



**Unraveling the Diverse Roles of Cellular Metabolites Using Advanced
Computational Approaches**

By:

Aayushi Mittal

PhD19203

Under the Supervision of:

Dr. Gaurav Ahuja

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

September, 2025



**Unraveling the Diverse Roles of Cellular Metabolites Using Advanced
Computational Approaches**

By:

Aayushi Mittal

PhD19203

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

September, 2025

Certificate

This is to certify that the thesis titled “**Unraveling the Diverse Roles of Cellular Metabolites Using Advanced Computational Approaches**” being submitted by **Ms. Aayushi Mittal** 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 her 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.

September 2025

A handwritten signature in black ink, appearing to read 'Gaurav Ahuja', with a stylized flourish at the end.

Dr. Gaurav Ahuja

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

Acknowledgment

First and foremost, I extend my deepest gratitude to my advisor, Dr. Gaurav Ahuja. His scientific insights, guidance and expertise have been instrumental in shaping this thesis, encouraging me to explore new ideas and refine my research. The journey has been intense and, at times, challenging, but through it all, I have grown immensely as a researcher. I also appreciate his patience and understanding during the times I struggled, allowing me the space to find my way while offering valuable insights when needed. I am also grateful for the opportunities he provided me to work in different research environments, both nationally and internationally, which broadened my scientific perspective and enriched my research journey.

I am also immensely thankful to IIIT Delhi for providing an excellent research environment that fostered learning, collaboration, and the occasional moment of scientific clarity. My department has been instrumental in this journey, and I appreciate the faculty, administrative staff, and fellow researchers who have made this place a hub for curiosity and discovery. A special thanks to the IT department for ensuring that our computational experiments ran smoothly (or at least as smoothly as one can hope in research).

This research would not have been possible without financial support, and I sincerely thank my funding agency, CSIR, for providing the resources necessary to carry out my work. Their support enabled me to fully dedicate myself to research, access essential datasets and tools, and present my findings at conferences.

I extend my gratitude to my internal examiners, Dr. Debarka Sengupta and Dr. Vibhor Kumar, for their insightful feedback and thought-provoking questions, which helped refine my work and challenged me to see my research from different perspectives. Their meticulous review and suggestions have strengthened this thesis, and I deeply appreciate the time and effort they put into evaluating it.

A huge shout-out to my labmates and friends who have made this journey far more enjoyable. Science is best done in good company, and I have been lucky to have shared this experience with an amazing group of people. The countless discussions, troubleshooting and de-bugging sessions, and shared moments of triumph (and frustration) have been an integral part of this Ph.D. A special mention to one labmate and friend, you know who you are, who has been like a brother to me. From our shared passions for anime and Asian food to late-night work marathons and even late-night dinners after submissions, this Ph.D. would have been far less bearable without him. Thank you for the endless conversations, the unwavering support, and for making even the most stressful days a little more fun.

To my two best friends, Vedika and Shifa, thank you for keeping me sane. You have been my escape from the chaos of research, my reality check, and my much-needed source of laughter. Whether it was listening to my rants or simply reminding me that life exists outside the lab, your support has meant the world.

Finally, to my family, thank you for everything. To my mother, for her constant encouragement and motivation, reminding me why I started this journey in the first place. To my father, whose quiet strength and unwavering faith in me have been a source of immense support. To my brother, for being the voice of reason when I got too caught up in my own head. And to my entire family, for their love, patience, and, yes, for constantly asking when I would finally be "done." Your support has been my foundation, and this thesis is as much yours as it is mine.

To everyone who has been part of this journey, thank you. It has been long, exhausting, and at times chaotic, but I wouldn't have made it through without each of you.

Aayushi Mittal

A handwritten signature in black ink, appearing to read 'Aayushi' followed by a stylized flourish and a period.

Abstract

Cellular metabolites, traditionally seen as intermediates in metabolic pathways, perform diverse roles that extend into cellular signaling, homeostasis, and disease progression. This study deciphers the multifaceted roles of metabolites as internal threats, signaling modulators, and regulators of complex cellular networks, employing a multidisciplinary approach that combines computational modeling, deep learning frameworks, and experimental validations.

In one of the objectives, we developed a novel computational framework, Metabokiller, to predict the carcinogenic potential of metabolites by integrating biochemical properties such as oxidative stress induction, genomic instability, anti-apoptotic activity, and epigenetic alterations. By leveraging a robust ensemble learning approach, Metabokiller identified a range of endogenous metabolites with carcinogenic potential. Experimental validation of two predicted metabolites, 4-nitrocatechol and 3,4-dihydroxyphenylacetic acid, demonstrated their ability to induce genotoxicity, mutagenesis, and malignant transformation in yeast and human cell lines. These findings underscore the significant yet underexplored role of metabolites as internal carcinogenic threats, advancing the predictive capabilities of computational models for toxicology and drug safety.

In another objective involving the field of olfactory and GPCR signaling, two advanced platforms, OdoriFy and EvOlf, were developed to decode ligand-receptor interactions with high precision. OdoriFy is an artificial intelligence-driven platform designed to decode the molecular interactions between human odorant receptors and their ligands. Through its four predictive engines, Odorant Predictor, Odor Finder, OR Finder, and Odorant-OR Pair Analysis, OdoriFy enables the systematic identification and characterization of odorant-receptor interactions, offering mechanistic insights through explainable artificial intelligence. Extending beyond odorant research, EvOlf, a unified deep learning framework, was developed to predict ligand interactions across a wide spectrum of GPCRs, including both odorant and non-odorant receptors. With 105,235 GPCR-ligand interaction data points spanning multiple species, EvOlf bridges evolutionary diversity and functional specificity, highlighting the role of GPCRs as key mediators of metabolite signaling and their potential as therapeutic targets.

Further investigations into the regulatory roles of metabolites revealed their capacity to modulate GPCR signaling pathways. In *Saccharomyces cerevisiae*, the ability of endogenous metabolites, including zymosterol, lanosterol, and ubiquinone-6, to act as allosteric regulators of the pheromone receptor Ste2p was explored. These metabolites were found to suppress pheromone-induced programmed cell death (PCD) by targeting the Ste2p-Gpa1p interface, a critical node in GPCR signaling. Experimental approaches, including growth kinetics, cell viability assays, shmoo formation, and MAPK activation,

consistently demonstrated the ability of these metabolites to attenuate Ste2p-mediated signaling. Site-directed mutagenesis of key residues at the Ste2p-Gpa1p interface further confirmed the mechanistic basis of metabolite-mediated regulation, underscoring their specificity and functional importance.

This work takes a multi-faceted approach by integrating computational and experimental methodologies to uncover the diverse roles of metabolites in health and disease. It not only advances our understanding of how metabolites function as internal threats and signaling regulators but also highlights their therapeutic potential in targeting complex biological systems. By decoding the biochemical and molecular mechanisms underlying metabolite interactions, these works lay the groundwork for future studies aimed at exploiting metabolites as biomarkers, therapeutic agents, and tools for predictive toxicology, opening new avenues in cancer biology, chemosensory research, and GPCR-based drug discovery.

List of Publications ([link](#))

Thesis Related Publications (5)

1. Mohanty Sanjay Kumar*, **Aayushi Mittal***, Namra*, Aakash Gaur, Subhadeep Duari, Saveena Solanki, Anmol Kumar Sharma, et al. 2025 “Deep Learning Reveals Endogenous Sterols as Allosteric Modulators of the GPCR-Gα Interface” *eLife* 14:RP106397 (**co-first author; reviewed pre-print**)
2. **Mittal, Aayushi**, Sanjay Kumar Mohanty, and Mudit Gupta. 2023 “EvOlf: A Multimodal Deorphanization Approach for Therapeutically Relevant Odorant Receptors.” *Chemical Senses*. bjad041.010. (**first author; abstract**)
3. **Mittal, Aayushi**, and Gaurav Ahuja. 2023. “Advancing Chemical Carcinogenicity Prediction Modeling: Opportunities and Challenges.” *Trends in Pharmacological Sciences* 44 (7): 400–410. (**first author; co-corresponding author; opinion**)
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). (**first author**)
5. Gupta, Ria*, **Aayushi Mittal***, Vishesh Agrawal*, Sushant Gupta, Krishan Gupta, Rishi Raj Jain, Prakriti Garg, et al. 2021. “OdoriFy: A Conglomerate of Artificial Intelligence-Driven Prediction Engines for Olfactory Decoding.” *The Journal of Biological Chemistry* 297 (2): 100956. (**co-first author**)

Other Publications (12)

1. 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 (2): 115270. (**contributing author; resource**)
2. 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. (**co-first author; technical report**)
3. 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. (**co-first author; review**)
4. Bhattacharya, Namrata*, Swagatam Chakraborti*, Krishan Gupta*, **Aayushi Mittal***, Debajyoti Sinha, Colleen Nelson, Tanmoy Chakraborty, Gaurav Ahuja, and Debarka Sengupta. 2022.

“Simultaneous Visualization of Cells and Marker Genes from scRNA-Seq Studies.” *bioRxiv*. (co-first author; pre-print)

5. Goel, Mansi*, **Aayushi Mittal***, Vijaya R. Jain, Akshay Bharadwaj, Shivani Modi, Gaurav Ahuja, Ankur Jain, and Kaushal Kumar. 2023 “Integrative Functions of the Hypothalamus: Linking Cognition, Emotion and Physiology for Well-Being and Adaptability” *Annals of Neurosciences*. 2024;0(0). (co-first author; review)
6. 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. (contributing author)
7. 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)
8. 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). (co-first author)
9. 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)
10. Kalra, Siddhant*, **Aayushi Mittal***, Manisha Bajoria, Tripti Mishra, Sidrah Maryam, Debarka Sengupta, and Gaurav Ahuja. 2021. “Challenges and Possible Solutions for Decoding Extranasal Olfactory Receptors.” *The FEBS Journal* 288 (14): 4230–41. (co-first author; viewpoint)
11. 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)
12. 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. (co-first author)

Table of Contents

Certificate	2
Acknowledgment	3
Abstract	5
List of Publications (link)	7
Thesis Related Publications (5)	7
Other Publications (12)	7
Table of Contents	9
List of Figures	12
List of Tables	14
Chapter 1 - Introduction	1
1.1 Objective of the Thesis	5
1.2 Organization of the Chapters	5
Chapter 2 - Decoding the Functional Role of Cellular Metabolites as Endogenous Carcinogens	7
2.1 Introduction	7
2.2 Materials and Methods	10
2.2.1 Data Compilation	10
2.2.2 Model Building	12
2.2.3 Strains and Growth Conditions	13
2.2.4 Comet Assay for Genotoxicity Evaluation	13
2.2.5 Mutagenesis Assay	14
2.2.6 Cell Death Assay	15
2.2.7 Apoptosis assay	15
2.2.8 Reactive Oxygen Species (ROS) Quantification	16
2.2.9 Selection of Canavanine-Resistant Mutants with Accelerated Cell Division	16
2.2.10 RNA Isolation	17
2.2.11 RNA Sequencing Analysis	17
2.2.12 Mutation Analysis	18
2.2.13 Soft-Agar Assay	18
2.2.14 Data and Code Availability	18
2.3 Results	19
2.3.1 Metabokiller: A Novel Approach to Carcinogenicity Prediction	19
2.3.2 Metabokiller Outperforms Other Models in Carcinogenicity Prediction	27
2.3.3 Identification of Potential Endogenous Metabolic Threats by Metabokiller	29
2.3.4 Experimental Validation of Metabokiller Predictions.	32
2.3.5 Transcriptomic Analysis of 4NC and DP-Induced Mutants	39
2.3.6 4NC and DP Induce Malignant Transformation in Human Cells.	41
2.4 Discussion	42
Chapter 3 - Identification of Endogenous Activators of Cellular GPCRs Using Deep Learning Approaches	46

3.1 Introduction	46
3.2 Materials and Methods	49
3.2.1 OdoriFy: Methods and Implementation	49
3.2.1.1 Data Compilation	49
3.2.1.2 Prediction Engine: Odorant Predictor	50
3.2.1.3 Prediction Engine: Odor Finder, OR Finder, and Odorant-OR Pair Analysis	51
3.2.1.4 Implementation of the frontend	53
3.2.2 EvOlf: Methods and Implementation	54
3.2.2.1 Data Compilation	54
3.2.2.2 Feature Extraction	54
3.2.2.3 Data Processing and Model Building	55
3.2.2.4 Deep Learning Architectures	56
3.2.2.5 Data and Code Availability	59
3.3 Results	60
3.3.1 Development of an AI model to Predict Human-Specific Odorant-OR Pairs	60
3.3.1.1 OdoriFy: An AI-Driven Platform for Human Olfactory Research	60
3.3.1.2 Odorant Predictor: A Dual Function Tool for Odorant Identification and Explanation	63
3.3.1.3 Predicting Odorant-Odorant Receptor Interactions Using OdoriFy	66
3.3.1.4 Comprehensive Investigation of Odorant-OR Interactions Using OdoriFy	67
3.3.1.5 OdoriFy Synergizes with the Structure-Based OR-Ligand Interaction Method and Outperforms Other Prediction Tools	69
3.3.2 Generation of Unified, Evolutionary-Driven GPCR-Ligand Pair Prediction Models Using Deep Learning Approach	73
3.3.2.1 EvOlf: Data-Driven Profiling of GPCR-Ligand Interactions	73
3.3.2.2 Multi-dimensional Feature Integration for GPCR-Ligand Interaction Prediction in EvOlf	77
3.3.2.3 Comprehensive GPCR-Ligand Interaction Prediction Using Multi-Layered Neural Architectures	79
3.3.2.4 Unified Models for GPCR-Ligand Interactions: Leveraging Diversity in Receptor Classes and Species	83
3.3.2.5 Unified High-Dimensional Embeddings for GPCR-Ligand Prediction	86
3.3.2.6 EvOlf Outperforms Conventional Methods in GPCR-Ligand Interaction Prediction	88
3.4 Discussion	90
Chapter 4 - Dissecting the Role of Endogenous Metabolites as Key Allosteric Regulators of GPCR Signaling in Yeast	94
4.1 Introduction	94
4.2 Materials and Methods	96
4.2.1 Yeast Strains	96
4.2.2 Pre-loading of Yeast Cells with Metabolite	97
4.2.3 Cell Viability Assay using Growth Kinetics	97
4.2.4 Cell Viability Assay Using Propidium Iodide	97

4.2.5 Cell Viability Assay Using FUN TM 1 Staining	98
4.2.6 Mating Assay	99
4.2.7 Phospho-MAPK Activity-Based Western Blot	99
4.2.8 Transgenic Reporter Assay	101
4.2.9 Site-Directed Mutagenesis	102
4.2.10 Cellular Morphometric Assay (Shmoo formation)	103
4.3 Results	103
4.3.1 Elevated metabolic levels selectively inhibit GPCR signaling	103
4.3.2 Site-Directed Mutants Abrogate Metabolite-Mediated Rescue Phenotype	109
4.4 Discussion	113
Chapter 5 - Conclusion and Summary	116
References	118

List of Figures

Chapter 2 - Decoding the Functional Role of Cellular Metabolites as Endogenous Carcinogens	7
Figure 2.1 Dataset Composition for Biochemical Property-Based Models in Metabokiller.	19
Figure 2.2 Chemical Heterogeneity and Functional Group Enrichment in Biochemical Properties of Carcinogenicity	20
Figure 2.3 Workflow Detailing Metabokiller Functionalities.	22
Figure 2.4 Performance Metrics and Validation of Biochemical Property-Based Models	24
Figure 2.5 Integrative Framework of Metabokiller: Leveraging Biochemical Properties for Robust Carcinogenicity Prediction	25
Figure 2.6 Metabokiller's Robust Performance on External Testing Datasets	26
Figure 2.7 Metabokiller Outperforms Other Prediction Methods.	27
Figure 2.8 Uncovering Carcinogenic Potential of Human Metabolites through Metabokiller	29
Figure 2.9 Metabokiller Unravels Potential Oncometabolites.	32
Figure 2.10 Metabokiller's Insights into Carcinogenic Potential of 4NC and DP	33
Figure 2.11 Experimental Validation of Genotoxicity and Mutagenic Potential of 4NC and DP	34
Figure 2.12 Mechanistic Insights into 4NC-induced Genotoxicity	36
Figure 2.13 4NC and DP Trigger an Anti-apoptotic Response in Yeast.	37
Figure 2.14 RNA Sequencing Reveals Differential Transcriptomic Profiles Induced by 4NC and DP Treatments.	39
Figure 2.15 Gene Ontology Analysis Highlights Pathways Impacted by 4NC and DP	40
Figure 2.16 4NC and DP Trigger Malignant Transformation of Human Cells	42
Chapter 3 - Identification of Endogenous Activators of Cellular GPCRs Using Deep Learning Approaches	46
Figure 3.1 Graphical User Interface of OdoriFy webserver.	63
Figure 3.2 Odorant Predictor allows rapid classification of user-supplemented chemicals into odorant or non-odorants.	65
Figure 3.3 Odorant Predictor allows explainability at the sub-structural levels.	66
Figure 3.4 A unified deep neural network-based prediction model supports Odor Finder, OR Finder, and Odorant-OR Pair Analysis	69
Figure 3.5 Cross-validation of the top predicted agonists/non-agonists of OR1A1 using an orthogonal structure-based approach.	72
Figure 3.6 EvOlf Features a Diverse GPCR-Ligand Dataset Across Multiple Species and Receptor Classes	74
Figure 3.7 GPCR and Ligand Diversity in the EvOlf Dataset	76
Figure 3.8 Multifaceted Data Representation for GPCR-Ligand Interaction Modeling in EvOlf	78
Figure 3.9 Comparative Analysis of Deep Learning Architectures for GPCR-Ligand Interaction Prediction	82
Figure 3.10 Integrating ORs and Non-ORs Improves Predictive Power	84
Figure 3.11 Evaluating the Impact of Multi-Species Data on GPCR-Ligand Interaction Prediction	85
Figure 3.12 EvOlf Generates Joint Embeddings for a Comprehensive GPCR-Ligand Interaction Space	

Figure 3.13 EvOlf Outperforms Existing Tools in GPCR-Ligand Predictions	89
---	----

Chapter 4 - Dissecting the Role of Endogenous Metabolites as Key Allosteric Regulators of GPCR Signaling in Yeast **94**

Figure 4.1 Growth Curve Profiles of Metabolite Pre-treated Yeast	104
Figure 4.2 Exogenous Metabolites Treatment Resists α -Factor-Induced Programmed Cell Death	107
Figure 4.3 Metabolite-Induced Modulation of Yeast Mating and Ste2p Signaling	108
Figure 4.4 Impact of Site-Directed Mutations and Metabolites on Ste2p Signaling and α -Factor-Induced PCD	110
Figure 4.5 Effects of Ste2 Mutants on Shmoo Formation in Response to α -Factor.	112

List of Tables

Chapter 2 - Decoding the Functional Role of Cellular Metabolites as Endogenous Carcinogens	7
Table 2.1 Summary of datasets used for the development and evaluation of Metabokiller.	12
Table 2.2 Performance of Feature and Model Combinations for Biochemical Properties	23
Chapter 4 - Dissecting the Role of Endogenous Metabolites as Key Allosteric Regulators of GPCR Signaling in Yeast	94
Table 4.1 The primers used in PCR amplification for pRS306-P _{FUS1} -eGFP-CYC1	101
Table 4.2 The primers used in cloning and site-directed mutagenesis	103

Chapter 1 - Introduction

Deoxyribonucleic acid, or DNA, is the molecular repository of a cell's genetic information. It is a linear polymer that gets passed along from one generation to the next, carrying instructions programmed in its nucleotide sequence to form all biomolecules and other cellular components. The groundbreaking discovery of the DNA's double-helical structure by Watson and Crick has revolutionized our understanding of cellular structure and function, paving the way for modern biology (Watson and Crick 1953). This discovery also suggested that DNA could replicate itself, explaining how genetic information can be copied and transmitted across generations. The discoveries of messenger RNA (mRNA) and transfer RNA (tRNA), along with the deciphering of the genetic code, provided further insight into how the information from DNA can be translated into proteins. All these advances, along with many more, led to the conceptualization of the popular theory of the central dogma of life. Proposed by Francis Crick in 1958, the central dogma of life describes the flow of genetic information in cells from DNA to mRNA to protein. It has been a cornerstone in our understanding of cellular biology.

However, this dogma inadvertently sidelined a critical class of biomolecules: metabolites (Schreiber 2005). As end products of enzymatic reactions and intermediates in metabolic pathways, metabolites are not mere bystanders but active participants that influence cellular behavior, signaling, and phenotypic outcomes. Positioned at the intersection of biochemical processes, metabolites act as central nodes, orchestrating cellular homeostasis, energy production, and regulatory networks (Guijas et al. 2018; Rinschen et al. 2019). Despite their pivotal role, their functional diversity and mechanistic contributions remain poorly understood compared to nucleic acids and proteins (Guijas et al. 2018).

Metabolites actively interact with all three major families of macromolecules, DNA, RNA, and proteins, modulating various cellular processes (Schreiber 2005). Metabolites can affect DNA regulation through several mechanisms influencing epigenetic modifications, and chromatin structure, for example, acetyl-CoA, a central metabolite at the intersection of several catabolic pathways, influences gene expression by providing acetyl groups for histone acetylation, modifying histone proteins and thereby affecting chromatin dynamics (Carthew 2021). Similarly, the role of S-adenosylmethionine (SAM) as a crucial methyl donor for histone and DNA methylation, reinforces the interplay between metabolism and gene regulation (Carthew 2021). In the realm of RNA, metabolites play crucial roles in riboswitch regulation, and RNA modification affecting its stability, processing, and function (Winkler and Breaker 2003). The relationship between metabolites and proteins is perhaps the most extensively studied, with metabolites acting as enzyme substrates, cofactors or inhibitors, allosteric modulators, and disruptors of protein-protein interactions. Post-translational modifications, such as succinylation mediated by

succinyl-CoA and glycosylation facilitated by UDP-glucose, exemplify how metabolites influence protein structure and function (Dennis and Brewer 2013; Hart et al. 2011).

Certain oncometabolites that accumulate aberrantly in cancer have emerged as key players in tumorigenesis. Notable examples include 2-hydroxyglutarate, fumarate, and succinate, which disrupt cellular signaling and epigenetic regulation, leading to uncontrolled proliferation (reviewed in (M. Yang, Soga, and Pollard 2013)). 2-Hydroxyglutarate, produced by mutant isocitrate dehydrogenase, inhibits α -ketoglutarate-dependent dioxygenases, altering histone and DNA methylation and driving malignant transformation. Similarly, fumarate and succinate accumulate in tumors with defective fumarate hydratase or succinate dehydrogenase, stabilizing hypoxia-inducible factor and promoting angiogenesis and metabolic reprogramming (reviewed in (M. Yang, Soga, and Pollard 2013)). These metabolites not only contribute to cancer initiation and progression but also serve as potential diagnostic biomarkers and therapeutic targets.

Metabolites also play a crucial role in G protein-coupled receptor (GPCR) signaling, which regulates a broad range of physiological processes (Husted et al. 2017). Many GPCRs function as metabolite sensors, responding to various metabolic intermediates and products, including short-chain fatty acids (SCFAs), medium-chain fatty acids (MCFAs), long-chain fatty acids (LCFAs), hydroxy-carboxylic acids, and TCA cycle intermediates (Strassheim et al. 2021; Husted et al. 2017). For example, free fatty acid receptors GPR40, GPR41, and GPR43 respond to fatty acids of different chain lengths, while GPR81 is activated by lactate, and GPR91 by succinate (Melhem et al. 2019; Oliveira de Souza, Sun, and Oh 2021). These metabolite-sensing GPCRs are expressed in diverse tissues and influence numerous physiological processes, such as energy homeostasis, immune regulation, gut function, and adipose tissue function (Duncan et al. 2023). The activation of these receptors modulates insulin secretion, glucose homeostasis, lipid metabolism, inflammation, gut motility, and hormone secretion (M. Zhang et al. 2024). For instance, SCFAs like propionate, acetate, and butyrate activate GPR43 (FFA2 receptor), GPR41 (FFA3 receptor), and GPR109A, influencing immune cell differentiation and trafficking (reviewed in (Pardella et al. 2022; Tan et al. 2017; Husted et al. 2017)). In the context of metabolic disorders, metabolite-sensing GPCRs have been recognized as potential drug targets for preventing and treating conditions such as diabetes, dyslipidemia, and obesity (Husted et al. 2017).

The chemical diversity within metabolites is vast and extraordinary, encompassing a wide range of structures and functions. Metabolites include nucleotides, carbohydrates, amino acids, fatty acids, and various secondary metabolites (A.-H. Zhang et al. 2020, 2019; Fu et al. 2020). For example, lipids exhibit remarkable heterogeneity due to differences in chain length, saturation, and headgroup composition

(Budin et al. 2014), while secondary metabolites display functional specialization, such as acting as antibiotics or signaling molecules (Damavandi, Shojaei, and Esfahani 2023; Saad, Abdelgaleil, and Shiono 2021). This chemical diversity underpins their functional versatility and adaptability across biological systems. This diversity allows metabolites to participate in numerous cellular processes, from energy production and biosynthesis to signaling and defense mechanisms.

Evolutionary conservation of cellular metabolites is a fascinating aspect of metabolism. Studies have shown that the degree of conservation varies significantly across different metabolic pathways and individual metabolites (Bar-Even et al. 2011). Core metabolic processes, such as carbohydrate metabolism, energy production, amino acid synthesis, and nucleotide metabolism, tend to be highly conserved across species (Liska et al. 2023; Peregrín-Alvarez, Sanford, and Parkinson 2009; Peregrin-Alvarez, Tsoka, and Ouzounis 2003). This conservation reflects the essential nature of these pathways for basic cellular functions (Noor et al. 2010; Liska et al. 2023). On the other hand, secondary metabolic pathways often exhibit greater variability and species-specific adaptations (Weng and Noel 2012; Liska et al. 2023). This divergence in secondary metabolism likely reflects the diverse ecological niches and environmental pressures faced by different organisms (Timasheva et al. 2024; Liska et al. 2023). Interestingly, the degree of metabolite conservation can be predicted by factors such as metabolite abundance, essentiality, and association with human diseases (Liska et al. 2023). This pattern of conservation and divergence in metabolic pathways and metabolite concentrations provides valuable insights into the evolutionary forces shaping cellular metabolism and the functional importance of different metabolic processes (Braakman and Smith 2012; Ebenhöf and Heinrich 2001).

Decoding metabolite functions is crucial for a holistic understanding of cellular physiology and disease mechanisms. Metabolites serve as direct readouts of cellular states, reflecting the combined influences of genetic, environmental, and lifestyle factors (Kell and Oliver 2016; Clish 2015). Understanding metabolite functions can lead to the identification of novel biomarkers for disease diagnosis and prognosis, as well as potential therapeutic targets (Newgard 2017). Metabolomics has emerged as a powerful tool for discovering small-molecule metabolite signatures that provide valuable insights into metabolic targets and disease mechanisms (Newgard 2017; Patti, Yanes, and Siuzdak 2012; Wishart 2019). For instance, metabolomics has been used to identify altered abundance of metabolites and affected metabolic pathways in a disease-specific manner, such as characterizing glutamate accumulation as a new diagnostic opportunity in breast cancer (Budeczies et al. 2015; Armitage and Barbas 2014). The approach has also been applied to investigate cellular activity and characterize interactions between organisms and their environment, leading to the discovery of new disease risk biomarkers (Holmes et al. 2008). Furthermore, metabolomics-driven approaches have been employed to identify therapeutic targets

through various methods, including dose-response metabolomics, stable isotope-resolved metabolomics, and multi-omics integration (Fan et al. 2009; Kaddurah-Daouk, Kristal, and Weinshilboum 2008; C. H. Johnson, Ivanisevic, and Siuzdak 2016). These approaches enable the discovery of enzyme targets and clarification of drug mechanisms of action by analyzing metabolic pathway alterations (Pan et al. 2024).

Quantitative estimates of metabolites have expanded greatly facilitated by the advent of metabolomics databases, providing comprehensive repositories of metabolite information (Fiehn, Barupal, and Kind 2011; Wishart et al. 2022). The Human Metabolome Database (HMDB), now in its fifth version, catalogs over 217,000 annotated metabolites, offering detailed insights into metabolite structures, concentrations, biological roles, and associated pathways, serving as invaluable resources for metabolomics research (Wishart et al. 2022). Complementary to this, integrative platforms like the Kyoto Encyclopedia of Genes and Genomes (KEGG) and genome-scale metabolic models (GEMs) have advanced our understanding of metabolic networks. KEGG provides pathway maps and molecular interaction networks, aiding in omics data interpretation (Kanehisa and Goto 2000), while GEMs offer mathematical models for *in silico* prediction of metabolic behaviors, with applications in biotechnology, metabolic engineering, and systems medicine (Passi et al. 2021). These tools have become fundamental in systems biology, integrating genomics, proteomics, and metabolomics to decode complex metabolic networks (Haas et al. 2017; Joyce and Palsson 2006). Their impact extends to drug discovery, personalized medicine, and disease biomarker identification, significantly advancing our understanding of cellular metabolism and its role in health and disease (Patti, Yanes, and Siuzdak 2012; Wishart 2016).

Despite significant progress in metabolomics research, decoding the functions of cellular metabolites remains a formidable challenge due to several factors (C. H. Johnson and Gonzalez 2012). Unlike genes and proteins, which are directly encoded by the genome, metabolites lack direct genetic templates, making their origins and functions harder to trace. The complexity of metabolic networks, with their intricate interconnections and regulatory mechanisms, also makes it difficult to isolate and understand the role of individual metabolites (Guimerà and Nunes Amaral 2005; Förster et al. 2003; Guijas et al. 2018). The dynamic nature of metabolite concentrations, which can change rapidly in response to various stimuli like environmental conditions, further complicates analysis (Tang 2011; Rinschen et al. 2019; Guijas et al. 2018). The vast chemical diversity of metabolites, ranging from small polar molecules to complex lipids, necessitates the use of multiple analytical platforms, each with its own set of limitations (Mirnaghi and Caudy 2014). Furthermore, the interpretation of metabolomic data in the context of physiological and pathological states is complicated by the multifactorial nature of many diseases and the heterogeneity of patient populations (Rinschen et al. 2019; Guijas et al. 2018). The integration of metabolomics data with other omics data, such as genomics and proteomics, adds another layer of complexity to the analysis

(Pinu, Goldansaz, and Jaine 2019; Rinschen et al. 2019). Additionally, the lack of standardized protocols for sample collection, preparation, and data analysis across different laboratories hinders the reproducibility and comparability of metabolomics studies (Rinschen et al. 2019; Guijas et al. 2018). Overcoming these challenges requires continued development of advanced analytical techniques, improved bioinformatics tools, and standardized approaches to metabolomics research.

1.1 Objective of the Thesis

This thesis aims to address key open questions surrounding the multifaceted roles of cellular metabolites in physiology and pathophysiology. While metabolites have been traditionally viewed as metabolic intermediates or end products, their emerging roles as internal threats, signaling modulators, and regulators of cellular processes remain underexplored. The main objective of this work is to decode the diverse functional roles of cellular metabolites through an integrative approach, combining experimental and computational methodologies. Specifically, this thesis investigates three major aspects: (1) the potential of endogenous metabolites to act as internal threats to genomic stability and cellular integrity, (2) their interactions with GPCRs, including both odorant and non-odorant receptors, and (3) their capacity to function as intracellular allosteric regulators in GPCR-mediated signaling pathways. By integrating computational predictions with experimental validation, this thesis provides a holistic understanding of the roles of metabolites in health and disease, laying a foundation for therapeutic applications targeting metabolite-mediated pathways. This work not only advances fundamental knowledge of cellular metabolism and signaling but also highlights the translational potential of metabolomics in predictive toxicology, drug discovery, and personalized medicine.

1.2 Organization of the Chapters

This thesis is organized into five chapters, each building upon the previous to comprehensively address the diverse roles of cellular metabolites in physiology and pathophysiology through computational and experimental approaches. The organization of the chapters is as follows:

Chapter 1 introduces the foundational concepts of the central dogma and highlights the critical yet underappreciated role of metabolites in cellular processes. It discusses the chemical diversity, evolutionary conservation, and functional importance of metabolites in modulating cellular phenotypes, signaling, and disease progression. The chapter also outlines the major challenges in decoding metabolite functions and introduces the objectives of this thesis.

Chapter 2 focuses on the functional role of metabolites as potential internal threats. It presents the development of Metabokiller, a computational framework designed to predict carcinogenic metabolites by integrating key biochemical properties such as oxidative stress, genomic instability, and epigenetic modifications. This chapter also delves into experimental validations of predicted cellular metabolites, showcasing their carcinogenic potential.

Chapter 3 explores the interactions between metabolites and GPCRs, focusing on their roles as signaling modulators. This chapter describes the development of OdoriFy, an AI-driven platform for predicting odorant interactions with human-specific olfactory receptors, and EvOlf, a unified, evolutionary-driven framework for predicting GPCR-ligand interactions across odorant and non-odorant GPCRs using deep learning. The chapter highlights how computational models enable the identification of novel metabolite-receptor pairs, bridging gaps in understanding receptor diversity, ligand specificity, and evolutionary relationships. By integrating data across species and receptor classes, this chapter provides a comprehensive perspective on the roles of metabolites as signaling molecules in both olfactory and broader physiological contexts.

Chapter 4 investigates the role of endogenous metabolites as intracellular allosteric regulators of GPCR signaling using the yeast *Saccharomyces cerevisiae* as a model system. This chapter explores how metabolites like zymosterol, lanosterol, and ubiquinone-6 modulate the Ste2p receptor, influencing pheromone-induced programmed cell death and mating phenotypes. The experimental work integrates site-directed mutagenesis, and cell viability assays among others to elucidate the molecular mechanisms by which these metabolites function as intracellular modulators of GPCR activity.

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, combining computational modeling and experimental validation to unravel the diverse roles of cellular metabolites.

Chapter 2 - Decoding the Functional Role of Cellular Metabolites as Endogenous Carcinogens

2.1 Introduction

A eukaryotic cell is a complex ecosystem, constantly interacting with a diverse array of compounds throughout its lifecycle (Rappaport 2018; Farland et al. 2019; Swenberg et al. 2011). These interactions, involving both exogenous and endogenous factors, can significantly impact cellular processes and homeostasis. While external agents such as radiation and chemical mutagens are well-recognized threats, a cell's own metabolites can also pose substantial risks to genomic stability and cellular function (Wishart 2019). The compounds with the ability to induce malignant transformation are designated as carcinogens. Carcinogens can alter the genome, disrupt cellular surveillance mechanisms, or interfere with key regulatory processes, ultimately leading to cancer development (Rappaport 2018; Farland et al. 2019; Luch 2005). Notably, only 5-10% of cancers are hereditary, while the majority arise from exposure to carcinogens, underscoring the critical importance of understanding these compounds and their mechanisms of action (Anand et al. 2008).

The International Agency for Research on Cancer (IARC) has established a comprehensive framework for identifying potential human carcinogens, known as Group 1 carcinogens. This framework outlines ten key characteristics that serve as indicators of carcinogenic potential. The ten characteristics are electrophilicity or metabolic activation; genotoxicity; altered DNA repair response or genomic instability; epigenetic modification; oxidative stress induction; chronic inflammation; immunosuppression; modulation of receptor-mediated effects; immortalization induction; and alteration of cell proliferation, cell death, or nutrient supply. The presence of one or more of these traits is sufficient to raise concerns about a compound's carcinogenic potential (Smith et al. 2016). This nuanced approach reflects the complex and multifaceted nature of carcinogenesis, providing researchers and regulatory bodies with a structured method for evaluating the potential cancer-causing properties of various substances.

Despite the established framework for identifying carcinogens, the process of carcinogen identification remains a significant challenge in both experimental and computational domains. Experimentally, the primary method for identifying potential carcinogens involves long-term bioassays in animal models, typically rodents. These studies, while valuable, are labor-intensive, time-consuming, and expensive. Moreover, they raise ethical concerns due to the large number of animals required for each experiment (National Research Council (US) Committee on Comparative Toxicity of Naturally Occurring Carcinogens 1996; Cogliano et al. 2004). The translation of results from animal studies to human risk

assessment is not always straightforward, as species-specific differences in metabolism and physiology can lead to discrepancies in carcinogenic responses (Cogliano et al. 2004; Van Norman 2019). Furthermore, the constrained exposure duration in these studies may not accurately reflect the long-term, low-dose exposures often encountered in human environments (Van Norman 2019). Additionally, the ever-expanding space of compounds, with development of new drugs and industrial by-products, necessitates more rapid and cost-effective screening methods.

Given the inherent limitations of traditional experimental methods, there is a growing interest in leveraging computational approaches, particularly artificial intelligence (AI) and machine learning (ML) techniques, to predict and identify potential carcinogens. These computational approaches offer several advantages, including expediting the ongoing efforts to unravel the molecular and chemical mechanisms underlying carcinogenicity and addressing ethical concerns associated with conventional carcinogen testing. Although not a substitute for experimental validation, these approaches, particularly AI models, enable the rapid screening of millions of compounds, substantially narrowing the pool of candidates requiring experimental evaluation. Additionally, these tools can serve as early warning systems for identifying potential carcinogenicity in existing drugs and those under development (Bourcier et al. 2015).

Among computational approaches, quantitative structure-activity relationship (QSAR) models are the most widely used for carcinogenicity prediction. By analyzing molecular descriptors and biological activity data, QSAR models seek to identify patterns and relationships that explain how chemical structures influence biological outcomes like carcinogenicity using various statistical and mathematical approaches (Dearden 2017). Despite their widespread application, QSAR methods face several significant challenges. QSAR models often rely on limited training data, leading to poor predictive performance for new compounds. These models also often have restricted applicability domains, focusing on specific chemical classes such as nitro compounds (Morales et al. 2006), aromatic amines (Benigni et al. 2000; Franke et al. 2001) or polycyclic aromatic hydrocarbons (Zhou, Dai, and Gu 2003), thus limiting their utility for diverse chemical spaces. Furthermore, models based solely on structural or physicochemical properties often lack sufficient information on complex carcinogenic mechanisms.

A significant limitation of many predictive models is their heavy reliance on genotoxicity or mutagenicity as primary indicators for carcinogenicity, leading to suboptimal performance when assessing non-genotoxic or mildly genotoxic carcinogens. While a strong correlation exists between genotoxicity and carcinogenicity, this relationship alone is insufficient to elucidate the diverse mechanisms of action exhibited by all carcinogens. Non-genotoxic carcinogens, which do not directly interact with DNA, exert their carcinogenic effects through alternative mechanisms, such as altering gene expression, disrupting

cellular signaling pathways, or modulating the tumor microenvironment (Hernández et al. 2009). Notable examples of non-genotoxic carcinogens include 1,4-dichlorobenzene, a tumor promoter (Butterworth, Aylward, and Hays 2007); 17 β -estradiol, an endocrine modifier (Liehr 2000); 2,3,7,8-tetrachlorodibenzo-p-dioxin, a receptor mediator (Knerr and Schrenk 2006); and cyclosporine, an immunosuppressant (Hojo et al. 1999). These compounds demonstrate carcinogenic activity despite lacking the direct DNA-damaging properties typical of genotoxic agents. Conversely, some genotoxic compounds do not necessarily lead to carcinogenesis. For instance, N-nitroso-2-hydroxy morpholine exhibits genotoxic properties but is not carcinogenic, highlighting the complex relationship between DNA damage and cancer development (Hernández et al. 2009). This discrepancy underscores the limitations of relying solely on genotoxicity or mutagenicity as a predictor of carcinogenic potential.

Given the vastness and complexity of the chemical space, these linear approaches often struggle to generalize effectively to unseen compounds with diverse structural and functional properties. This limitation underscores the need for more sophisticated modeling techniques that integrate diverse biochemical and mechanistic features, moving beyond genotoxicity to account for the multifaceted nature of carcinogenicity and capture complex, non-linear relationships between chemical structures and their biological effects.

Despite the potential of artificial intelligence (AI) in carcinogenicity prediction, several challenges hinder its broader application. One of the primary limitations is the scarcity of large, high-quality datasets, particularly those with human-relevant annotations, which are essential for training robust models. Additionally, many advanced AI approaches, such as deep learning, function as "black boxes," offering limited interpretability and making their predictions difficult to explain. Generalizability is another concern, as models trained on specific chemical classes or cancer types often fail to perform well on novel compounds or diverse cancer subtypes. Furthermore, the integration of biological knowledge and mechanistic insights into carcinogenesis remains a challenge, as most models primarily identify data-driven patterns without incorporating underlying biological pathways. Current models also struggle with predicting non-genotoxic carcinogens, which promote cancer through mechanisms other than direct DNA damage. Finally, the reliance on rodent-based datasets introduces translational challenges, as these models may not accurately capture the complexities of human carcinogenesis, limiting their relevance in human-focused risk assessments (Bourcier et al. 2015).

Of particular interest in this chapter is the role of cellular metabolites as potential internal threats. While much attention has been given to exogenous carcinogens, recent research has revealed that endogenous metabolites can also play a significant role in carcinogenesis. These cellular metabolites, produced

through normal physiological processes, can act as double-edged swords - essential for life yet potentially harmful when dysregulated. Some endogenous metabolites have been identified as oncometabolites, capable of driving distinct cancers through various mechanisms. Key examples of oncometabolites include 2-hydroxyglutarate, succinate, fumarate, and lactate, among others (Dando et al. 2019). The accumulation of these oncometabolites can lead to alterations in cellular metabolism, epigenetic regulation, and signaling pathways, ultimately leading to tumorigenesis (Dando et al. 2019). For instance, 2-hydroxyglutarate, produced by mutant isocitrate dehydrogenase enzymes, can inhibit α -ketoglutarate-dependent dioxygenases, leading to widespread epigenetic changes, altering gene expression patterns, interfering with normal cellular differentiation processes, and promoting oncogenesis (Ježek 2020; M. Yang, Soga, and Pollard 2013; Dang et al. 2010).

Here, we introduce Metabokiller, a novel computational framework designed to predict carcinogenicity by employing an ensemble classification approach that integrates key biochemical properties commonly associated with carcinogens. These properties include oxidative stress, genomic instability, anti-apoptotic activity, electrophilicity, epigenetic alterations, and proliferative potential, collectively providing a comprehensive basis for carcinogenicity prediction. Through its application, Metabokiller identified a range of human metabolites with potential carcinogenic properties, including both previously characterized and uncharacterized compounds. Among the predictions, two metabolites, 4-nitrocatechol (4NC) and 3,4-dihydroxyphenylacetic acid (DP), were selected for further investigation. Their predicted carcinogenicity was experimentally validated through a series of functional assays and deep RNA sequencing, demonstrating the robustness and reliability of Metabokiller in uncovering biologically relevant carcinogens and advancing research in the field.

2.2 Materials and Methods

2.2.1 Data Compilation

To construct the ensemble model for carcinogenicity prediction, referred to as Metabokiller, both biological and chemical properties associated with carcinogens were utilized. These properties included the potential to induce a proliferative response, cellular oxidative stress, epigenetic modifications, genomic instability, and anti-apoptotic effects. Additionally, due to the inherent electrophilic nature of many carcinogens, electrophilic potential was selected as a critical property.

