene Groups in 13 Placental Mammals.” *Genome Research* 24 (9): 1485–96.
Niimura, Yoshihito, and Masatoshi Nei. 2005. “Evolutionary Dynamics of Olfactory Receptor Genes in Fishes and Tetrapods.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (17): 6039–44.
Nirenberg, Marshall. 2004. “Historical Review: Deciphering the Genetic Code--a Personal Account.” *Trends in Biochemical Sciences* 29 (1): 46–54.
Nishibata, Yoshihiko, and Akiko Itai. 1991. “Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation.” *Tetrahedron* 47 (43): 8985–90.
Nozaki, Yuji, and Takamichi Nakamoto. 2018. “Predictive Modeling for Odor Character of a Chemical Using Machine Learning Combined with Natural Language Processing.” *PloS One* 13 (6): e0198475.
O’Boyle, Noel M., Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. 2011. “Open Babel: An Open Chemical Toolbox.” *Journal of Cheminformatics* 3 (1): 33.
O’Callaghan, Katie, Athan Kuliopulos, and Lidija Covic. 2012. “Turning Receptors on and off with Intracellular Pepducins: New Insights into G-Protein-Coupled Receptor Drug Development.” *The Journal of Biological Chemistry* 287 (16): 12787–96.
O’Hayre, Morgan, Maria S. Degese, and J. Silvio Gutkind. 2014. “Novel Insights into G Protein and G Protein-Coupled Receptor Signaling in Cancer.” *Current Opinion in Cell Biology* 27 (April):126–35.
Oldham, William M., and Heidi E. Hamm. 2008. “Heterotrimeric G Protein Activation by G-Protein-Coupled Receptors.” *Nature Reviews. Molecular Cell Biology* 9 (1): 60–71.
Ostrov, Nili, Miguel Jimenez, Sonja Billerbeck, James Brisbois, Joseph Matragrano, Alastair Ager, and Virginia W. Cornish. 2017. “A Modular Yeast Biosensor for Low-Cost Point-of-Care Pathogen Detection.” *Science Advances* 3 (6): e1603221.
Öz-Arslan, Devrim, Melis Yavuz, and Beki Kan. 2024. “Exploring Orphan GPCRs in Neurodegenerative Diseases.” *Frontiers in Pharmacology* 15 (June):1394516.
Páll, Szilárd, Artem Zhmurov, Paul Bauer, Mark Abraham, Magnus Lundborg, Alan Gray, Berk Hess, and Erik Lindahl. 2020. “Heterogeneous Parallelization and Acceleration of Molecular Dynamics Simulations in GROMACS.” *The Journal of Chemical Physics* 153 (13): 134110.
Pearlman, David A., and Mark A. Murcko. 1993. “CONCEPTS: New Dynamic Algorithm Forde Novo Drug Suggestion.” *Journal of Computational Chemistry* 14 (10): 1184–93.
Pirhadi, Somayeh, Jocelyn Sunseri, and David Ryan Koes. 2016. “Open Source Molecular Modeling.” *Journal of Molecular Graphics & Modelling* 69 (September):127–43.
Poynton, M. R., B. M. Choi, Y. M. Kim, I. S. Park, G. J. Noh, S. O. Hong, Y. K. Boo, and S. H. Kang. 2009. “Machine Learning Methods Applied to Pharmacokinetic Modelling of Remifentanil in Healthy Volunteers: A Multi-Method Comparison.” *The Journal of International Medical Research* 37 (6): 1680–91.
Ramsundar, B., P. Eastman, E. Feinberg, J. Gomes, K. Leswing, A. Pappu, M. Wu, and V. Pande. 2019. “DeepChem: Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry.” *Materials Science and Biology. Https://github. Com/deepchem/deepchem (accessed Aug 8, 2017)*.
Ranzani, Marco, Vivek Iyer, Ximena Ibarra-Soria, Martin Del Castillo Velasco-Herrera, Mathew Garnett, Darren Logan, and David J. Adams. 2017. “Revisiting Olfactory Receptors as Putative Drivers of Cancer.” *Wellcome Open Research* 2 (February):9.