The training datasets for six independent models representing these properties were compiled in a multi-step process from literature sources and multiple databases, collectively annotated as MK_{TN}. Data

curation, filtering, and cross-validation were performed manually to ensure the authenticity and reliability of the datasets. In the first step, data were gathered from all available literature, online databases, and web servers relevant to the biological and chemical properties of interest. In the second step, compounds were categorized into pro-activity (class 1) and anti- or no-activity (class 0) groups for each of the six models. The third step involved rigorous filtering, cross-validation, and rechecking of the collected data to eliminate contradictory entries. During data curation, compounds exhibiting contradictory activities at different concentrations were excluded to prevent conflicting training datapoints and ensure consistent biochemical profiles across datasets. After these stringent screening steps, the curated dataset consisted of 12,957 class 1 compounds and 22,711 class 0 compounds across the six biochemical properties. For the genomic instability model, 12,127 compounds were included, comprising 8,931 class 0 and 3,196 class 1 compounds. The dataset for electrophilic properties contained 1,714 compounds, with 746 in class 0 and 968 in class 1. The epigenetic modification dataset consisted of 15,145 compounds, equally distributed between 7,576 class 0 and 7,569 class 1 compounds. The anti-apoptotic property dataset included 798 compounds, divided into 525 class 0 and 273 class 1. The proliferative property dataset contained 225 compounds, with 123 in class 0 and 102 in class 1. Lastly, the oxidative stress dataset comprised 5,659 compounds, with 4,810 in class 0 and 849 in class 1. Detailed information about these datasets, including their respective compounds, is available through the Zenodo repository.

In addition to these datasets, a list of experimentally validated carcinogens and non-carcinogens was compiled from multiple independent sources. These compounds were categorized into four datasets: Metabokiller External Training Data (MK_{ETn}), Metabokiller Testing Data 1 (MK_{Ts1}), Metabokiller Testing Data 2 (MK_{Ts2}), and an independent dataset (I.D.). Detailed descriptions of the compounds in these datasets are available in the table below (**Table 2.1**).

For all compounds, Simplified Molecular Input Line Entry System (SMILES) representations were extracted using the PubChem Identifier Exchange Service. These SMILES were then converted into canonical SMILES format using OpenBabel (O’Boyle et al. 2011) to ensure standardization and consistency in downstream analyses. This comprehensive and systematically curated dataset served as the foundation for building the Metabokiller model, enabling robust and reliable carcinogenicity prediction.

Dataset	Model	Acronym	# Carcinogens	# Non-carcinogens	# Total	Sources
Training data	Metabokiller	MK _{ETn}	303	163	466	CCRIS, ISSCAN
Testing dataset 1	Metabokiller	MK _{Ts1}	37	24	61	CPDBAS
Testing dataset 2	Metabokiller	MK _{Ts2}	65	35	100	Multiple sources
Combined Testing dataset 1+2	Metabokiller	MK _{Ts1+2}	102	59	161	Multiple sources
Independent dataset	Comparison	I.D.	82	28	110	Multiple sources

Table 2.1 Summary of datasets used for the development and evaluation of Metabokiller.

The table summarizes the datasets utilized for training and testing the Metabokiller model and for comparison with other tools. It includes details on the dataset type (training or testing), the corresponding model or purpose, dataset acronym, the number of carcinogens and non-carcinogens, the total number of compounds, and the sources from which the data were curated.

2.2.2 Model Building

The Metabokiller ensemble model relies on six independent predictive models, each designed to represent a distinct biochemical property associated with carcinogens. These biochemical models collectively define the predictive space for carcinogenicity. The training datasets for each model were categorized into two classes based on known activity: class 0, representing compounds with anti-activity or no-activity, and class 1, representing compounds with pro-activity.

Three distinct feature-extraction methods were evaluated for the model: bioactivity-based descriptors from the Signaturizer library, chemistry-based molecular descriptors generated using Mordred software, and graph-based features obtained through the DeepChem library. Multiple classification algorithms were tested for model construction, including random forest, multilayer perceptron (MLP), k-nearest neighbor (KNN), support vector machine (SVM), stochastic gradient descent (SGD), logistic regression (LR), GraphConvModel, AttentiveFP (AFP), graph convolution network (GCN), and graph attention network (GAT). Feature selection was performed using the Boruta algorithm (Kursa, Jankowski, and Rudnicki 2010), which identifies the most relevant features by iteratively ranking and filtering less significant variables. To address class imbalances in the training datasets, upsampling and downsampling techniques were applied using the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2011), ensuring balanced representation across classes. Combinations of these feature extraction methods and classification algorithms were systematically explored, resulting in the development of twelve distinct models for each biochemical property using default parameters.

The best-performing classifier for each biochemical property was selected based on the performance metrics. The model was further optimized through random-grid-search hyperparameter tuning to obtain the best-performing and most stable hyperparameters. The optimized models were then trained with the best hyperparameters and subjected to rigorous validation using both 20-fold bootstrapping and 10-fold cross-validation techniques. This comprehensive approach ensured the stability and reliability of the individual biochemical models.

To construct the Metabokiller ensemble model, a dataset comprising experimentally validated bonafide carcinogens and non-carcinogens (annotated as MK_{ETn}) was used. The MK_{ETn} dataset was projected onto all six biochemical models, and the prediction probabilities for pro-activity were extracted. These probabilities served as input features for building the ensemble model. Among the classification algorithms (RF, MLP, SVM, KNN, and GBM) evaluated for the ensemble, gradient boosting machine (GBM)-based models demonstrated superior predictive performance. To further enhance the GBM-based ensemble model, random-grid hyperparameter tuning was performed to optimize its parameters. To mitigate potential data biases, twenty distinct GBM-based models were generated using bootstrapping, and a majority voting strategy was implemented for final carcinogenicity predictions. This method leveraged consensus across multiple models, reducing the influence of individual model variability.

2.2.3 Strains and Growth Conditions

The BY4741 strain (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) of *Saccharomyces cerevisiae* was used for all experimental assays. For apoptosis-related experiments, *Δfis1* knockout (*MATa fis1::kanMX4 Δhis3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) was used. Unless specified otherwise, yeast cells were grown at 30 °C with constant shaking at 200 revolutions per minute (rpm) in yeast extract-peptone-dextrose (YPD) medium comprising 1% yeast extract, 2% peptone, and 2% dextrose. For solid media, 1.5% agar was added to the YPD medium to prepare plates.

2.2.4 Comet Assay for Genotoxicity Evaluation

To assess genotoxicity, the comet assay was performed using yeast secondary cultures grown to an optical density of 0.6-0.7 at 600 nm (OD₆₀₀). Yeast cells were exposed to varying concentrations (0.1 μM, 1 μM, 10 μM, and 100 μM) of 4-nitrocatechol (4NC; Sigma-Aldrich, N15553) and 3,4-dihydroxyphenylacetic acid (DP; Sigma-Aldrich, 11569). Hydroxyurea was used as a positive control, while solvent (water) only treatments served as the negative control. The yeast cells were grown in the presence of metabolite for 90 min at 30 °C at 200 r.p.m. Post-treatment, equal number of cells from each condition were pelleted via centrifugation and washed twice with pre-cooled phosphate-buffered saline (PBS). The cell pellet was

resuspended in 0.3% agarose prepared in S-buffer (1 M sorbitol, 25 mM KH₂PO₄, pH 6.5), followed by the addition of 40 units of zymolyase enzyme (Sigma-Aldrich, L2524) to degrade the cell wall. This suspension (80 μ L) was spread onto glass slides pre-coated with 0.8% agarose and incubated for 30 minutes at 30 °C. The slides were subsequently treated with freshly prepared lysis buffer (30 mM NaOH, 1 M NaCl, 0.05% SDS, 50 mM EDTA, 10 mM Tris-HCl, pH 10) for 2 hours, followed by three 20-minute washes with electrophoresis buffer (30 mM NaOH, 10 mM EDTA, 10 mM Tris-HCl, pH 10). Electrophoresis was performed for 15 minutes at 70 mV/cm. After electrophoresis, the slides were incubated in the neutralization buffer (10 mM Tris-HCl, pH 7.4) for 10 mins. Following neutralization, the slides were dehydrated in graded ethanol solutions (76% and 96%) for 10 minutes each, and stained with Giemsa for 20 minutes. The slides were then washed thrice with 1 $\times$ PBS Tween (0.1% Tween20) for 10 min each to remove excessive dye. The comets were visualized in a brightfield microscope. Comet tail lengths were measured for approximately 500 randomly selected cells per condition using ImageJ software.

2.2.5 Mutagenesis Assay

To evaluate mutagenic potential, yeast cells were grown overnight. Cultures were normalized based on optical density measurements at 600 nm (OD₆₀₀), and equal numbers of cells were inoculated into 96-well plates containing 100 μ L of YPD medium along with respective metabolites (4NC and DP) at final concentrations of 0.1 μ M, 1 μ M, and 10 μ M each, alongside solvent-only control. A total of 52 biological replicates were set up for each condition. To minimize evaporation, the 96-well plates were sealed with Breathe-Easy sealing membranes (Sigma-Aldrich, Z380059-1PAK). Plates were incubated at 30 °C at 200 rpm, and the medium was replenished with fresh medium 24 h after incubation with the aforementioned metabolites or vehicles. After 48 h incubation, cell count was estimated by measuring OD₆₀₀, which was further used to normalize the frequency of Can^R mutants. For the quantification of mutagenesis, a spot assay was performed. Aliquots from each condition were plated on synthetic complete (SC) medium supplemented with canavanine (60 μ g/mL) and containing 2% dextrose, 6.7 g/L yeast nitrogen base without amino acids, 20 mg/L histidine, 120 mg/L leucine, 20 mg/L methionine, 20 mg/L uracil, 20 mg/L adenine, and 2% agar. The plates were incubated at 30 °C for 2 to 3 days, after which the number of colonies was counted. The colony counts were normalized with OD₆₀₀ values from the earlier cell density measurements to compute OD₆₀₀-normalized mutant counts. These normalized counts were then used to calculate the fold change in mutagenesis under each experimental condition. For the estimation of mutational frequencies, the normalized mutant counts were divided by the total number of cells present in the culture for each condition.

A similar experimental setup was used for rescue experiments, where cells were treated with 4NC at a final concentration of 10 μ M, either in the presence or absence of a ROS scavenger. In these rescue experiments, l-ascorbic acid was used as the ROS scavenger at a final concentration of 50 μ M to evaluate its ability to mitigate the mutagenic effects of 4NC.

2.2.6 Cell Death Assay

Propidium iodide staining was used to quantitatively estimate cell death in treated and untreated conditions. Yeast cells were cultured overnight at 30 °C with shaking at 200 rpm in YPD medium. An equal number of cells from primary culture were inoculated into 96-well deep plates and treated with respective metabolites (4NC and DP) at final concentrations of 0.1 μ M, 1 μ M, and 10 μ M each for 12 hours (n = 8 biological replicates). Aliquots were taken after 9 and 12 hours of incubation for staining. Propidium iodide staining was performed in 96-well fluorescence plates at a final concentration of 5 μ g/mL, using aliquots collected after 9 and 12 hours of incubation. The samples were incubated in the dark for 15 minutes. Heat-killed cells served as positive controls. The fluorescence was measured using Biotek Synergy HTX multi-mode reader at excitation and emission wavelengths of 530/25nm and 590/25nm, respectively.

2.2.7 Apoptosis assay

Wild-type BY4741 and *Δfis1* knockout yeast strain (*MATa fis1::kanMX4 Δhis3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) were cultured overnight at 30 °C with shaking at 200 rpm in YPD medium. The cultures were normalized based on OD₆₀₀, and an equal number of cells from both wild-type and *Δfis1* strains were inoculated into 96-well plates containing 100 μ L of fresh YPD medium. Cells were treated with the metabolites 4NC and DP at a final concentration of 10 μ M and incubated for 24 hours, with 8 biological replicates each. Cells treated with 199 mM acetic acid for 200 minutes served as a positive control, while untreated cells served as a negative control. Following incubation, serial 1:10 dilutions of the cultures were prepared in distilled water. A spot assay was then performed by plating aliquots onto synthetic complete (SC) medium agar plates. The SC medium contained essential nutrients and supported yeast growth while allowing colony formation (Teng and Hardwick 2009). The plates were incubated at 30 °C for 2 to 3 days. Colony-forming units (CFU) per milliliter were calculated using the following formula:

$$CFU/mL = \frac{\text{Number of colonies} \times \text{Dilution Factor}}{\text{Volume of culture plated}}$$

2.2.8 Reactive Oxygen Species (ROS) Quantification

Intracellular ROS levels were quantified using 2',7'-dichlorofluorescein diacetate (DCFH-DA) staining. Yeast cultures were treated with metabolites at specified concentrations for 12 hours. DCFH-DA was added to a final concentration of 10 μ M, and fluorescence was measured after 30 minutes. Hydrogen peroxide-treated cells served as positive controls. The levels of intracellular reactive oxygen species (ROS) were quantified using 2',7'-dichlorofluorescein diacetate (DCFH-DA) staining. Yeast cells were initially cultured overnight at 30 °C with shaking at 200 rpm in YPD medium. With primary culture as inoculum, equal cells were transferred to 96-well deep plates and exposed to the metabolites 4NC and DP at final concentrations of 0.1 μ M, 1 μ M, and 10 μ M. Incubation was carried out for 12 hours at 30 °C with 8 biological replicates per condition. DCFH-DA staining was performed in 96-well fluorescence plates on aliquots taken after 9 and 12 hours of incubation. Hydrogen peroxide (10 mM) was included as a positive control. DCFH-DA dye (85048, SRL) was added to each well at a final concentration of 10 μ M. The plates were incubated in the dark for 30 minutes to allow the dye to react with intracellular ROS. Fluorescence measurements were recorded using a Biotek Synergy HTX multi-mode reader. The excitation and emission wavelengths were set to 485/20 nm and 525/20 nm, respectively.

2.2.9 Selection of Canavanine-Resistant Mutants with Accelerated Cell Division

The BY4741 strain was cultured at 30 °C with shaking at 200 rpm in YPD medium. The overnight culture was used as an inoculum to ensure consistent cell density across conditions. Cells were transferred to microcentrifuge tubes containing YPD medium supplemented with the metabolites 4NC and DP at final concentrations of 0.1 μ M, 1 μ M, and 10 μ M. A negative control without metabolites was included. Cultures were incubated for 72 hours at 30 °C, with the medium being refreshed every 24 hours by centrifuging the cells, discarding the spent medium, and resuspending the cells in fresh YPD. After the 72-hour incubation, cells from each condition were plated onto synthetic complete (SC) medium agar plates containing canavanine at a concentration of 30 μ g/mL. This step served as the first round of selection to identify canavanine-resistant colonies. Plates were incubated at 30 °C until colonies formed.

Canavanine-resistant colonies were then transferred to YPD medium containing canavanine but without the metabolites, allowing for a second selection based on growth characteristics. This growth assay identified colonies displaying accelerated growth kinetics. RNA was isolated from the canavanine-resistant mutants that exhibited accelerated growth for further analysis. The extracted RNA underwent quality control checks to ensure its suitability for downstream applications, including library

preparation and sequencing. This process enabled the identification of transcriptomic changes associated with canavanine resistance and accelerated growth under experimental conditions.

2.2.10 RNA Isolation

Yeast cells (BY4741) were grown at 30 °C at 200 rpm in synthetic complete medium containing canavanine (30 µg/mL) in biological triplicate and harvested and resuspended in 500µl of 1× PBS. Zymolyase (40 units) treatment was performed by incubating for 30 min at 30 °C. After zymolyase treatment, mechanical lysis was performed using glass beads. The resulting lysate was transferred to fresh microcentrifuge tubes (MCTs) and centrifuged. The cell pellet was resuspended in 500 µL of TRIzol reagent (Ambion Life Technologies, 15596026). Five hundred microliters of chilled chloroform were added to each tube, followed by 15 min of mechanical agitation and 5 min of incubation at room temperature. After this, samples were centrifuged at 12,000g for 15 min at 4 °C. The upper aqueous phase was carefully transferred to a fresh MCT containing 250 µL of pre-chilled isopropanol. To enhance RNA precipitation, 1 µL of glycogen was added, followed by a 30 minute incubation at -80 °C. Nucleic acid was pelleted down by centrifugation at 12,000g for 15 min at 4 °C. The pellet was washed with 70% ethanol at 5,000g for 5 min at 4 °C. After air-drying, the pellet was resuspended in RNase-free water.

To remove DNA contamination, samples were treated with DNase at 37 °C for 15 min. DNase inactivation was performed by heating the samples at 65 °C with 5 mM EDTA for 10 min. The purified RNA was finally resuspended in RNase-free water and stored at -80 °C for downstream applications.

2.2.11 RNA Sequencing Analysis

After sequencing, quality control of the raw paired-end sequencing read files was conducted using MultiFastQ to ensure data integrity and accuracy. The processed reads were mapped to the *Saccharomyces cerevisiae* reference genome (ENSEMBL R64-1-1; GCA_000146045.2) using the align function of the Rsubread package (version 2.6.4) (Liao, Smyth, and Shi 2019). This mapping process generated alignment files in BAM format. Prior to mapping, trimming was performed on both ends of the sequencing reads to remove low-quality bases and adapters. The inbuilt function of Rsubread was used to detect short insertion and deletion (indel) mutations within the sequencing data and the output was stored in a variant call format (VCF) format. An expression matrix containing the raw read counts for each sample was obtained using the featureCounts function. Normalization and analysis of the read count matrix were performed using the RUVg function of the RUVSeq package (version 1.24.0). This step removed unwanted technical variations from the RNA sequencing data. Differential gene expression analysis was conducted using the DESeq2 package (version 1.30.1) (Love, Huber, and Anders 2014), with

a fold-change cutoff of 1.5 and a statistical significance threshold of $P < 0.05$. Functional Gene Ontology (GO) analysis of differentially expressed genes was carried out using the GO Slim Mapper tool of Saccharomyces Genome Database (SGD) and Metascape. Normalized read count matrix is available at the Zenodo repository, whereas the raw FASTQ files are accessible at ArrayExpress.

2.2.12 Mutation Analysis

Variant call format (VCF) files containing insertion and deletion (indel) mutation information were generated using the Rsubread package, as discussed above. The variant information from three replicates was merged, and a single VCF file was generated using VCFtools (v.0.1.16-1) (Danecek et al. 2011). To ensure the reliability of the detected mutations, only variants supported by at least five sequencing reads were selected for downstream analysis. Cross-comparison of VCF files was performed between treatment groups exposed to 3,4-dihydroxyphenylacetic acid (DP) or 4-nitrocatechol (4NC) and the negative control (untreated cells). Variants identified in the control group were excluded, and only mutations unique to each treatment condition were retained for further analysis. The mutations were segregated into frameshift (insertion or deletion) or substitution using the SIFT 4G Annotator using R64 (sacCer3) as a reference genome (Vaser et al. 2016).

2.2.13 Soft-Agar Assay

BEAS-2B human epithelial cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin. Cells were maintained in a humidified incubator at 37 °C with 5% CO₂. For the soft-agar assay, a cell suspension of BEAS-2B cells was prepared at a density of 1×10^4 cells per well. The cell suspension was added to a top agar solution containing 0.3% agar, with a final volume of 0.5 mL per well. This suspension was layered over a base of 0.5% solidified agar. The overnight culture was treated with 10 μ M concentrations of either 3,4-dihydroxyphenylacetic acid (DP) or 4-nitrocatechol (4NC). The cells were supplemented with DMEM complete medium twice until colonies appeared after 3 weeks. After the formation of colonies, staining was performed using 0.1% crystal violet dissolved in 10% ethanol to visualize the colonies. Stained colonies were imaged using a light microscope equipped with a 4X objective lens and a brightfield filter for documentation and analysis.

2.2.14 Data and Code Availability

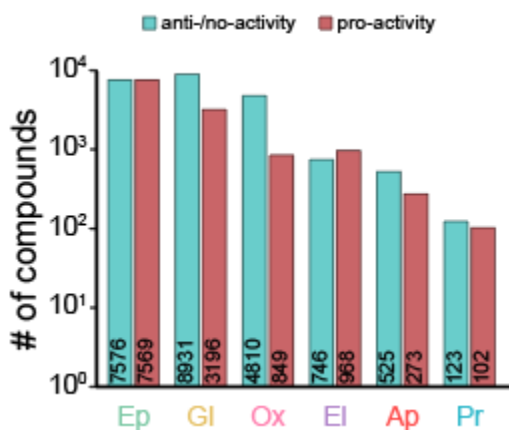
Raw RNA sequencing files are available on ArrayExpress (accession E-MTAB-11179). Processed datasets, including compound descriptors and source information, are hosted on GitHub

(<https://github.com/theahuja-lab/Metabokiller>) and Zenodo (<https://doi.org/10.5281/zenodo.6683106>). Code for the machine learning models is accessible on GitHub and Zenodo.

2.3 Results

2.3.1 Metabokiller: A Novel Approach to Carcinogenicity Prediction

Cancer cells possess a multitude of traits that enable them to survive, display uncontrollable growth, and metastasize. Some of these key features include their ability to proliferate indefinitely, disregarding normal cellular signals that limit growth. They can also evade apoptosis, the programmed cell death mechanism, allowing them to persist even when damaged. They are often electrophilic, producing reactive molecules that damage cellular components, leading to increased DNA mutations. This process contributes to genomic instability, where mutations accumulate rapidly, driving cancer's progression. Epigenetic changes, such as altered DNA methylation or histone modification, reprogram gene expression in ways that deactivate tumor suppressors and activate oncogenes, further supporting cancer growth. Additionally, cancer cells experience high oxidative stress from an overload of reactive oxygen species, which fuels mutations and enables metabolic shifts that help them thrive in hostile conditions. Together, these characteristics allow cancer cells to evade the body's defenses and grow unchecked. These distinct biochemical properties were utilized in Metabokiller, a novel ensemble-based approach for carcinogenicity prediction, to capture the multifaceted nature of carcinogens.

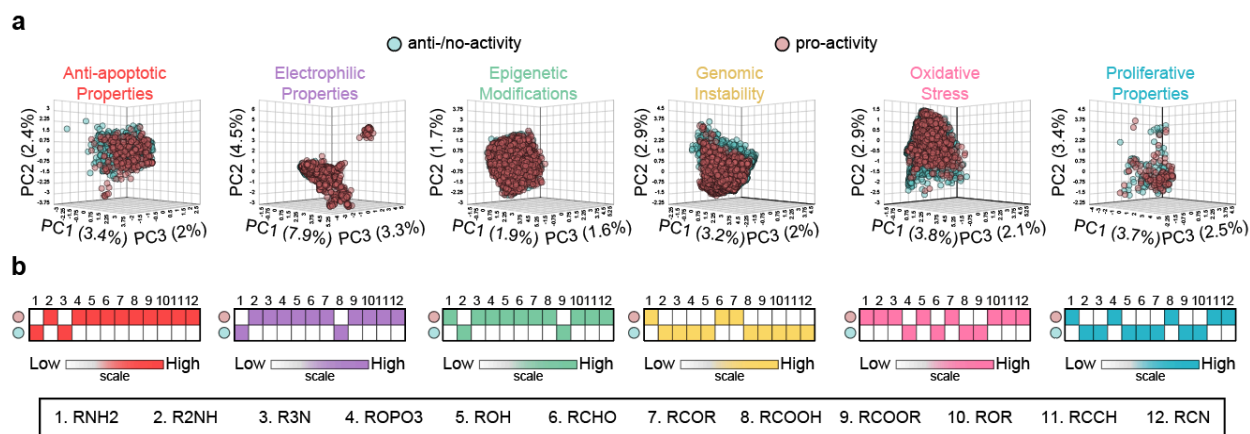


Adapted and modified from Mittal A. 2022 et al., Nature Chemical Biology (permission granted)

Figure 2.1 Dataset Composition for Biochemical Property-Based Models in Metabokiller.

Bar graph depicting the number of compounds used to build six independent biochemical models of Metabokiller. Turquoise and red bars indicate compounds with anti-/no- and pro-activities in the indicated conditions, respectively.

To build metabokiller, the training datasets were manually curated and compiled, containing information about compounds that are experimentally reported to impact cellular proliferative properties (Pr), genomic instability (GI), oxidative state (Ox), epigenetic landscape (Ep), anti-apoptotic properties (Ap), and electrophilic properties (El). These six independent datasets were segregated into pro- (Class 1) and anti-/no- (Class 0) activity categories, collectively encompassing 35,668 compounds (annotated as MK_{Tn}) (**Figure 2.1**). To gain deeper insights into the chemical heterogeneity between classes, principal component analysis (PCA) was performed using extended-connectivity fingerprints (ECFP) (Rogers and Hahn 2010) as features. This analysis revealed a high degree of chemical heterogeneity in most cases, as evidenced by the spread and lack of clear clustering in the PCA plots. The wide distribution of data points across the principal component space indicates significant variations in the chemical features of molecules, suggesting diverse structural attributes and functional group compositions (**Figure 2.2a**). The inter-class comparison of functional groups revealed distinct patterns of enrichment or de-enrichment across the datasets, shedding light on the chemical diversity and biases in functional group distributions. For instance, alcohols (ROH), aldehydes (RCHO), and ketones (RCOR) exhibited relatively balanced distributions across classes, whereas functional groups associated with anti-apoptotic activity, oxidative stress, electrophilicity, and epigenetic modifications were predominantly enriched in Class 1 (**Figure 2.2b**).

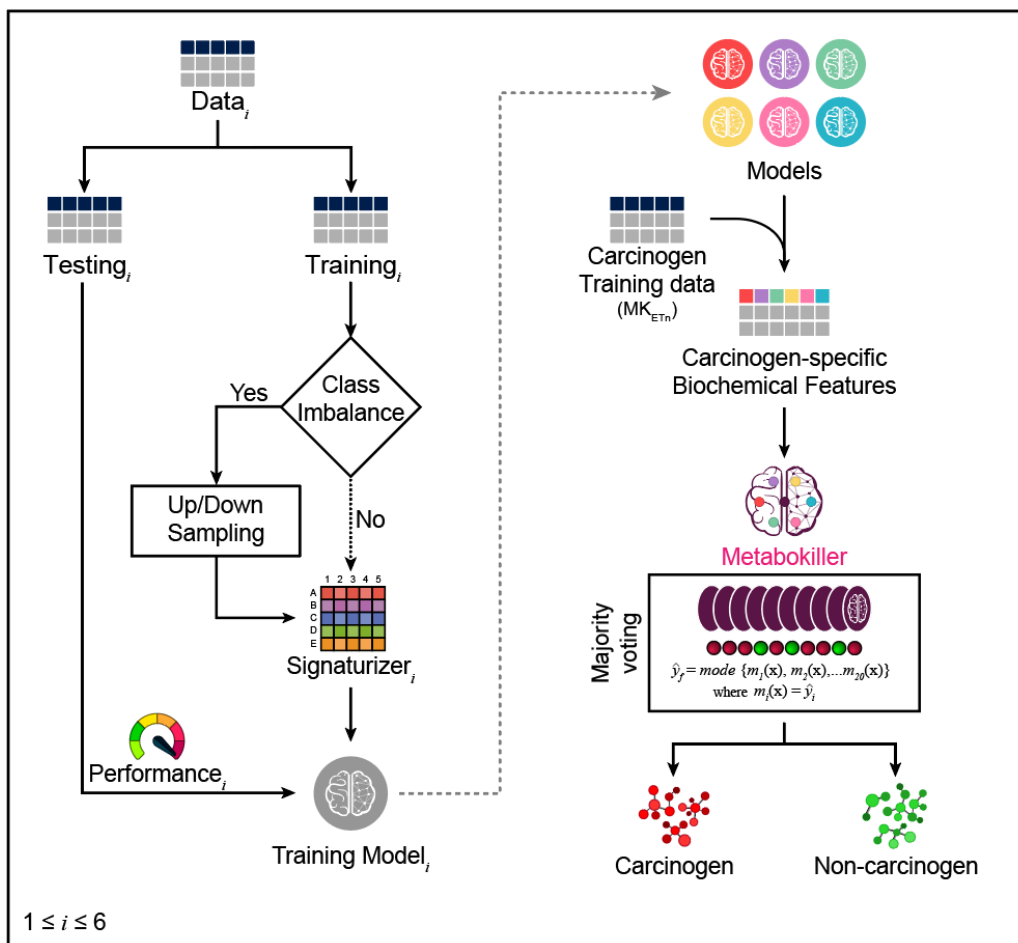


Adapted and modified from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.2 Chemical Heterogeneity and Functional Group Enrichment in Biochemical Properties of Carcinogenicity

(a) PCA depicting chemical heterogeneity of the two classes (pro- and anti-/no-) in the indicated datasets. **(b)** Heat maps depicting 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; and RCN, nitrile) in both classes.

To systematically investigate these six biochemical properties, we built six independent binary classification models. To identify the most effective modeling approach, three distinct feature extraction methodologies were evaluated: bioactivity-based descriptors derived using the Signaturizer library (Bertoni et al. 2021); chemistry-based molecular descriptors generated by the Mordred (Moriwaki et al. 2018); and graph-based features obtained through the DeepChem library. These feature sets were chosen to comprehensively capture the biochemical, physicochemical, and structural properties of the molecules. A range of machine learning algorithms, like random forest (RF), multilayer perceptron (MLP), k-nearest neighbor (KNN), support vector machine (SVM), stochastic gradient descent (SGD), and logistic regression (LR), and deep learning algorithms like GraphConvModel (GCM), Attentive FP (AFP), graph convolution network (GCN), and graph attention network (GAN) were implemented. By systematically evaluating the performance of these models, the aim was to identify the optimal combination of feature extraction techniques and algorithms for each classification task (**Figure 2.3**).



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.3 Workflow Detailing Metabokiller Functionalities.

Schematic representation depicting the step-by-step workflow used to build all six individual biochemical models and the ensemble model (Metabokiller). Up/down sampling approach was used to counteract the class imbalance. Signaturizer library was used to generate bioactivity features. Hyperparameter tuning was performed to obtain the best-performing model parameters. The ensemble model (Metabokiller) was built using biochemical features of experimentally validated carcinogens/non-carcinogens generated using six models. The majority voting method was used to assign the final carcinogenicity status.

A generic workflow for model development was used to systematically process and analyze the biochemical datasets. The initial step involved the random splitting of the compiled datasets into training and testing subsets, with the training set used for model learning and the testing set reserved for independent evaluation to assess model performance. Next, molecular features were extracted from the SMILES (Simplified Molecular Input Line Entry System) representations of the molecules. Feature extraction was performed using Signaturizer, Mordred, and Deepchem libraries. Following this, feature selection was carried out using the Boruta algorithm (Kursa, Jankowski, and Rudnicki 2010) that identifies the most relevant features for prediction by iteratively ranking and removing less important variables. This ensured the retention of only highly variable, minimally correlated, and influential features, thereby improving model interpretability and reducing the risk of overfitting. To address the issue of class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was applied (Chawla et al. 2011). This technique generates synthetic examples in the minority class by interpolating between existing samples, thereby balancing the class distribution and enhancing the ability of the models to correctly classify underrepresented classes. An array of classification algorithms, spanning both traditional machine learning methods and deep learning approaches, was then implemented for model development. Each classifier was trained using the processed training dataset, leveraging the selected features and the balanced class distributions. Finally, all models were rigorously evaluated on the unseen testing dataset to assess their predictive performance. Metrics such as accuracy, precision, recall, F1-score, and AUC-ROC were used to evaluate the models' performance and identify the most effective approaches (**Figure 2.3**).

This workflow was used to construct twelve classification models for each of the six biochemical properties. The best performing and thus selected models for downstream analysis were KNN for anti-apoptotic activity, MLP for electrophilic activity, SVM for epigenetic modifications, RF for genomic instability, MLP for oxidative stress, and RF for proliferation. A notable trend observed during the model evaluation process was the superior performance of models built using bioactivity-based descriptors generated by the Signaturizer library over those based on chemistry-based molecular descriptors

(Mordred) and graph-based features (DeepChem) (**Table 2.2**). To further enhance the performance and reliability of the selected base models, we implemented a systematic optimization process involving randomized-search-based hyperparameter tuning. After selecting the best hyperparameters for each model, we tested these hyperparameters with rigorous validation using two complementary techniques: bootstrapping and tenfold cross-validation to assess the consistency and reliability of the optimized models across different data splits and sampling conditions (**Figure 2.4**).

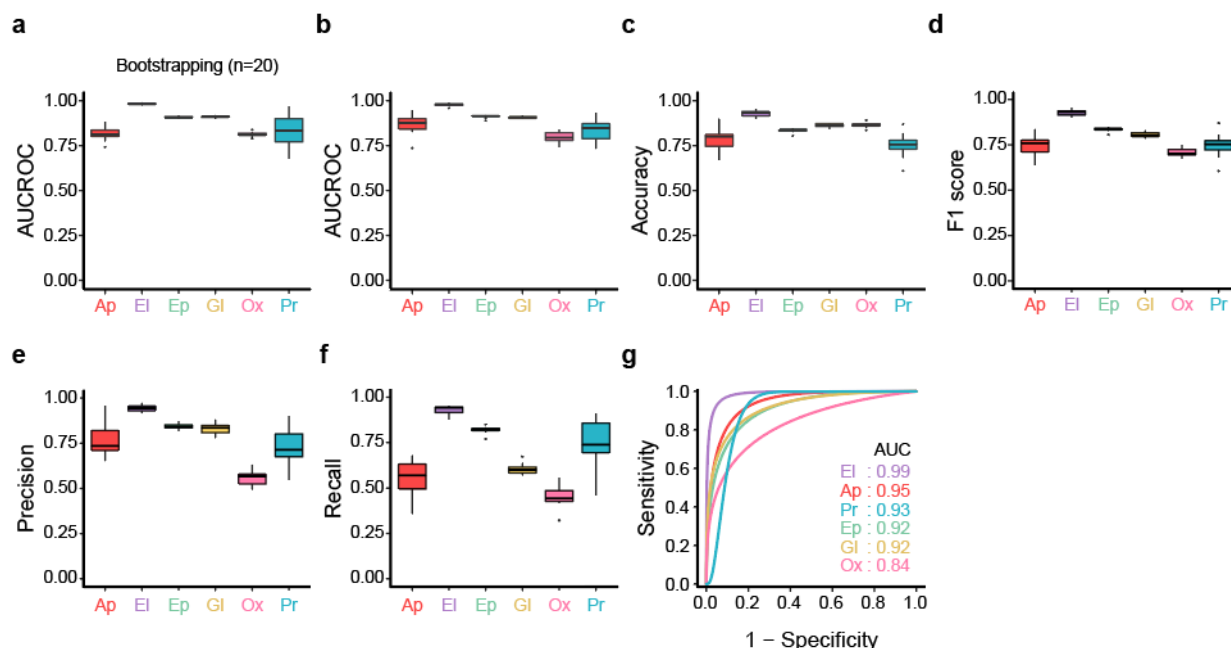
		Ap		EI		Ep		GI		Ox		Pr	
Features	Models	Accuracy	AUCROC	Accuracy	AUCROC	Accuracy	AUCROC	Accuracy	AUCROC	Accuracy	AUCROC	Accuracy	AUCROC
Signaturizer	RF	0.78	0.86	0.91	0.97	0.82	0.91	0.86	0.91	0.89	0.83	0.70	0.78
Signaturizer	MLP	0.76	0.84	0.94	0.98	0.84	0.91	0.84	0.91	0.86	0.81	0.65	0.73
Signaturizer	KNN	0.79	0.81	0.88	0.96	0.82	0.88	0.87	0.89	0.81	0.80	0.60	0.73
Signaturizer	SVM	0.67	0.86	0.93	0.99	0.84	0.92	0.87	0.92	0.82	0.80	0.67	0.78
Mordred	MLP	0.72	0.72	0.90	0.97	0.50	0.50	0.83	0.85	0.57	0.73	0.58	0.75
Mordred	KNN	0.67	0.65	0.84	0.91	0.61	0.65	0.80	0.82	0.79	0.65	0.65	0.78
Mordred	SGD	0.36	0.52	0.71	0.72	0.50	0.50	0.73	0.50	0.48	0.64	0.74	0.73
Mordred	LR	0.68	0.72	0.54	0.83	0.50	0.57	0.73	0.58	0.77	0.73	0.74	0.77
Graph	GCM	0.60	0.55	0.87	0.95	0.71	0.75	0.85	0.88	0.75	0.70	0.60	0.73
Graph	AFP	0.69	0.66	0.88	0.96	0.71	0.75	0.84	0.87	0.78	0.70	0.58	0.61
Graph	GCN	0.70	0.67	0.89	0.96	0.75	0.79	0.84	0.88	0.79	0.71	0.60	0.69
Graph	GAT	0.67	0.61	0.84	0.94	0.68	0.70	0.76	0.86	0.78	0.69	0.58	0.67

Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Table 2.2 Performance of Feature and Model Combinations for Biochemical Properties

Accuracy and AUCROC values for machine learning models trained with Signaturizer (bioactivity-based), Mordred (chemistry-based), and graph-based features. Columns represent performance on anti-apoptotic (Ap), electrophilic (EI), epigenetic (Ep), genomic instability (GI), oxidative stress (Ox), and proliferative properties (Pr).

After this, we developed an ensemble model, termed Metabokiller designed to predict carcinogenicity by integrating the outputs from these six classification models. This ensemble approach leveraged the prediction probability scores from the individual models, each representing a distinct biochemical property, under a unified machine-learning framework. To build Metabokiller, we carefully curated a dataset comprising 466 experimentally validated carcinogens and non-carcinogens from the ISSCAN and CCRIS databases, annotated as MK_{ETn}. This dataset was used for training and evaluation of the ensemble. Importantly, the MK_{ETn} dataset was designed to have no overlap with the MK_{Tn} dataset, which was used in the training of the six biochemical property-based models, ensuring that the training and validation phases were free of dataset-dependent biases (**Figure 2.5a**). Analysis of the dataset revealed that carcinogens within MK_{ETn} exhibited non-redundancy and diversified biochemical properties within the dataset, as demonstrated by their divergent directions of the arrows in the eigenspace (**Figure 2.5b**).



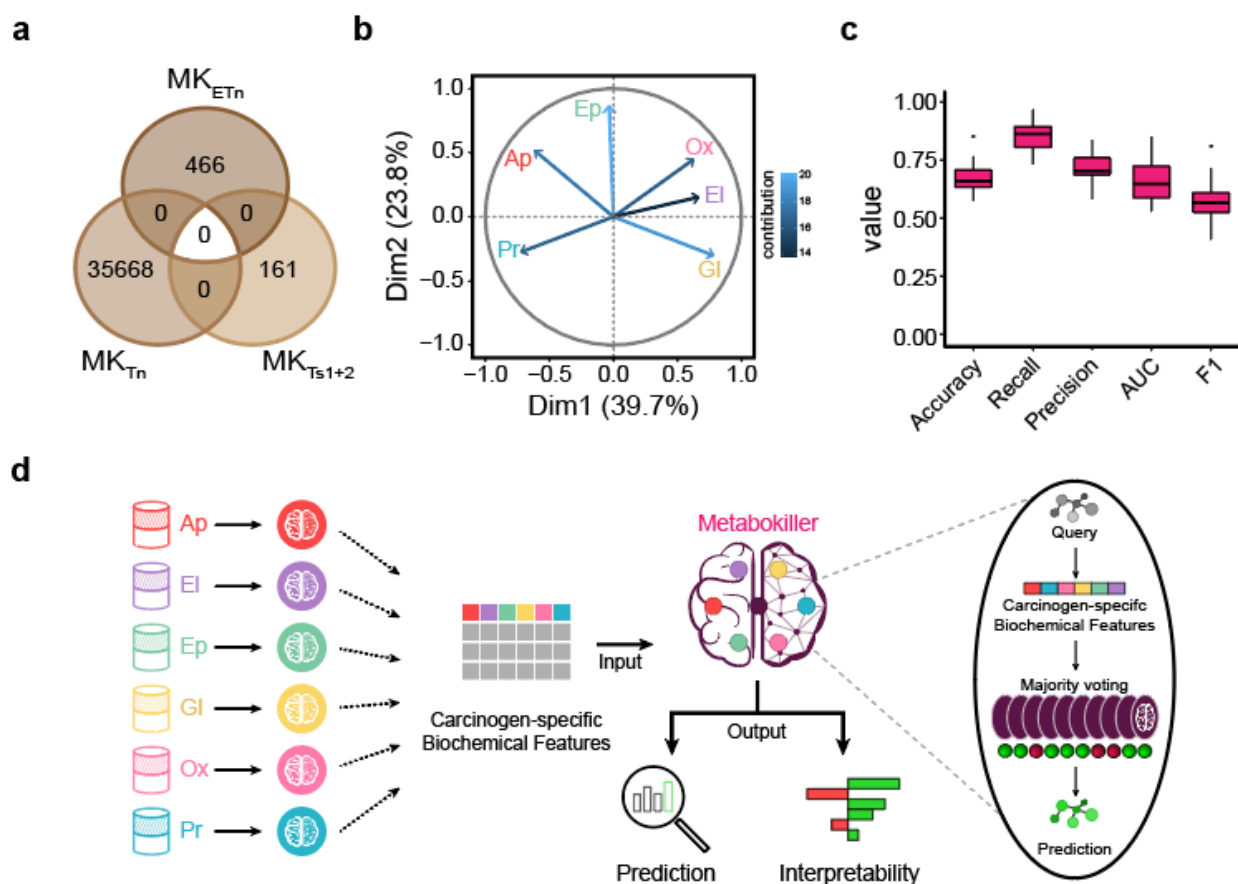
Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.4 Performance Metrics and Validation of Biochemical Property-Based Models

(a) Box plot depicting the AUCROC values of the bootstrapping ($n = 20$ repetitions) of the indicated models. **(b-f)** Box plots depicting the AUCROC, accuracy, F1 Score, precision, and recall of the indicated models as inferred from the 10-fold cross-validation. **(g)** Area under the curve (AUC) plots representing the performance of the best models obtained using 10-fold cross-validation of the indicated best models.

After data curation, we used the following methodology to develop Metabokiller. For each compound in this dataset, the representation of the six key biochemical properties (anti-apoptosis, electrophilicity, epigenetic modifications, genomic instability, oxidative stress, and pro-proliferation) was computed based on the outputs of their respective classification models. These prediction probabilities were then utilized as input features for the ensemble model, capturing the complex interplay between biochemical properties relevant to carcinogenicity. To identify the most effective classification algorithm for the ensemble, we evaluated six different machine-learning methods. Among these, the gradient boosting machine (GBM) emerged as the most suitable choice, exhibiting superior precision and recall values, critical metrics for accurately identifying carcinogens while minimizing false positives and false negatives. After completing hyperparameter tuning to optimize the GBM model, a total of twenty distinct GBM models were constructed using a bootstrapping technique (**Figure 2.5c**). These models were then integrated into an ensemble framework that uses a majority voting method for the carcinogenicity prediction of query compounds. The majority voting strategy ensures that the ensemble leverages the consensus across

multiple independent models, thereby minimizing the impact of individual model biases and improving classification accuracy (Figure 2.5d).