Rees, Stephen, Dwight Morrow, and Terry Kenakin. 2002. “GPCR Drug Discovery Through the Exploitation of Allosteric Drug Binding Sites.” *Receptors and Channels*. https://doi.org/10.3109/10606820214640.
Reyes-Alcaraz, Arfaxad, Emilio Y. Lucero Garcia-Rojas, Richard A. Bond, and Bradley K. McConnell. 2020. “Allosteric Modulators for GPCRs as a Therapeutic Alternative with High Potential in Drug Discovery.” *Molecular Pharmacology*. https://doi.org/10.5772/intechopen.91838.
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?: Explaining

the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939778.
Rockman, Howard A., Walter J. Koch, and Robert J. Lefkowitz. 2002. "Seven-Transmembrane-Spanning Receptors and Heart Function." *Nature* 415 (6868): 206–12.
Romero, Guillermo, Mark von Zastrow, and Peter A. Friedman. 2011. "Role of PDZ Proteins in Regulating Trafficking, Signaling, and Function of GPCRs: Means, Motif, and Opportunity." *Advances in Pharmacology (San Diego, Calif.)* 62:279–314.
Rosenbaum, Daniel M., Søren G. F. Rasmussen, and Brian K. Kobilka. 2009. "The Structure and Function of G-Protein-Coupled Receptors." *Nature* 459 (7245): 356–63.
Roth, Bryan L., John J. Irwin, and Brian K. Shoichet. 2017. "Discovery of New GPCR Ligands to Illuminate New Biology." *Nature Chemical Biology* 13 (11): 1143–51.
Ruiz Tejada Segura, Mayra, Eman Abou Moussa, Elisa Garabello, Thiago S. Nakahara, Melanie Makhlouf, Lisa S. Mathew, Filippo Valle, et al. 2021. "A 3D Transcriptomics Atlas of the Mouse Olfactory Mucosa." *bioRxiv*. https://doi.org/10.1101/2021.06.16.448475.
Salmaso, Veronica, and Stefano Moro. 2018. "Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview." *Frontiers in Pharmacology* 9 (August):923.
Sanchez-Lengeling, Benjamin, Jennifer N. Wei, Brian K. Lee, Richard C. Gerkin, Alán Aspuru-Guzik, and Alexander B. Wiltschko. 2019. "Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1910.10685.
Santos, Rita, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologa, Anneli Karlsson, et al. 2017. "A Comprehensive Map of Molecular Drug Targets." *Nature Reviews. Drug Discovery* 16 (1): 19–34.
Saraiva, Luis R., Gaurav Ahuja, Ivan Ivandic, Adnan S. Syed, John C. Marioni, Sigrun I. Korsching, and Darren W. Logan. 2015. "Molecular and Neuronal Homology between the Olfactory Systems of Zebrafish and Mouse." *Scientific Reports* 5 (1): 11487.
Saraiva, Luis R., Ximena Ibarra-Soria, Mona Khan, Masayo Omura, Antonio Scialdone, Peter Mombaerts, John C. Marioni, and Darren W. Logan. 2015. "Hierarchical Deconstruction of Mouse Olfactory Sensory Neurons: From Whole Mucosa to Single-Cell RNA-Seq." *Scientific Reports* 5 (1): 18178.
Saraiva, Luis R., Fernando Riveros-McKay, Massimo Mezzavilla, Eman H. Abou-Moussa, Charles J. Arayata, Melanie Makhlouf, Casey Trimmer, et al. 2019. "A Transcriptomic Atlas of Mammalian Olfactory Mucosae Reveals an Evolutionary Influence on Food Odor Detection in Humans." *Science Advances* 5 (7): eaax0396.
Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33 (5): 495–502.
Scholz, Paul, Benjamin Kalbe, Fabian Jansen, Janine Altmueller, Christian Becker, Julia Mohrhardt, Benjamin Schreiner, Guenter Gisselmann, Hanns Hatt, and Sabrina Osterloh. 2016. "Transcriptome Analysis of Murine Olfactory Sensory Neurons during Development Using Single Cell RNA-Seq." *Chemical Senses* 41 (4): 313–23.