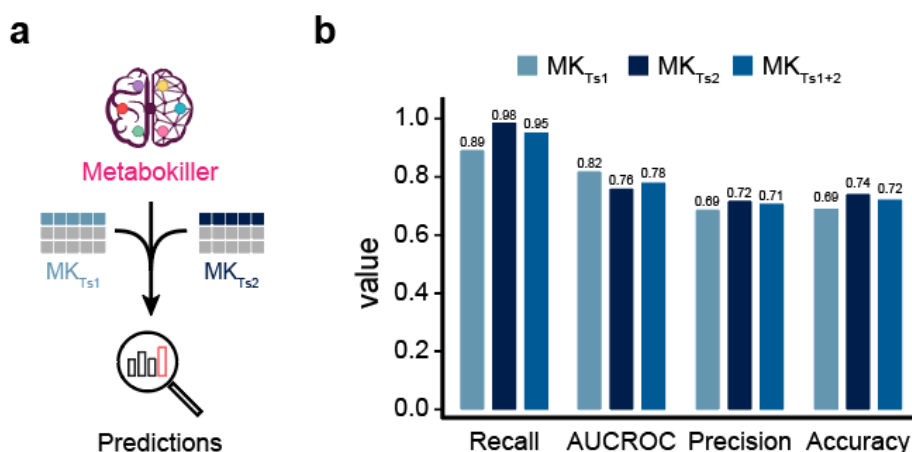
Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.5 Integrative Framework of Metabokiller: Leveraging Biochemical Properties for Robust Carcinogenicity Prediction

(a) Venn diagram depicting overlaps between the datasets used to build six independent biochemical property-based models (annotated as MK_{Tn}), the training dataset used to build the Metabokiller ensemble model (annotated as MK_{ETn}), and the external testing data (annotated as MK_{Ts1+2}). (b) Variables factor map (PCA) depicting the direction and contribution of all the six variables (individual models) representing the experimentally validated carcinogens (MK_{ETn}) in the Eigenspace. (c) Box plot depicting the model performance of the twenty Gradient Boosting Machine (GBM)-based models generated using bootstrapping technique ($n = 20$ repetitions). (d) Schematic representation depicting Metabokiller architecture. Metabokiller relies on six distinct carcinogen-centric biochemical properties for carcinogenicity prediction. Metabokiller offers two main modules, prediction and interpretability.

To comprehensively evaluate the performance of Metabokiller on unseen external datasets (MK_{Ts1} and MK_{Ts2}), a list of experimentally validated and chemically diverse carcinogens and non-carcinogens was compiled from multiple independent sources. Metabokiller demonstrated highly competitive performance

metrics, highlighting the robustness and reliability of its ensemble approach. Notably, Metabokiller demonstrated consistently high precision (0.69, 0.72, and 0.71) and recall values (0.89, 0.98, and 0.95) across MK_{Ts1} , MK_{Ts2} , and the combined dataset (MK_{Ts1+2}) datasets (**Figure 2.6**). These results indicate Metabokiller's strong capability to accurately identify carcinogens while maintaining a low rate of false negatives, which is particularly critical for applications in predictive toxicology and public health. Together, these metrics underscore the model's balance between sensitivity and specificity, making it a reliable tool for carcinogenicity prediction.



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.6 Metabokiller's Robust Performance on External Testing Datasets

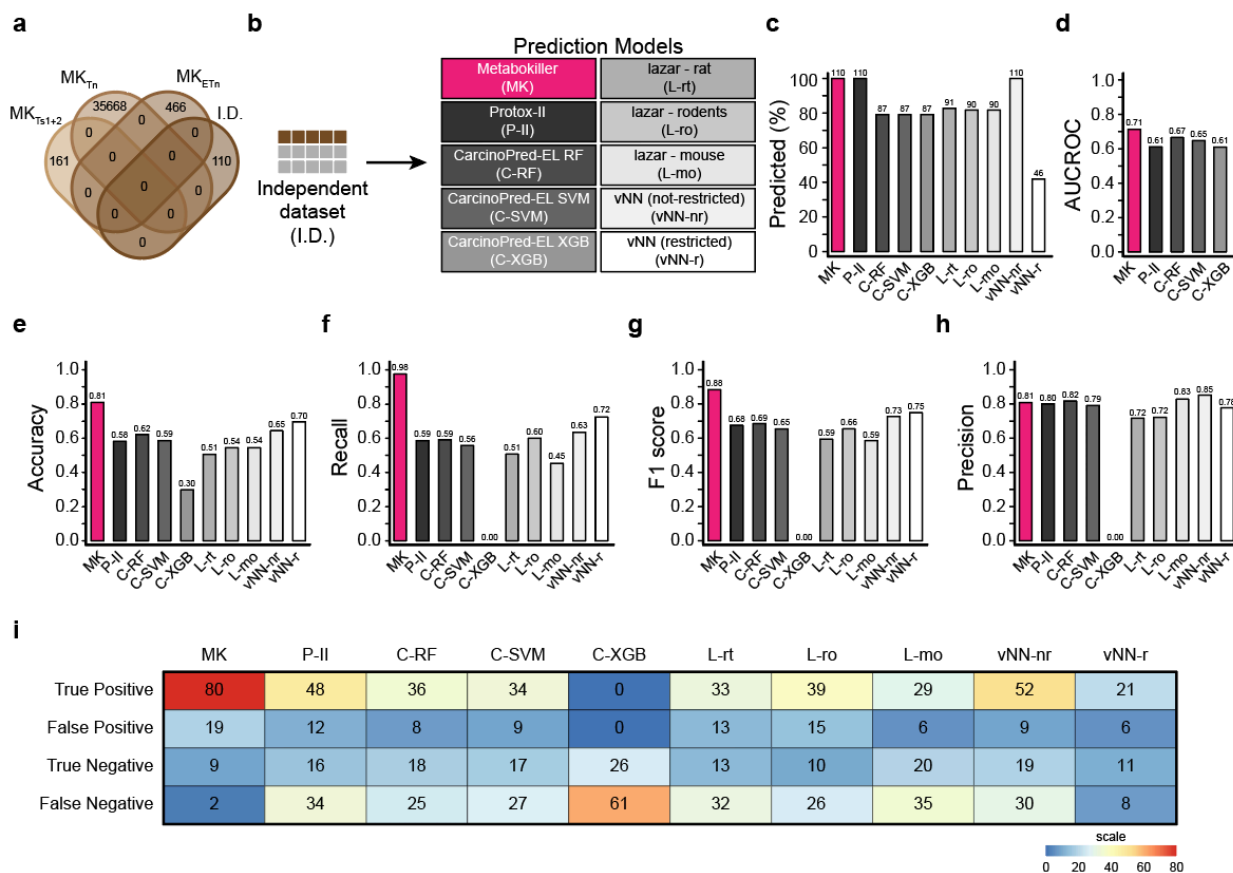
(a) Schematic representation of the projection of the external testing datasets containing information about experimentally validated carcinogens and non-carcinogens (annotated as MK_{Ts1} and MK_{Ts2}) on Metabokiller. (b) Bar graph depicting the indicated performance estimation metrics of Metabokiller on external testing datasets.

In summary, Metabokiller represents a novel and effective approach for carcinogenicity prediction, demonstrating high predictive performance on unseen datasets. Unlike traditional models that rely on limited datasets of known carcinogens and non-carcinogens, Metabokiller is designed to integrate and analyze diverse biochemical properties associated with cancer, allowing for a more comprehensive evaluation of carcinogenic potential. This unique framework not only enhances its predictive accuracy but also provides a level of biological interpretability that is often lacking in other machine learning-based tools. By combining its robust prediction accuracy with a strong emphasis on interpretability, Metabokiller not only serves as a reliable tool for carcinogenicity prediction but also provides a framework for understanding the biochemical basis of carcinogenicity. This dual capability makes it a powerful resource for applications in the public health domain, predictive toxicology, and drug development, where accurate identification of carcinogenic compounds is essential. These results further

solidify Metabokiller as a robust and generalizable model for carcinogenicity prediction, bridging the gap between experimental validation and computational predictions.

2.3.2 Metabokiller Outperforms Other Models in Carcinogenicity Prediction

To evaluate the predictive capabilities of Metabokiller in comparison to other widely used tools, we conducted a comprehensive comparative analysis using an independent dataset (I.D.) of experimentally validated carcinogens and non-carcinogens. To ensure that the evaluation was unbiased and reflective of the model's ability to generalize to novel compounds, this dataset was meticulously curated to exclude any overlapping compounds with those used in training or testing phases of Metabokiller models (MK_{Tn}, MK_{ETn}, and MK_{Ts1+2}), or with the datasets used for training the other tools/models included in the comparison (Figure 2.7a).



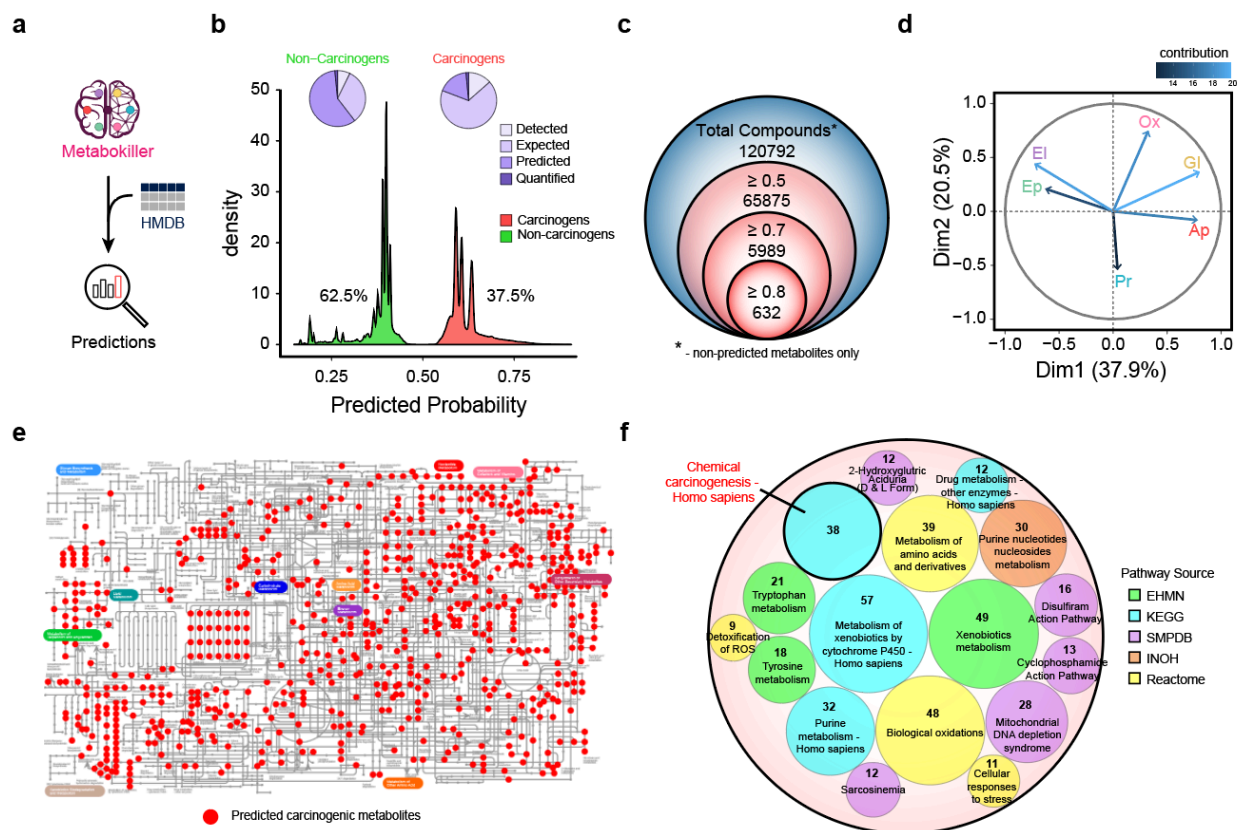
Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.7 Metabokiller Outperforms Other Prediction Methods.

(a) Venn diagram depicting non-overlapping compounds selected for the cross-model performance evaluation (annotated as independent dataset (ID)). (b) Diagram depicting the strategy used to cross-compare the performance of Metabokiller with indicated best-practice methods. (c) Percentage bar graph depicting the number of compounds

predicted by the indicated methods. **(d-h)** Bar graphs depicting the **(d)** AUCROC, **(e)** accuracy, **(f)** recall, **(g)** F1 score, and **(h)** precision values of the indicated models on the independent dataset. **(i)** Heatmap depicting the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions on the Independent Dataset for indicated methods/tools.

For this comparison, Metabokiller was tested alongside nine other well-established tools/methods for carcinogenicity prediction. These included Protox-II (Banerjee et al. 2018), CarcinoPred-EL (RF, SVM, and XGB) (L. Zhang et al. 2017), lazar (rat, rodents, and mouse) (Maunz et al. 2013), and versatile variable nearest neighbor (vNN-r for restricted and vNN-nr for non-restricted) (Schyman et al. 2017) (**Figure 2.7b**). These tools represent a diverse array of computational approaches commonly employed in the field. Unlike Metabokiller, which incorporates probabilistic outputs from six biochemical properties to provide a holistic prediction, these tools rely solely on limited datasets of validated (non-)carcinogens (Protox-II, lazar, and CarcinoPred-EL) or (non-)mutagens (vNN). Such reliance on limited data can constrain the predictive scope and accuracy of these tools, particularly when applied to chemically diverse or novel compounds.



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.8 Uncovering Carcinogenic Potential of Human Metabolites through Metabokiller

(a) Scheme depicting the *in silico* screening of HMDB metabolites using Metabokiller. **(b)** Density plots depicting the distribution of Metabokiller prediction probabilities of HMDB metabolites. Red and green represent carcinogens and non-carcinogens, respectively. Pie charts at the top represent relative proportions of HMDB metabolites, annotated as detected, expected, predicted, and quantified. **(c)** Venn diagram depicting predicted carcinogenic human metabolites, further segregated based on prediction probability cutoffs. **(d)** Variables factor map depicting the contribution of all the six individual models in predicting carcinogenic metabolites from HMDB (probability cutoff ≥ 0.5). **(e)** Projection of the predicted carcinogens (indicated as red dots; probability cutoff ≥ 0.7) on the human metabolic space, achieved using iPath Web Server. **(f)** Circle packing chart representing metabolic pathway enrichment results of the predicted carcinogenic metabolites (probability cutoff ≥ 0.7) using Metabokiller. The pathway information is gathered from the Edinburgh Human Metabolic Network (EHMN), the Integrating Network Objects with Hierarchies (INOH), Reactome, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Small Molecule Pathway Database (SMPDB).

Comparative analyses across all methods revealed that Metabokiller outperformed its counterparts in most performance metrics, showcasing its robustness and reliability. Metabokiller achieved the highest recall (0.98), allowing it to accurately identify true positives (carcinogens) with minimal false negatives. Similarly, Metabokiller attained an accuracy of 0.81, demonstrating its ability to correctly classify both carcinogens and non-carcinogens across the query dataset. In contrast, the other models demonstrated lower recall and accuracy scores, with some failing to make predictions for certain query compounds, further highlighting their limitations (**Figure 2.7c-h**).

This superior performance of Metabokiller can be attributed to its unique design, which integrates diverse biochemical properties such as anti-apoptotic activity, electrophilicity, oxidative stress, epigenetic modifications, genomic instability, and proliferation. This holistic approach enables Metabokiller to integrate key chemical and functional features that are critical for carcinogenicity prediction, enabling it to accurately classify a wide range of compounds with varying chemical structures and mechanisms of action. Unlike traditional models that rely solely on limited datasets of known carcinogens and non-carcinogens, Metabokiller leverages a multidimensional framework that reflects the complexity of carcinogenicity, distinguishing it as a robust and reliable tool for predictive carcinogenicity.

2.3.3 Identification of Potential Endogenous Metabolic Threats by Metabokiller

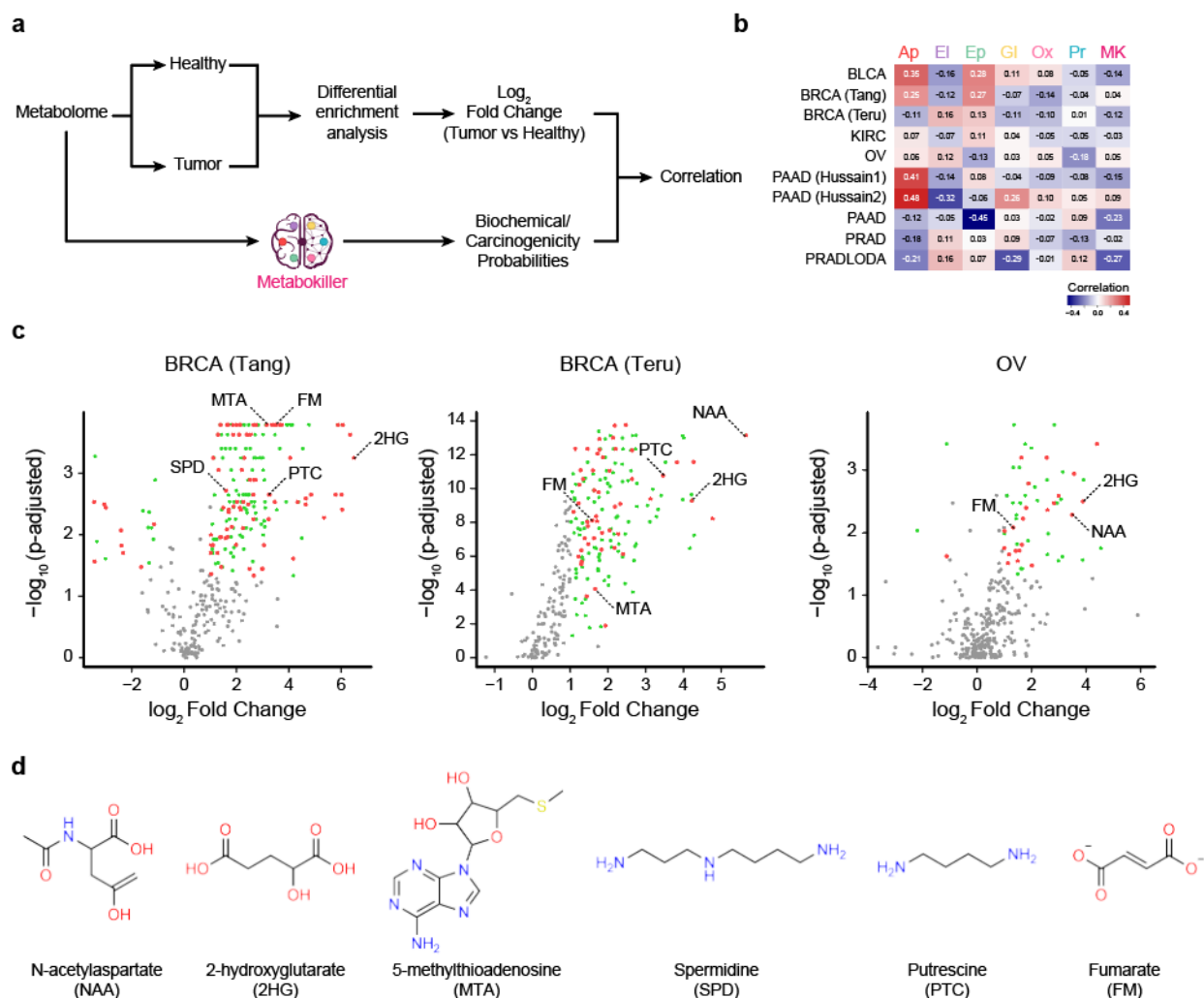
Metabokiller was next employed to explore the potential carcinogenicity of endogenous cellular metabolites. To achieve this, we conducted a large-scale *in silico* screening of the human metabolome by projecting all human metabolites cataloged in the Human Metabolome Database (HMDB) (Wishart et al. 2007) onto Metabokiller and calculating their carcinogenicity probabilities (**Figure 2.8a**). HMDB is a

comprehensive repository of small molecule metabolites found in the human body, accounting for a total of 217,921 metabolites. Given that HMDB also contains predicted metabolites, the downstream analysis was focused exclusively on non-predicted metabolites. Remarkably, the analysis revealed that approximately 37.5% of the metabolites analyzed were predicted to exhibit carcinogenic properties using a probability threshold of 0.5 (**Figure 2.8b, c**). Applying more stringent probability cut-offs, 5,989 metabolites with carcinogenic probabilities of 0.7 or greater and 632 metabolites with probabilities of 0.8 or greater were identified (**Figure 2.8c**). These stringent thresholds provided a narrower subset of metabolites with a high likelihood of possessing carcinogenic properties, allowing for more targeted exploration of their roles in cellular processes.

To understand the relationships between the biochemical properties associated with these predicted endogenous carcinogenic metabolites, their prediction probability vectors were computed and compared within the eigenspace (**Figure 2.8d**). This approach allowed us to assess the extent to which the six individual biochemical properties contribute to the carcinogenicity predictions. This analysis demonstrated that all six biochemical properties comprehensively and non-redundantly captured the carcinogenic feature space, emphasizing the robustness and versatility of the biochemical-based framework employed in Metabokiller. Furthermore, when these predicted carcinogenic metabolites (probability cutoff ≥ 0.5) were mapped to pathway space, selective enrichment in specific metabolic pathways was observed (**Figure 2.8e**). An over-representation analysis focusing on metabolites with probabilities ≥ 0.7 revealed enrichment in pathways related to tyrosine and tryptophan metabolism, nucleotide metabolism, drug metabolism, xenobiotics, and oxidative stress, among others. Notably, this unbiased functional analysis also identified 38 previously characterized carcinogenic metabolites, validating the predictive accuracy of Metabokiller (**Figure 2.8f**). Examples of predicted carcinogenic metabolites from HMDB included melphalan, sulfur mustard, 1,2-dichloroethane, chloral, thiotepa, N-nitrosodimethylamine, and 2,2,2-trichloroethanol. These metabolites are known for their roles in DNA damage, mutagenesis, and disruption of cellular homeostasis, which are hallmark mechanisms of carcinogenesis.

As a complementary case study, we used Metabokiller to assess the functional relevance of known oncometabolites. To this end, a recently integrated cancer metabolomics dataset was re-analyzed to identify metabolites enriched or depleted in tumors (Reznik et al. 2018). Using differential enrichment analysis, the $\log_2$ fold changes of tumor-associated metabolites were calculated to determine their relative abundance compared to non-tumor conditions. Next, all identified tumor-associated metabolites were projected onto Metabokiller to calculate their carcinogenicity prediction probabilities (**Figure 2.9a**). A dataset-specific correlation analysis between $\log_2$ fold change values and carcinogenicity probabilities

revealed a weak negative correlation overall, suggesting that while some metabolites are enriched in tumors, their predicted carcinogenicity may not necessarily increase proportionally (**Figure 2.9b**). To further explore this relationship, an in-depth investigation of metabolites significantly enriched in tumors, defined by a $\log_2$ fold change ≥ 1 and p-value < 0.05 , was performed, which identified a subset of previously reported oncometabolites, including 2-hydroxyglutarate, succinate, fumarate, N-acetyl aspartate, adenylosuccinate, and polyamines such as putrescine, spermidine, and spermine (Dando et al. 2019), further validating the predictive capabilities of Metabokiller and its relevance for uncovering biologically significant carcinogenic metabolites (**Figure 2.9c, d**). In summary, by integrating both the biological and chemical properties of carcinogens, Metabokiller enabled the prediction of a subset of human metabolites with carcinogenic potential.



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.9 Metabokiller Unravels Potential Oncometabolites.

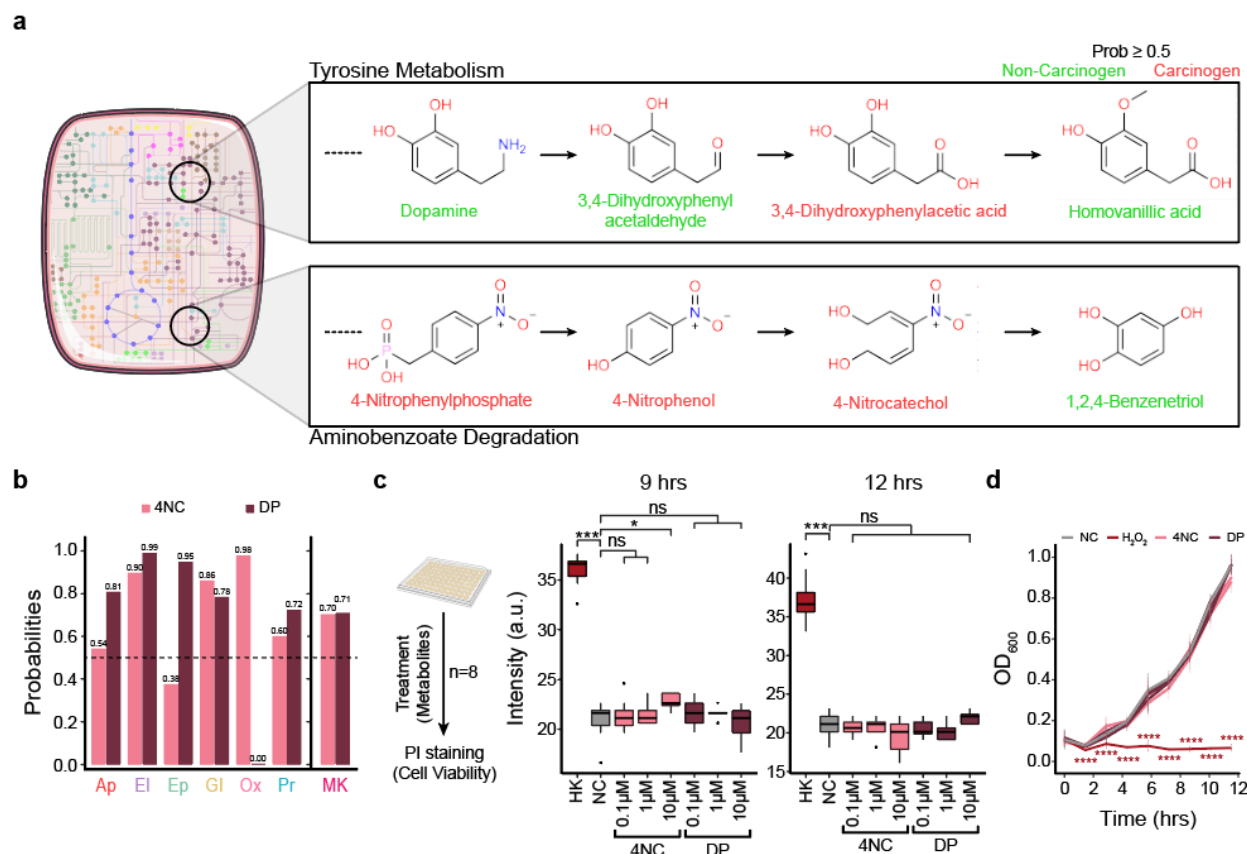
(a) Schematic representation of the steps involved in processing pan-cancer metabolomics dataset. Of note, Pearson correlation was computed between $\log_2$ fold change (tumor vs healthy) and biochemical/carcinogenicity probabilities. **(b)** Heatmap detailing the correlation values further segregated based on cancer type. **(c)** Volcano plots depicting the differentially enriched/de-enriched metabolites in the indicated cancer datasets. Gray dots highlight the metabolites that do not qualify for the enrichment cutoff ($\log_2$ fold change ≥ 1 or ≤ -1 , and p -value (adjusted) < 0.05), and green and red dots represent the metabolites that qualify for the enrichment cutoff and are predicted as non-carcinogenic and carcinogenic by Metabokiller respectively. The p -value was computed using two-sided Mann-Whitney U test and corrected using Benjamini-Hochberg method. **(d)** Structural information of some of the well-characterized oncometabolites reported in the literature and predicted by Metabokiller.

2.3.4 Experimental Validation of Metabokiller Predictions.

To validate the predictions made by Metabokiller, two previously uncharacterized human metabolites, 4-nitrocatechol (4NC) and 3,4-dihydroxyphenylacetic acid (DP), were selected for further experimental analysis. DP is a known metabolic intermediate in the tyrosine metabolism pathway, while 4NC is involved in aminobenzoate degradation (**Figure 2.10a**). Interestingly, DP was predicted to be a non-carcinogen by all alternative methods except Metabokiller, which classified it as a potential carcinogen with a prediction probability of 0.71. This discrepancy highlights the sensitivity of Metabokiller in capturing carcinogenic features overlooked by other models. Analysis of the individual biochemical properties contributing to the prediction revealed that DP's carcinogenic potential is associated with its high genotoxicity, electrophilic nature, pro-proliferative and anti-apoptotic properties, and its ability to alter the epigenetic landscape. Interestingly, its probability score for inducing oxidative stress was below the threshold, suggesting that DP may not contribute to carcinogenicity through reactive oxygen species (ROS) production. Similarly, 4NC was predicted as a potential carcinogen by Metabokiller with a prediction probability of 0.70. Unlike DP, 4NC exhibited high scores across all individual models except for its potential to alter the epigenetic landscape (**Figure 2.10b**).

High concentrations of certain biomolecules can often have pleiotropic effects, indirectly impairing key cellular processes and potentially confounding the interpretation of experimental results. To evaluate whether the concentrations of 4NC and DP used in the downstream assays affected the survival and physiology of *Saccharomyces cerevisiae*, a propidium iodide-based cell viability assay was conducted. In this assay, yeast cells were exposed to various concentrations of 4NC and DP (0.1 μM , 1 μM , 10 μM). Following treatment, the cells were stained with propidium iodide, a dye that selectively binds to nucleic acids in cells with compromised membranes, allowing for the quantification of non-viable cells. The results showed no significant increase in cell death across all tested concentrations of either metabolite,

indicating that neither 4NC nor DP caused substantial cytotoxic effects under the conditions used (**Figure 2.10c, d**).



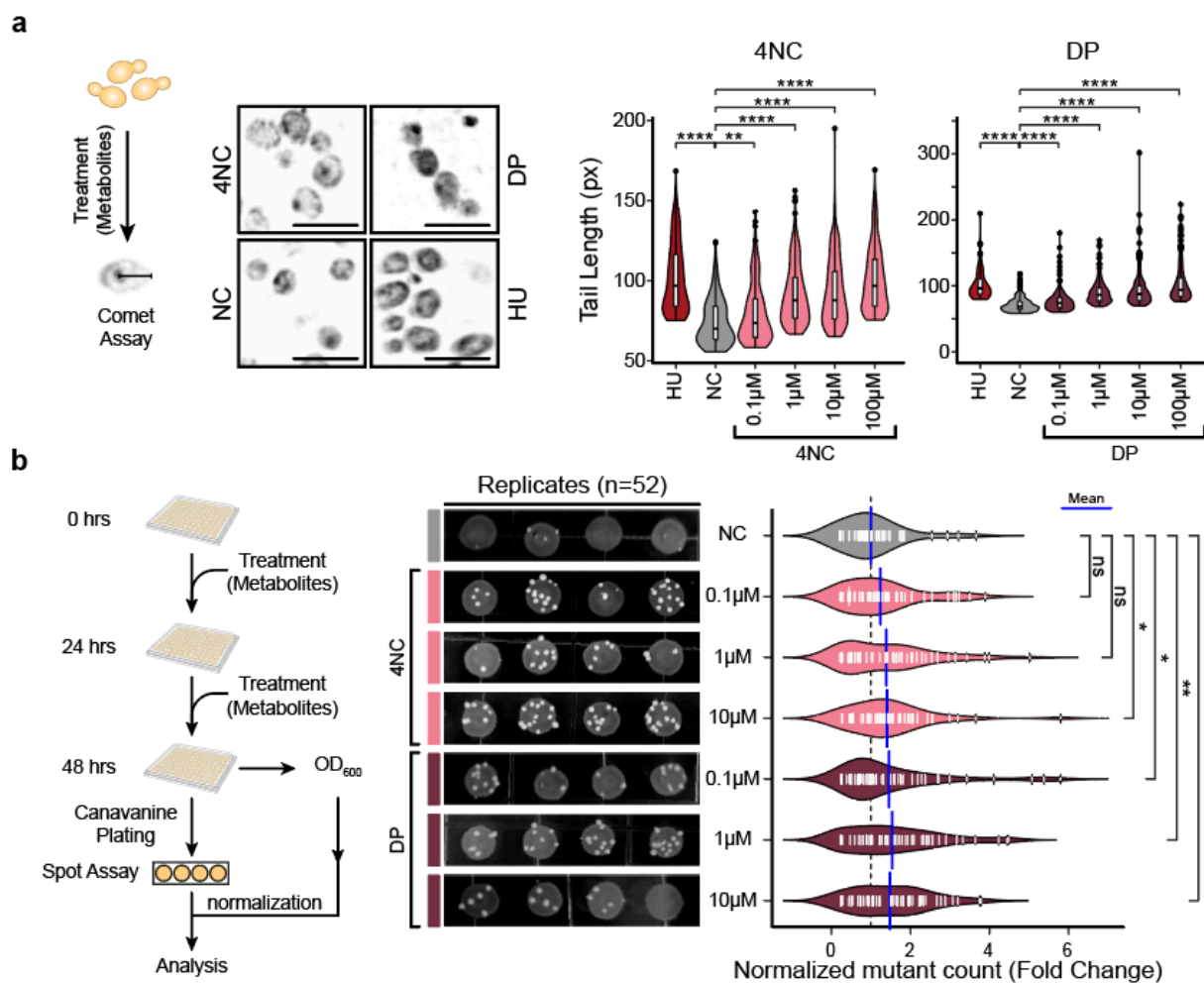
Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.10 Metabokiller's Insights into Carcinogenic Potential of 4NC and DP

(a) Schematic representation highlighting the predicted-carcinogenic metabolic intermediates of the tyrosine metabolism pathway and aminobenzoate degradation pathway. **(b)** Bar graph denoting Metabokiller predictions for 4NC and DP. **(c)** Box plots depicting the fluorescence intensity of propidium iodide staining indicating cell viability in the indicated conditions ($n = 8$ biological replicates) after 9 hours (left) and 12 hours (right) of treatment. Of note, heat-killed (HK) yeast cells were used as a positive control. Two-sided Mann-Whitney U test was used to compute statistical significance between the test conditions and the negative control. **(d)** Growth curve profiles of the treated and untreated wild-type yeast during transient exposure with the indicated conditions ($n = 8$ biological replicates with technical duplicates). Data points represent mean $\pm$ SD. Two-sided Student's t -test was used to compute statistical significance between the positive (H_2O_2 treated yeast cells) and negative control (untreated yeast cells).

To evaluate the potential of 4NC and DP to induce genotoxicity and genomic instability, we performed a single-cell gel electrophoresis assay, commonly referred to as the comet assay, a widely accepted method for detecting DNA breaks and lesions. The experiments were conducted using *Saccharomyces cerevisiae*

(budding yeast) as a model organism. Yeast cells were exposed to varying concentrations of 4NC and DP to assess the dose-dependent effects of these metabolites. Following exposure, lysed cells were stained with Giemsa dye for visualization of comet-like tail structures indicative of DNA damage. Hydroxyurea (HU), a known DNA repair inhibitor with mutagenic and genotoxic properties, was used as a positive control. Quantitative analysis of comet tail lengths revealed a statistically significant, dose-dependent increase in DNA damage in cells treated with 4NC and DP, indicating their potential to induce genomic instability (**Figure 2.11a**). These results align with the carcinogenic potential predicted by Metabokiller and provide experimental evidence supporting the genotoxicity of these metabolites.



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.11 Experimental Validation of Genotoxicity and Mutagenic Potential of 4NC and DP

(a) Micrographs (left) depicting comet assay results. NC and hydroxyurea (HU) represent negative and positive controls, respectively. Violin plots (right) depicting the distribution of comet tail lengths ($n = 3$ biological replicates). Scale, $10\ \mu\text{m}$. **(b)** Schematic (left) depicting mutagenesis experiments workflow. Spot assay micrographs (middle)

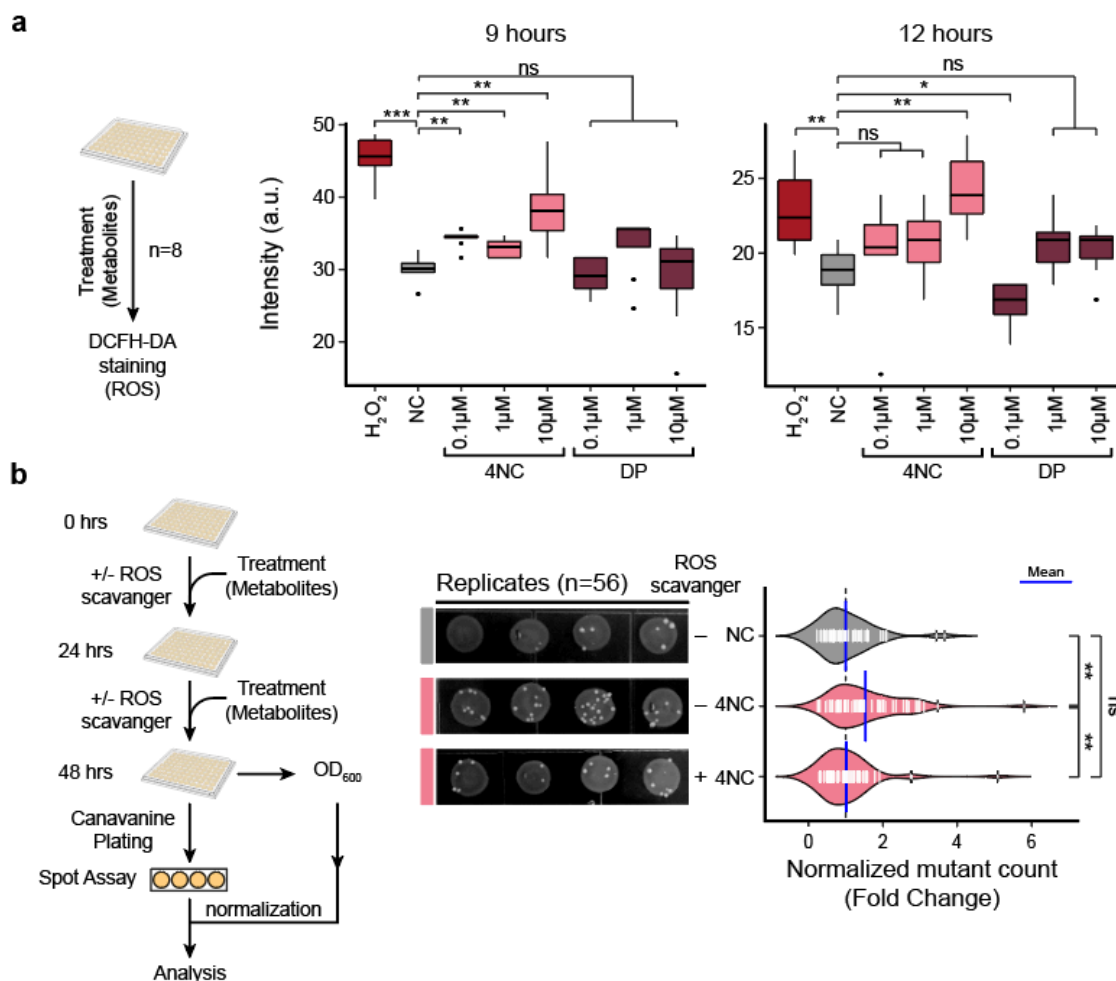
depicting canavanine-resistant mutants ($n = 52$ biological replicates). Bean plot (right) depicting the distribution of normalized frequency counts of canavanine-resistant mutants.

To further investigate the mutagenic and genotoxic properties of 4NC and DP, we conducted mutagenesis assays to determine whether these metabolites can induce genetic mutations. Since the training dataset for the genomic instability predictions in Metabokiller includes compounds that have been screened using the Ames test, 4NC and DP might also exhibit mutagenic potential. To test this hypothesis, a high-throughput mutagenesis assay using *Saccharomyces cerevisiae* as a model organism was performed, screening for canavanine-resistant mutants, which serve as an indicator of mutational events in the *CAN1* gene. Yeast cells were treated with different concentrations of 4NC and DP alongside negative vehicle controls to determine the frequency of mutagenesis. The results showed a significant increase in the frequency of canavanine-resistant mutants in treated conditions compared to the controls. Normalized mutational frequencies for untreated conditions revealed a mean mutant count of 6.4 per 10^7 cells in 100 μ l of culture. In treated conditions, both 4NC and DP exhibited markedly higher mutant counts, with higher mean mutant counts at elevated concentrations, further supporting their mutagenic potential (**Figure 2.11b**). Taken together, this dose-dependent increase in mutational frequency provides compelling experimental evidence for the mutagenic and genotoxic potential of both DP and 4NC, consistent with Metabokiller's genomic instability predictions.

To investigate whether the observed genotoxic effects of 4NC were associated with the induction of oxidative stress, we examined its ability to generate reactive oxygen species (ROS) in *Saccharomyces cerevisiae*. Metabokiller predictions indicated that 4NC, but not DP, had the potential to induce elevated ROS levels (**Figure 2.10b**). To experimentally validate this prediction, a fluorometric assay using the ROS-sensitive dye 2',7'-dichlorofluorescein diacetate (DCFH-DA) was used to quantify the oxidative stress. The results demonstrated a selective and concentration-dependent increase in ROS levels in yeast cells treated with 4NC, whereas no significant ROS production was observed in cells treated with DP (**Figure 2.12a**). These findings strongly align with the prediction probabilities generated by Metabokiller, reinforcing the model's accuracy in identifying oxidative stress as a potential mechanism underlying 4NC's genotoxicity.

To establish a direct link between elevated ROS levels and the genotoxic effects of 4NC, we conducted mutagenesis rescue experiments by co-incubating yeast cells with 4NC and a ROS scavenger, L-ascorbic acid. The L-ascorbic acid neutralizes ROS, thereby mitigating oxidative stress. In these experiments, we observed a significant reduction in the number of canavanine-resistant mutants in cells treated with the L-ascorbic acid compared to cells treated with 4NC alone. This decrease in mutation frequency strongly

suggests that the genotoxic effects of 4NC are potentially mediated through ROS generation, providing mechanistic insight into the carcinogenic potential of 4NC as predicted by Metabokiller (**Figure 2.12b**).