Serizawa, Shou, Kazunari Miyamichi, and Hitoshi Sakano. 2004. "One Neuron-One Receptor Rule in the Mouse Olfactory System." *Trends in Genetics: TIG* 20 (12): 648–53.
———. 2005. "Negative Feedback Regulation Ensures the One Neuron-One Receptor Rule in the Mouse Olfactory System." *Chemical Senses* 30 Suppl 1 (January):i99–100.
Severin, F. F., and A. A. Hyman. 2002. "Pheromone Induces Programmed Cell Death in S. Cerevisiae." *Current Biology: CB* 12 (7): R233–35.
Sharma, Kanika, Gaurav Ahuja, Ashiq Hussain, Sabine Balfanz, Arnd Baumann, and Sigrun I. Korsching. 2016. "Elimination of a Ligand Gating Site Generates a Supersensitive Olfactory Receptor." *Scientific Reports* 6 (1): 28359.
Shaw, William M., Hitoshi Yamauchi, Jack Mead, Glen-Oliver F. Gowers, David J. Bell, David Öling,

Niklas Larsson, Mark Wigglesworth, Graham Ladds, and Tom Ellis. 2019. “Engineering a Model Cell for Rational Tuning of GPCR Signaling.” *Cell* 177 (3): 782–96.e27.
Shen, Lingling, Jian Fang, Lulu Liu, Fei Yang, Jeremy L. Jenkins, Peter S. Kutchukian, and He Wang. 2024. “Pocket Crafter: A 3D Generative Modeling Based Workflow for the Rapid Generation of Hit Molecules in Drug Discovery.” *Journal of Cheminformatics* 16 (1): 33.
Shen, Min-Yi, and Andrej Sali. 2006. “Statistical Potential for Assessment and Prediction of Protein Structures.” *Protein Science: A Publication of the Protein Society* 15 (11): 2507–24.
Shen, Siyuan, Chang Zhao, Chao Wu, Suyue Sun, Ziyan Li, Wei Yan, and Zhenhua Shao. 2023. “Allosteric Modulation of G Protein-Coupled Receptor Signaling.” *Frontiers in Endocrinology* 14 (February):1137604.
Sicho, M., X. Liu, D. Svozil, and G. J. P. van Westen. 2021. “GenUI: Interactive and Extensible Open Source Software Platform for de Novo Molecular Generation and Cheminformatics.” *Journal of Cheminformatics* 13 (1): 73.
Sinha, Dona, Priyanka Saha, Anurima Samanta, and Anupam Bishayee. 2020. “Emerging Concepts of Hybrid Epithelial-to-Mesenchymal Transition in Cancer Progression.” *Biomolecules* 10 (11). https://doi.org/10.3390/biom10111561.
Skalic, Miha, Davide Sabbadin, Boris Sattarov, Simone Sciabola, and Gianni De Fabritiis. 2019. “From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design.” *Molecular Pharmaceutics* 16 (10): 4282–91.
Smith, Jeffrey S., Robert J. Lefkowitz, and Sudarshan Rajagopal. 2018. “Biased Signalling: From Simple Switches to Allosteric Microprocessors.” *Nature Reviews. Drug Discovery* 17 (4): 243–60.
Sokolov, S. S., K. V. Galkina, E. A. Litvinova, D. A. Knorre, and F. F. Severin. 2020. “The Role of LAM Genes in the Pheromone-Induced Cell Death of S. Cerevisiae Yeast.” *Biochemistry. Biokhimiia* 85 (3): 300–309.
Song, Kun, Xinyi Liu, Wenkang Huang, Shaoyong Lu, Qiancheng Shen, Lu Zhang, and Jian Zhang. 2017. “Improved Method for the Identification and Validation of Allosteric Sites.” *Journal of Chemical Information and Modeling* 57 (9): 2358–63.