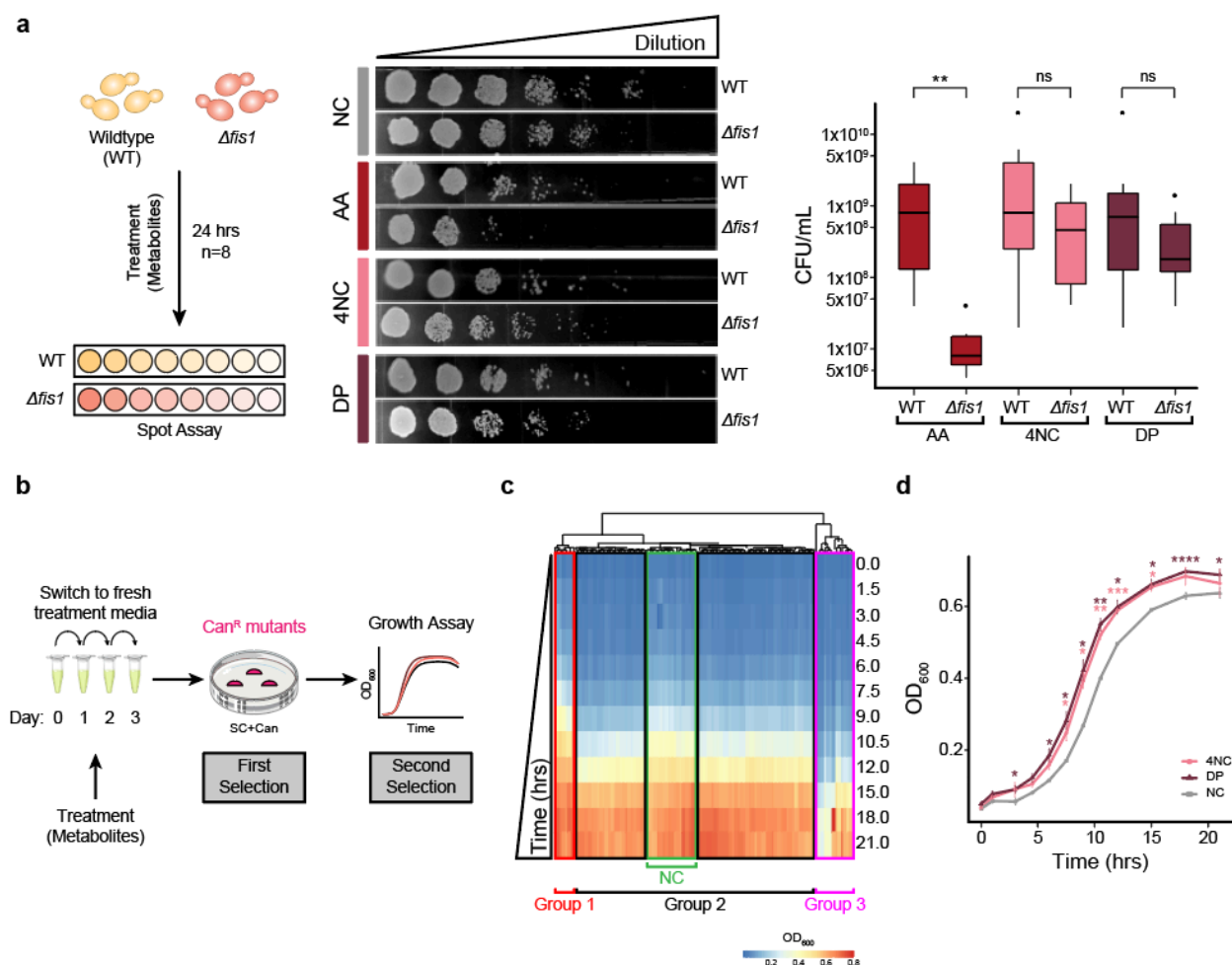


Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.12 Mechanistic Insights into 4NC-induced Genotoxicity

(a) Schematic diagram on the left depicting the experimental workflow used to measure reactive oxygen species (ROS) using DCFH-DA dye-based assay. Box plots on the right depicting the results of reactive oxygen species (ROS) levels inferred using DCFH-DA dye-based assay in the indicated conditions ($n = 8$ biological replicates). ROS levels were measured 9 hours and 12 hours post-incubation. Notably, hydrogen peroxide (H_2O_2) treated yeast cells were used as a positive control. Two-sided Mann-Whitney U test was used to compute statistical significance between the test conditions and the negative control. a.u., arbitrary units. **(b)** Schematic (left) depicting mutagenesis rescue experiment workflow. Micrographs (middle) depicting the rescue of mutagenic effects of 4NC by ROS scavenger ($n = 56$ biological replicates). Bean plot (right) depicting the distribution of normalized frequency counts of canavanine-resistant mutants.

To further assess the predictions of Metabokiller, the ability of 4NC and DP to induce anti-apoptotic responses was evaluated. Using a yeast-based apoptosis assay, wild-type yeast cells and $\Delta fis1$ knockout strains were treated with 4NC and DP for 24 hours. Acetic acid, a known apoptosis inducer, was used as a positive control. In wild-type yeast cells, treatment with DP, 4NC, and acetic acid resulted in significant levels of cell death, indicative of an apoptotic response. However, a marked reduction in cell death was observed in $\Delta fis1$ knockout cells treated with DP or 4NC, compared to those treated with acetic acid (**Figure 2.13a**). This reduction suggests that both 4NC and DP activate anti-apoptotic pathways, allowing the cells to partially evade apoptosis despite apoptotic stimuli, a property not observed with acetic acid. These experimental results were consistent with Metabokiller's predictions, further validating its interpretability and robustness in identifying mechanisms of carcinogenicity.



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.13 4NC and DP Trigger an Anti-apoptotic Response in Yeast.

(a) Schematic (left) depicting the experimental design of the apoptosis assay. Micrographs (middle) depicting the DP and 4NC-induced anti-apoptotic response. NC and acetic acid (AA) represent negative and positive controls,

respectively. Box plot (right) highlighting colony forming units per milliliter (CFU/mL) in the indicated conditions ($n = 8$ biological replicates). Two-sided Mann-Whitney U test was used to compute statistical significance. **(b)** Scheme depicting the sequential selection of canavanine-resistant yeast mutants with distinct growth profiles. **(c)** Heat map depicting the growth profiles of the metabolite-induced canavanine-resistant mutants. Mutants were clustered based on their growth profiles. Three major clusters were obtained with either accelerated (Group 1), decelerated (Group 3) or normal growth (Group 2). The green-colored rectangle in the middle indicates the growth profiles of the untreated cells (NC). **(d)** Growth curve profiles of wild-type untreated (NC), 4NC-treated, and DP-treated mutants with accelerated growth under optimal growth conditions ($n = 3$ and 8 biological replicates for metabolite-treated and NC, respectively). Error bars represent standard errors. Student's t -test were used to compute statistical significance.

Given the prediction by Metabokiller that 4NC and DP as potential carcinogens, an investigation was conducted to determine whether mutations induced by these metabolites could alter genes associated with carcinogenesis. The study employed a two-step selection procedure to identify and characterize metabolite-induced mutants. In the first selection step, yeast cells exposed to 4NC and DP were screened for canavanine resistance. This resistance is indicative of mutations in the *CAN1* gene, confirming the mutagenic potential of the tested metabolites. A total of 167 canavanine-resistant mutants were obtained, representing a population of cells that had undergone mutagenic events. In the second selection step, these mutants were further analyzed to assess alterations in growth patterns through a growth kinetics assay. To provide a baseline comparison, randomly selected wild-type clones, which were not subjected to the first selection step, were also included in the analysis (**Figure 2.13b**).

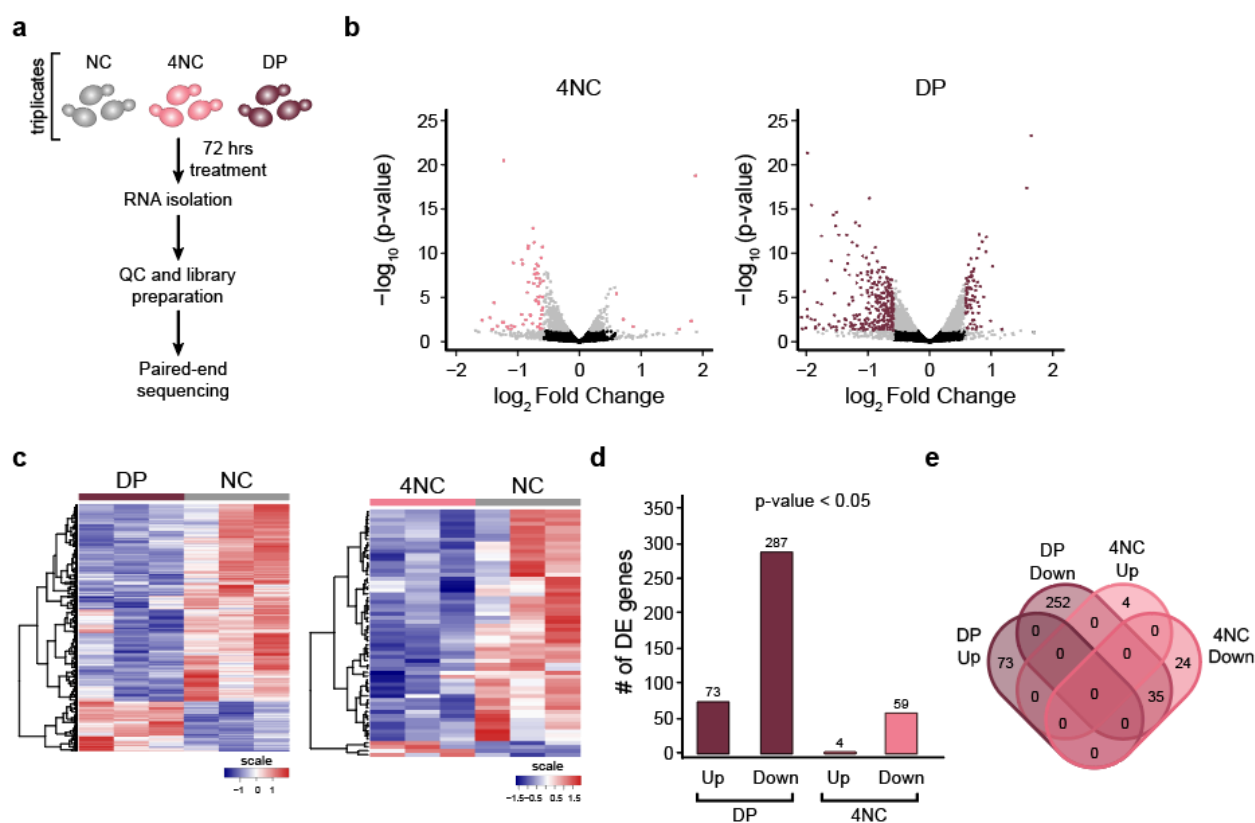
Growth profiles were obtained for all mutants and wild-type clones, and hierarchical clustering based on growth kinetics revealed three distinct phenotypic groups. The first group displayed accelerated growth compared to wild-type cells, suggesting potential mutations in genes that enhance cellular proliferation. The second group included mutants with growth patterns similar to those of wild-type yeast, indicating mutations with negligible effects on cellular proliferation and overall growth kinetics. The third group consisted of mutants with decelerated growth, suggesting detrimental mutations that impaired cellular fitness or metabolic efficiency (**Figure 2.13c, d**). Collectively, these findings demonstrate that the mutagenic effects of 4NC and DP result in random genetic alterations that can drive diverse phenotypic changes, supporting the carcinogenic potential of these metabolites as predicted by Metabokiller.

These findings demonstrate the power of Metabokiller's multi-property approach, which leverages biochemical diversity to identify subtle carcinogenic traits. The detailed probability breakdown for each property allows for a mechanistic understanding of how these metabolites might contribute to carcinogenesis, providing a robust basis for experimental follow-up studies. By accurately predicting the

carcinogenic potential of DP and 4NC and explaining their contributions through a property-specific lens, Metabokiller offers a significant advantage over traditional models, reinforcing its utility in metabolomic carcinogenicity analysis. These predictions pave the way for experimental validation of their effects on cellular systems and their potential roles in cancer biology.

2.3.5 Transcriptomic Analysis of 4NC and DP-Induced Mutants

To investigate the molecular mechanisms impaired by the treatment of 4NC and DP, paired-end deep RNA sequencing was performed on both wild-type and mutant yeast cells in biological triplicates. The mutants selected for this analysis were canavanine-resistant (identified in the first selection step) and exhibited accelerated growth kinetics (identified in the second selection step), mimicking the proliferative behavior observed in cancer cells. This approach provided a comprehensive view of the transcriptomic alterations induced by these metabolites (**Figure 2.14a**).



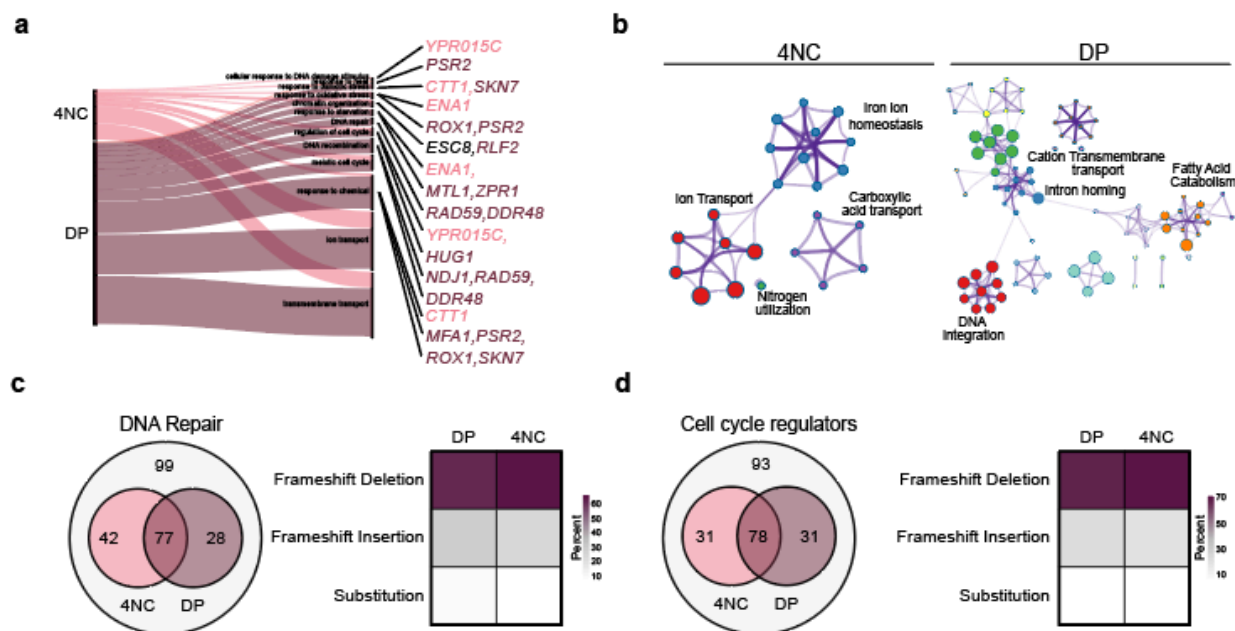
Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.14 RNA Sequencing Reveals Differential Transcriptomic Profiles Induced by 4NC and DP Treatments.

(a) Schematic representation depicting the experimental design of RNA sequencing, featuring the group information, treatment duration, and the sequencing parameters. **(b)** Volcano plot indicating the differentially expressed genes (DEGs) between the treated and untreated conditions. p-value was computed using Wald test and corrected using

Benjamini-Hochberg method. **(c)** Heat maps representing the expression profile of differentially expressed genes in DP and untreated (NC) conditions (left) and 4NC and NC conditions (right). **(d)** Bar graph depicting the number of DEGs after fold-change cutoff of 1.5, and a $p < 0.05$ in the indicated conditions. p -values were computed using Wald test and corrected using the Benjamini-Hochberg method. **(e)** Venn diagram depicting the number of differentially expressed genes shared among 4NC and DP (as compared to NC).

Differential gene expression analysis of the normalized read counts revealed a substantial number of downregulated genes in both DP- and 4NC-treated mutants compared to wild-type controls (**Figure 2.14b-d**). Interestingly, while no overlap was observed in the sets of upregulated genes between the two treatment conditions, 35 genes were commonly downregulated, suggesting certain shared molecular targets between the two but through distinct mechanisms (**Figure 2.14e**). To gain deeper insights into the functional pathways affected by these downregulated genes, a Gene Ontology slim term analysis was conducted. The analysis revealed significant enrichment of impaired genes in pathways associated with DNA repair (*RAD59* and *DDR48*), chromatin organization genes (*ESC8* and *RLF2*), DNA recombination (e.g., *NDJ1*, *RAD59*, and *DDR48*), and cell cycle regulation (e.g., *YPR015C* and *HUG1*). Genes involved in cellular stress responses were also significantly altered, indicating a broad disruption of essential cellular processes (**Figure 2.15a, b**).



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.15 Gene Ontology Analysis Highlights Pathways Impacted by 4NC and DP

(a) Gene Ontology (GO) slim mapper analysis depicting the association of the differentially expressed genes in the indicated ontologies. **(b)** Metascape-based Functional Gene Ontology analysis identified the involvement of

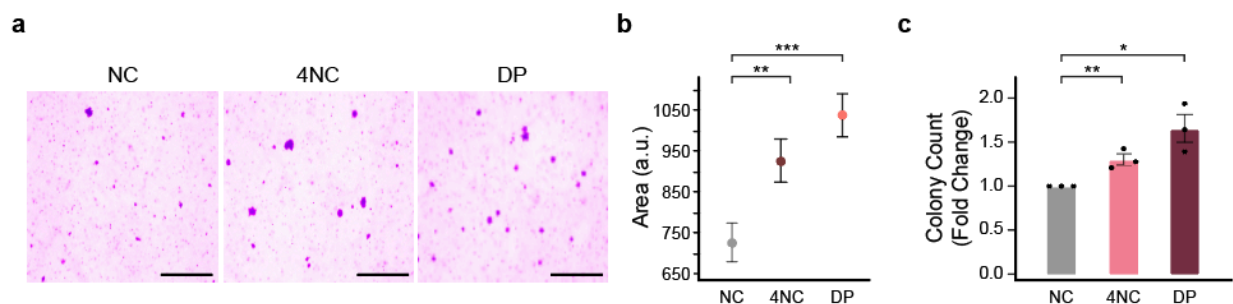
differentially expressed genes in the indicated prominent biological processes. (c, d) The Venn diagrams and their accompanying heat maps on the right side collectively depicting the number of mutated genes (insertions, deletions, and substitution mutations) involved in DNA repair (c) or cell cycle regulation (d). The heat maps on the right further segregate the identified mutations as frameshift (insertion or deletion) and substitution

To further investigate the mutational landscape, insertion-deletion (indel) mutations were analyzed from RNA sequencing data. Using raw FASTQ files, several indel mutations were identified in the *CAN1* gene of mutants, validating the initial selection step based on canavanine resistance. Applying conservative filtering criteria (supporting read counts ≥ 1), additional mutations were detected in multiple protein-coding genes involved in critical pathways such as DNA repair and cell cycle regulation (**Figure 2.15c, d**). These findings provide molecular evidence to explain the observed phenotypic responses in mutants treated with DP and 4NC, reinforcing the connection between genotoxicity and impaired cellular processes.

2.3.6 4NC and DP Induce Malignant Transformation in Human Cells.

To evaluate the carcinogenic effects of 4-nitrocatechol (4NC) and 3,4-dihydroxyphenylacetic acid (DP) on human cells, a soft-agar assay was conducted using BEAS-2B normal human epithelial cells. The soft-agar assay is a well-established functional test for detecting malignant transformation. The results of this assay demonstrated that both 4NC and DP induced malignant transformation in BEAS-2B cells. Metabolite-treated cells exhibited robust colony formation in the soft agar, indicative of their capacity for anchorage-independent growth and providing direct evidence of their carcinogenic properties. These findings not only validate the predictions made by Metabokiller but also provide critical insights into the molecular basis of the carcinogenic potential of these metabolites.

Microscopic imaging revealed clear differences between control (NC), 4NC-treated, and DP-treated cells. In control conditions, relatively few colonies were observed, and their sizes were small, indicating limited anchorage-independent growth. In contrast, cells treated with 4NC and DP exhibited significantly more colonies with markedly larger sizes, indicative of malignant transformation. This visual evidence highlights the transformative effects of these metabolites. Quantification of the colony size revealed a significant increase in mean colony area for both 4NC- and DP-treated cells compared to control conditions. The fold change in the number of transformed colonies was also assessed. Both 4NC and DP significantly increased the number of transformed colonies compared to the control, with DP inducing the highest fold change (**Figure 2.16**). The results from the soft-agar assay provide strong support for the carcinogenic potential of 4NC and DP. The ability of 4NC and DP to induce robust colony formation in BEAS-2B cells demonstrates their capacity to alter normal cellular behavior and promote oncogenic transformation.



Adapted and modified with permission from Mittal A. et al., 2022, Nature Chemical Biology

Figure 2.16 4NC and DP Trigger Malignant Transformation of Human Cells

(a) Representative micrographs depicting the results of malignant transformation assay. Scale bar, 100 μ m. **(b)** Mean-whisker plot depicting the size of the malignantly transformed colonies in the indicated conditions. Data presented as mean $\pm$ SEM. Number of data points are 722 (NC), 673 (4NC), and 792 (DP) from 3 independent experiments. **(c)** Bar graphs represent the fold increase in the number of transformed colonies in the indicated conditions. Data presented as mean $\pm$ SEM ($n = 3$ independent experiments). **(b, c)** Two-sample Student's t -test was used to compute the statistical significance between test conditions and the negative control (NC).

In conclusion, Metabokiller demonstrates exceptional utility as a predictive tool for carcinogenicity, leveraging a robust framework based on diverse biochemical properties to identify and characterize metabolites with potential carcinogenic effects. The model exhibits high accuracy and recall on unseen datasets, outperforming widely used prediction methods, and its property-specific interpretability provides mechanistic insights into carcinogenesis. Experimental validations for two predicted carcinogens, 4-nitrocatechol and 3,4-dihydroxyphenylacetic acid, revealed their ability to induce genotoxicity, mutagenesis, oxidative stress, anti-apoptotic responses, and malignant transformation. These findings are supported by transcriptomic and mutational analyses, which identified disruptions in pathways regulating genomic stability, DNA repair, and cell cycle progression. The integration of molecular, cellular, and functional assays underscores the broad applicability of Metabokiller in decoding metabolite-induced carcinogenic mechanisms and advancing the understanding of cancer biology. This comprehensive approach highlights the potential of Metabokiller as a critical resource for predictive toxicology, drug discovery, and public health, setting the foundation for future studies aimed at characterizing carcinogenic risks in diverse chemical spaces.

2.4 Discussion

The integrity of a cellular genome is constantly threatened by both intrinsic and extrinsic factors, making it highly vulnerable to damage and disruption (Lutz and Fekete 1996). While significant attention has

been devoted to understanding the chemical composition and mechanisms of exogenous compounds that compromise genomic fidelity, recent evidence underscores the substantial threat posed by endogenous metabolites (Lutz and Fekete 1996; Rattray et al. 2018). Regardless of their source, the DNA alterations induced by these compounds have been implicated in various pathological conditions, including cancer and aging (Hoeijmakers 2009; Ahuja et al. 2019). Accurate evaluation of the carcinogenic potential of such compounds is, therefore, crucial in reducing the global cancer burden. Despite stringent regulatory mandates for carcinogenicity testing, numerous drugs approved by regulatory agencies, such as the Food and Drug Administration, have later been identified as carcinogens, necessitating their withdrawal from the market (Siramshetty et al. 2016). This highlights the pressing need for robust predictive frameworks to ensure safety during drug development and minimize the risks of unintentional carcinogenic exposure.

Computational approaches, particularly quantitative structure-activity relationship models, have been employed for pre-screening carcinogenic compounds (L. Zhang et al. 2017; Fjodorova et al. 2010; Morales et al. 2006; Zhou, Dai, and Gu 2003; Ruiz et al. 2008). These models offer a faster and cost-effective alternative to traditional experimental methods. However, the predictive performance of these models is often suboptimal when applied to unseen data. This limitation is primarily attributed to the scarcity of experimentally validated datasets available for model training. Furthermore, many of these models rely heavily on a limited set of features, such as structural similarity or genotoxicity, which fail to capture the multifaceted nature of carcinogenicity mechanisms.

To address these challenges, a novel approach was developed in which carcinogenicity predictions are based on the biochemical properties associated with known carcinogens. The ensemble model, Metabokiller, integrates six biochemical properties: proliferative response, oxidative stress, epigenetic alterations, genomic instability, anti-apoptotic effects, and electrophilic potential, to provide a comprehensive framework for carcinogenicity prediction. This model demonstrated superior performance compared to other widely used methods, achieving higher prediction accuracy and robustness. Additionally, Metabokiller incorporates explainable artificial intelligence features, enabling a detailed understanding of the contributions of each biochemical property to its predictions. This level of interpretability is a significant advancement in the field, addressing a major limitation of traditional "black-box" predictive models.

One of the key applications of Metabokiller was its use in a large-scale computational screening of the human metabolome, derived from the Human Metabolome Database, for carcinogenic potential. This analysis identified a significant number of metabolites with predicted carcinogenicity, including both well-characterized and previously uncharacterized compounds. Functional analysis of these predicted

carcinogens revealed selective enrichment in pathways such as tyrosine and tryptophan metabolism, nucleotide metabolism, drug metabolism, xenobiotics, and oxidative stress pathways. Among the predicted carcinogenic metabolites, 38 were previously characterized as carcinogens, reinforcing the robustness and reliability of the model in capturing carcinogenic biochemical features.

The application of Metabokiller also provided novel insights into tumor-associated metabolites, identifying a large number of previously uncharacterized oncometabolites. In addition to these discoveries, the model flagged several well-characterized oncometabolites, including 2-hydroxyglutarate, succinate, fumarate, adenylosuccinate, and polyamines such as putrescine, spermidine, and spermine (Dando et al. 2019). Functional studies of these metabolites highlight their critical roles in cancer biology. For example, fumarate accumulation, due to loss of fumarate hydratase function, has been linked to renal cell carcinoma. Similarly, succinate accumulation resulting from succinate dehydrogenase dysfunction has been implicated in tumor growth and repopulation following radio- and chemotherapy (Dando et al. 2019). 2-hydroxyglutarate has been shown to rewire cancer cell metabolism, induce oxidative stress, and modulate chromatin remodeling enzymes, further illustrating its multifaceted role in carcinogenesis (Ježek 2020). Despite the critical roles of these oncometabolites, only a limited number have been thoroughly characterized, emphasizing the need for tools like Metabokiller to expand the understanding of their functions and mechanisms of action.

It is important to note, however, that carcinogenic potential is not solely determined by a compound's intrinsic biochemical properties. Dose and exposure patterns play a crucial role, with carcinogen dose-response relationships often displaying nonlinear characteristics such as biphasic or hormetic effects, where low doses may be protective while high doses prove harmful (Fukushima et al. 2005). Moreover, inter-species metabolic differences (e.g., cytochrome P450 variability (Rodriguez-Antona and Ingelman-Sundberg 2006; Lewis, Ioannides, and Parke 1998), ADME properties) and route-of-administration effects can profoundly alter biological outcomes even at identical nominal doses. Unfortunately, the systematic incorporation of dose-dependency into predictive models remains severely constrained by the scarcity and heterogeneity of concentration-dependent carcinogenicity data, with most available datasets relying on binary classifications rather than detailed quantitative measures. When concentration data are available, they often come from different experimental contexts with varying exposure durations, dose regimens, endpoints and sensitivities, making meaningful integration extremely challenging.

While Metabokiller offers several advantages, it also has certain limitations. The current model does not account for additional carcinogenic properties such as the induction of chronic inflammation, inhibition of

senescence, cellular transformation, modulation of growth factors, alteration of signaling pathways, and immunosuppression (Smith et al. 2016). These properties, although critical in carcinogenesis, are excluded due to the limited availability of comprehensive data in existing literature. Furthermore, the model does not consider the metabolic activation of inert compounds by cellular enzymes, which can lead to carcinogenicity (Schmidt 2014). Another limitation is the lack of dose-dependency information as discussed above, since Metabokiller does not incorporate dosage as a factor in its predictions. Additionally, unlike toxicogenomic-based prediction models, Metabokiller does not account for tissue specificity, which could influence the carcinogenic potential of compounds (Gusenleitner et al. 2014). Lastly, the current version of the model focuses solely on carcinogenicity prediction and does not assess other toxicity endpoints such as oral or liver toxicity.

Despite these limitations, Metabokiller represents a significant advancement in carcinogenicity prediction by addressing the challenges posed by limited training data and providing a biochemically enriched explanation for its predictions. Its high recall value ensures accurate identification of true carcinogens, making it a reliable tool for health-related applications. The large-scale screening of the human metabolome using Metabokiller not only validates its predictive power but also opens new avenues for functional metabolomics, offering insights into the roles of cancer-associated metabolites in disease progression. The interpretability features of Metabokiller further enhance its utility, enabling researchers to uncover the biochemical basis of carcinogenicity and identify potential therapeutic targets. This integrative and transparent approach positions Metabokiller as a valuable resource in predictive toxicology and cancer research.

Chapter 3 - Identification of Endogenous Activators of Cellular GPCRs Using Deep Learning Approaches

3.1 Introduction

Metabolites play pivotal roles in both physiological and pathological processes, acting as key intermediates in metabolic pathways and as signaling molecules that influence cellular behavior. Their ability to modulate various biological responses stems from interactions with specific receptor systems, notably G protein-coupled receptors (GPCRs) (Husted et al. 2017). Understanding the diverse roles of metabolites requires a comprehensive exploration of these interactions, as they provide critical insights into the biochemical networks that govern cellular function, and can help uncover novel regulatory mechanisms and therapeutic opportunities, bridging gaps in our knowledge of cellular signaling and metabolic regulation.

GPCRs constitute a large and diverse superfamily of membrane proteins. These serve as molecular sensors to external stimuli, bridging the gap between the extracellular environment and cellular responses by orchestrating a complex network of signalling pathways that are fundamental to life processes. GPCRs detect a diverse array of extracellular signals including light, odors, hormones, neurotransmitters, peptides, lipids, and other physical and chemical stimuli including metabolites, thus mediating a vast array of physiological processes, including sensory perception, hormonal regulation, neurotransmission, and immune responses among others (M. Zhang et al. 2024; D. Yang et al. 2021). Acting as gatekeepers of cellular communication, these versatile and ubiquitously expressed receptors have garnered immense interest as drug targets. It is estimated that approximately 35% of currently marketed drugs target GPCRs, underscoring their critical importance in drug discovery and development (M. Zhang et al. 2024; D. Yang et al. 2021).

GPCRs are classified into several distinct classes based on their structural and functional characteristics, reflecting the diversity of their ligands, signalling mechanisms, and biological roles. The most widely recognized classification system divides GPCRs into six main classes: A, B, C, D, E, and F (Kolakowski 1994; Horn et al. 2003). Class A (Rhodopsin), Class B (Secretin and adhesion), Class C (Glutamate), Class D (Fungal GPCRs), Class E (cAMP receptors), and Class F (Frizzled/Smoothed GPCRs). Class A, also known as the Rhodopsin-like family, is the largest, diverse and most extensively studied class, accounting for over 80% of all GPCRs. It includes receptors for various ligands such as peptides, biogenic amines, and lipid-like substances (Fridmanis et al. 2007). Class A GPCRs can be broadly divided into two major categories: odorant and non-odorant receptors. These two subclasses, while sharing structural

similarities, have evolved distinct functional specializations that reflect their diverse roles in physiological processes. Non-odorant GPCRs encompass a diverse range of receptors involved in various physiological processes across multiple organ systems. This group includes receptors for neurotransmitters, hormones, and other signaling molecules, and can be further categorized based on the type of ligand they interact with, such as aminergic, peptide, protein, lipid, and nucleotide receptors (M. Zhang et al. 2024; D. Yang et al. 2021). Non-odorant GPCRs are widely distributed throughout the body and are involved in numerous functions. In contrast, odorant receptors, a subset of olfactory receptors, constitute the largest gene family within the human GPCR repertoire, accounting for approximately half of all GPCR genes (Aloni, Olender, and Lancet 2006; Reed 2004; Hayden and Teeling 2014). These receptors are the key drivers of odor perception as they contribute to the first critical step of odorant detection (Buck and Axel 1991). ORs are characterized by their ability to recognize multiple odorants, and conversely, individual odorants can activate multiple odorant receptors, enabling a combinatorial coding scheme for odor perception (Bushdid et al. 2014; Gerkin and Castro 2015; Meister 2015).

ORs are increasingly being recognized for their roles in diverse biological processes beyond their traditional function in olfaction. Their extra-nasal expression has been implicated in diverse biological processes such as muscle regeneration, sperm chemotaxis, appetite regulation, bronchial constriction and dilation, inflammation, and energy metabolism (reviewed in (Lee, Depoortere, and Hatt 2019)). High-throughput RNA sequencing studies have revealed that OR expression varies significantly across human tissues, with some ORs showing low expression in most tissues and others exhibiting highly selective expression in one or two specific tissues (Feldmesser et al. 2006). Studies on primary and cultured skin cells (melanocytes and keratinocytes) revealed the expression of multiple ORs, which includes OR2AT4, OR2A4, OR2A7, OR51B5, and OR51E2 (Busse et al. 2014; Tsai et al. 2017; Gelis et al. 2016). Functional dissection using their cognate ligands revealed their involvement in regulating key cellular processes such as proliferation, cytokinesis, cellular migration, and regeneration (Lee, Depoortere, and Hatt 2019; Maßberg et al. 2015; Busse et al. 2014; Tsai et al. 2017).

In the context of disease, transcriptomic studies have demonstrated the upregulation of certain ORs in cancers, such as OR51E2 in prostate cancer and OR2B6 in breast cancer (Masjedi, Zwiebel, and Giorgio 2019; Weber et al. 2018; Neuhaus et al. 2009). OR51E2, in particular, stands out as a promising biomarker and therapeutic target, with its ligand-mediated activation inducing anti-proliferative responses and cellular senescence in prostate cancer cells (Sanz et al. 2014). Similarly, other extranasal ORs, such as OR2AT4 have shown therapeutic potential, with activators like Brahmanol and Sandalore aiding wound healing and promoting hair growth (Busse et al. 2014; Chéret et al. 2018). These findings highlight the

versatility and functional importance of ORs in various physiological and pathological processes, extending beyond their traditional role in smell perception.

Recent advancements in GPCR structural biology has facilitated the use of *in silico* methodologies, like molecular docking to model receptor-ligand interactions, and identify novel binding sites. Combining docking with molecular dynamics simulations further improves accuracy, providing deeper insights into receptor-ligand interactions (Casadó and Casadó-Anguera 2023). These approaches also support virtual screening of large compound libraries, enabling the identification of new leads based on binding affinity, novelty, and structural diversity (Lyu et al. 2019; Ballante et al. 2020; Stein et al. 2020). On the other hand, LBDD relies on the knowledge of known active ligands to predict and design new compounds, employing methods such as quantitative structure-activity relationships (QSAR), pharmacophore modeling, similarity searching, machine learning and deep learning algorithms (Warszycki et al. 2021; Jang et al. 2016; Wang et al. 2021; Bilsland et al. 2015; Rifaioğlu et al. 2020; Castleman et al. 2022; Mansouri and Judson 2016). This approach is particularly useful when structural information about the target GPCR is limited.

Recent advancements in machine learning (ML) and deep learning (DL) approaches have significantly enhanced GPCR-ligand prediction, leading to the development of several notable models (Seo et al. 2018; Oh et al. 2022; Yadav et al. 2022). Notable models in the literature include pdCSM-GPCR (Velloso, Ascher, and Pires 2021), AiGPro (Brahma et al. 2025), GPCR_LigandClassify.py (Ahmed et al. 2021), DeepGPCR (H. Zhang et al. 2024), and GPCR-IPL score (Kumar, Teli, and Kim 2024). These models employ a range of sophisticated computational techniques such as graph-based signatures, supervised machine learning, convolutional neural networks, and attention mechanisms to predict GPCR-ligand interactions, classify compounds, and estimate binding affinities, thus significantly improving ligand screening and GPCR deorphanization. While similar progress has been made in the domain of odorant receptors (ORs), with models focusing on predicting odorant-receptor interactions (Ollitrault et al. 2024; Sharma et al. 2022; Bushdid et al. 2018; Jabeen et al. 2021; Cong et al. 2022; Caballero-Vidal et al. 2020; A. Gupta et al. 2021), the vast majority of GPCR-based ligand prediction efforts disproportionately focus on non-odorant GPCRs. Despite ORs being the largest subclass of GPCRs, they have remained largely underrepresented in computational GPCR research. To the best of our knowledge, there is a significant lack of studies that incorporate ORs within broader GPCR ligand-prediction models. Given the emerging therapeutic relevance of ORs beyond olfaction, there is a need for integrative modeling approaches that consider the full spectrum of GPCRs, including both ORs and non-ORs to provide a more comprehensive understanding of receptor-ligand interactions across diverse receptor families.

Here, we introduce EvOlf, a cutting-edge multimodal deep-learning framework designed to predict novel interactions between ligands and a broad range of GPCRs, including ORs. Unlike most existing models that represent the data using only a single type of descriptor, such as chemistry-based descriptors for ligands and conventional sequence-based descriptors for receptors, EvOlf integrates multiple high-dimensional descriptors representing various feature properties of ligands and receptors to generate low-dimensional embeddings that accurately represent ligand-receptor pairs. This innovative approach is trained on an extensive dataset comprising over 100,000 interaction data points across 24 mammalian species to help diversify the data and detail the underlying logic involved in GPCR-ligand interaction. EvOlf is an extension of OdoriFy, a comprehensive web server for deep neural network-based prediction of human odorant-receptor interactions. It includes four prediction engines: Odorant Predictor, Odor Finder, OR Finder, and Odorant-OR Pair Analysis, that enable the prediction of odorant status, identification of responsive ORs for given odorants, and prediction of putative odorants for user-supplied wild-type or mutant OR protein sequences. Additionally, OdoriFy incorporates explainable artificial intelligence (XAI) modules, which highlight key structural elements of predicted odorants or ORs at the atomic or amino acid levels, enhancing interpretability. Together, OdoriFy and EvOlf represent a significant advancement in understanding GPCR function, with applications ranging from olfaction and metabolism to drug discovery and receptor-targeted therapeutics.

3.2 Materials and Methods

3.2.1 OdoriFy: Methods and Implementation

3.2.1.1 Data Compilation

The training data for the four prediction engines within OdoriFy - Odor Finder, OR Finder, and Odorant-OR Pair Analysis - were manually curated from published studies available up to March 2020, identified through systematic searches on PubMed. Given that the performance of machine learning-based classification and prediction models heavily depends on the quality of the input dataset, a rigorous, multiphase data curation process was implemented to ensure the reliability and accuracy of the data. The data compilation process was carried out in three distinct phases to minimize errors and resolve discrepancies. In the first phase, all articles reporting OR interaction data were downloaded, with an emphasis on publications that detailed known agonist and non-agonist interactions with ORs. Each selected publication was thoroughly reviewed, and interaction details were manually extracted and compiled into an initial dataset. This step ensured that no relevant interactions were overlooked and that the dataset was comprehensive. In the second phase, the extracted data were subjected to cross-validation

by an independent team to ensure accuracy and consistency. This phase focused on identifying and resolving ambiguous or conflicting data points while retaining verified interaction data. Of note, entries that exhibited different response profiles (agonist versus non-agonist) at different concentrations were excluded from the dataset to maintain consistency in the binary classification framework and avoid confounding effects of concentration-dependent response variability. The third and final phase involved a manual verification of all interactions by all contributing authors to further reduce the likelihood of errors. The three-phase process ensured a high-confidence dataset, free from inconsistencies and redundancies, with a total of 679 agonist-OR interaction pairs and 5474 non-agonist-OR interaction pairs.

To ensure the reliability of the dataset used for the Odorant Predictor classification model, which evaluates odorant properties in user-supplied chemicals, similar rigorous data validation steps were conducted. The training and testing datasets for the Odorant Predictor were compiled from two independent and widely recognized repositories: PubChem and “The Good Scents Company” database. The data curation process began with the extraction of chemical compounds classified as either odorants or non-odorants from the aforementioned repositories. Initial filtering steps were conducted to remove redundant and conflicting entries. All the chemical entries were subjected to the same multiphase verification process used in the preparation of other OdoriFy datasets. This included independent cross-validation by multiple team members and a final collective review to ensure accuracy and consensus. These steps were critical for ensuring the integrity and reliability of the dataset. The final curated dataset comprised 5003 odorants and 857 non-odorants, providing a high-confidence dataset for model training and testing. Each chemical compound was represented using the Simplified Molecular Input Line Entry System (SMILES) format. To address potential inconsistencies arising from the use of various SMILES subformats, the OPSIN tool (Lowe et al. 2011) was employed for uniform conversion.

3.2.1.2 Prediction Engine: Odorant Predictor

The architecture of the Odorant Predictor module is based on Transformer-CNN (Karpov, Godin, and Tetko 2020). Transformer-CNN with the default settings was used for the model generation. The input data were formatted as per the authors' recommendations. One of the key features of Transformer-CNN is that it allows augmentation of the SMILES for training and inference, thereby enabling the construction of robust models even with smaller datasets. In the case of Odorant Predictor, the model was built using the quantitative structure-activity (property) relationship workflow, an inbuilt function in Transformer-CNN. The interpretability of the model at the substructure level was performed using the layerwise relevance propagation method. Layerwise relevance propagation method is an explanation algorithm used in tandem with neural networks to evaluate and allow the propagation of relevance from

the last layer to the input layer (Bach et al. 2015). With this approach, the most relevant features can be identified. To build and test the accuracy of the prediction model and avoid training data biases, we performed an iterative process (100 times) in which each cycle involves the random splitting of total data into training (80%) and testing (20%) sets, followed by model building and its performance evaluation. Of note, the performance of each individual model was computed using standard parameters such as accuracy, balanced accuracy, Cohen's kappa score, precision, recall, the area under the curve of receiver operating characteristic (AUC-ROC), and F1 score. Accuracy is described as the proportion of correct predictions by the model. However, accuracy alone as a measure of model performance may be misleading when exposed to an unequal number of samples in each class, a problem commonly regarded as the class imbalance problem. These limitations can be overcome by the use of balanced metrics that control for class imbalance. Precision provides information on the proportion of positive predictions that are true positives, whereas the recall parameter provides information on the proportion of the true positives that are correctly predicted. F1 score is the harmonic mean of precision and recall. The AUC-ROC provides information about how well the model can distinguish between the classes. The higher the value of AUC-ROC (toward 1), the higher is the capability of the underlying model to distinguish between the classes. Cohen's kappa score compares the accuracy of the model while controlling for the hypothetical scenario of random classification.

3.2.1.3 Prediction Engine: Odor Finder, OR Finder, and Odorant-OR Pair Analysis

The model underpinning Odor Finder, OR Finder, and Odorant-OR Pair Analysis was built using PyTorch, a Python-based deep learning toolkit. Of note, a single model was used at the back end that supports each of these prediction engines. The input datasets used for model building constitute protein sequences associated with the ORs, SMILES of their agonists/non-agonists, and the resultant activation status (1 = activation; 0 = non-activation). For building and testing the efficacy of the model, we followed a similar strategy of iteration and model performance estimation as discussed previously for the Odorant Prediction model. Notably, the input dataset was first filtered through Odorant Predictor to remove any non-odorants from the data. Importantly, the majority of the agonists-OR pairs (643 of 679) and non-agonists-OR pairs (5390 of 5474) in the dataset qualified for odorant properties (prediction probability cutoff > 0.5) and therefore were used for the downstream model building. In the case of the Odorant-OR Pair Analysis prediction engine, the input data were randomly split into training (80%) and testing (20%) datasets. The former dataset was used for model building, whereas the latter was used for performance estimation. Notably, the odorant-OR interaction data contained a higher number of non-agonists as compared to agonists. To circumvent this, we oversampled the minority class using Resample from the sklearn python package to ensure equal contribution from both classes. The OR

sequences and their respective agonists and non-agonists SMILES from both datasets were One-Hot encoded (conversion of a sequence of symbols into bit vectors) using the following strategy. One-Hot encoding of the receptor sequence was done by taking 26 capital letters and one special character (ABCDEFGHIJKLMNOPQRSTUVWXYZ), and One-Hot encoding of SMILES was done for the following 77 characters:

`()+-. /-0123456789=#@$ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]abcdefghijklmnopqrstuvwxyz^`

Two independent Bidirectional Long Short-Term Memory (BiLSTM) networks were applied on the One-Hot encoded vectors of SMILES and receptor sequence, respectively. Both these individual layers were then passed through the linear layers of the network, and their resulting vectors were concatenated. This was followed by hidden layers with rectified linear unit activation functions and Softmax as the output layer. We used dropout to prevent model overfitting. Hyperparameter tuning using grid search revealed the optimal lengths for SMILES and OR, which are 75 and 315, respectively. For both SMILES and OR sequences, 50 turned out to be the optimal number for hidden layers. All our model performance measurements were carried out on bootstrapped training and validation datasets generated by random sampling.

OR finder identifies the potential responsive ORs for the query odorant/SMILES. For this, instead of searching the entire dataset to fetch the user-supplied SMILES and return their cognate receptors, we narrowed our search space by applying this search function on a subset of dataset SMILES, which are structurally similar SMILES to the user query. This was achieved using the Tanimoto similarity metric for fingerprinting and computing the molecular similarity using RDKit (http://www.rdkit.org/RDKit_Overview.pdf) (<https://github.com/rdkit/rdkit>). After obtaining the limited set of OR sequences, we calculated their binding probabilities with the query SMILES by BiLSTM-based model (discussed previously). In addition to this approach, we also employed a brute force approach wherein we predicted the binding nature of the query SMILES with each unique receptor sequence available in our manually curated database. This step checks the potential binding of the input SMILES with all human ORs and returns the top K ORs ($10 \geq K \geq 1$), along with the information about their protein sequences and binding probabilities. It is undoubtedly slower but theoretically more accurate and more comprehensive than similarity-based prediction methodology (discussed previously) (Jiménez-Luna, Grisoni, and Schneider 2020; Sundararajan, Taly, and Yan 2017). Of note, the prediction model was trained on both the mutant and wildtype ORs; however, in the case of OR Finder, the prediction engine only checks for binding with wildtype receptors (i.e., ~400 unique human ORs). The output of the OR

finder provides a list of potential ORs along with their binding probabilities for the user-provided odorants.

In the Odor Finder prediction engine, which enables the identification of odorant molecules for wildtype or mutant human ORs, we used the aforementioned brute force approach for scanning our database. Of note, the underlying model used here is the same as the one we used for the Odorant-OR Pair Analysis. Outputs of each prediction engine are then converted into comma-separated values (CSVs) and displayed to users. We implemented an interpretability module for each prediction engine to gain insights into the “why” of a prediction. Our interpretability module returns the positively and negatively contributing atoms with shades of green and red, respectively. The interpretability module is built on the integrated gradients, which provide a visual representation of the importance of the input features contributing to the model's prediction (Captum: <https://github.com/pytorch/captum>) (Kokhlikyan et al. 2020). Such approaches have been extensively used in the drug discovery domain, sentiment analysis, natural language processing, image classifications, and so on. For substructure analysis and highlighting atoms, we used the RDkit Python library (http://www.rdkit.org/RDKit_Overview.pdf) that draws molecules and colors them individually according to their relevance in the model's predictions.

3.2.1.4 Implementation of the frontend

OdoriFy is built on the open-source cascading style sheets framework, utilizing the widely popular bootstrap with Popper.js, and jQuery, javascript library. Apart from providing an overall theme to the Web server, advanced techniques such as lazy loading of style sheets/images and rendering minified HTML have been used to ensure fast loading times. Given an anticipated load of requests on the Web server and possible security threats, instead of processing an uploaded CSV on the server side, the parsing happens on the client side via Papa Parse.js (<https://github.com/mholt/PapaParse>). As a CSV is uploaded, each row of the CSV is processed using papaparse.js, based on the prediction engine. For Odorant Predictor and OR Finder, the desired CSV consisting of a single column with SMILES in each row is required. For Odor Finder, the desired CSV consists of two columns of FASTA rows, the first column being the FASTA header and the other column being the receptor sequence. Odorant-OR Pair Analysis, on the other hand, requires an additional column of SMILES in its input CSV file. For populating the results on the front-end and achieving maximum possible modularity while creating the views, Django templating has been used to render the pages with the context being received from the Django controller (version 3.1.5). The contact form parameters are received using a get request, and then using a simple mail transfer protocol, an acknowledgment mail is sent to the user, assuring that their query will be processed as soon as possible.

3.2.2 EvOlf: Methods and Implementation

3.2.2.1 Data Compilation

The dataset for EvOlf was meticulously compiled to establish a robust foundation for predicting GPCR-ligand interactions across odorant and non-odorant GPCRs. For odorant GPCRs, the dataset for EvOlf builds upon the data curated for OdoriFy, which focused on human olfactory receptors and their interactions. This foundation was significantly expanded by manually curating interaction data for humans and additional mammalian species from published literature and specialized databases, such as the Molecule to Olfactory Receptor Database (M2OR) (Lalis et al. 2024). For non-odorant GPCRs, we sourced the interaction data from the GPCR-Ligand Association database (GLASS, v2019), specifically selecting active and inactive interactions with K_d values (Chan et al. 2015). This expansion provided a more comprehensive representation of GPCR-ligand interactions, spanning both evolutionary and functional contexts. Interactions were classified into agonist (Class 1) and non-agonist (Class 0) categories based on available experimental evidence. Similar to OdoriFy, each data point was meticulously reviewed to ensure accuracy, remove duplicates, and resolve inconsistencies. Of note, entries that exhibited different response profiles (agonist versus non-agonist) at different concentrations were excluded from the dataset to maintain consistency in the binary classification framework and avoid confounding effects of concentration-dependent response variability. The EvOlf dataset includes 105,235 GPCR-ligand interaction data points, categorized into 45,226 agonist pairs and 60,009 non-agonist pairs, representing a balanced perspective on receptor-ligand dynamics. This dataset spans 24 mammalian species, capturing evolutionary diversity and allowing the study of both conserved and species-specific GPCR functionalities. Additionally, the inclusion of wild-type and mutant receptor data facilitates the exploration of structural variations and their functional consequences.

3.2.2.2 Feature Extraction

Unlike most existing models that represent the data using only a single type of descriptor, such as chemistry-based descriptors for ligands and conventional sequence-based descriptors for receptors, EvOlf integrates multiple high-dimensional descriptors representing various feature properties of ligands and receptors to generate low-dimensional embeddings that accurately represent ligand-receptor pairs. We applied five advanced methodologies to extract meaningful, non-redundant features capturing the biochemical properties of each ligand, with each method contributing a unique layer of information for the model. Starting with the SMILES notation for each ligand, which was converted into canonical SMILES via Openbabel, we processed it through Signaturizer for bioactivity-based features (3200

features) (Bertoni et al. 2021), ChemBERTa-2 (DeepChem/ChemBERTa-77M-MLM) for contextual embeddings from transformer architecture (384 embeddings), with SMILES longer than 512 characters removed (Chithrananda, Grand, and Ramsundar 2020; Ahmad et al. 2022), Mol2Vec for molecular substructures (300 embeddings) (Jaeger, Fulle, and Turk 2018), Graph2Vec for graph representations of molecular structure (128 embeddings) (Narayanan et al. 2017), and Mordred for chemical descriptors (1492 features) (Moriwaki et al. 2018). Both 2D and 3D descriptors were calculated from Mordred. For GPCR receptors, we extracted protein descriptors from sequences through four methodologies: ProtR for conventional protein descriptors (9920 features) (Xiao et al. 2015), ProtBERT (Rostlab/prot_bert_bfd) embeddings for biological context from protein sequences (1024 features), with sequences longer than 1024 amino acids removed (Elnaggar et al. 2022), ProtT5 (Rostlab/prot_t5_xl_half_uniref50-enc) for capturing complex sequence patterns (1024 features) (Elnaggar et al. 2022), MathFeature for mathematical properties of protein sequences (8155 features) (Bonidia et al. 2022).

3.2.2.3 Data Processing and Model Building

The data processing pipeline was used to prepare ligand and GPCR feature sets for model training, validation, and testing. To begin, the dataset was split into training (80%), validation (10%), and test (10%) sets. The ligand and GPCR feature sets were pre-processed independently. First, the features were standardized using StandardScaler to normalize their distributions. Separate scalers were trained and applied for each feature type, and the scaler models were saved for consistent preprocessing during testing. Next, dimensionality reduction was applied using Principal Component Analysis (PCA), reducing each feature set to 128 principal components. PCA models were also saved for subsequent use in validation and testing. After preprocessing, the training dataset was upsampled using Synthetic Minority Oversampling Technique (SMOTE) to balance the class distribution (Chawla et al. 2011). All the ligand and GPCR features were concatenated to form a unified input vector for oversampling. After SMOTE, the augmented data was reshaped back into its original tensor-based structure, ensuring compatibility with the model. To build the base model, five distinct deep-learning architectures (described below) were implemented and evaluated for the prediction task. Each architecture was designed to capture intricate features of GPCR-ligand interactions, incorporating layers such as transformers, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks. These architectures were trained on the processed training data, and their performances were evaluated using separate testing data, with metrics such as balanced accuracy, F1-score, and AUC-ROC curves. Architecture 5 demonstrated superior performance across all evaluation metrics and was subsequently chosen for downstream analysis. The selected architecture was further validated using 10-fold cross-validation to assess its generalizability and robustness. This step ensured that the model was not overfitting and performed consistently across

various data splits. Finally, the entire dataset was utilized to retrain the model, incorporating all available data points to maximize predictive accuracy. The resulting model, termed the EvOlf Model, represents the culmination of iterative optimization and rigorous validation, providing a reliable tool for predicting GPCR-ligand interactions.

3.2.2.4 Deep Learning Architectures

To develop a robust model for predicting GPCR-ligand interactions, five distinct deep-learning architectures were designed, each leveraging unique methodologies for embedding generation and feature integration. All architectures for ligands and GPCRs use the same input pre-processed feature files, derived from multiple descriptor generation techniques. These features are fed into independent pipelines for ligands and GPCRs.

While each architecture employs unique strategies for embedding generation and feature integration, they share a common modular structure. Each architecture comprises two primary components: embedding generators and fusion modules, as well as an additional classification module, which remains the same in all architectures. The embedding generator module features two separate but identical pipelines: one for ligands and one for GPCRs. These pipelines are designed to independently process the feature files of ligands and receptors, transforming different features into high-dimensional embeddings that encode essential biochemical and structural properties. These embeddings are generated through various combinations of transformers, convolutional layers, and recurrent networks, depending on the architecture. These methodologies enable the pipelines to capture both local and global features, providing a comprehensive representation of ligands and GPCRs in a shared latent space.

Following the embedding generation, the outputs from the ligand and GPCR pipelines are integrated in the fusion module. This module combines the embeddings to form a joint representation of the ligand-receptor pair. While the specific fusion strategies vary across architectures, ranging from simple concatenation to sophisticated transformer-based operations, the goal remains the same: to capture the complex dependencies and interactions between ligands and GPCRs. This joint representation is subsequently passed to the feed-forward network, where the model predicts whether the ligand-GPCR pair is an agonist (Class 1) or non-agonist (Class 0). All models were trained for 100 epochs with a batch size of 64, using the Adam optimizer and CrossEntropyLoss as the loss function. A learning rate scheduler (ReduceLROnPlateau) dynamically adjusted the learning rate when validation loss plateaued for two consecutive epochs. The following subsections detail the specific designs and methodologies of each architecture.

3.2.2.4.1 Architecture 1

Architecture 1 begins by concatenating multiple input feature sets for ligands into a single representation. The first stage of processing involves passing the ligand input through a 2D convolutional layer (in_channels=1, out_channels=10) with kernel sizes of 5 and 4 for ligands and GPCRs, respectively. This convolutional layer extracts localized spatial patterns within the ligand features. The output is then processed by an LSTM layer (input_size=124, hidden_size=128, num_layers=2, batch_first=True, dropout=0.2). The final hidden state from the LSTM serves as the compact embedding representation for the ligand. A parallel pipeline is used for receptor features. The embeddings generated by the ligand and receptor pipelines are concatenated to form a fused embedding vector which is then passed into a fully connected feed-forward network with a dropout rate of 0.2. The final layer of a softmax activation function is applied, yielding probabilities for the two interaction classes: agonist (Class 1) and non-agonist (Class 0). The model was trained with a learning rate of 5×10^{-6} .

3.2.2.4.2 Architecture 2

The architecture begins by concatenating the input features derived from multiple sources, which are then passed through a transformer module (d_model=128, nhead=4, num_encoder_layers=2, num_decoder_layers=2, batch_first=True) to capture contextual relationships between the features. The transformer output is normalized using layer normalization and processed through a one-dimensional convolutional layer (out_channels = 1, kernel_size=2, groups=1) with 5 and 4 in channels for ligands and GPCRs respectively, to refine spatial patterns and extract key feature representations. The processed outputs from the ligand and receptor pipelines are then combined through concatenation around dim=1. The combined embeddings undergo a mean-reduction operation to create a fused embedding that captures the interaction dynamics between ligands and receptors. To prevent overfitting during training, a dropout layer with a dropout probability of 0.4 is applied to the fused embeddings before classification. This fused embedding is passed through a feed-forward network comprising three linear layers with ReLU activations. A final layer of softmax function is used to generate probabilities for binary classification into agonist or non-agonist interactions. The model was trained with a learning rate of 1×10^{-5} .

3.2.2.4.3 Architecture 3

This architecture consists of two identical fusion pipelines for ligands and receptors, each leveraging a Transformer-LSTM combination. Each pipeline begins by concatenating multiple feature inputs for ligands and receptors, respectively. These concatenated feature sets are passed through a Transformer layer (d_model=128, nhead=4, num_encoder_layers=2, num_decoder_layers=2, batch_first=True). This

module captures complex relationships within the input features through self-attention mechanisms. The Transformer output is subsequently fed into an LSTM (input_size=128, hidden_size=128, batch_first=True). The final hidden state of the LSTM is passed through a fully connected layer, generating compact embeddings for ligands and receptors, respectively. The ligand and receptor embeddings generated by the separate pipelines are fused into a unified representation. This fused vector is expanded into a three-channel tensor by stacking the individual ligand and receptor embeddings alongside the fused representation. The resulting 2D tensor is then processed through a series of one-dimensional convolutional layers to extract hierarchical spatial features. The convolutional layers include two stages: the first with 32 filters and the second with 64 filters, both using a kernel size of 3 and a stride of 1, followed by max-pooling layers with a kernel size of 2. The convolutional output is flattened and passed to a fully connected feed-forward network with ReLU activation. Dropout rate of 0.2 is employed in the classifier to prevent overfitting. Xavier initialization is applied to all linear and convolutional layers, and orthogonal initialization is used for the LSTM weights to ensure efficient training. The softmax function is applied to predict the interaction classes: agonist (Class 1) or non-agonist (Class 0). The model was trained with a learning rate of 5×10^{-6} and weight decay of 1×10^{-4} .

3.2.2.4.4 Architecture 4

This architecture includes two identical pipelines for ligands and receptors, each consisting of a Transformer module, a convolutional layer, and an LSTM. The feature vectors for ligands and receptors are first concatenated from multiple input features, five for ligands and four for receptors. These concatenated feature sets are passed through separate Transformer layers (d_model=128, nhead=4, num_encoder_layers=2, num_decoder_layers=2, batch_first=True). The Transformer blocks apply self-attention to model intra-feature relationships, enabling the extraction of meaningful patterns. The Transformer outputs are normalized using layer normalization before being passed to a 1D convolutional layer (out_channels=100, kernel_size=2) with 5 and 4 input channels for ligands and GPCRs, respectively. The convolutional operation enhances spatial feature extraction by focusing on local dependencies within the feature vectors. Following the convolutional layer, the features are further processed by an LSTM module (input_size = 127, hidden_size = 128, num_layers = 2, batch_first = True, dropout = 0.2). The final hidden state of the LSTM, representing the most informative temporal features, serves as the embedding for ligands and receptors, respectively. The ligand and receptor embeddings are concatenated along the first dimension to form a fused representation. This fused vector is mean-reduced to ensure that both embeddings contribute equally to the final prediction. To improve generalization, a dropout layer with a rate of 0.2 is applied to the resultant fused vector before it is passed to the feed-forward network, followed by a softmax activation function to generate class probabilities for

agonist (Class 1) and non-agonist (Class 0) interactions. All linear layers in the classifier are initialized using Xavier initialization, with biases set to zero to ensure stable weight updates during training. Similarly, LSTM weights are initialized using Xavier initialization for weights and zero initialization for biases. The model was trained with a learning rate of 1×10^{-6} .

3.2.2.4.5 Architecture 5

In Architecture 5, the pipeline begins by concatenating the input features for ligands and GPCRs, which are independently processed through a transformer module with self-attention ($d_model=128$, $nhead=4$, $num_encoder_layers=2$, $num_decoder_layers=2$, $batch_first = True$) to capture contextual relationships. This output is normalized and passed through one-dimensional convolutional layers ($out_channels = 10$, $kernel_size = 2$) to refine spatial patterns with 5 and 4 in channels for ligands and GPCRs, respectively, followed by LSTM layers ($input_size = 127$, $hidden_size = 128$, $num_layers = 2$, $batch_first = True$, $dropout = 0.1$) that extract sequential dependencies. The final hidden states of the LSTMs serve as the ligand and GPCR embeddings, respectively. These embeddings are then fused using a dedicated fusion transformer ($d_model=128$, $nhead=4$, $num_encoder_layers=2$, $num_decoder_layers=2$, $batch_first=True$), which utilizes the concept of inter-attention and models the complex interdependencies between ligands and GPCRs. The output from the fusion transformer is normalized and passed through a fully connected feed-forward network consisting of dropout (0.2) and ReLU activations, culminating in a binary classification output. A softmax layer is applied at the end to convert the logits into probabilities, allowing for the final classification and confidence scoring of interaction predictions. The architecture's weights are initialized using Xavier initialization for linear and convolutional layers, with biases set to zero to ensure stability during training. The model was trained with a learning rate of 1×10^{-5} .

3.2.2.5 Data and Code Availability

The EvOlf dataset used to train the models in this study is provided in the supplementary tables and is accessible as a user-friendly Shiny database for interactive exploration of GPCR-ligand interactions at <https://the-ahuja-lab.shinyapps.io/EvOlf/>. The code supporting the findings of this study is available in the project's GitHub repository at <https://github.com/the-ahuja-lab/EvOlf>. The repository includes the full implementation of all deep learning architectures, along with data preprocessing scripts and model training pipelines. Additionally, we provide a Google Colab notebook to enable users to generate predictions using EvOlf without requiring a local setup.

3.3 Results

3.3.1 Development of an AI model to Predict Human-Specific Odorant-OR Pairs

Olfactory receptors (ORs) are a diverse and highly specialized subset of G protein-coupled receptors (GPCRs) that play a critical role in detecting and discriminating a vast array of odorants. Despite their importance, studying ORs and their interactions with odorants has been challenging due to the sheer diversity of potential ligand-receptor pairs and the lack of comprehensive computational tools to facilitate this exploration. Addressing these challenges, OdoriFy was developed as an advanced open-source platform that integrates deep neural network-based prediction models with explainable artificial intelligence functionalities.

3.3.1.1 OdoriFy: An AI-Driven Platform for Human Olfactory Research

OdoriFy is a sophisticated open-source web server that integrates deep neural network-based prediction models with explainable artificial intelligence (AI) functionalities. It has been specifically developed to serve as a comprehensive platform for researchers studying the chemical basis of olfaction. Its primary objective is to streamline and centralize the analysis of chemical interactions involved in olfactory processes, providing a unified solution for various computational tasks related to odorant research. OdoriFy includes several functional modules, each tailored to address distinct aspects of olfactory research. The Odorant Predictor module facilitates the classification of compounds as odorants or non-odorants based on their chemical structure. The Odor Finder module is designed to identify agonists for human odorant receptors (ORs), enabling researchers to predict the interactions of specific compounds with target receptors. Additionally, the OR Finder module identifies ORs responsive to user-specified compounds. The Odorant-OR Pair Analysis module enables detailed studies of specific odorant-receptor pairs, further assisting in characterizing receptor-chemical interactions. The graphical user interface is designed for ease of use, enabling researchers to efficiently identify potential agonists for specific ORs or validate the odorant properties of selected compounds (**Figure 3.1**).

The Odorant Predictor module within OdoriFy facilitates the classification of chemical compounds as odorants or non-odorants. Users can input query chemicals in the SMILES (Simplified Molecular Input Line Entry System) format, which provides a standardized textual representation of chemical structures. This module supports two methods of data input: uploading a file containing comma-separated variables or directly pasting SMILES strings into a designated input window on the web interface. To further enhance user convenience, OdoriFy incorporates a direct link to the OPSIN (Open Parser for Systematic IUPAC Nomenclature) web server. This integration allows users to easily convert compound structures

into the required SMILES format. This feature streamlines the workflow for researchers who may be less familiar with the SMILES notation, ensuring broad accessibility to the tool. The output of the Odorant Predictor includes hyperlinked tabular data that classifies chemicals as odorants or non-odorants, provides prediction probabilities, and highlights decision-determining atoms using interpretable AI functionalities (**Figure 3.1**).

The OR Finder prediction engine in OdoriFy is designed to identify odorant receptors (ORs) that may bring to the user-specified chemical compounds. Similar to the Odorant Predictor module, the OR Finder accepts chemical inputs in the SMILES format. The OR Finder module offers two distinct modes of analysis: rapid testing and normal testing. Rapid testing prioritizes computational speed, making it suitable for preliminary assessments or large-scale screening studies. However, this method sacrifices some accuracy for efficiency. In contrast, the normal testing mode provides higher predictive accuracy, making it the preferred choice for detailed and reliable analyses. The output of the OR Finder module shares similarities with that of the Odorant Predictor. Results are presented in a hyperlinked tabular format, including the prediction probability of the predicted cognate ORs, along with their FASTA sequences. Furthermore, the module identifies the specific amino acids within the OR sequences that are critical for the receptor's interaction with the chemical ligands using explainable artificial intelligence techniques (**Figure 3.1**).

The Odor Finder module in OdoriFy is designed to predict odorant molecules that interact with user-provided odorant receptors (ORs), including both wild type and mutated variants. Users can submit OR sequences in FASTA format, with flexibility in data submission, allowing sequences to be input either directly into a designated text field or uploaded as a file. Odor Finder also allows users to customize the number of top predicted interacting odorants included in the output. By default, the module identifies and reports the top five odorants, but this threshold can be adjusted based on the specific needs of the study. The prediction results generated by Odor Finder are presented in a format consistent with other prediction engines within OdoriFy, ensuring a standardized user experience (**Figure 3.1**).

The Odorant-OR Pair Analysis module within OdoriFy enables a detailed investigation of specific interactions between a chemical ligand and an odorant receptor (OR). Users can input the FASTA sequence of the receptor of interest, along with the corresponding chemical query in SMILES format. The resulting predictions include interaction probabilities along with additional contextual information as provided by other modules. To facilitate data management and ease of access, OdoriFy allows users to download the output from all prediction engines as a single compressed zip file. Additionally, all modules offer the option to receive results via email. This functionality enhances user convenience, particularly for

computationally intensive analyses that may require extended processing times (**Figure 3.1**). This functionality positions OdoriFy as a robust and versatile tool for advancing research in olfactory science .

[illegible]

Adapted and modified with permission from Gupta R. & Mittal A.* et al., 2022, The Journal of Biol. Chemistry*

Figure 3.1 Graphical User Interface of OdoriFy webserver.

A snippet of the graphical user interface of prediction engines supported by OdoriFy. It also contains information about the required format of the input data for each prediction engine, along with descriptions of the possible outputs from the server.

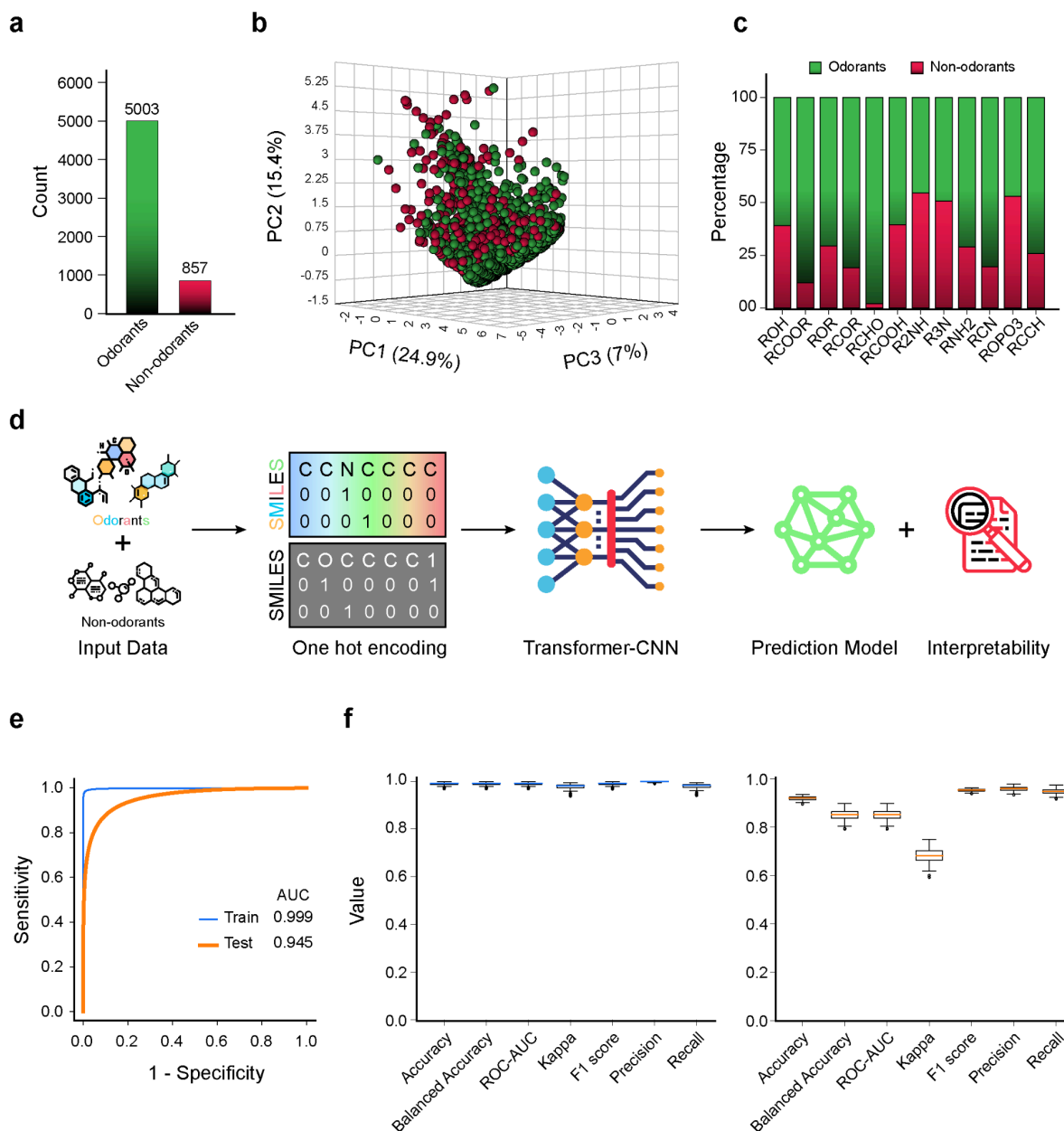
3.3.1.2 Odorant Predictor: A Dual Function Tool for Odorant Identification and Explanation

To develop the Odorant Predictor module, a comprehensive dataset of odorant and non-odorant molecules was meticulously compiled, refined, and validated from the literature and publicly available databases. The dataset underwent iterative refinement and verification to ensure its integrity and suitability for use in the predictive model. The final dataset of 5860 compounds included 5003 molecules classified as odorants and 857 as non-odorants (**Figure 3.2a**). This process was essential for developing a robust and reliable computational framework for odorant classification.

To evaluate the chemical diversity within the dataset, the molecules were transformed into atom pair fingerprints using the ChemmineR package (Cao et al. 2008). Atom pair fingerprints represent the properties of atoms and molecules as numerical vectors, capturing structural and chemical features that are essential for computational analysis. These fingerprints served as input features for principal component analysis (PCA), a statistical method used to reduce dimensionality and visualize data. PCA was employed to gain insights into the structural variation between the two classes of molecules. The 3d PCA revealed substantial chemical heterogeneity between odorants and non-odorants, highlighting the diverse molecular architectures associated with each category (**Figure 3.2b**). Further analysis focused on the functional group composition of the dataset. This investigation revealed a higher enrichment of certain functional groups, such as ethers, esters, and aldehydes, among the odorant molecules. These findings align with previously reported trends in the literature (Brower and Schafer 1975; B. A. Johnson et al. 2005), suggesting that these specific functional groups contribute to the physicochemical properties that define odorant activity (**Figure 3.2c**).

For the development of the Odorant Predictor model, we used the Transformer-CNN framework (Karpov, Godin, and Tetko 2020), a state-of-the-art computational approach for quantitative structure-activity and structure-property relationship modeling. The input for the model consisted of molecular structures provided in the form of SMILES strings. These SMILES were transformed into numerical embeddings using a one-hot encoding method, which converts molecular descriptors into a structured and machine-readable format. The processed embeddings were subsequently used to train the predictive model (**Figure 3.2d**).

The Odorant Predictor model demonstrated exceptional predictive performance, as evidenced by an area under the receiver operating characteristic curve (AUC-ROC) value of 0.945 when tested on an independent dataset. This high AUC-ROC value indicates the model's strong ability to differentiate between odorant and non-odorant molecules, highlighting its robustness and reliability. The model's performance was further corroborated by additional evaluation metrics, including balanced accuracy, Cohen's kappa, F1 score, precision, and recall (**Figure 3.2e, f**). These metrics collectively confirmed the model's accuracy, precision, and overall reliability in predicting odorant classification across diverse molecular structures.

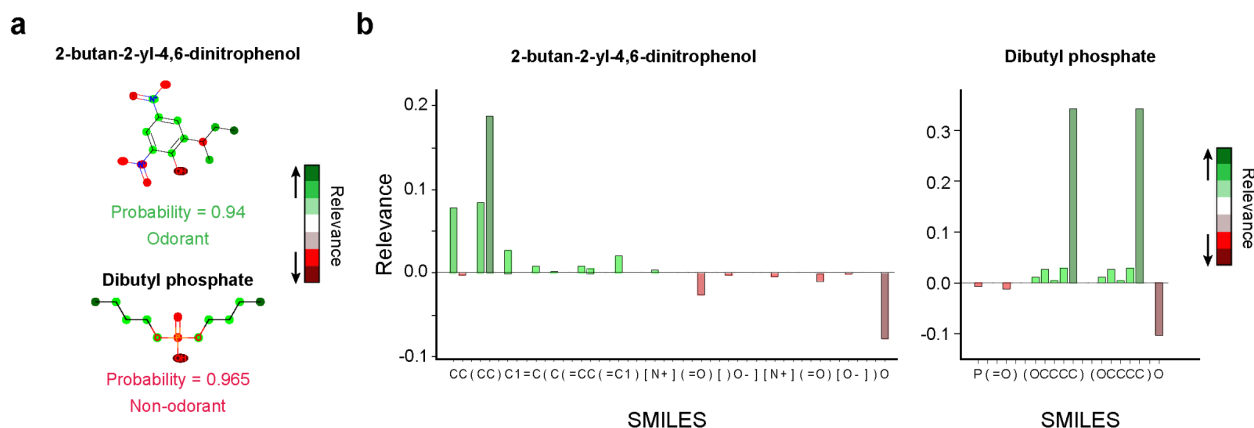


Adapted and modified with permission from Gupta R*. & Mittal A.* et al., 2022, The Journal of Biol. Chemistry

Figure 3.2 Odorant Predictor allows rapid classification of user-supplemented chemicals into odorant or non-odorants.

(a) Bar plot depicting the number of known odorants and non-odorants in the input dataset used to build the prediction model. (b) Principal component analysis (PCA) on the atom pair fingerprints of known odorants and non-odorants computed using the ChemmineR Bioconductor package. The plot depicts the chemical heterogeneity of odorants and non-odorants along with the three principal components (PC1, PC2, and PC3). (c) Bar graph representing the relative abundance of 12 prominent functional groups (de)enriched in known odorants and non-odorants, collectively describing the relative composition of the functional groups in the input dataset. Of note, the functional group estimation was performed using the ChemmineR package. (d) A schematic representation of the deep learning architecture used to build the model for the Odorant predictor engine. (e) AUC plot representing the performance of the best model in classifying the odorants and non-odorants on the training (blue) and testing (orange) datasets. (f) box plot representing the distribution of 100 random iterations of the key metrics collectively describing the model performance on the training (left) and testing (right) datasets.

The built-in interpretability module of the Transformer-CNN framework further enhanced the utility of the Odorant Predictor by providing insights into the decision-making criteria of the predictive model. This module offers atomic-level explanations of the classification process, identifying specific atomic features and molecular substructures that contributed to the model's predictions and are critical for distinguishing odorants from non-odorants. By elucidating the structural determinants of odorant functionality, the interpretability module adds a layer of transparency and scientific rigor to the computational analysis (Figure 3.3). Thus, the Odorant Predictor module is a powerful tool that not only facilitates the accurate classification of odorant molecules but also provides valuable insights into the chemical and structural properties that define odorant behavior, making it a significant asset for hypothesis-driven research and experimental design.



Adapted and modified with permission from Gupta R*. & Mittal A.* et al., 2022, The Journal of Biol. Chemistry

Figure 3.3 Odorant Predictor allows explainability at the sub-structural levels.

(a) Colored chemical structures representing the relative contribution of the highlighted atoms in the classification decision. The green and red colors represent the highest and lowest contribution thresholds, respectively. The respective odorant status along with prediction probabilities are mentioned below each molecule **(b)** Bar graphs indicating the relevance of each element of the input SMILES in model classification decisions in the indicated examples.

3.3.1.3 Predicting Odorant-Odorant Receptor Interactions Using OdoriFy

To evaluate the potential of deep neural networks for predicting interactions between odorant receptors (ORs) and their ligands, a comprehensive dataset of known agonists and non-agonists for various tested ORs was compiled from publicly available sources (**Figure 3.4a**). This dataset underwent rigorous curation to ensure its quality and reliability. Redundant and conflicting data points were carefully identified and removed, with further refinement involving filtering the dataset through the Odorant Predictor engine where only those molecules were selected that were classified as odorants with a probability score exceeding 0.5. This step ensured that the dataset was focused on molecules with a high likelihood of interacting with ORs. This refinement process helped minimize noise and enhance the dataset's overall accuracy. The final dataset included 5474 agonists-OR pairs and 679 non-agonists-OR pairs (**Figure 3.4b**). A functional group analysis of the refined dataset revealed significant heterogeneity in molecular structures, with similar levels of enrichment observed for individual functional groups among both agonists and non-agonists. Importantly, none of the confirmed agonists or non-agonists contained specific functional groups such as RCCH (alkynes) or ROPO3 (phosphates), suggesting that these may not play a significant role in modulating OR-ligand interactions (**Figure 3.4c**).

Using this refined dataset, a predictive model was developed employing a bidirectional long short-term memory (BiLSTM) architecture, combined with linear transformation layers to capture the sequential and structural features of the input data. To optimize model performance, hyperparameter tuning was conducted. This process determined the optimal sequence lengths for the SMILES and OR inputs to be 75 and 315 characters, respectively. Additionally, the optimal number of hidden layers was identified as 50 for both SMILES and OR sequences. To mitigate the inherent class imbalance in the dataset, an upsampling approach was implemented, increasing the representation of agonist-OR data points in the training dataset. The model training process employed an iterative strategy consisting of 100 steps, where the input data were randomly split into training (80%) and testing (20%) subsets. The model's performance was evaluated at each iteration using the unseen testing data. One-hot encoding was performed on both the OR sequences and the chemical inputs to prepare the data for neural network

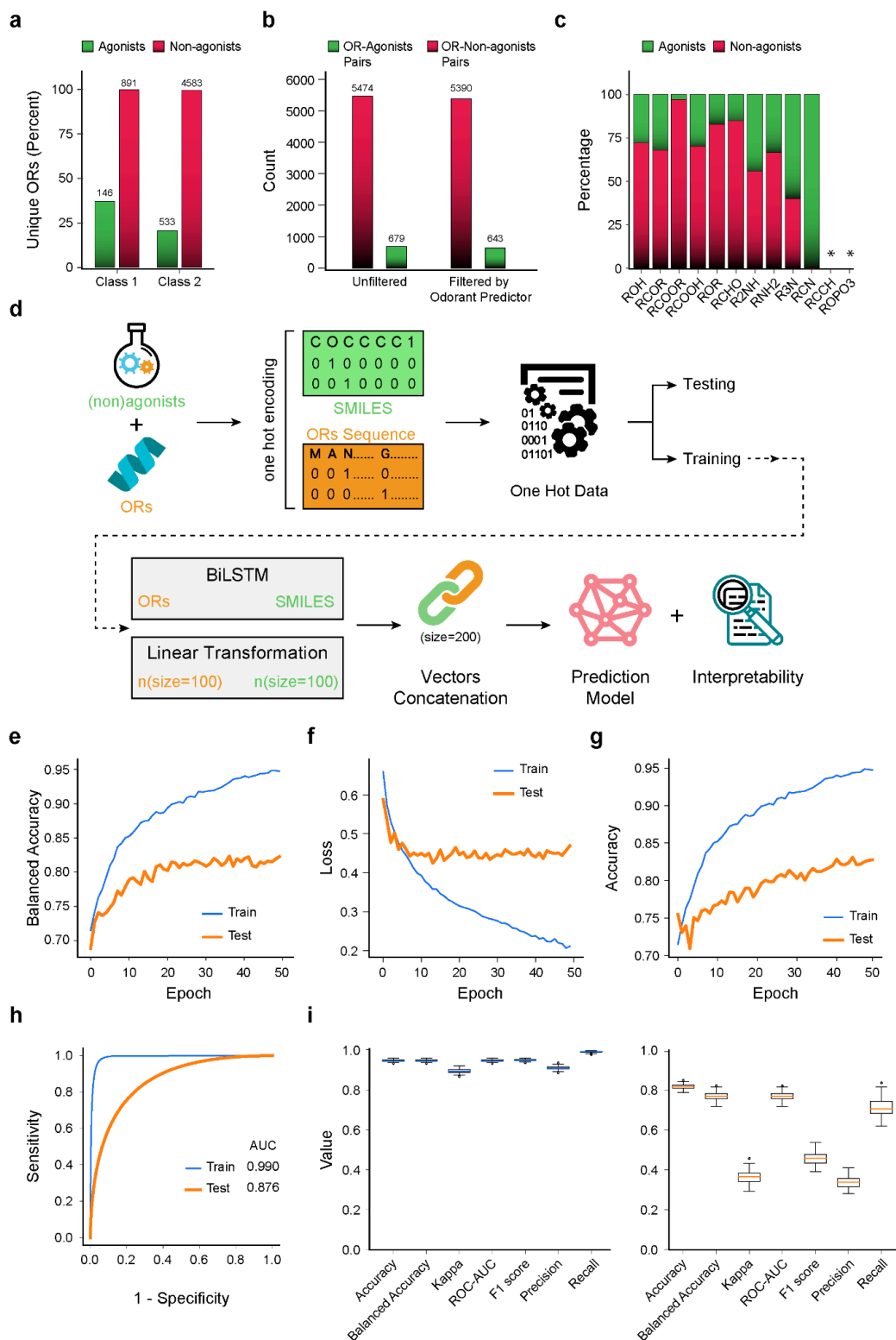
processing. The BiLSTM architecture was applied with a dropout rate of 0.5 to mitigate overfitting, and a linear transformation of size 100 was employed for feature dimensionality reduction. The resultant fixed-size vectors were concatenated and used for model training (**Figure 3.4d**). The resulting predictive model exhibited strong performance, achieving an AUC-ROC value of 0.876 on the testing dataset, demonstrating the model's strong ability to distinguish between agonist-OR and non-agonist-OR classes. Additional performance metrics, including precision, recall, F1 score, and balanced accuracy, were also evaluated and fell within acceptable ranges, further validating the reliability and robustness of the model (**Figure 3.4e-i**).

Similar to the Odorant Predictor, the Captum library was used to provide insights at the amino acid level for ORs, identifying residues that are pivotal for ligand interaction. This interpretability ensures that predictions are not only accurate but also scientifically meaningful, offering a deeper understanding of the molecular determinants underlying receptor-ligand interactions.

This optimized predictive model served as the foundation for the OR Finder, Odor Finder, and Odorant-OR Pair Analysis modules. These modules leverage the predictive capabilities of the base model to address various aspects of OR-odorant interactions, from identifying responsive receptors to characterizing specific receptor-ligand pairs. By combining one of the most comprehensive and heterogeneous datasets of human OR-odorant interactions with state-of-the-art deep learning methodologies, a high-performance prediction model was developed, providing a powerful tool for predicting and testing OR-odorant and OR-non-odorant interactions with high accuracy and reliability.

3.3.1.4 Comprehensive Investigation of Odorant-OR Interactions Using OdoriFy

A key distinguishing feature of the OdoriFy web server is its capability to accept multiple inputs for the comprehensive prediction and validation of odorant-receptor (OR) interactions. The OR Finder module accepts chemical inputs in the form of SMILES representations of odorant molecules. Upon processing these inputs, it returns a list of predicted ORs along with their interaction probabilities. These probabilities provide a quantitative measure of the model's confidence in the predicted interactions between the submitted odorant and the identified ORs. Additionally, the module incorporates advanced explainability features that highlight specific atoms within the input odorant molecule, indicating their relevance in the decision-making process of the model. Similarly, the predicted ORs are analyzed at the amino acid level, with each residue assigned a score reflecting its relative importance in the interaction prediction. This dual-level interpretability offers valuable insights into both the ligand and receptor components, enabling a deeper understanding of the molecular determinants that influence OR-odorant interactions.