Spehr, Marc, and Steven D. Munger. 2009. “Olfactory Receptors: G Protein-Coupled Receptors and beyond.” *Journal of Neurochemistry* 109 (6): 1570–83.
Spiegel, Jacob O., and Jacob D. Durrant. 2020. “AutoGrow4: An Open-Source Genetic Algorithm for de Novo Drug Design and Lead Optimization.” *Journal of Cheminformatics* 12 (1): 25.
Sprenger, K. G., Vance W. Jaeger, and Jim Pfaendtner. 2015. “The General AMBER Force Field (GAFF) Can Accurately Predict Thermodynamic and Transport Properties of Many Ionic Liquids.” *The Journal of Physical Chemistry. B* 119 (18): 5882–95.
Sriram, Krishna, and Paul A. Insel. 2018. “G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs?” *Molecular Pharmacology* 93 (4): 251–58.
Stevens, Raymond C., Vadim Cherezov, Vsevolod Katritch, Ruben Abagyan, Peter Kuhn, Hugh Rosen, and Kurt Wüthrich. 2013. “The GPCR Network: A Large-Scale Collaboration to Determine Human GPCR Structure and Function.” *Nature Reviews. Drug Discovery* 12 (1): 25–34.
Stornaiuolo, Mariano, Agostino Bruno, Lorenzo Botta, Giuseppe La Regina, Sandro Cosconati, Romano Silvestri, Luciana Marinelli, and Ettore Novellino. 2015. “Endogenous vs Exogenous Allosteric Modulators in GPCRs: A Dispute for Shuttling CB1 among Different Membrane Microenvironments.” *Scientific Reports* 5 (October):15453.
Tan, Kun, Samantha H. Jones, Blue B. Lake, Jennifer N. Dumdie, Eleen Y. Shum, Lingjuan Zhang, Song Chen, et al. 2020. “The Role of the NMD Factor UPF3B in Olfactory Sensory Neurons.” *eLife* 9 (August). https://doi.org/10.7554/eLife.57525.
Tan, Longzhi, Qian Li, and X. Sunney Xie. 2015. “Olfactory Sensory Neurons Transiently Express Multiple Olfactory Receptors during Development.” *Molecular Systems Biology*, December. https://doi.org/10.15252/msb.20156639.
Tan, Longzhi, and Xiaoliang Sunney Xie. 2018. “A Near-Complete Spatial Map of Olfactory Receptors in the Mouse Main Olfactory Epithelium.” *Chemical Senses* 43 (6): 427–32.

Temin, H. M., and S. Mizutani. 1970. “RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus.” *Nature* 226 (5252): 1211–13.
Teng, Xinchen, and J. Marie Hardwick. 2009. “Reliable Method for Detection of Programmed Cell Death in Yeast.” *Methods in Molecular Biology* 559:335–42.
Tetko, Igor V., Johann Gasteiger, Roberto Todeschini, Andrea Mauri, David Livingstone, Peter Ertl, Vladimir A. Palyulin, et al. 2005. “Virtual Computational Chemistry Laboratory--Design and Description.” *Journal of Computer-Aided Molecular Design* 19 (6): 453–63.
Tian, Chuan, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migues, John Bickel, et al. 2020. “ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution.” *Journal of Chemical Theory and Computation* 16 (1): 528–52.
Tian, Wei, Chang Chen, Xue Lei, Jieling Zhao, and Jie Liang. 2018. “CASTp 3.0: Computed Atlas of Surface Topography of Proteins.” *Nucleic Acids Research* 46 (W1): W363–67.
Tiedje, Christopher, Daniel G. Holland, Ursula Just, and Thomas Höfken. 2007. “Proteins Involved in Sterol Synthesis Interact with Ste20 and Regulate Cell Polarity.” *Journal of Cell Science* 120 (Pt 20): 3613–24.
Topiol, Sid. 2018a. “Current and Future Challenges in GPCR Drug Discovery.” *Methods in Molecular Biology* 1705:1–21.
———. 2018b. “New Opportunities for GPCR Allosteric Modulators.” *Future Medicinal Chemistry*. https://doi.org/10.4155/fmc-2017-0313.
Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86.
Trott, Oleg, and Arthur J. Olson. 2010. “AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.” *Journal of Computational Chemistry* 31 (2): 455–61.
Vamathevan, Jessica, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, et al. 2019. “Applications of Machine Learning in Drug Discovery and Development.” *Nature Reviews. Drug Discovery* 18 (6): 463–77.
Vassart, Gilbert, and Sabine Costagliola. 2011. “G Protein-Coupled Receptors: Mutations and Endocrine Diseases.” *Nature Reviews. Endocrinology* 7 (6): 362–72.
Velazhahan, Vaithish, Ning Ma, Gáspár Pándy-Szekeres, Albert J. Kooistra, Yang Lee, David E. Gloriam, Nagarajan Vaidehi, and Christopher G. Tate. 2021. “Structure of the Class D GPCR Ste2 Dimer Coupled to Two G Proteins.” *Nature* 589 (7840): 148–53.
Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. “Graph Attention Networks.” *Https://openreview.net › Forumhttps://openreview.net › Forum*. https://openreview.net/pdf?id=rJXMpikCZ.
Venkatakrishnan, A. J., Xavier Deupi, Guillaume Lebon, Christopher G. Tate, Gebhard F. Schertler, and M. Madan Babu. 2013. “Molecular Signatures of G-Protein-Coupled Receptors.” *Nature* 494 (7436): 185–94.
Volkamer, Andrea, Daniel Kuhn, Friedrich Rippmann, and Matthias Rarey. 2012. “DoGSiteScorer: A Web Server for Automatic Binding Site Prediction, Analysis and Druggability Assessment.” *Bioinformatics* 28 (15): 2074–75.
Wang, Hao, and Weitao Yang. 2018. “Force Field for Water Based on Neural Network.” *Journal of Physical Chemistry Letters* 9 (12): 3232–40.
Wang, Shiwei, Juan Xie, Jianfeng Pei, and Luhua Lai. 2023. “CavityPlus 2022 Update: An Integrated Platform for Comprehensive Protein Cavity Detection and Property Analyses with User-Friendly Tools and Cavity Databases.” *Journal of Molecular Biology* 435 (14): 168141.
Weis, William I., and Brian K. Kobilka. 2018. “The Molecular Basis of G Protein-Coupled Receptor Activation.” *Annual Review of Biochemistry* 87 (1): 897–919.

Westhuizen, Emma T. van der, Celine Valant, Patrick M. Sexton, and Arthur Christopoulos. 2015. "Endogenous Allosteric Modulators of G Protein–Coupled Receptors." *The Journal of Pharmacology and Experimental Therapeutics* 353 (2): 246–60.
Wettschureck, Nina, and Stefan Offermanns. 2005. "Mammalian G Proteins and Their Cell Type Specific Functions." *Physiological Reviews* 85 (4): 1159–1204.
Wishart, D., Y. D. Feunang, A. Marcu, Anchi Guo, Kevin Y. H. Liang, R. Vázquez-Fresno, Tanvir Sajed, et al. 2017. "HMDB 4.0: The Human Metabolome Database for 2018." *Nucleic Acids Research* 46 (D1): D608–17.
Wold, Eric A., and Jia Zhou. 2018. "GPCR Allosteric Modulators: Mechanistic Advantages and Therapeutic Applications." *Current Topics in Medicinal Chemistry* 18 (23): 2002–6.
Wolf, F. Alexander, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. 2019. "PAGA: Graph Abstraction Reconciles Clustering with Trajectory Inference through a Topology Preserving Map of Single Cells." *Genome Biology* 20 (1): 59.
Wolock, Samuel L., Romain Lopez, and Allon M. Klein. 2019. "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data." *Cell Systems* 8 (4): 281–91.e9.
Wootten, Denise, Arthur Christopoulos, Maria Marti-Solano, M. Madan Babu, and Patrick M. Sexton. 2018. "Mechanisms of Signalling and Biased Agonism in G Protein-Coupled Receptors." *Nature Reviews. Molecular Cell Biology* 19 (10): 638–53.