Adapted and modified with permission from Gupta R*. & Mittal A.* et al., 2022, The Journal of Biol. Chemistry

Figure 3.4 A unified deep neural network-based prediction model supports Odor Finder, OR Finder, and Odorant-OR Pair Analysis

(a) Bar plot depicting the proportion of human class I and II ORs with the information on known agonists (green) and/or non-agonists (red). The absolute count of known agonist and non-agonist-OR pairs available for each class is mentioned above each bar in numerics. **(b)** Bar graph depicting the shift in the number of OR-agonists and OR-non-agonist pairs in the initially compiled dataset, before and after filtering via Odorant Predictor. **(c)** Bar graph representing the relative abundance of the 12 prominent functional groups enriched in the bonafide agonists and non-agonists, collectively describing the functional group composition of the input dataset. The functional groups were computed using ChemmineR, a Bioconductor package. The asterisk indicates the absence of the indicated functional group. **(d)** A schematic representation of the unified deep learning architecture used to build the model to provide backend support to three prediction engines, namely Odorant Predictor, Odor Finder, OR Finder, and Odorant-OR Pair Analysis. **(e-g)** The line plots depicting the increase in **(e)** balanced accuracy, **(f)** loss, and **(g)** accuracy over epochs for training and testing datasets where blue and orange colors represent training and testing data, respectively. **(h)** AUC plot representing the performance of the model in classifying the agonists and non-agonists on the training and test dataset. **(i)** Box plots representing the distribution of 100 random iterations of the key metrics collectively describing the potency of the model for training (left) and testing (right) datasets.

The Odor Finder module provides analogous functionality but is designed to accept OR sequences as inputs, submitted in FASTA format. Instead of predicting responsive receptors, this module identifies potential interacting odorant molecules and returns their corresponding interaction probabilities. As with the OR Finder, the Odor Finder module includes explainability features at the atomic level for odorants and the amino acid level for OR sequences. This parallel approach ensures consistency across the predictive engines while addressing distinct research needs, whether focusing on receptors or ligands. OdoriFy includes the Odorant-OR Pair Analysis module, which allows users to evaluate specific OR-odorant pairs for potential interaction. Collectively, the suite of prediction engines within OdoriFy represents a powerful conglomerate of artificial intelligence-driven tools for the investigation of OR-odorant interactions. Its multifaceted approach, encompassing input flexibility, robust predictive accuracy, and detailed interpretability, provides researchers with a comprehensive platform for studying the complex molecular dynamics of olfactory systems. This versatility makes OdoriFy an essential resource for advancing the field of olfactory research and understanding receptor-ligand interactions.

3.3.1.5 OdoriFy Synergizes with the Structure-Based OR-Ligand Interaction Method and Outperforms Other Prediction Tools

An alternative approach to prediction-based computational methods for studying odorant receptor (OR) and odorant interactions is protein structure-based ligand interaction analysis, typically performed through

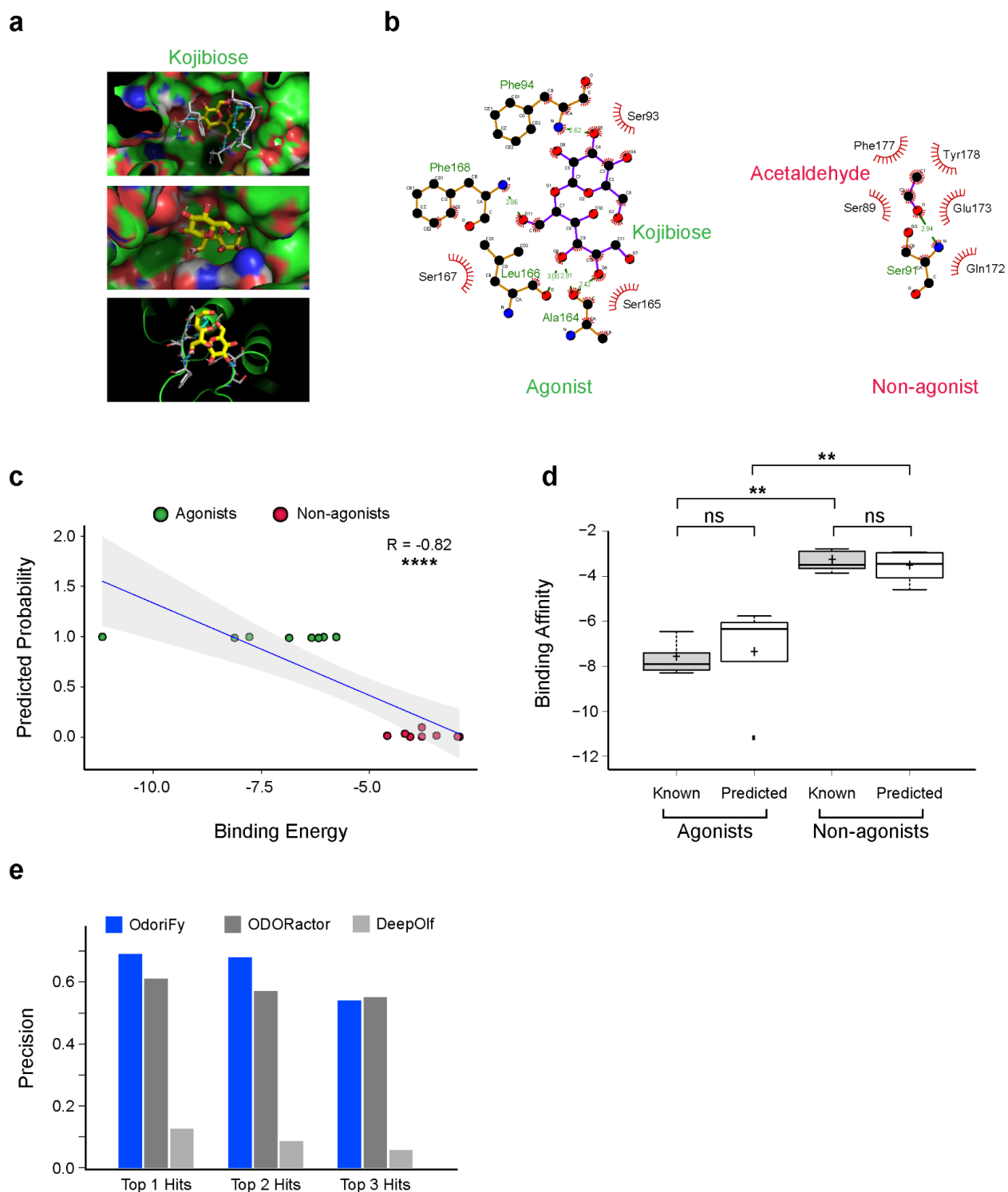
molecular docking. However, a major limitation of this technique to ORs is the absence of experimentally resolved structures of mammalian ORs in the Protein Data Bank (PDB) (Berman et al. 2000). To address this gap, computational modeling using fold recognition methods, followed by molecular dynamics (MD) simulations, has become a widely adopted strategy for predicting stable OR structures. Once a stable receptor model is obtained, dynamic docking can be employed to analyze interactions between the computationally derived OR structures and potential odorant molecules.

To evaluate the validity of OdoriFy's predictions, a comparative analysis was performed using OR1A1, a broadly tuned human OR, as a representative receptor. A stable structure of OR1A1 was recently modeled in our lab (A. Gupta et al. 2021), using the fold recognition method with GPCR-I-TASSER, a specialized tool for modeling G-protein-coupled receptors (J. Zhang et al. 2015). This initial model was further refined using molecular dynamics simulations performed with the GROMACS software package to ensure structural stability and biological relevance (Abraham et al. 2015).

To cross-verify OdoriFy's predictions, eight agonists and eight non-agonists were selected based on their high and low interaction probabilities predicted by OdoriFy, respectively. These compounds were subjected to dynamic docking within the predicted active site of the OR1A1 model. To provide a baseline for comparison, five known agonists and five known non-agonists of OR1A1 were included as positive and negative controls. The docking analysis revealed a significant inverse correlation between the binding energies obtained through molecular docking and the interaction probabilities predicted by OdoriFy for both agonists and non-agonists. This strong correlation suggests a high degree of synergy between the prediction-based approach of OdoriFy and the structure-based docking method, thereby reinforcing the reliability of the computational predictions. Interestingly, no significant differences in binding energies were observed between the known and predicted agonists or non-agonists of OR1A1, suggesting the robustness of the structure-based method (**Figure 3.5a-d**). These results further validate the reliability of OdoriFy's predictive framework in identifying meaningful OR-odorant interactions.

To assess how OdoriFy compares with existing tools, an independent validation dataset from the curated interactions was tested across OdoriFy, ODORactor (Xinyi Liu et al. 2011), and DeepOlf (Sharma et al. 2022). These models were selected due to their established use in odorant-receptor interaction predictions. The comparative analysis primarily focused on model precision due to functional limitations inherent in ODORactor and DeepOlf, which restricted a broader evaluation of other performance metrics. Under stringent evaluation conditions (Top-K hits with $K \leq 2$), OdoriFy outperformed both alternative methods, demonstrating superior predictive accuracy. When the evaluation criteria were relaxed to include the top three predictions, OdoriFy exhibited comparable performance to ODORactor while still outperforming

DeepOlf (**Figure 3.5b**). These results suggest that OdoriFy maintains a high level of predictive accuracy even under less stringent conditions.



Adapted and modified with permission from Gupta R*. & Mittal A.* et al., 2022, The Journal of Biol. Chemistry

Figure 3.5 Cross-validation of the top predicted agonists/non-agonists of OR1A1 using an orthogonal structure-based approach.

(a) The micrograph depicts the interaction of OR1A1 protein with its agonist Kojibiose inside the binding pocket of a three-dimensional OR1A1 protein structure. **(b)** Lig plot highlighting the key amino acids within the binding pocket that are predicted to be interacting with a known agonist, Kojibiose (left) and nonagonist acetaldehyde (right). **(c)** Scatter plot describing the relationship between predicted OR1A1 agonists (indicated in green) and non-agonists (indicated in red) with the binding energies obtained using AutoDock. **(d)** Box plot representing the distribution of the binding energies for the known and predicted agonists and non-agonists. Mann-Whitney U test was used for statistical analysis, where ns represents nonsignificant, whereas ** represents $p \leq 0.01$. **(e)** bar plot representing the comparative performance of OdoriFy, ODORactor, and DeepOlf in precisely predicting the top n hits for the input SMILES/ligands.

The benchmarking study underscores the competitive advantage of OdoriFy over existing tools, particularly in scenarios that demand high precision and confidence in top-ranked predictions. This superior performance can be attributed to OdoriFy's advanced deep-learning architecture, robust training on a comprehensive dataset, and incorporation of interpretability features that enhance its reliability and usability in olfactory research. By outperforming alternative tools in stringent validation scenarios, OdoriFy establishes itself as a state-of-the-art resource for the prediction and analysis of odorant-receptor interactions.

3.3.2 Generation of Unified, Evolutionary-Driven GPCR-Ligand Pair Prediction Models Using Deep Learning Approach

Metabolites play critical roles in regulating a wide range of physiological processes, serving not only as intermediates in metabolic pathways but also as signaling molecules that interact with GPCRs to modulate cellular responses. Understanding these interactions is essential for unraveling the complex biochemical networks that govern both health and disease (Husted et al. 2017). However, the study of metabolite-GPCR interactions has been hindered by an incomplete understanding of the diverse receptor subtypes involved in metabolite signaling. ORs, the largest subfamily of GPCRs, are often overlooked in broader GPCR analyses due to their specialized role in olfaction and the challenges of studying them outside sensory contexts. Yet, their extranasal expression has been linked to diverse functions, including cell migration, immune regulation, and cancer progression (Sanz et al. 2014; Orecchioni, Matsunami, and Ley 2022; Weber et al. 2018). Ignoring ORs in metabolite-GPCR interaction studies risks overlooking critical aspects of metabolite signaling and its systemic effects. To gain a more comprehensive understanding of metabolite roles, it is imperative to expand the scope of research to include both ORs and non-ORs. Integrating ORs into the broader GPCR framework allows for the discovery of conserved mechanisms, novel ligand-receptor interactions that are critical for understanding metabolite-mediated signaling, and insights into the evolutionary pressures shaping GPCR functionality. To bridge this gap, we developed OdoriFy as a dedicated platform for investigating odorant-OR interactions. However, its focus on the olfactory domain underscored the need for a broader tool to explore GPCR-ligand interactions beyond odorant research. Building on the foundation established by OdoriFy, we developed EvOlf, a unified, evolutionary-driven GPCR-ligand pair prediction model built using a deep learning approach. EvOlf leverages evolutionary relationships and structural diversity across GPCRs by encompassing both ORs and non-ORs to provide a robust framework for predicting GPCR-ligand interactions. EvOlf lays the foundation for advancing our understanding of metabolite functions by serving as an essential first step in accurately predicting their interactions with GPCRs.

3.3.2.1 EvOlf: Data-Driven Profiling of GPCR-Ligand Interactions

To ensure that EvOlf can accurately predict GPCR-ligand interactions across a broader evolutionary and functional landscape, its training data was significantly expanded beyond the dataset used for building OdoriFy. The original dataset, focused on human ORs and their corresponding odorant interactions, was enriched with data from additional species to capture the evolutionary diversity of ORs. In addition to broadening the scope of ORs, the dataset was further extended to include non-odorant GPCRs to reflect the full structural and functional diversity of the GPCR superfamily (**Figure 3.6**). This extension

incorporated GPCRs involved in diverse physiological processes, capturing interactions with a wide range of ligands such as peptides, hormones, lipids, metabolites, and small molecules. The training dataset for EvOlf comprises a total of 105,235 GPCR-ligand interaction data points, making it one of the largest and most comprehensive resources for studying GPCR-ligand interactions. It includes 48,557 interactions involving odorant GPCRs (ORs) and 56,678 involving non-odorant GPCRs (nonORs), ensuring a relatively balanced representation of both receptor subtypes. Of these interactions, 45,226 are agonistic in nature, indicating a positive receptor response upon ligand binding, while the other 60,009 represent non-agonistic interaction. For model building, these interactions were divided into two classes: agonist pairs were assigned to Class 1, while non-agonist pairs were assigned to Class 0. As expected, being the primary focus for drug discovery and other physiological studies, human-specific interactions dominate the dataset, accounting for 80,627 data points. Additionally, 24,608 data points are derived from 23 other mammalian species, showcasing its evolutionary diversity. Among non-human mammalian species, rodents contribute substantially, with mice and rats accounting for 17,574 and 16,176 data points, respectively, reflecting their importance as model organisms in biomedical research. Other species, including primates, domestic animals, and smaller mammals, add valuable diversity (**Figure 3.6**). This rich and diverse dataset highlights both conserved and species-specific GPCR functions, providing a robust foundation for building predictive models that can unravel the complex roles of GPCRs in physiological and pathological processes.

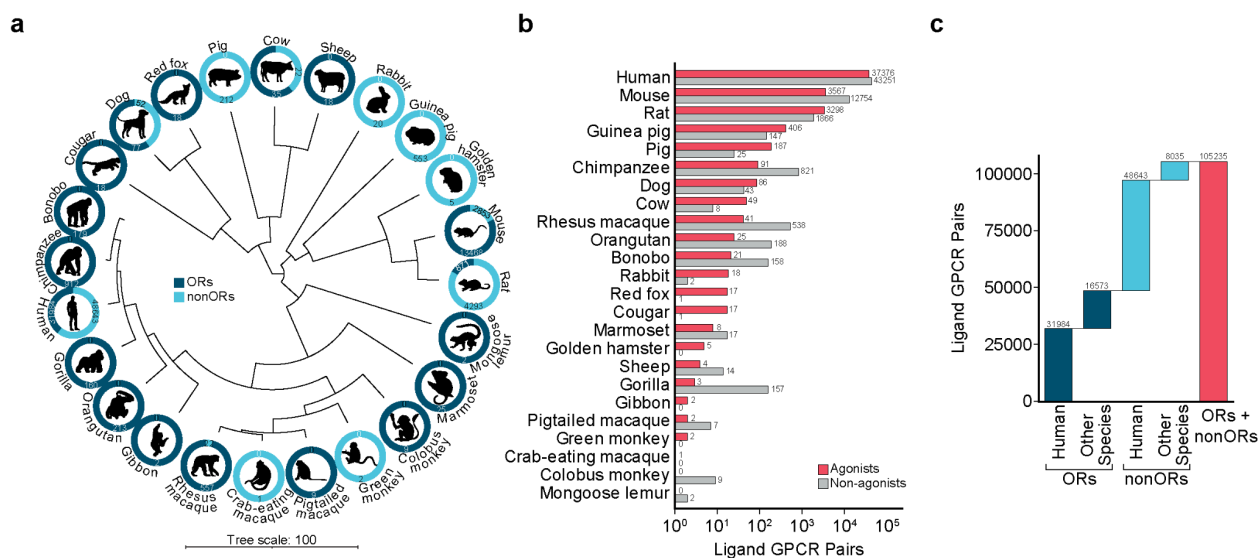


Figure 3.6 EvOlf Features a Diverse GPCR-Ligand Dataset Across Multiple Species and Receptor Classes

(a) A phylogenetic tree depicting the species included in the dataset, with the number of ligand-GPCR pairs available for ORs and non-ORs per species indicated as donut plots. Dark blue represents ORs, while light blue represents non-ORs. (b) Bar graph representing the number of agonists (red) and non-agonists (gray) GPCR-ligand

pairs recorded for each species. (c) A waterfall plot illustrating the contribution of different species (human and non-human) and receptor types (ORs and non-ORs) to the final integrated dataset used for EvOlf, culminating in the full dataset (ORs+nonORs) that captures the diversity of GPCR-ligand interactions across species.

The EvOlf dataset includes a total of 1,191 receptors, spanning both ORs and non-ORs. Notably, it also includes a significant number of receptor mutants, primarily associated with ORs, further enriching the dataset by enabling the study of structural variations and their functional consequences. The availability of such a large number of receptors, both wild-type and mutant, ensures a comprehensive representation of receptor diversity (**Figure 3.7a**). The principal component analysis (PCA) plot of all the GPCRs in the EvOlf dataset reveals the structural and functional diversity between ORs and nonORs. The separation along the principal components indicates that while there is some overlap between the two receptor classes, they occupy distinct regions in the feature space. This suggests that ORs and nonORs possess both shared and unique characteristics that contribute to their functional roles and ligand specificity. Interestingly, the clusters for each receptor class are relatively compact, occupying a restricted portion of the overall receptor space. However, when considered together, ORs and nonORs significantly expand the receptor space, reflecting the extensive diversity and functional versatility of the GPCR superfamily (**Figure 3.7b**). This observation underscores the importance of integrating both receptor types to capture the full range of GPCR-ligand interactions and their underlying complexities. Receptors such as CNR2, CNR1, S1PR1, OPRM1, and S1PR3 are among the most represented in the dataset, which are associated with critical physiological pathways, including cannabinoid signaling, sphingolipid signaling, and opioid regulation (Rios, Gomes, and Devi 2006; Doyle et al. 2020; S. Liu et al. 2021) (**Figure 3.7c**). We also looked at the conservation within and between ORs and nonORs, using TwinCons alignment scores (Penev et al. 2021). We observed more positive TwinCons scores, suggesting a high level of conservation within each group in addition to conservation between the groups, along with some near-zero values, which corresponds to variability within and between groups (**Figure 3.7d**).

The EvOlf dataset contains a total of 38,214 unique ligands, with 1,160 ligands associated with ORs and 37,091 with non-ORs (**Figure 3.7e**). There is minimal overlap between the two groups, with only 37 ligands shared. Similar to the case of receptors, the PCA plot for ligands illustrates the distribution of ligands across ORs and non-ORs, highlighting clear differences between the two classes. While there is some overlap, the distinct clustering suggests that the chemical properties of ligands interacting with ORs and non-ORs differ significantly (**Figure 3.7f**). Integrating the ligands of ORs and non-ORs leads to covering substantially broader chemical space, further adding to the diversity of the ligands. Notable ligands in the dataset include allyl phenyl acetate, cis-3-hexen-1-ol, androstadienone, and (Z)-4-undecenal, which are among the most represented (**Figure 3.7g**). On comparing functional groups

between OR and non-OR ligands, we observe that OR ligands show a notable enrichment for functional groups commonly associated with odorants, such as aldehydes (RCHO), esters (RCOOR), and alcohols (ROH), while non-OR ligands exhibit a higher representation of functional groups such as amines (RNH₂, R₂NH, R₃N) and phosphates (ROPO₃). These differences underline the chemical diversity of the ligand datasets and the distinct interaction profiles of ORs and non-ORs (**Figure 3.7h**). This diverse ligand representation provides a robust foundation for studying receptor-ligand interactions and highlights the importance of integrating both receptor types to capture the full spectrum of GPCR functionality.

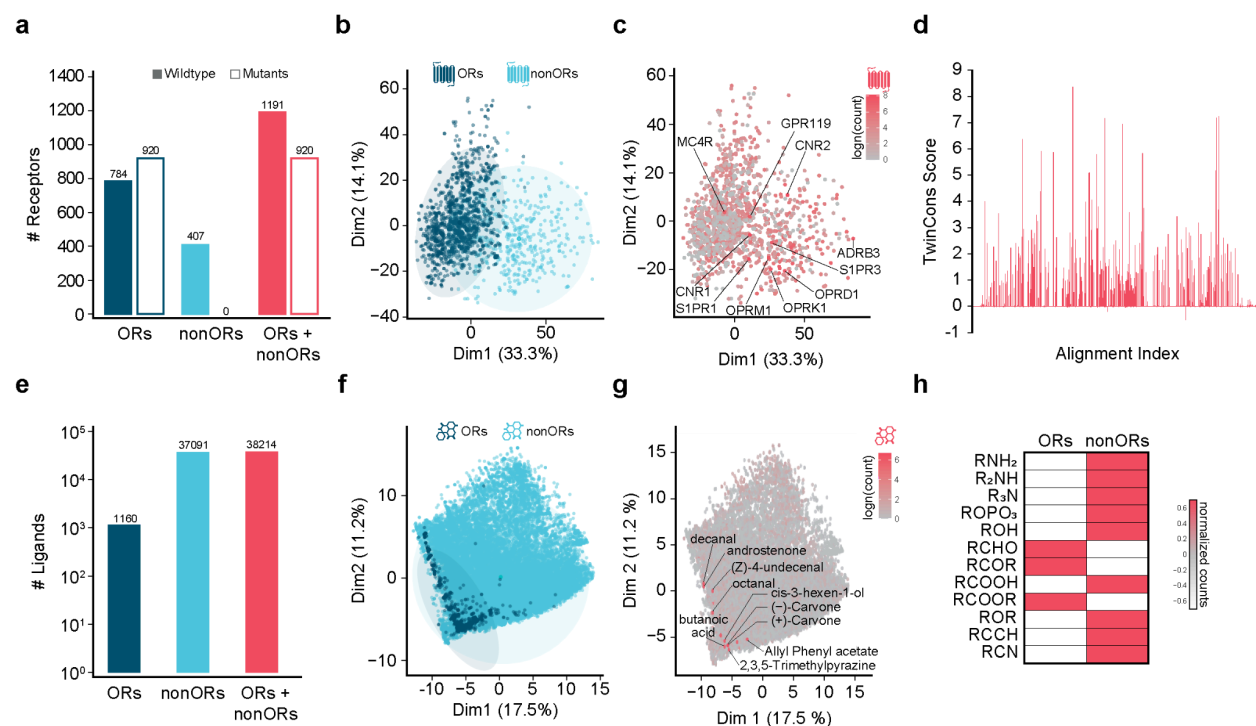


Figure 3.7 GPCR and Ligand Diversity in the EvOlf Dataset

(a) Bar plot showing the number of receptors present in the dataset, categorized into ORs, non-ORs, and the combined pan GPCR dataset (ORs + non-ORs), with wildtype and mutant receptor counts indicated. **(b)** Principal component analysis (PCA) plot, made using AmorProt fingerprint, showing the distribution of ORs and non-ORs, highlighting distinct clustering patterns between receptor classes. **(c)** A PCA plot of GPCRs, with receptor colored based on the log-transformed counts of available interaction data, indicating the most well-characterized versus less-characterized receptors in the dataset. **(d)** TwinCons conservation scores plotted against the alignment index, illustrating the evolutionary conservation of GPCR sequences. **(e)** Bar plot showing the number of ligands in the dataset available for each receptor class (ORs, non-ORs, and the pan GPCR dataset), illustrating the ligand availability and diversity within different receptor categories. **(f)** PCA plot of ligand chemical space demonstrating the clustering patterns between ligands interacting with ORs and non-ORs. Atom Pair Fingerprints from the ChemmineR package were used as features. **(g)** PCA plot of ligands, with ligands colored based on the log-transformed counts of available interaction data, highlighting the most well-characterized ligands in the dataset.

dataset. **(h)** Heatmap of normalized functional group distributions between the ligands available for ORs and non-ORs, highlighting differences in chemical preferences.

3.3.2.2 Multi-dimensional Feature Integration for GPCR-Ligand Interaction Prediction in EvOlf

Unlike most existing models that rely on a single type of descriptor, such as chemistry-based descriptors for ligands or conventional sequence-based descriptors for receptors, EvOlf integrates multiple high-dimensional descriptors to comprehensively capture the diverse feature properties of both ligands and receptors. EvOlf then generates low-dimensional embeddings that accurately represent ligand-receptor pairs, enabling more robust predictions of GPCR-ligand interactions. By combining diverse feature extraction methodologies, EvOlf captures a multi-dimensional representation of the functional, biochemical, structural, and contextual properties of the input data, which is critical for addressing the complexity and diversity inherent in GPCR-ligand interactions.

To achieve this, we applied five advanced methodologies to extract meaningful, non-redundant features from ligands, each contributing unique layers of information to the model. Starting with the SMILES (Simplified Molecular Input Line Entry System) representation of each ligand, we processed it through the following techniques: Signaturizer, which captures bioactivity-based features (Bertoni et al. 2021); ChemBERTa-2, a transformer-based architecture, which generates contextual embeddings that account for molecular context and semantics (Chithrananda, Grand, and Ramsundar 2020; Ahmad et al. 2022); Mordred, a widely used tool for chemical descriptors, which extracts features covering physicochemical, topological, and geometric properties of each ligand (Moriwaki et al. 2018); Mol2Vec, which encodes molecular substructures into embeddings, focusing on the chemical building blocks of each molecule (Jaeger, Fulle, and Turk 2018); and Graph2Vec, which leverages graph-based representations to capture the molecular structure (Narayanan et al. 2017). Together, these methodologies provide a rich and multi-dimensional representation of ligand characteristics, ensuring that all relevant functional, chemical, and structural information is captured (**Figure 3.8a, b**).

For GPCR receptors, we employed four advanced feature extraction techniques to obtain protein descriptors that capture the functional and structural complexity of the receptor sequences. ProtR was used to extract conventional protein descriptors, generating features that provide a broad overview of the biochemical properties of each receptor (Xiao et al. 2015). ProtBERT, a transformer-based model, generated embeddings that account for biological context derived from the receptor sequences, capturing nuances in sequence function and interaction potential (Elnaggar et al. 2022). ProtT5, another transformer-based model, contributed embeddings focusing on capturing complex sequence patterns and higher-order relationships between amino acid residues (Elnaggar et al. 2022). Finally, MathFeature

extracted features representing the mathematical properties of protein sequences, providing a unique perspective on sequence structure and dynamics (Bonidia et al. 2022) (**Figure 3.8a, c**).

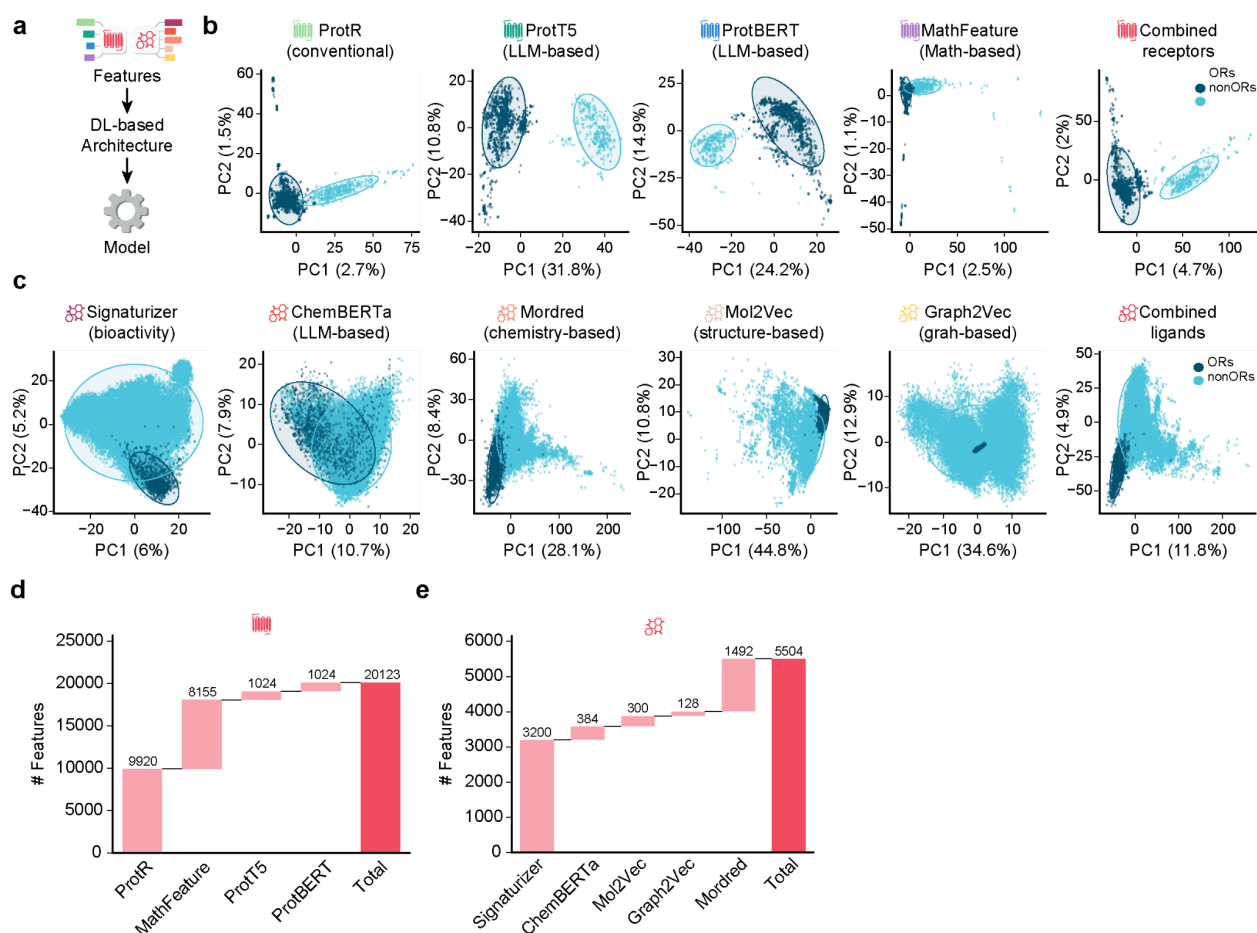


Figure 3.8 Multifaceted Data Representation for GPCR-Ligand Interaction Modeling in EvOlf

(a) Schematic representation (top left) of the feature extraction and deep-learning-based modeling pipeline in EvOlf. Receptor and ligand features are extracted using diverse computational approaches that serve as inputs to a deep-learning architecture, which processes and integrates them into joint embeddings for downstream GPCR-ligand interaction prediction. (b-c) Principal Component Analysis (PCA) plots illustrating the feature representations of (b) GPCRs and (c) ligands generated using various featurizing methods. Each plot shows the distribution of olfactory receptors (ORs, dark blue) and non-olfactory receptors (nonORs, light blue) in the feature space. The rightmost column represents the combined embeddings obtained by integrating all feature sets. (d) Waterfall plot depicting the number of molecular features extracted for GPCRs from different feature extraction techniques. (e) Waterfall plot showing the number of features extracted for ligands from indicated protein-based features.

By integrating these diverse feature sets, EvOlf captures the full range of biochemical and structural properties of both ligands and receptors, providing a comprehensive and nuanced understanding of their

interactions. This multidimensional approach not only enhances the predictive power of the model but also allows for the exploration of previously underappreciated features that contribute to GPCR-ligand binding. The ability to combine such diverse feature extraction methodologies highlights the versatility and sophistication of EvOlf, making it a robust tool for studying GPCR-mediated signaling and ligand-receptor interactions. This integrative approach lays the groundwork for uncovering novel interactions and mechanisms underlying GPCR functionality across diverse biological contexts.

3.3.2.3 Comprehensive GPCR-Ligand Interaction Prediction Using Multi-Layered Neural Architectures

After completing the feature extraction stage, the focus shifted to the next critical phase: employing deep learning to predict GPCR-ligand interactions. To achieve this, we developed and tested five distinct deep learning architectures, each incorporating unique methodologies for feature processing, integration, and classification. To ensure fair evaluation, all architectures were compared using the same testing dataset, and their performance was rigorously assessed across multiple metrics. The dataset was split into training (80%), validation (10%) and testing (10%) sets. The training dataset underwent extensive pre-processing. Each feature file was scaled and subjected to principal component analysis (PCA) to reduce dimensionality while retaining essential information. Importantly, the scaling and PCA models were saved during training and later applied to the testing data to maintain consistency between training and testing datasets. Additionally, the training data was upsampled to address class imbalance using SMOTE, ensuring that both agonist and non-agonist pairs were adequately represented during training. This pre-processing pipeline laid the groundwork for robust and reliable model evaluation.

The five architectures share foundational similarities, such as independent pipelines for ligand and receptor processing, transformers for capturing contextual relationships, and convolutional layers for refining spatial patterns (**Figure 3.9a**). However, they differ in how they process, integrate, and classify ligand and receptor features. Of note, each model was trained for 100 epochs.

Architecture 1 adopts a hybrid convolution-recurrent encoding pipeline to process ligand and receptor features. Ligand features are concatenated and passed through a 2D convolutional layer to extract spatial patterns before being reshaped and processed by an LSTM network, capturing sequential dependencies. Similarly, receptor features undergo convolutional processing followed by an LSTM module. The final hidden states of both LSTM networks form compact representations of the ligand and receptor features, which are then concatenated into a joint embedding and passed through fully connected layers with dropout for classification. While this setup effectively captures localized molecular motifs and sequential

dependencies, the fusion mechanism relies on simple concatenation, which does not explicitly model interactions between ligand and GPCR features (**Figure 3.9a**).

Architecture 2 integrates ligand and receptor features using separate transformer-based modules followed by 1D convolutional layers to extract meaningful representations. Ligand features are processed through a transformer encoder-decoder architecture to capture long-range dependencies and contextual relationships, followed by layer normalization and a 1D convolutional layer for further refinement. A similar pipeline is applied to receptor features, ensuring parallel processing for both data types. The outputs from these pipelines are concatenated, and a mean-pooling operation derives a joint embedding, which is then passed through fully connected layers with dropout for classification. By leveraging the attention mechanism of transformers, Architecture 2 enhances representational power over Architecture 1's LSTM-based encoding, allowing dynamic feature weighting and improving contextual feature extraction. Additionally, the 1D convolutional layer refines spatial patterns while maintaining localized feature representations. However, the fusion strategy relies on a simple mean-reduction operation, which, while reducing feature dimensionality, may lead to the loss of crucial ligand-GPCR interaction details (**Figure 3.9b**).

Architecture 3 combines transformers, LSTMs, and convolutional layers to extract and fuse features from both ligands and receptors, effectively capturing both local and global dependencies. Ligand and receptor features are first processed independently using transformer modules to capture contextual relationships, followed by LSTM layers to model sequential dependencies. The outputs from the LSTMs are fused into a joint embedding, which, along with the individual ligand and receptor embeddings, is passed through a series of 1D convolutional layers to refine the feature representation further. Unlike its predecessors, Architecture 3 introduces a more expressive fusion mechanism by expanding the fusion process into a three-channel tensor representation, where individual ligand and receptor embeddings are stacked alongside their fused representation. This enables hierarchical spatial processing via convolutional layers, allowing the model to extract multi-resolution patterns from the fused feature space. Additionally, LSTM-based feature transformation post-fusion enhances the model's ability to preserve sequential dependencies. The final feature set is flattened and passed through a fully connected classifier for interaction prediction. While this architecture significantly improves fusion expressiveness, its reliance on fixed feature alignment during fusion may limit its adaptability to varying ligand-receptor interaction patterns (**Figure 3.9c**).

Architecture 4 employs a hybrid Transformer-Convolution-LSTM approach to process ligand and receptor features. Ligand and receptor features are first passed through transformer modules to capture

contextual and relational information, followed by layer normalization to stabilize training dynamics. The transformer outputs are then refined using convolutional layers to preserve spatial patterns before being processed by LSTMs, which model sequential dependencies and generate compact representations. Unlike Architecture 3, where transformers operate in isolation during embedding generation, Architecture 4 integrates self-attention within both the embedding and fusion stages, allowing ligand and receptor descriptors to adaptively reweigh feature importance based on learned attention scores. The addition of layer normalization post-transformer processing stabilizes training dynamics, reducing the risk of vanishing or exploding gradients in deep networks. The final outputs from the ligand and receptor pipelines are combined into a joint representation using a stacking and averaging operation, then passed through a fully connected classifier with dropout for interaction prediction. While this architecture enhances feature fusion by incorporating multiple extraction techniques, its reliance on mean-reduction in the fusion step still limits its ability to fully exploit multi-scale dependencies between ligands and GPCRs (**Figure 3.9d**).

Architecture 5 represents the most architecturally sophisticated design, implementing a multi-stage self-attention mechanism across all three computational blocks. Ligand and receptor features are independently passed through transformer modules to capture contextual relationships, followed by layer normalization and convolutional layers to refine spatial patterns. These refined features are further processed through LSTMs to model sequential dependencies, with the final hidden states representing compact embeddings for ligands and receptors. Unlike prior architectures that rely on static feature concatenation, Architecture 5 introduces a dedicated fusion transformer, a cross-attention module that dynamically models fine-grained ligand-receptor interactions in a shared latent space. This mechanism ensures that ligand embeddings are contextually aware of receptor-specific variations and vice versa, capturing complex interaction patterns that previous architectures overlook. The fused representation undergoes a final layer normalization step before being passed through a fully connected classifier with dropout for interaction prediction. The classification module benefits from this enriched feature space, enabling high-confidence predictions through a deep MLP with dropout-driven stochastic regularization. Additionally, the architecture is optimized using Xavier initialization, orthogonal weight regularization, and a ReduceLROnPlateau scheduler, ensuring stability during gradient-based optimization. By leveraging multi-stage self-attention across all computational blocks and introducing a transformer-driven fusion strategy, Architecture 5 offers superior model expressivity, feature interpretability, and classification robustness, making it the optimal framework for GPCR-ligand binding prediction (**Figure 3.9e**).

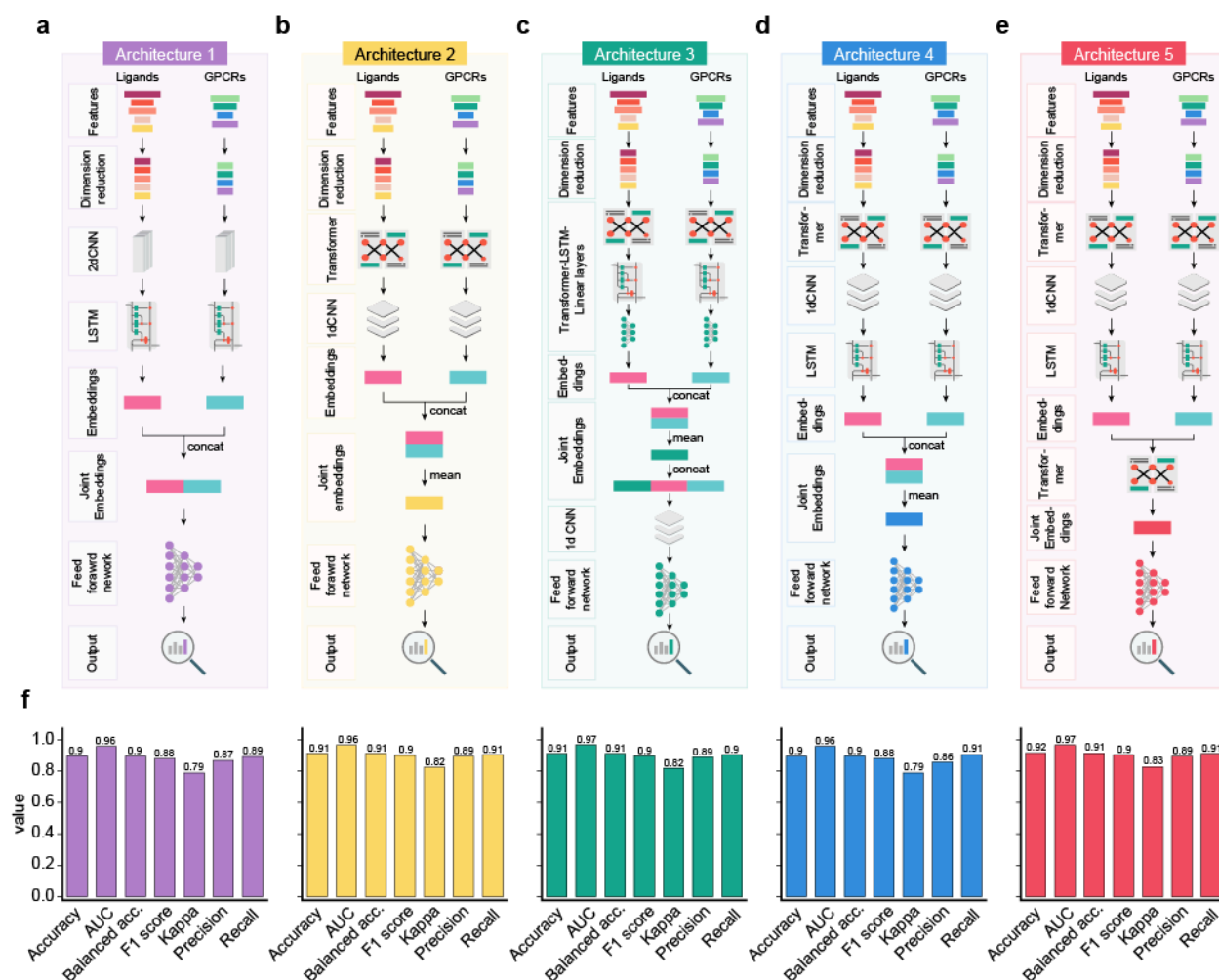


Figure 3.9 Comparative Analysis of Deep Learning Architectures for GPCR-Ligand Interaction Prediction

(a–e) Schematic representations of the five deep learning architectures: (a) Architecture 1, (b) Architecture 2, (c) Architecture 3, (d) Architecture 4, and (e) Architecture 5, designed for GPCR-ligand interaction prediction. Each architecture incorporates different feature processing techniques, including dimensionality reduction, convolutional layers, LSTMs, and transformer-based embeddings. Ligand and GPCR features undergo distinct transformations before being integrated into joint embeddings, which serve as input to a final classification network. (f) Bar plot showing the performance of the five architectures on the testing dataset across key evaluation metrics, including accuracy, AUC, balanced accuracy, F1-score, Cohen's kappa, log loss, precision, and recall.

Upon comparing the architectures using metrics such as accuracy, area under the curve (AUC), F1 score, and recall, Architecture 5 consistently outperformed the others, including an accuracy of 0.92, a balanced accuracy of 0.91, and an AUC of 0.97, demonstrating superior performance on the testing dataset (**Figure 3.9f**). Its combination of transformers, convolutional layers, LSTMs, and a fusion transformer enabled it to comprehensively capture the complex dependencies in GPCR-ligand interactions, making it the most effective model for this task.