Wu, Victoria, Huwate Yeerna, Nijiro Nohata, Joshua Chiou, Olivier Harismendy, Francesco Raimondi, Asuka Inoue, Robert B. Russell, Pablo Tamayo, and J. Silvio Gutkind. 2019. "Illuminating the Onco-GPCRome: Novel G Protein-Coupled Receptor-Driven Oncocrine Networks and Targets for Cancer Immunotherapy." *The Journal of Biological Chemistry* 294 (29): 11062–86.
Xiong, Zhaoping, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, et al. 2020. "Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism." *Journal of Medicinal Chemistry* 63 (16): 8749–60.
Xue, Li C., João Pglm Rodrigues, Panagiotis L. Kastritis, Alexandre Mjj Bonvin, and Anna Vangone. 2016. "PRODIGY: A Web Server for Predicting the Binding Affinity of Protein-Protein Complexes." *Bioinformatics* 32 (23): 3676–78.
Xu, Mingyuan, Ting Ran, and Hongming Chen. 2021. "De Novo Molecule Design Through the Molecular Generative Model Conditioned by 3D Information of Protein Binding Sites." *Journal of Chemical Information and Modeling* 61 (7): 3240–54.
Yang, Dehua, Qingtong Zhou, Viktorija Labroska, Shanshan Qin, Sanaz Darbalaei, Yiran Wu, Elita Yuliantie, et al. 2021. "G Protein-Coupled Receptors: Structure- and Function-Based Drug Discovery." *Signal Transduction and Targeted Therapy* 6 (1): 1–27.
Yap, Chun Wei. 2011. "PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints." *Journal of Computational Chemistry* 32 (7): 1466–74.
Yuan, Huating, Min Yan, Guanxiong Zhang, W. Liu, Chunyu Deng, Gaoming Liao, Liwen Xu, et al. 2018. "CancerSEA: A Cancer Single-Cell State Atlas." *Nucleic Acids Research* 47 (D1): D900–908.
Yuan, Yaxia, Jianfeng Pei, and Luhua Lai. 2011. "LigBuilder 2: A Practical de Novo Drug Design Approach." *Journal of Chemical Information and Modeling* 51 (5): 1083–91.
———. 2020. "LigBuilder V3: A Multi-Target de Novo Drug Design Approach." *Frontiers in Chemistry* 8 (February):142.
Zha, Jinyin, Qian Li, Xinyi Liu, Weidong Lin, Tingting Wang, Jiacheng Wei, Ziliang Zhang, et al. 2023. "AlloReverse: Multiscale Understanding among Hierarchical Allosteric Regulations." *Nucleic Acids Research* 51 (W1): W33–38.
Zhang, Jian, Jianyi Yang, Richard Jang, and Yang Zhang. 2015. "GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome." *Structure (London, England: 1993)* 23 (8): 1538–49.
Zhang, Nan-Nan, Drew D. Dudgeon, Saurabh Paliwal, Andre Levchenko, Eric Grote, and Kyle W. Cunningham. 2006. "Multiple Signaling Pathways Regulate Yeast Cell Death during the Response to

Mating Pheromones." *Molecular Biology of the Cell* 17 (8): 3409–22.
Zhang, Ping, Lidija Covic, and Athan Kuliopulos. 2015. "Pepducins and Other Lipidated Peptides as Mechanistic Probes and Therapeutics." *Methods in Molecular Biology* 1324:191–203.
Zhavoronkov, Alex, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, et al. 2019. "Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors." *Nature Biotechnology* 37 (9): 1038–40.
Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 8 (January):14049.
Zhou, Yiwei, Jiyong Meng, Chanjuan Xu, and Jianfeng Liu. 2021. "Multiple GPCR Functional Assays Based on Resonance Energy Transfer Sensors." *Frontiers in Cell and Developmental Biology* 9 (May):611443.
Zundert, G. C. P. van, J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. de Vries, and A. M. J. J. Bonvin. 2016. "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes." *Journal of Molecular Biology* 428 (4): 720–25.