3.3.2.4 Unified Models for GPCR-Ligand Interactions: Leveraging Diversity in Receptor Classes and Species

To evaluate the effectiveness of integrating diverse receptor data, we compared the performance of three models trained on distinct subsets of the GPCR dataset: (i) the Pan GPCRs model trained on a combined dataset of both ORs and non-ORs (ORs+nonORs), (ii) the ORs-only model trained exclusively on odorant receptor data, and (iii) the non-ORs-only model trained exclusively on non-odorant receptor data (**Figure 3.10a**). With this comparison, we aimed to determine whether training on a combined dataset provides superior generalizability across all receptor types, challenging the conventional approach of focusing on either ORs or non-ORs in isolation.

The results unequivocally demonstrate the superiority of the Pan GPCRs model over the specialized models trained on single receptor classes. While the ORs-only model excelled on OR-specific testing data (AUC = 0.89), it performed poorly when tested on non-OR testing datasets (AUC for nonORs = 0.55 and ORs+nonORs = 0.57), highlighting its limited scope. Similarly, the non-ORs-only model performed well on non-OR-specific testing data (AUC = 0.94) but failed to generalize to OR testing data (AUC for ORs = 0.54 and ORs+nonORs = 0.75), indicating its domain-specific nature. This domain dependency in specialized models reflects a significant limitation in capturing the broader structural and functional diversity inherent in GPCR-ligand interactions. In contrast, the Pan GPCRs model, trained on the integrated dataset of both ORs and non-ORs, consistently outperformed the ORs-only and non-ORs-only models across all testing datasets. Notably, the combined model achieved results comparable to the specialized models on their respective domains while significantly outperforming them in cross-domain evaluations (AUC for ORs = 0.89, nonORs = 0.94, ORs+nonORs = 0.97). For instance, while the ORs-only and non-ORs-only models showed sharp declines in metrics such as balanced accuracy, F1-score, and kappa when evaluated on cross-domain datasets, the pan GPCRs model demonstrated remarkable stability, achieving high scores across the board (**Figure 3.10b, c**). This demonstrates that the pan GPCRs model effectively captures the unique and shared features of both receptor types, enabling it to generalize well across diverse receptor families. This underscores the importance of adopting a holistic approach to GPCR-ligand interaction modeling, particularly for applications requiring broad-spectrum predictions across diverse receptor families.

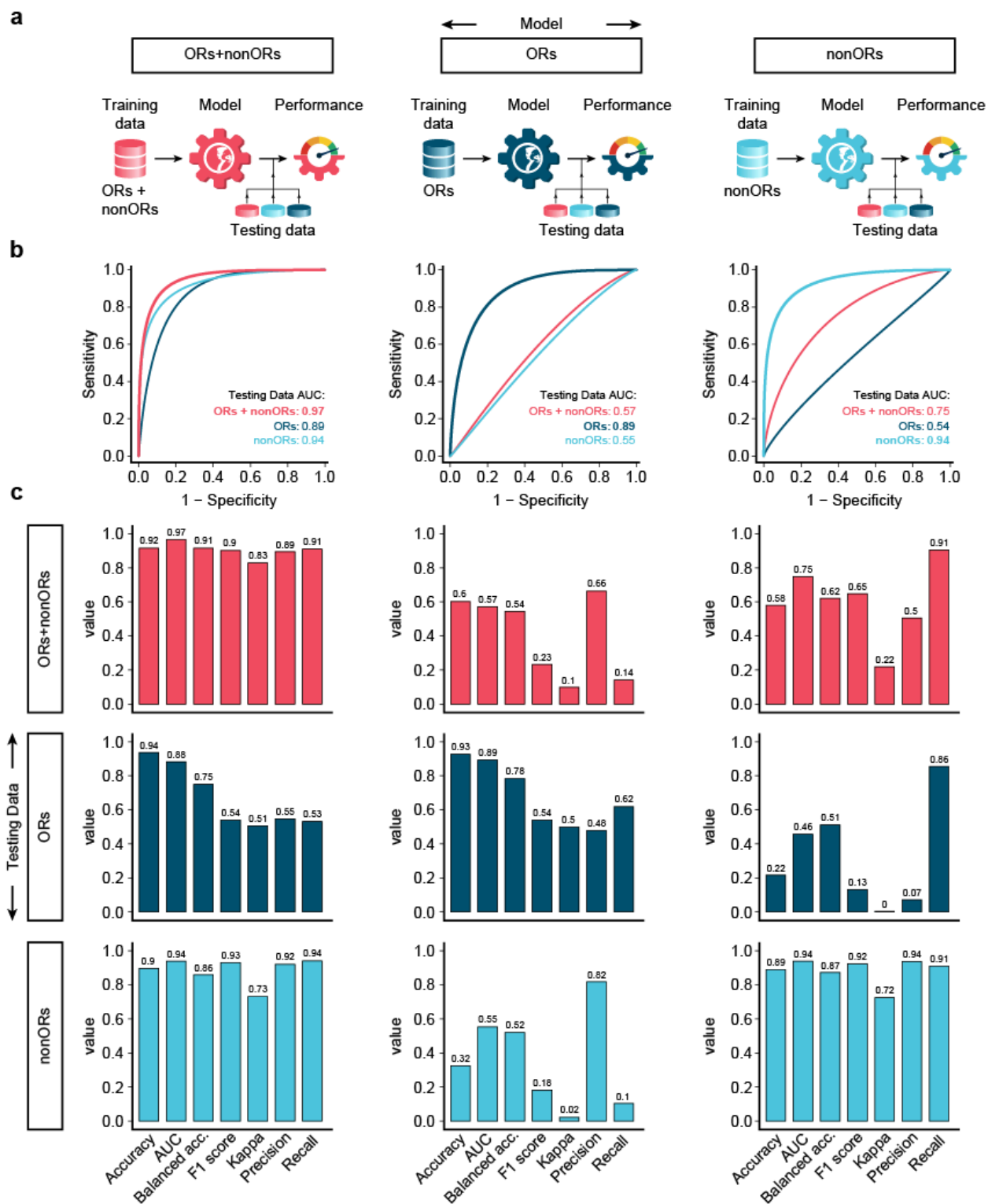


Figure 3.10 Integrating ORs and Non-ORs Improves Predictive Power

(a) Schematic representation of the training and evaluation framework used to assess the impact of integrating receptor classes. Models were trained separately on datasets containing pan GPCRs (ORs+nonORs), ORs alone,

and nonORs alone. Each trained model was then evaluated on different testing datasets to assess generalization performance. **(b)** Receiver Operating Characteristic (ROC) curves for models trained on pan GPCRs, ORs-alone, and nonORs-alone datasets. The Area Under the Curve (AUC) values indicate that the combined pan GPCR model outperforms the individual receptor models in cross-domain predictions. **(c)** Bar Plots show the performance metrics, including accuracy, AUC, balanced accuracy, F1-score, Cohen's kappa, precision, and recall for each model across different testing datasets.

To further assess the utility of integrating data from multiple species in training, we designed a comparative study to evaluate whether leveraging data from other species confers any benefit for predicting ligand-receptor interactions, particularly for human receptors. To this end, two models were trained using Architecture 5, one model was trained on a dataset comprising of GPCRs (ORs and nonORs; referred to as pan GPCRs) from multiple mammalian species, while the other model utilized a dataset containing only human-specific GPCRs (hORs and non-hORs; referred to as human GPCRs). Both models were tested across six datasets: pan GPCRs (ORs+nonORs), human GPCRs (hORs+non-hORs), ORs, nonORs, hORs, and non-hORs (**Figure 3.11a**).

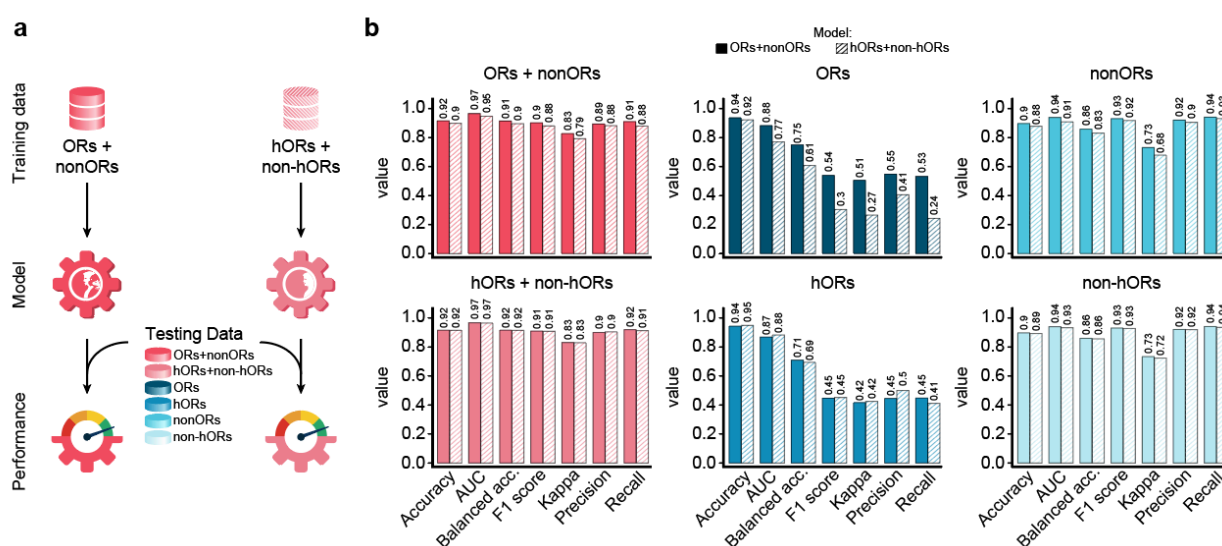


Figure 3.11 Evaluating the Impact of Multi-Species Data on GPCR-Ligand Interaction Prediction

(a) Schematic representation of the training and evaluation setup to assess the impact of multi-species data on predictive performance. Models were trained on either a pan-GPCR dataset (ORs + nonORs) or a human-only dataset (hORs + non-hORs) and tested across six different datasets. **(b)** Bar plots showing the performance of the models trained on pan-GPCRs (solid bars) and human-only GPCRs (striped bars) on the indicated testing data.

The results indicate that the performance of the model trained on pan GPCRs was only marginally better than the model trained on human GPCRs across most testing datasets, indicating that the inclusion of non-human GPCRs adds a layer of diversity to the training data, thereby slightly enhancing the model's

generalization ability. This improvement, while modest in general, was particularly significant for the testing dataset focused on ORs, where the pan-GPCR model (AUC = 0.88, balanced accuracy = 0.75) outperformed the human GPCR model (AUC = 0.88, balanced accuracy = 0.61) by a noticeable margin. The superior performance on OR-specific data suggests that including GPCRs from multiple species allows the model to capture the greater variability inherent in OR-ligand interactions, which are often more diverse due to their specialized roles in olfactory signaling. For the other testing datasets, such as those focusing on non-ORs or combined datasets, the performance gap between the two models was minor (**Figure 3.11b**). This result implies that the human GPCR dataset is largely sufficient for capturing the patterns required for accurate prediction in these cases. However, the notable difference observed for OR-specific data demonstrates the value of integrating multi-species data for tasks that involve higher receptor diversity or when focusing on specific GPCR subtypes.

3.3.2.5 Unified High-Dimensional Embeddings for GPCR-Ligand Prediction

After finalizing Architecture 5 and pan GPCR integrated dataset capturing both receptor and species diversity, as the optimal model and training dataset for GPCR-ligand interaction prediction, we conducted a 10-fold cross-validation (CV) to robustly evaluate its generalization performance across the dataset (**Figure 3.12a, b**). The 10-fold cross-validation results demonstrated consistently high performance metrics across all folds. As depicted in the figure, metrics such as accuracy, AUC, balanced accuracy, F1 score, Cohen's kappa, precision, and recall all exhibited minimal variance, reflecting the stability and robustness of the model. Notably, the average accuracy and AUC values were 0.91 and 0.96, indicating that the model reliably distinguished between agonist and non-agonist GPCR-ligand interactions (**Figure 3.12c**). The consistency of these results across multiple folds underscores the effectiveness of Architecture 5 in capturing the complex dependencies between GPCRs and ligands. By leveraging 10-fold cross-validation, we ensured that the performance metrics were not influenced by any particular dataset split, providing robust evidence for the generalizability of the model to unseen data.

After cross-validation, we proceeded to train the final model using the entire dataset (100% of the data) for 100 epochs to fully leverage the richness and diversity of the dataset. This step ensured that the model had access to all available information for capturing the intricate dependencies between GPCRs and ligands, thereby enhancing its predictive power. This final model, trained on the complete dataset, is designated as EvOlf (**Figure 3.12a, d**).

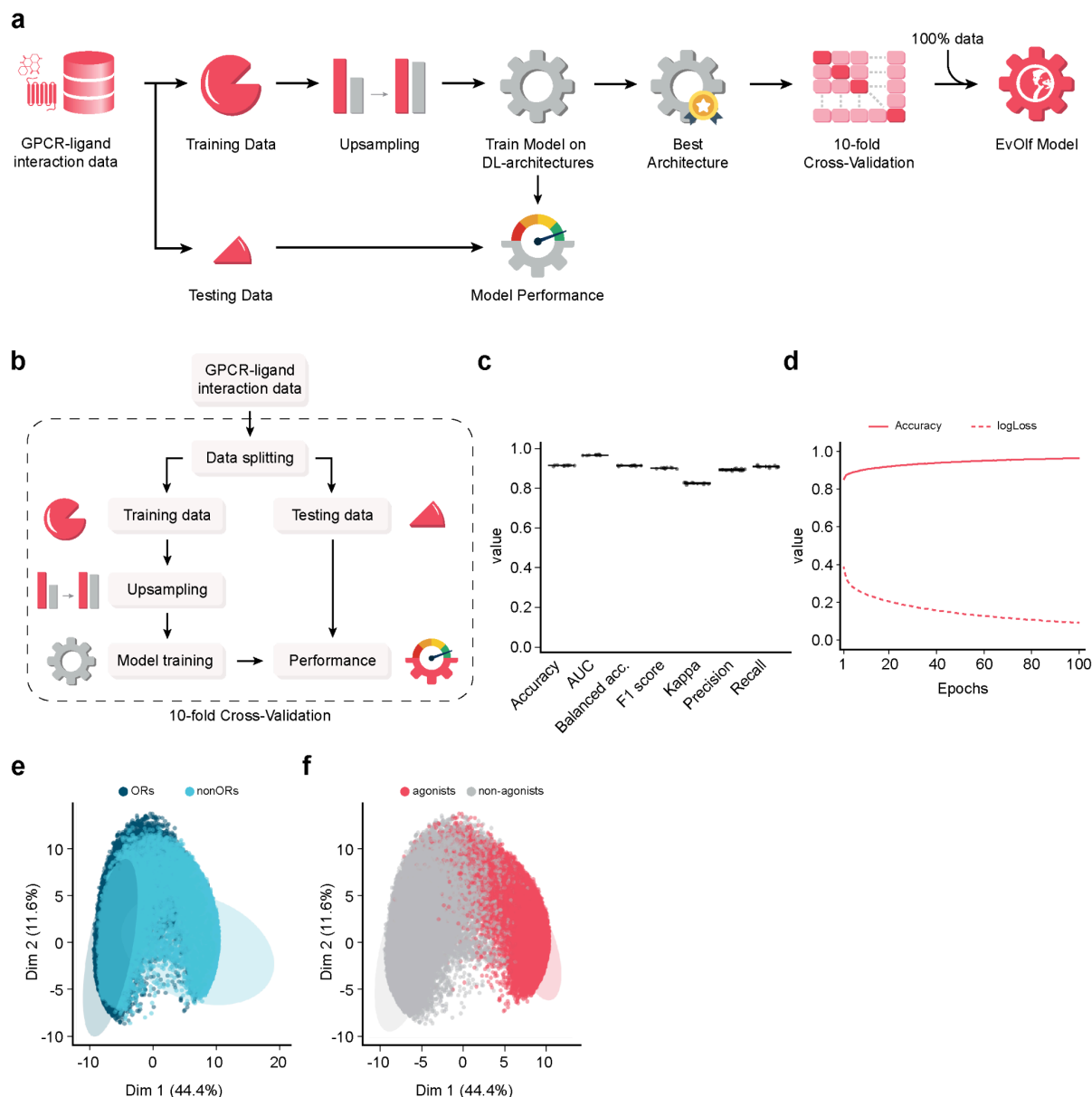


Figure 3.12 EvOlf Generates Joint Embeddings for a Comprehensive GPCR-Ligand Interaction Space

(a) Overview of the EvOlf model development pipeline. GPCR-ligand interaction data was split into training and testing sets, followed by upsampling of the training data to balance the dataset. Multiple deep-learning architectures were evaluated, and the best-performing architecture was selected for final training. The model was then subjected to 10-fold cross-validation before training on the full dataset to generate the final EvOlf model. **(b)** Schematic representation of the 10-fold cross-validation procedure used to assess the stability of the model hyperparameters. **(c)** Box plot showing performance metrics obtained from 10-fold cross-validation, demonstrating high and stable accuracy, AUC, balanced accuracy, F1-score, kappa, precision, and recall across all folds. **(d)** Training performance of the final EvOlf model trained on the complete pan-GPCR dataset. **(e)** PCA visualization of the GPCR-ligand interaction data using the joint embeddings provided by EvOlf, with ORs (dark blue) and nonORs

(light blue) forming separate clusters. **(f)** PCA visualization of the GPCR-ligand interaction data using the joint embeddings provided by EvOlf, with agonist (dark blue) and non-agonist (light blue) pairs forming separate clusters.

In addition to providing interaction probabilities, EvOlf generates joint embeddings that serve as comprehensive representations of ligand-receptor pairs. These embeddings capture intricate biochemical and structural relationships between ligands and receptors, offering a robust foundation for analyzing GPCR-ligand interactions beyond simple classification tasks. Using these embeddings, we explored the overall distribution of the data to gain further insights into the diversity and functional properties encapsulated by EvOlf. The embeddings effectively distinguish between receptor classes and interaction types. When categorized by receptor type, embeddings corresponding to ORs and nonORs form distinct clusters, with some degree of overlap (**Figure 3.12e**). This separation highlights the unique characteristics and specialized functions of each receptor class, while the overlap reflects shared biochemical and structural features underlying their functionality. Such clustering underscores the ability of EvOlf to capture the complexity and diversity of the GPCR superfamily. Similarly, when grouped by interaction type, the embeddings for agonist and non-agonist pairs show clear separation (**Figure 3.12f**). These distinctions validate the embeddings as accurate and meaningful representations of ligand-receptor interactions.

3.3.2.6 EvOlf Outperforms Conventional Methods in GPCR–Ligand Interaction Prediction

To evaluate the predictive performance of EvOlf in comparison to existing computational models, we benchmarked it against tools like AiGPro (Brahma et al. 2025) and pdCSM-GPCR (Velloso, Ascher, and Pires 2021). Many of the existing GPCR classification tools are outdated, rendering their installation cumbersome and error-prone (Ahmed et al. 2021). Some models require explicit structural data of GPCRs, while others are constrained by the need for domain-specific information (Kumar, Teli, and Kim 2024; H. Zhang et al. 2024; Yamane and Ishida 2023). Given these challenges, we selected two regression-based models: AiGPro and pdCSM-GPCR, for comparison with EvOlf.

We compiled an unseen dataset consisting of 3870 agonist-GPCR interaction data points for human and mouse models. However, all mouse data points were removed since AiGPro and pdCSM-GPCR are trained exclusively on human GPCRs. Additionally, both models are restricted to specific human GPCR subsets, with AiGPro and pdCSM providing predictions for 231 and 36 GPCRs, respectively. As a result, we further filtered out interactions involving GPCRs absent from their respective models, reducing the final dataset to 3622 interaction points for AiGPro and 607 for pdCSM-GPCR. For performance evaluation, we categorized predictions into binary outcomes based on a GPCR activity cut-off of 100 nM

and computed hit-and-miss rates for each model. EvOlf demonstrated superior predictive power, achieving a hit rate of 0.81 and a miss rate of 0.19, whereas AiGPro performed significantly worse with a hit rate of 0.28 and a miss rate of 0.72 (**Figure 3.13a**). Similarly, when compared to pdCSM-GPCR, EvOlf outperformed with a hit rate of 0.82 and a miss rate of 0.18 compared to pdCSM-GPCR's hit rate of 0.44 and miss rate of 0.56 (**Figure 3.13b**). These results underscore EvOlf's robust predictive capability across diverse GPCR-agonist interactions. In contrast to the narrow applicability of AiGPro and pdCSM-GPCR due to their specificity to certain GPCRs, EvOlf offers broader generalizability and greater accuracy without requiring extensive domain knowledge or structural inputs, making it a more accessible and versatile tool for GPCR interaction predictions.

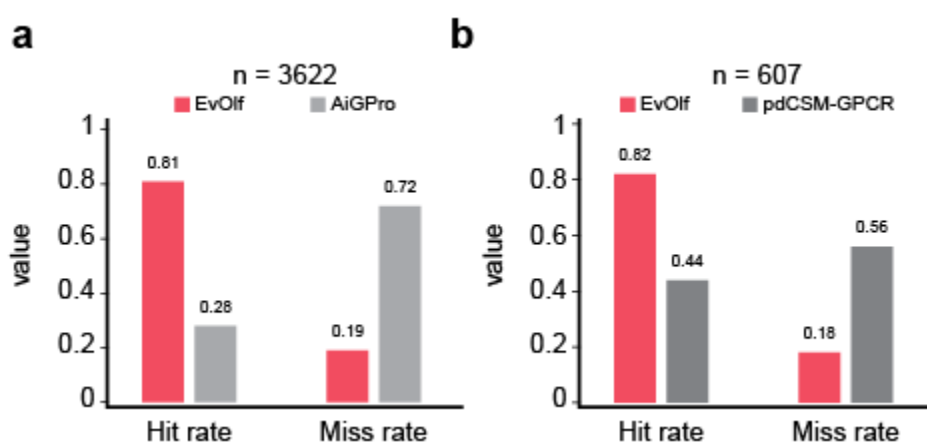


Figure 3.13 EvOlf Outperforms Existing Tools in GPCR-Ligand Predictions

(a) Bar plot comparing the performance of EvOlf against AiGPro on an unseen dataset of 3622 agonist-GPCR interactions. **(b)** Bar plot comparing the performance of EvOlf against pdCSM-GPCR on an unseen dataset of 607 agonist-GPCR interactions.

By utilizing the full spectrum of data, including both agonist and non-agonist interactions across ORs and non-ORs, EvOlf represents the culmination of our efforts to build a comprehensive and reliable tool for GPCR-ligand interaction prediction. This unified model serves as a critical resource for advancing our understanding of GPCR functionality, facilitating the discovery of novel interactions, and exploring the roles of metabolites in health and disease. EvOlf now stands as a robust, final implementation ready for application in diverse research and therapeutic contexts. EvOlf demonstrates state-of-the-art accuracy in predicting GPCR-ligand interactions, including those involving the often overlooked ORs. By coupling evolutionary-guided interactions with advanced deep learning techniques, EvOlf not only enhances the deorphanization process but also holds promise for uncovering new therapeutic targets and advancing our understanding of GPCR biology. This framework represents a significant step forward in computational

biology, offering a robust tool for researchers in the field of drug discovery and beyond. Taken together, the present work utilizes the power of Artificial Intelligence to deorphanize therapeutic-relevant GPCRs.

3.4 Discussion

The development of EvOlf and OdoriFy represents significant advancements in computational frameworks for understanding GPCR-ligand interactions, with each model tailored to address distinct challenges in the field. OdoriFy serves as a pioneering tool for chemosensory research, integrating four specialized prediction engines: Odorant Predictor, Odor Finder, OR Finder, and Odorant-OR Pair Analysis. These modules collectively allow researchers to validate odorant properties, identify responsive ORs, predict putative odorants for given ORs, and verify specific OR-odorant pairs. Importantly, OdoriFy incorporates explainable artificial intelligence, enabling users to decode predictions at both atomic and amino acid levels, a feature absent in other models such as ODORactor (Xinyi Liu et al. 2011), and DeepOlf (Sharma et al. 2022). Its high accuracy, content-rich interface, and customizable functionalities make it a valuable resource for OR deorphanization, prioritizing odorants for experimental validation and uncovering the molecular basis of odorant-OR interactions. OdoriFy's focus on the olfactory domain provides insights into human-specific chemosensory mechanisms and opens avenues for understanding odorant perception. For example, OdoriFy can help identify the physicochemical properties that define odorants, assist in designing synthetic odorants, and explore the molecular basis of odorant inhibition, a critical yet underexplored aspect of olfactory signaling. Moreover, it can predict endogenous ligands for extranasal ORs, such as OR51E2 in prostate cancer (Abaffy et al. 2018), underscores its relevance beyond traditional olfaction studies. At the same time, OdoriFy's focus on the olfactory system, while groundbreaking, highlighted the need for a more generalized framework to explore GPCR-ligand interactions beyond the olfactory domain.

EvOlf builds directly on the foundation established by OdoriFy, significantly expanding the scope of GPCR-ligand interaction studies to encompass both olfactory and non-olfactory GPCRs across multiple species. By integrating a diverse dataset of over 105,000 interaction data points, including wild-type and mutant receptors, EvOlf bridges the gap between specialized chemosensory research and generalized GPCR studies. This comprehensive dataset captures structural variations and functional diversity, enabling high accuracy predictions while shifting the focus from the sensory roles of odorants to the broader roles of metabolites. Metabolites, which often act as signaling molecules, interact with GPCRs to regulate diverse biological processes, including immune responses, energy metabolism, and neuronal signaling (Husted et al. 2017). EvOlf's predictive framework offers a powerful avenue for uncovering the roles of metabolites, particularly those that are less studied or unknown. By screening large chemical

libraries like the ZINC database (Tingle et al. 2023), EvOlf can facilitate the discovery of novel ligands for both orphan and known GPCRs, accelerating deorphanization efforts and identifying receptor-ligand pairs with potential therapeutic significance. Furthermore, metabolomics analyses often identify metabolites with unknown roles (Giera, Yanes, and Siuzdak 2022), while transcriptomic studies highlight GPCRs that are upregulated in pathological conditions such as cancer, inflammation, or neurological disorders (Zeng et al. 2021; Couvineau, Gomariz, and Tan 2020). By linking these datasets, EvOlf could predict interactions between disease-associated GPCRs and metabolites, providing a mechanistic understanding of how these receptors mediate disease processes and could reveal new therapeutic targets or biomarkers, advancing personalized medicine approaches.

A key strength of EvOlf is its ability to address the limitations of existing GPCR-ligand interaction models, which often suffer from over-specialization and limited generalizability. Many models, such as ODORactor (Xinyi Liu et al. 2011), DeepOlf (Sharma et al. 2022), AiGPro (Brahma et al. 2025), pdCSM-GPCR (Velloso, Ascher, and Pires 2021), GPCR_LigandClassify.py (Ahmed et al. 2021), DeepGPCR (H. Zhang et al. 2024), and GPCR-IPL score (Kumar, Teli, and Kim 2024), use graph-based molecular representations, convolutional neural networks (CNNs), and attention mechanisms to enhance predictions. However, most are designed specifically for either odorant or non-odorant GPCRs, making them less effective for studying the full diversity of the GPCR superfamily (Kumar, Teli, and Kim 2024; H. Zhang et al. 2024; Yamane and Ishida 2023; Velloso, Ascher, and Pires 2021; Brahma et al. 2025; Ollitrault et al. 2024; Sharma et al. 2022; Bushdid et al. 2018; Jabeen et al. 2021; Cong et al. 2022; Caballero-Vidal et al. 2020; A. Gupta et al. 2021; R. Gupta et al. 2021). Furthermore, some models are trained on species-specific datasets, restricting their generalizability to other species (Brahma et al. 2025; Velloso, Ascher, and Pires 2021). The neglect of evolutionary information and mutational variations is another significant drawback. GPCRs have evolved over millions of years, leading to variations in sequence and structure that can significantly impact ligand binding. Similarly, mutations, both natural and engineered, can alter receptor function and ligand specificity. Models that fail to account for these variations may produce inaccurate predictions, particularly for novel or less-studied GPCRs. Many models rely on structural data, but they are constrained by the limited availability of experimentally determined GPCR structures (Kumar, Teli, and Kim 2024; H. Zhang et al. 2024; Yamane and Ishida 2023). While predictive models like AlphaFold offer alternatives, errors in predicted structures can propagate and compound through downstream analyses, reducing reliability, especially for receptors without high-confidence structures. Furthermore, most existing GPCR-ligand interaction models are built only on GLASS (Chan et al. 2015), a highly heterogeneous dataset integrating data from diverse experimental sources with varying methodologies and assays, each with its own sensitivity and detection

limits (Kumar, Teli, and Kim 2024; H. Zhang et al. 2024; Yamane and Ishida 2023; Velloso, Ascher, and Pires 2021; Brahma et al. 2025). This variability poses a significant challenge for regression-based models, as differences in receptor expression levels, assay formats, and experimental conditions lead to inconsistencies in reported EC50 values. EvOlf overcomes these limitations by incorporating both odorant and non-odorant GPCRs, integrating evolutionary insights, and leveraging a large, diverse dataset that includes mutant receptors.

EvOlf provides not only joint embeddings that capture the comprehensive features of ligand-receptor pairs but also individual embeddings for ligands and receptors, creating a robust low-dimensional yet information-rich framework for accurate interaction prediction. The ligand-only embeddings can serve as a valuable resource for exploring the chemical space of small molecules, enabling the clustering, and classification of ligands based on shared structural or functional properties. Such embeddings could be instrumental in virtual screening for drug discovery, allowing researchers to identify ligands with similar binding profiles or pharmacological activities and help elucidate the physicochemical properties that govern ligand-receptor specificity. They also pave the way for the rational design of novel ligands optimized for therapeutic applications. Receptor-only embeddings, on the other hand, offer a powerful tool for investigating GPCR diversity and evolution. By clustering receptors based on their embeddings, researchers can identify subfamilies or evolutionary relationships that are not immediately evident from sequence-based analyses. These embeddings can also be leveraged to identify conserved structural motifs critical for receptor activity. Joint embeddings, which integrate ligand and receptor information, provide the most direct avenue for predicting interactions and inferring the roles of metabolites. They can be used to identify novel ligand-receptor pairs, prioritize targets for experimental validation, and map the molecular basis of ligand-receptor specificity. Additionally, these embeddings could help uncover patterns that link metabolite structures to specific signaling pathways, thereby providing insights into the physiological and pathological roles of metabolites. Together, these embeddings enable hypothesis generation, support exploratory research, and bridge computational predictions with experimental efforts, advancing our understanding of GPCR-mediated processes.

Despite their strengths, both EvOlf and OdoriFy have certain limitations that warrant consideration. While both models excel at predicting interactions, they do not directly account for key biological variables like ligand concentration-dependent effects, receptor allosteric modulation, or temporal dynamics, which are critical for understanding the full physiological impact of GPCR-ligand interactions. Furthermore, the reliance on computational predictions requires rigorous experimental validation to confirm the biological relevance of the inferred interactions. OdoriFy's focus on human-specific olfactory systems limits its utility for comparative evolutionary studies, and its exclusive emphasis on OR-ligand interactions leaves

other chemosensory receptor types unexplored. Although EvOlf expands upon OdoriFy by integrating data across multiple species and including non-olfactory GPCRs, its dataset remains biased toward human-specific interactions, potentially constraining its applicability to non-human systems and broader evolutionary contexts. Additionally, since both models are trained exclusively on terrestrial mammalian data, they are unsuitable for predicting aquatic GPCR-ligand interactions where waterborne odorants have fundamentally different physicochemical properties, solubility characteristics, and detection mechanisms shaped by distinct evolutionary pressures (Mollo et al. 2014). These limitations highlight the need for ongoing refinement and expansion of both tools to enhance their utility and scope.

In conclusion, OdoriFy and EvOlf represent complementary milestones in GPCR research. OdoriFy served as a foundational platform for studying OR-ligand interactions, setting the stage for the development of EvOlf. EvOlf extends this work by offering a unified, evolutionary-driven framework for studying GPCR-ligand interactions, with an emphasis on uncovering the diverse roles of metabolites in physiological and pathological contexts. Together, these models lay the groundwork for future advancements in systems biology, drug discovery, and metabolite research, marking a significant step toward unraveling the complex biochemical networks governed by GPCRs advancing our understanding of metabolite functions and their implications in health and disease.

Chapter 4 - Dissecting the Role of Endogenous Metabolites as Key Allosteric Regulators of GPCR Signaling in Yeast

4.1 Introduction

G protein-coupled receptors (GPCRs), the largest family of membrane proteins in eukaryotes, play a crucial role in cellular signaling. They are characterized by their seven transmembrane α -helical structure and their ability to transduce extracellular signals into intracellular responses. These signals may vary from physical stimuli like photons to chemical signals such as ions, lipids, neurotransmitters, hormones, and peptides (M. Zhang et al. 2024). Their dysregulation is associated with a wide range of diseases, and their prevalence and involvement with virtually every physiological process in the body have made them prime targets in drug development for therapeutic interventions (M. Zhang et al. 2024; D. Yang et al. 2021).

Traditionally, therapeutic approaches targeting GPCR activity have focused on the orthosteric or ligand binding sites, which are typically located in a pocket formed by the extracellular-facing regions of the transmembrane helices. This site is where endogenous ligands, such as hormones, neurotransmitters, lipids, peptides, and metabolites, as well as exogenous compounds like drugs and small molecules, bind to initiate or inhibit receptor activation. For instance, in the β_2 adrenergic receptor (β_2 AR), carazolol, an inverse agonist, binds in the orthosteric pocket (Bokoch et al. 2010). However, due to the high degree of sequence conservation among the orthosteric sites, it is often challenging to selectively target specific sub-types of GPCRs, leading to reduced specificity and an increased likelihood of off-target effects (Goupil, Laporte, and Hébert 2012). To address this limitation, the focus is slowly shifting to alternative strategies, such as targeting allosteric sites.

Allosteric sites are alternative binding sites on proteins, spatially distinct from the orthosteric or active site. Allosteric regulators may lead to alterations (positive, negative, or no effect) in the binding of orthosteric ligands to the receptor, thereby affecting subsequent downstream signaling (Christopoulos 2002). Significant advancements have been made in recent years in the identification and development of allosteric modulators targeting various GPCRs. Moreover, allosteric sites exhibit a high degree of sequence divergence across receptor subtypes, contributing to the high selectivity and reduced side effect profile of allosteric modulators (Christopoulos 2002). The positive allosteric modulator cinacalcet, targeting the calcium-sensing receptor (CaSR), represents the first GPCR allosteric modulator to receive regulatory approval. It is used to treat secondary hyperparathyroidism in end-stage renal disease (Block et al. 2004), albeit with some reported adverse effects (Garrick 2010).

Despite extensive research on synthetic exogenous compounds with allosteric properties, there is a notable lack of detailed knowledge concerning intracellular allosteric small molecules, their modes of regulation, and their functional significance in cellular physiology (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; Thal et al. 2018). These intracellular allosteric modulators range from simple ions to lipids to large protein complexes. For example, cholesterol and other sterols play a significant role as intracellular allosteric modulators of GPCRs (Jafurulla et al. 2019). Cholesterol has been shown to increase the thermal stability of rhodopsin and the β 2-adrenergic receptor (β 2AR) (Hanson et al. 2008), as well as enhance the binding affinities of some agonists at other GPCRs (Gimpl 2016). Crystal structures of the β 2AR have revealed three cholesterol binding sites (Hanson et al. 2008), with molecular dynamics simulations further indicating that cholesterol acts as an allosteric modulator by constraining the receptor's conformational flexibility (Manna et al. 2016).

The yeast mating response is orchestrated by a complex signaling network, with the G protein-coupled receptor Ste2p at its core, playing a crucial role in detecting and responding to mating pheromones. When the α -factor pheromone binds to Ste2p, it triggers a cascade of molecular events starting the activation of heterotrimeric G protein (Gpa1p), leading to the dissociation of $G\beta\gamma$ from $G\alpha$. This triggers a mitogen-activated protein kinase (MAPK) cascade, which culminates in the activation of transcription factors and the expression of mating-specific genes (Nakayama et al. 1988; Alvaro and Thorner 2016; Blumer, Reneke, and Thorner 1988).

The exposure of haploid yeast cells to low doses of mating pheromones, either due to unsuccessful mating or absence of a suitable partner, has been associated with the induction of programmed cell death (PCD) (Büttner et al. 2006; Severin and Hyman 2002; N.-N. Zhang et al. 2006; Carmona-Gutierrez et al. 2010). Mechanistically, pheromone-induced PCD is mediated through the Ste2/Ste3-dependent MAP kinase signaling pathway, which increases intracellular Ca^{2+} levels and mitochondrial activity, ultimately leading to the release of cytochrome c (Büttner et al. 2006; Severin and Hyman 2002; N.-N. Zhang et al. 2006; Carmona-Gutierrez et al. 2010). A successful mating event is marked by the absence of mating factor-induced PCD, implying that yeast may use this mechanism to remove infertile or compromised cells. This appears to favor the survival of diploid cells that perform meiotic recombination to boost genetic diversity over haploid cells.

It is well documented that exposure to elevated, non-physiological concentrations of α -factor results in PCD in a subset of the MATa cell population (~30 %) (N.-N. Zhang et al. 2006; Severin and Hyman 2002). Additionally, equivalent levels of α -factor result in varying PCD kinetics across different

laboratory strains; for instance, BY4741 shows greater resistance relative to W303 (N.-N. Zhang et al. 2006). Genetic screening studies have identified Yeast Suicidal Protein 1 (LAM1) and Yeast Suicidal Protein 2 (LAM2) as positive regulators of α -factor-induced PCD (Sokolov et al. 2020). While the exact mechanism driving these varying levels of cellular sensitivity and resistance to pheromone-induced PCD remains unclear, the potential existence of an unidentified innate rescue mechanism that confers resistance to MATa cells upon exposure to α -factor presents a promising area for further investigation.

Building upon these insights, this study aimed to explore the role of endogenous intracellular metabolites as potential modulators of Ste2p signaling, particularly in the context of α -factor-induced programmed cell death (PCD). Given the growing evidence that intracellular metabolites can act as allosteric regulators of GPCRs, we hypothesized that a subset of pheromone-resistant yeast cells might mitigate Ste2p-mediated PCD through metabolite-mediated modulation at the Ste2p-Gpa1p interface. To investigate this, we employed an AI-driven computational tool, Gcoupler to identify candidate metabolites predicted to interact with Ste2p, and assessed their functional relevance through a series of experimental approaches. These include growth kinetics, viability assessments, mating efficiency measurements, MAPK activation analysis, and reporter gene expression studies. Furthermore, site-directed mutagenesis targeting key residues identified as potential metabolite interaction sites allowed us to probe the molecular mechanism underlying this regulation. This study provides novel insights into the role of intracellular metabolites in GPCR signaling, shedding light on a previously unrecognized layer of regulatory complexity in the yeast mating pathway.

4.2 Materials and Methods

4.2.1 Yeast Strains

BY4741 (*MATa his3 Δ 1 leu2 Δ 0 met15 Δ 0 ura3 Δ 0*) and *ste2 Δ* (*MATa ste2 Δ ::kanMX his3 Δ 1 leu2 Δ 0 ura3 Δ 0 met15 Δ 0*) strains of *Saccharomyces cerevisiae* were used in all the experiments. For the mating assay, BY4742 strain (*MATa his3 Δ 1 leu2 Δ 0 lys2 Δ 0 ura3 Δ 0*) was also used. 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). For the preparation of solid media, 1.5% agar was added to the YPD medium. Both the site-directed yeast mutants (T155D, L289K, and S75A) as well as reconstituted *STE2* transgenes (rtWT) were grown in yeast extract-peptone-galactose-raffinose media (YPGR; 1% yeast extract, 2% peptone, 2% galactose, and 2% raffinose). The following composition of the SC agar plates was used for mating assay 2% glucose, 2% agar, 20 mg/L histidine, 120 mg/L leucine, 20 mg/L uracil, and 20 mg/L adenine, with 60 mg/L lysine or 20 mg/L methionine.

4.2.2 Pre-loading of Yeast Cells with Metabolite

Yeast cells were grown in primary culture, followed by secondary culture, in YPD medium at 30 °C at 200 rpm for 16 hours each. Ensuring an equal number of cells for both strains, 5 µL of secondary culture was inoculated into a 96-well plate containing 145 µL of YPD supplemented with respective metabolites. The tested metabolites included Coenzyme Q6 (CoQ6; 900150O, Avanti® Polar Lipids, Sigma-Aldrich), Zymosterol (ZST; 700068P, Avanti® Polar Lipids, Sigma-Aldrich), and Lanosterol (LST; L5768, Sigma-Aldrich), each tested at concentrations of 0.1 µM, 1 µM, and 10 µM. The cultures were then incubated at 30 °C with shaking at 200 rpm for 24 hours with multiple biological replicates. An equal volume of ethanol was added to the control group to negate the effects of the solvent.

4.2.3 Cell Viability Assay using Growth Kinetics

Wildtype and *ste2Δ* yeast strains were grown and pre-loaded with ZST, LST, and CoQ6 at concentrations of 1 µM, 1 µM, and 10 µM, respectively with multiple biological replicates. After pre-loading incubation, 5 µL of the pre-loaded cultures were inoculated into a 96-well plate containing 145 µL of YPD supplemented with respective metabolites (CoQ6, ZST, and LST) at the same concentrations. For the treated group, 10 µM α -factor (T6901, Sigma-Aldrich) was included, while DMSO (28580, SRL) was added in the untreated group as solvent control. The plates were sealed with breathe-easy membranes to prevent evaporation. The growth kinetics was performed in a Biotek Synergy HTX multi-mode reader for 16.5 hours at 30 °C with continuous orbital shaking at 559 CPM. Absorbance was measured at 600 nm every 30 minutes. The area under the curve (AUC) was calculated for each sample using the bayestestR package in R with default parameters. The AUC values were normalized to the optical density (OD600) of the cultures at time t=0.

4.2.4 Cell Viability Assay Using Propidium Iodide

Wildtype and *ste2Δ* yeast strains were grown and pre-loaded with ZST, LST, and CoQ6 at concentrations of 1 µM, 1 µM, and 10 µM, respectively. Following 24 hours of pre-loading, the α -factor was added to half the wells 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, 50 µL aliquots were taken from each well and mixed with 50 µL of staining solution (Propidium Iodide (PI) 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. An additional 50 μ L aliquot from the original plate was used to measure OD₆₀₀ following a 1:2 dilution with 1X PBS.

The fluorescence values were first normalized for data analysis with the blank normalized OD₆₀₀ values of the respective well. Cellular background fluorescence values of the unstained cells were subtracted from these normalized fluorescence values. The fold change between untreated solvent control and α -factor-treated samples was computed by dividing the fluorescence values of the treated samples by those of the corresponding controls. Mann Whitney U-test was used to measure the statistical significance.

For site-directed *STE2* mutants, the assay was performed similarly with specific modifications. The site-directed *STE2* mutants were pre-loaded with metabolite in yeast extract-peptone-galactose-raffinose (YPGR) medium instead of YPD.

4.2.5 Cell Viability Assay Using FUNTM1 Staining

Wildtype and *ste2 Δ* yeast strains were grown and pre-loaded with ZST, LST, and CoQ6 at concentrations of 1 μ M, 1 μ M, and 10 μ M, respectively. Following this 24-hour pre-loading, α -factor was added to a third of the wells at a final concentration of 30 μ M, acetic acid (85801, SRL) to another third at a final concentration of 199 mM, and DMSO to the remaining third to as a solvent control. After incubation, the cells were pelleted by centrifugation at 6000 rpm for 5 minutes at room temperature (RT) and washed with 1X PBS to remove residual media. The staining solution was freshly prepared in 1X GH solution (10 mM HEPES buffer, pH 7.2, 2% glucose) with FUNTM 1 Cell Stain (F7030, Invitrogen) at a final concentration of 5 μ M. The samples were incubated in the dark for 30 minutes, centrifuged at 6000 rpm for 5 minutes, and the pellet resuspended in 25 μ L of 1X GH solution. For microscopy, 5 μ L of the stained cell suspension was analyzed using a Nikon Eclipse Ci-L Fluorescence Microscope equipped with a Nikon DS-Qi2 camera. FITC and TRITC filters were used to capture green and red fluorescence, respectively, with an exposure time of 1 second each. The fluorescence images were analyzed using ImageJ software, and the number of cells displaying red and green fluorescence, indicative of cellular viability and vitality, were counted and normalized to the total cell count for each sample.

The assay for site-directed *STE2* mutants was conducted under identical conditions, with the exception that the yeast cells were grown in yeast extract-peptone-galactose-raffinose (YPGR) medium instead of YPD to induce *STE2* expression.

4.2.6 Mating Assay

BY4741 (*MATa*) yeast cells were grown and pre-loaded with ZST, LST and CoQ6 at final concentrations of 1 μ M, 1 μ M, and 10 μ M, respectively with 3 biological replicates each. After 24 hours of pre-loading, the cells were harvested and washed with fresh YPD at 6000 rpm for 5 minutes at RT to remove residual metabolites. The pellet was resuspended in fresh YPD, and cell density was equalized across all samples. For the mating assay, 100 μ L of the pre-treated BY4741 (*MATa*) cells were mixed with 100 μ L of untreated BY4742 cells (*MAT α*) in 96-well plates. The cells were thoroughly mixed, sealed with breathable membranes, and incubated at 30 °C with shaking at 200 rpm for 4 hours to allow mating. The *ste2 Δ* strain was used as a negative control. After mating, the cells were harvested and washed with distilled water at 6000 rpm for 5 minutes at RT to remove all media traces. Following this, the cells were resuspended in 200 μ L of distilled water, serial dilutions of 1:10 were prepared in distilled water, and 5 μ L culture was spotted on the SC-Met (His, Leu, Ade, Ura, Lys), SC-Lys (His, Leu, Ade, Ura, Met), and SC-Met-Lys (His, Leu, Ade, Ura) agar plates each in 3 technical replicates. The plates were incubated at 30 °C for two days.

After incubation, the colonies were counted in ImageJ in the 10⁴ dilutions, which had the best resolution of colonies. Mating efficiency was calculated using the following formula:

$$\text{Mating Efficiency} = \frac{\text{colonies in DD} \times \text{Dilution Factor}}{\text{colonies in SD} \times \text{Dilution Factor}} \times 100$$

Here, SD refers to single dropout media (SC-Met for *MAT α* and SC-Lys for *MATa*), and DD refers to double dropout media (SC-Met-Lys).

4.2.7 Phospho-MAPK Activity-Based Western Blot

The wild type (BY4741) strain was grown in primary culture, followed by secondary culture, in YPD medium at 30 °C at 200 rpm for 16 hours each. The secondary culture (30 μ L) was inoculated into conical centrifuge falcon tubes containing 1.5 mL of YPD with respective metabolites at concentrations of 1 μ M (ZST and LST) and 10 μ M (CoQ6) and grown for 24 hours at 30 °C at 200 rpm with 6 biological replicates, with ethanol as solvent control. After incubation, the cells were treated with 3 μ M of α -factor for 5 minutes with DMSO as solvent control. Post incubation, the cells were harvested at 6000 rpm for 5 minutes at RT. The pellet was resuspended in 400 μ L of chilled Lysis buffer (9.5 mL 1X PBS, 200 μ L of Protease Inhibitor Cocktail, 200 μ L of PMSF (20 mg/mL), 100 μ L of 10% Triton-X100) with an equal volume of glass beads. The cells were homogenized for five cycles of 30 seconds each at 3500 rpm, followed by 1 minute incubation in ice after each cycle. The beads were allowed to settle down with a

short spin, following which the lysate was transferred to a fresh vial. The lysate was centrifuged at 2000 rpm for 2 min at 4°C, and the supernatants were transferred to fresh vials, and 5 µl were used in a BCA Protein Assay (Thermo Scientific #23225) carried out according to the manufacturer's instructions. Absorbance values were compared against bovine serum albumin (BSA) standards. Protein concentration in the lysates was normalized to 470 µg/µL with lysis buffer, mixed at a 1:4 ratio with 4X sample loading buffer (2 mL 1M Tris pH 6.8, 0.8 g SDS, 4 mL 100% glycerol, 0.4 mL β-mercaptoethanol, 1 mL 0.5 M EDTA, 8 mg bromophenol blue), and heated at 99 °C for 10 minutes.

Ten milligrams of protein sample were loaded onto 12% SDS-PAGE gels and run in SDS electrophoresis buffer at 40 mA. The resolving gel was then transferred to methanol-activated PVDF membranes at 300 mA for 120 min in transfer buffer (3 g Tris, 14.4 g Glycine, 200 methanol, 800 mL water) at 4°C. Membranes were blocked in either 5% (w/v) BSA (for phospho-Fus3) or 5% (w/v) skim milk in 1X TBST for 2 hours at room temperature and washed three times for 5 minutes each with 1X TBST. The membranes were probed with antibodies to phospho-p44/42 (Cell Signalling #4370, 1:2000), Fus3 (Cell Signalling #9102, 1:1000), or 3-Phosphoglycerate Kinase (Pgk1; Invitrogen #459250, 1:5000) as a loading control in 5% BSA for 2 hours at room temperature with shaking. Blots were washed three times for 5 minutes each with 1X TBST and then incubated with horseradish peroxidase-conjugated secondary antibodies (goat anti-rabbit, Cell Signalling #7074, 1:5000; horse anti-mouse, Cell Signalling #7076, 1:5000) in 1X TBST for 2 hours at room temperature. Blots were washed three times with 1X TBST and visualized using Immobilon Forte Western HRP Substrate (Millipore #WBLUF0500) on a UVTech imaging system.

To assess multiple targets, the phospho-Fus3 antibody was removed by treatment with mild Stripping Buffer (15 g Glycine, 1 g SDS, 10 mL Tween20, pH 2.2, make up to 1 L with water) for 10 minutes at RT with shaking, discarding the buffer and reincubating for 10 min in fresh stripping buffer, then rinsed thoroughly with 1X TBST five times for 5 minutes each before re-blocking and probing with Fus3 antibody. Blots were stripped once again and re-probed for Pgk1 as a loading control.

ImageJ was used to perform the densitometric analysis of the blots. The raw densitometric values obtained for p-Fus3 and Fus3 were normalized with the loading controls, and the p-Fus3/Fus3 ratio was calculated for each condition. For the mutant samples, this ratio for the α-factor stimulated samples was further normalized with their corresponding unstimulated solvent controls.

4.2.8 Transgenic Reporter Assay

The pRS306- P_{FUS1} -eGFP-CYC1 plasmid was constructed by subcloning the eGFP gene under the control of the yeast *FUS1* promoter into plasmid pRS306 using Gibson assembly. The resulting plasmid was integrated into the *URA3* locus of BY4741 strain, generating a strain with the genotype *MATa his3Δ1 leu2Δ0 met15Δ0 ura3::pRS306-P_{FUS1}-eGFP-CYC1*.

Primer Name	Sequence (5'-3')
pFUS1 Forward	GTAAAACGACGGCCAGTGAGCTCAATCCTTCAATTTTCTGGCAACTTTTCTC
pFUS1 Reverse	TACTCATGGATCCTTTGATTTTCAGAACTTGATGGCTTATATC
eGFP Forward	CCATCAAGTTTCTGAAAATCAAAGGATCCATGAGTAAGGGCGAGGAGCTGTTAC CG
eGFP Reverse	GCGTGACATAACTAATTACATGACTCGAGTTACTTGTACAGCTCGTCCATGCCG
pRS306 Forward	CTCGAGTCATGTAATTAGTTATGTCACG
pRS306 Reverse	GAGCTCACTGGCCGTCGTTTAC

Table 4.1 The primers used in PCR amplification for pRS306- P_{FUS1} -eGFP-CYC1

The yeast strain containing the integrated pRS306- P_{FUS1} -eGFP-CYC1 construct was cultured in primary and secondary cultures in SC-Ura medium (2% glucose, 200 mg/L histidine, 300 mg/L leucine, 400 mg/L methionine, 200 mg/L tryptophan, and 100 mg/L adenine) at 30 °C at 200 rpm for 16 hours each. Following secondary culture, 5 μ L of cells were inoculated into 96-well plates containing 145 μ L of SC-Ura medium supplemented with the respective metabolites at concentrations of 1 μ M (ZST and LST) and 10 μ M (CoQ6). Ethanol was used as the solvent control. Cultures were incubated at 30 °C with shaking at 200 rpm for 24 hours, with three biological replicates for each condition. After incubation, α -factor (T6901, Sigma-Aldrich) was added at a final concentration of 30 μ M, with DMSO as a solvent control. The cells were co-stained with Propidium iodide at a final concentration of 5 μ g/mL with 15 minutes of incubation. The plates were incubated in the dark for a total of 1.5 hours. Post incubation, 10 μ L of the sample was used for fluorescence microscopy (Nikon A1R Confocal/TIRF/FCS/FLIM system). GFP (488 nm) and mCherry (561 nm) filters were used for measuring green and red fluorescence, respectively. The images were captured at a scan size of 1024 pixels, scan speed of ¼ frame/sec, and line average at a count of 4.

The data analysis was performed using ImageJ. Dead cells were excluded based on mCherry fluorescence intensity values exceeding 2500 intensity units, and cells with an area below 3 pixels were removed to reduce noise in the data. For each experimental condition, 500 cells were randomly selected for the calculation of Corrected Total Cell Fluorescence (CTCF) using the following formula:

$$CTCF = IntDen - (Area\ of\ selected\ cell * MF)$$

Here, IntDen represents the Integrated Density of the selected cell, and MF is the mean fluorescence of background readings.

4.2.9 Site-Directed Mutagenesis

The gene encoding wild-type *STE2* was PCR amplified from the genome of *Saccharomyces cerevisiae* and further cloned into plasmid pRS304 under galactose-inducible *GALI* promoter and CYC1 terminator to generate plasmid pRS304-P_{GALI}-STE2-CYC1 using Gibson assembly (Gibson et al. 2010, 2009). The mutants were generated by PCR amplifying the gene with primers consisting of respective mutations and cloned into plasmid pRS304 under *GALI* promoter to generate pRS304-P_{GALI}-STE2 S75A-CYC1, pRS304-P_{GALI}-STE2 T155D-CYC1, and pRS304-P_{GALI}-STE2 L289K-CYC1. Wild-type (rtWT) and mutant *STE2* were integrated by digesting the pRS304 vector with restriction enzyme BstXI to generate linearized plasmid and transformed into *Saccharomyces cerevisiae* BY4741 *ste2Δ* strain. The strains were confirmed for integration by isolating the genome and PCR amplification using locus-specific primers. The following mutants were generated in this work:

- (a) *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ste2Δ::kanMX TRP1::P_{GALI}-STE2-CYC1*
- (b) *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ste2Δ::kanMX TRP1::P_{GALI}-STE2 S75A-CYC1*
- (c) *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ste2Δ::kanMX TRP1::P_{GALI}-STE2 T155D-CYC1*
- (d) *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ste2Δ::kanMX TRP1::P_{GALI}-STE2 L289K-CYC1*

Primer Name	Sequence (5' - 3')
STE2 FWD overlap	CTTTAACGTCAAGGAGGGATCCATGTCTGATGCGGCTCCTTCATTG
SDM1 STE2 FWD	GTCATGTGGATGACATCGAGAGCTAGAAAAACGCCGATT
SDM1 STE2 REV	AAATCGGCGTTTTTCTAGCTCTCGATGTCATCCACATGAC
SDM4 STE2 FWD	TTTCAGATAAAAGTTATTTTCGACGGCGACAACCTTCAAAAGGATA
SDM4 STE2 REV	TATCCTTTTGAAGTTGTCGCCGTCGAAAATAACTTTTATCTGAAA

SDM5 STE2 FWD	ACATTACTTGCTGTATTGTCTAAACCATTATCATCAATGTGGGCC
SDM5 STE2 REV	GGCCCACATTGATGATAATGGTTTAGACAATACAGCAAGTAATGT
STE2 REV overlap	CATAACTAATTACATGACTCGAGTCATAAATTATTATTATCTTCAGTCCAGAACTT TCTG
CYC.F	ATAATAATTTATGACTCGAGTCATGTAATTAGTTATGTCACGCTTAC
GAL1 REV BAMHI	GCCGCATCAGACATGGATCCCTCCTTGACGTTAAAGTATAGAGG

Table 4.2 *The primers used in cloning and site-directed mutagenesis*

4.2.10 Cellular Morphometric Assay (Shmoo formation)

The reconstituted *STE2* wild-type (rtWT) and mutants (T155D, L289K, S75A) were grown in primary culture, followed by secondary culture, in YPGR medium at 30 °C at 200 rpm for 16 hours each. Ensuring an equal number of cells for all, 5 µL of secondary culture was inoculated into a 96-well plate containing 145 µL of YPD in the presence or absence of respective metabolites at desired concentrations (T155D - 1 µM LST, L289K - 1 µM ZST and S75A - 10 µM CoQ6) and grown for 24 hours at 30°C at 200 rpm with 4 biological replicates. An equal volume of ethanol was added to the control group to negate the effects of the solvent. After 24 hours, α -factor (T6901, Sigma-Aldrich) was added at a final concentration of 10 µM, with DMSO as solvent control, and incubated for 1 hour. Post incubation, 5 µL of the sample was used for phase contrast microscopy (Nikon Eclipse Ci-L Fluorescence Microscope with Nikon DS-Qi2 camera). For each sample, the number of cells with shmoos was counted using ImageJ and normalized with the total cell counts.

4.3 Results

4.3.1 Elevated metabolic levels selectively inhibit GPCR signaling

The yeast mating pathway is a well-characterized cellular signaling pathway (Nakayama et al. 1988; Alvaro and Thorner 2016; N.-N. Zhang et al. 2006; Carmona-Gutierrez et al. 2010). To investigate the identity and mode of action of endogenous metabolites that may influence the GPCR-G protein interface, the pheromone-activated G protein-coupled receptor Ste2p was employed as a model system. To this end, we used the tool Gcoupler to predict the metabolites capable of targeting the Ste2p-Gpa1p interface and modulate GPCR signaling (Mohanty et al. 2023). Gcoupler is an AI-driven computational toolkit that

employs an integrative approach combining *de novo* ligand design, statistical methods, Graph Neural Networks, and Bioactivity-based ligand prioritization for predicting high-affinity ligands. Gcoupler predicted zymosterol (ZST), ubiquinone 6 (CoQ6), and lanosterol (LST) as the top metabolites with high binding probabilities to the intracellular Ste2p-Gpa1p interface. To evaluate the functional implications of these metabolites, we exogenously elevated the levels of ZST, CoQ6, and LST in the wild-type strain and assessed their effects on α -factor-induced programmed cell death (PCD) and Ste2p signaling using distinct techniques.

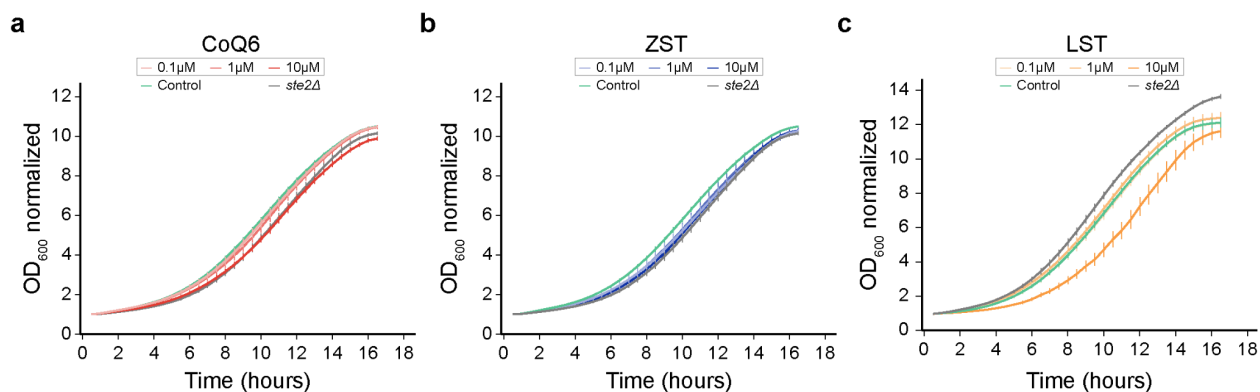


Figure 4.1 Growth Curve Profiles of Metabolite Pre-treated Yeast

Growth curve profiles of the wild-type (control), *ste2Δ*, and wild-type yeast pre-treated with the indicated metabolites.

First and foremost, to assess any potential cytotoxic effects or metabolic disruptions caused by the exogenously elevated metabolite, we monitored the growth profiles of the yeast cells. With the exception of the 10 μ M concentration of LST, no significant alterations in growth profiles were observed across the tested conditions (**Figure 4.1**). The absence of growth disturbances in most cases suggests that the selected concentrations of ZST, CoQ6, and LST are within physiologically relevant ranges and did not induce non-specific effects on cellular processes.

Yeast cells were pre-loaded with the selected metabolites for 24 hours prior to evaluating their effects on α -factor-induced programmed cell death (PCD). Following this, we assessed the cells for changes in PCD induction using three distinct approaches: growth profile analysis, cellular viability measurement via propidium iodide (PI) staining, and cellular viability and vitality assessment using FUNTM1 staining (Teng and Hardwick 2009; Van Zandycke et al. 2003). Growth kinetics-based PCD screening involves the calculation and comparison of the area under the growth profile curve (AUC). In wild-type cells treated with α -factor, we observed a significant reduction in AUC values, consistent with PCD induction. By contrast, no significant changes in AUC values were detected in cells pre-loaded with the metabolites or

in the *ste2Δ* mutant strain (**Figure 4.2a**). This result highlights the potential role of these metabolites in interfering with the signaling cascade initiated by Ste2p, thereby preventing α -factor-induced PCD.

Another experimental workflow used for assessing α -factor-induced programmed cell death (PCD) was based on a propidium iodide (PI)-based cell viability fluorometric assay. In α -factor-treated control cells, a significant increase in the fold change in PI fluorescence was observed, reflecting increased cell death. However, in cells pre-loaded with metabolites, the PI fluorescence values remained comparable to untreated controls, indicating a rescue from α -factor-induced PCD. Notably, the *ste2Δ* mutant, which lacks the pheromone receptor, also showed no significant increase in PI fluorescence, further confirming the Ste2p-dependent nature of the response. Heat-killed cells were used as a positive control (**Figure 4.2b**).

We also employed the FUNTM1 staining assay to assess and compare cell vitality and viability across different conditions using fluorescence microscopy (Van Zandycke et al. 2003). This method enabled the distinction of metabolically active and viable cells, characterized by faint green fluorescence with red cytoplasmic spots. As an additional control, we used acetic acid, a well-established exogenous stimulus for PCD, to compare the specificity of the α -factor-induced PCD response (Martins et al. 2019; Guedes et al. 2022). Quantitative analysis revealed that cells preloaded with ZST, CoQ6, and LST exhibited a significant rescue from α -factor-induced PCD compared to untreated wild-type cells. This observation highlights the protective effect of these metabolites in mitigating cell death under α -factor treatment. Conversely, *ste2Δ* cells, which lack the Ste2p receptor, did not show α -factor-induced PCD. Interestingly, the protective effects of metabolite preloading were observed exclusively under α -factor-induced conditions and were not evident when acetic acid was used to induce PCD (**Figure 4.2c, d**). This indicates that the rescue effect is specific to the α -factor signaling pathway rather than a generalized suppression of PCD.

To assess whether the observed modulation of Ste2p signaling by the metabolites extends beyond exogenous pheromone exposure assays, we performed a mating assay using MAT α and MAT a yeast strains. This assay serves as a natural proxy to evaluate Ste2p signaling during the yeast mating process (Bardwell 2005). MAT a cells pre-loaded with metabolites, including ZST, CoQ6, and LST, were incubated with MAT α cells to measure mating efficiency. The *ste2Δ* strain (MAT a) was used as a negative control (Hartwell 1980). The results indicated a decline in mating efficiency in metabolite-pre-loaded MAT a cells compared to control cells, consistent with the modulation of Ste2p signaling observed in previous pheromone-induced assays. This suggests that the metabolites not only influence exogenous pheromone signaling but also impact the natural mating process, highlighting their regulatory role in

Ste2p-mediated signaling pathways. Interestingly, CoQ6 did not show a significant decline in mating efficiency, indicating a distinct response compared to the other tested metabolites (**Figure 4.3a**). This discrepancy could be attributed to differences in α -factor concentrations, as natural mating conditions involve lower endogenous levels of α -factor compared to the artificially elevated conditions used during metabolite pre-loading.

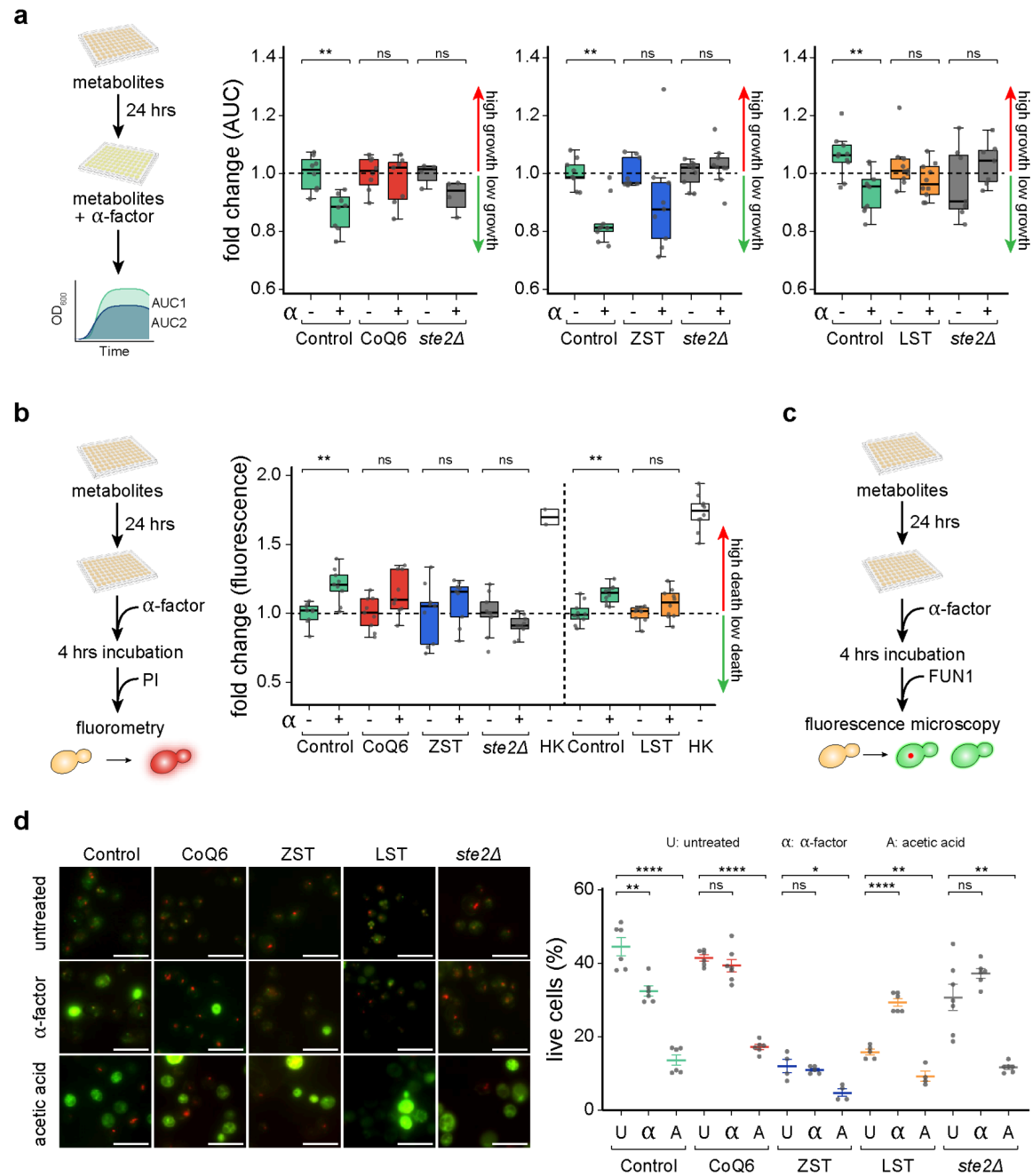


Figure 4.2 Exogenous Metabolites Treatment Resists α -Factor-Induced Programmed Cell Death

(a) Schematic diagram (left) depicting the workflow opted to monitor α -factor-induced programmed cell death (PCD) in the wild-type yeast cells pre-loaded for 24 hours with ubiquinone 6 (CoQ6), zymosterol (ZST), and lanosterol (LST) at three different concentrations. Changes in the PCD were quantified using the area under the curve (AUC) of the growth kinetics with respect to their respective controls. The box plot on the right depicts the rescue from the α -factor-induced PCD in the indicated conditions ($n=8$ or 9 biological replicates). The y-axis represents the fold change of the AUC values with respect to their respective controls. Mann Whitney U test was used to calculate statistical significance. **(b)** The schematic diagram depicting the experimental workflow used to quantify α -factor-induced programmed cell death using a propidium iodide (PI)-based cell viability fluorometric assay. Box plot depicting the rescue from the α -factor-induced PCD in the indicated conditions as inferred using PI-based cell viability fluorometric assay ($n = 9$ or 10 biological replicates; Heatkilled = 2). The y-axis represents the fold change of the PI fluorescence values with respect to their respective controls. Mann Whitney U test was used to calculate statistical significance. **(c)** Schematic representation of the experimental approach used to measure cell vitality and viability using microscopy-based FUNTM1 staining. **(d)** Representative micrographs (left) depicting the FUNTM1 staining results in the indicated conditions, Scale 10 μ m. Mean-whisker plot (right) depicting the relative proportion of the vital and viable yeast cells observed using FUNTM1 staining in the indicated conditions ($n = 3$ biological replicates). A Student's *t*-test was used to compute statistical significance. Error bars represent the SEM. Asterisks indicate statistical significance, whereas ns represents non-significance.

To further investigate the effects of metabolite preloading on pheromone signaling, we evaluated the mitogen-activated protein kinase (MAPK) pathway through phospho-Fus3 (p-Fus3) detection using Western blotting (Shellhammer et al. 2019). Fus3 is a key kinase in the MAPK cascade that is activated during α -factor-induced signaling, and its phosphorylation status serves as a reliable readout for pathway activation. Following α -factor treatment, we observed a significant increase in phosphorylated Fus3 levels (pFus3/Fus3) in untreated control cells, consistent with robust activation of the MAPK pathway. However, cells pre-treated with ZST and LST exhibited a statistically significant reduction in phosphorylated Fus3 levels compared to control cells exposed to α -factor. CoQ6 did not yield reliable results (data not shown). Total Fus3 levels remained constant across all conditions, confirming that the observed effects were specific to changes in phosphorylation and not due to differences in protein expression. Pgk1 was used as a loading control to ensure equal protein loading across samples (**Figure 4.3b**). These findings suggest that metabolite preloading can modulate the intensity of α -factor-induced MAPK activation, further supporting their regulatory role in Ste2p-mediated signaling pathways.

To further validate the observed rescue effects, we utilized a transgenic approach by expressing Enhanced Green Fluorescent Protein (eGFP) under the control of the Fus1 promoter (P_{FUS1}) and the CYC1 terminator. This reporter system is a well-established method for quantifying α -factor-mediated Ste2p

signaling. In this assay, eGFP expression serves as an indicator of Fus1 promoter activity, which is directly linked to α -factor-induced signaling via Ste2p (Shellhammer et al. 2019). Quantitative analysis of the eGFP expression revealed robust eGFP expression in wild-type cells treated with α -factor, indicating active Ste2p signaling. In contrast, cells pre-incubated with ZST, CoQ6, or LST showed no significant increase in eGFP expression compared to untreated controls, regardless of α -factor treatment. These results suggest that the tested metabolites effectively inhibited the activation of the Fus1 promoter, thereby suppressing α -factor-induced Ste2p signaling (**Figure 4.3c**). These results provide strong evidence that the tested metabolites can suppress α -factor-induced Ste2p signaling, likely through modulation of the Ste2p-Gpa1p interface. This inhibition of signaling is consistent across different experimental approaches. The consistency across different experimental approaches reinforces the conclusion that the tested metabolites can interfere with pheromone-induced Ste2p signaling and PCD, likely by influencing the Ste2p-Gpa1p interaction.

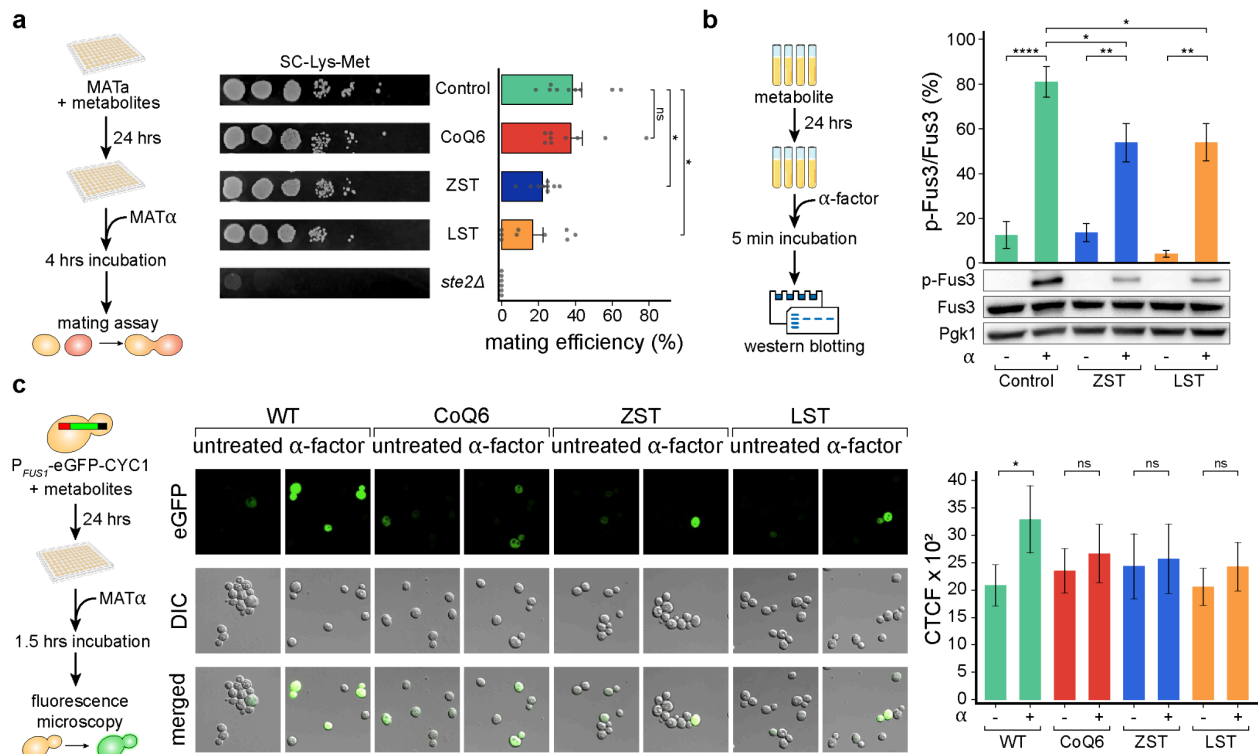


Figure 4.3 Metabolite-Induced Modulation of Yeast Mating and Ste2p Signaling

(a) Schematic representation (left) of the experimental design for the mating assay ($n = 3$ biological replicates, each with three technical replicates). MATa yeast cells were pre-loaded with the metabolites and then mated with MATα cells to evaluate the mating efficiency. Representative micrographs in the middle qualitatively depict the mating efficiency in the indicated conditions. The bar plots on the right depict the mating efficiency (mean $\pm$ SEM) in the

indicated conditions. Student's *t*-test was used to compute statistical significance. **(b)** Schematic representation (left) depicting the experimental design of phospho-MAPK activity-based Western blot. Barplots (right) depicting the *p*-Fus3 levels (mean $\pm$ SEM) in the indicated conditions. Error bars represent SEM. A Student's *t*-test was used to compute statistical significance. Asterisks indicate statistical significance, whereas ns represents non-significance. **(c)** Schematic representation (left) of the experimental approach used to measure the fluorescence in *P_{FUS1}-eGFP-CYC1* yeast cells. Representative micrographs (middle) depicting the eGFP expression in the yeast cells in the indicated conditions. Scale 20 μ m. Barplot (right) depicting the Corrected Total Cell Fluorescence (CTCF) value (mean $\pm$ SEM) in the indicated conditions. A Student's *t*-test was used to compute statistical significance. Asterisks indicate statistical significance, whereas ns represents non-significance.

4.3.2 Site-Directed Mutants Abrogate Metabolite-Mediated Rescue Phenotype

To investigate the molecular mechanism underpinning metabolite-mediated regulation of Ste2p signaling, site-directed mutagenesis was used to generate Ste2p mutants based on computational predictions. Gcoupler identified critical residues at the Ste2p-Gpa1p interface likely to interact with the selected metabolites. To experimentally confirm the role of these specific mutations in mediating the rescue response against α -factor-induced PCD *in vivo*, we generated missense mutants through site-directed mutagenesis. The targeted mutations included S75A, L289K, and T155D, designed to alter residues identified as critical for metabolite binding.

We analyzed the growth profiles of the reconstituted wild-type (rtWT) and the missense mutants T155D, L289K, and S75A over a 16-hour period to assess whether the introduced mutations influenced cellular growth under standard conditions. The results indicate that all strains, including rtWT and the three mutants, exhibited similar growth kinetics (**Figure 4.4a**). This suggests that the introduced mutations did not impact the overall fitness or viability of the cells under standard laboratory growth conditions.

To investigate whether the site-directed missense mutants exhibit a rescue response to α -factor-induced PCD, we performed a fluorometry-based propidium iodide (PI) assay. The reconstituted wild-type *STE2* transgene (rtWT) in a *Δste2* background served as a control. Upon α -factor exposure, the rtWT cells displayed significant cell death, consistent with wild-type behavior. Similarly, the mutants T155D and L289K demonstrated a notable increase in PI fluorescence, indicating higher cell death levels comparable to α -factor-treated rtWT cells. The role of metabolites in modulating Ste2p signaling was assessed by pre-loading the mutants with LST, ZST, or CoQ6 prior to α -factor treatment. In the rtWT cells, metabolite pre-treatment conferred a rescue response characterized by reduced PI fluorescence and decreased cell death, consistent with previous observations. However, in the T155D and L289K mutants, this rescue phenotype was completely absent, even with metabolite preloading. This suggests that the T155D and

L289K mutations disrupt critical residues at the intracellular Ste2p-Gpa1p interface necessary for metabolite-mediated modulation of signaling. Interestingly, the S75A mutant also failed to exhibit a rescue response upon metabolite pre-treatment (**Figure 4.2b**). Unlike the T155D and L289K mutants, this lack of response may be attributed to substantial structural alterations in Ste2p caused by the S75A mutation, which could interfere with both α -factor-induced signaling and metabolite-mediated effects.

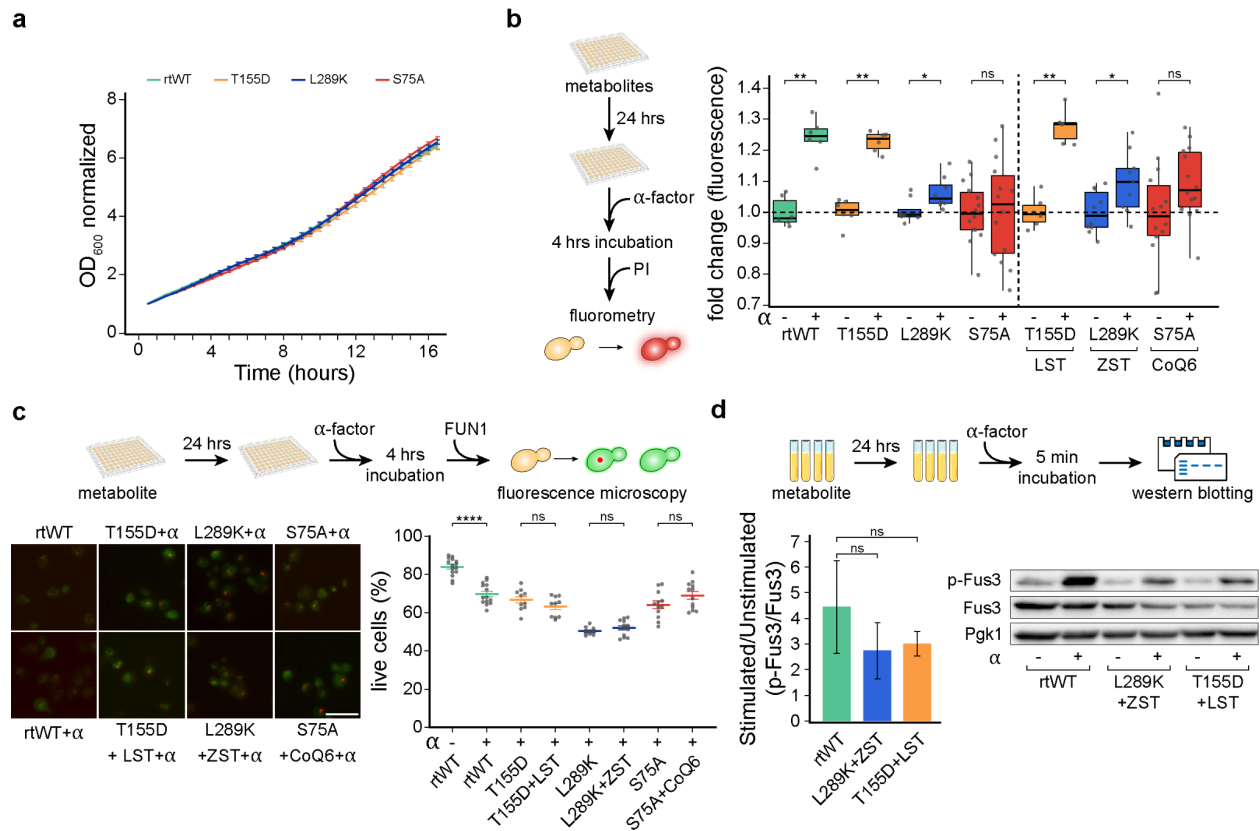


Figure 4.4 Impact of Site-Directed Mutations and Metabolites on Ste2p Signaling and α -Factor-Induced PCD

(a) Growth curve profiles of the reconstituted wild-type (rtWT) and indicated missense site-directed mutants. **(b)** The schematic diagram depicting the experimental workflow used to quantify α -factor-induced PCD in generated site-directed missense mutants (T155D, L289K, S75A) alongside reconstituted wild-type STE2 (rtWT), using a PI-based cell viability fluorometric assay. The box plot on the right depicting the increase in the relative proportion of dead cells upon α -factor exposure (left) and the loss of rescue phenotype from the α -factor-induced PCD (right) in the indicated conditions as inferred using PI-based cell viability fluorometric assay. The y-axis represents the fold change of the PI fluorescence values with respect to their respective controls. Mann Whitney U test was used to calculate statistical significance. **(c)** Schematic representation (top) of the experimental approach used to measure cell vitality and viability using microscopy-based FUNTM1 staining. Representative micrographs (below) depicting the FUNTM1 staining results in the indicated conditions, Scale 10 μ m. Mean-whisker plot depicting the relative proportion of the vital and viable yeast cells observed using FUNTM1 staining in the indicated conditions ($n = 4$ biological replicates). Error bars represent the SEM. A Student's t -test was used to compute statistical significance.

(d) Schematic representation (up) depicting the experimental design of phospho-MAPK activity-based Western blot. Barplots (down) depicting the p-Fus3 levels (mean $\pm$ SEM) in the indicated conditions. The y-axis represents the p-Fus3/Fus3 ratio for the stimulated condition normalized by its corresponding unstimulated sample. Error bars represent SEM. A Student's *t*-test was used to compute statistical significance. Asterisks indicate statistical significance, whereas ns represents non-significance.

To further evaluate the impact of Ste2p signaling dysregulation upon α -factor exposure, we employed an alternative approach using FUNTM1 staining to assess cell viability and vitality. The α -factor treatment significantly reduced the proportion of viable cells in rtWT. In contrast, the mutants T155D, L289K, and S75A failed to exhibit a rescue response upon α -factor exposure, as evidenced by the lack of significant changes in the relative proportions of viable cells upon metabolite pre-treatment (**Figure 4.4c**).

We further evaluated the impact of point mutations on α -factor-mediated Ste2p signaling by analyzing the phosphorylation levels of Fus3 (p-Fus3). The p-Fus3 levels were assessed using Western blotting, and the results were represented as the p-Fus3/Fus3 ratio normalized to the unstimulated condition for each sample. In the reconstituted wild-type *STE2* transgene (rtWT) background, α -factor treatment induced a significant increase in the levels of phosphorylated Fus3, indicative of robust MAPK activation. However, the T155D and L289K mutants pre-treated with ZST or LST exhibited a marked absence of rescue response. Upon α -factor exposure, the p-Fus3 levels in these mutants were significantly elevated (**Figure 4.4d**). This suggests that the mutations may disrupt the ability of the metabolites to modulate Ste2p-mediated signaling, resulting in uninhibited MAPK activation. Notably, the S75A mutant exhibited variability in the p-Fus3 levels, lacking a consistent pattern in response to metabolite pre-treatment or α -factor exposure (data not shown). These findings suggest that the introduced mutations, particularly T155D and L289K, interfere with critical interactions at the Ste2p-Gpa1p interface, thereby preventing metabolite-mediated modulation of Ste2p signaling. The abrogation of the rescue phenotype highlights the importance of these residues in facilitating the metabolite-induced inhibition of downstream MAPK signaling. These results provide further evidence of the mechanistic role of metabolites in modulating GPCR-mediated signal transduction.

We further assessed the impact of metabolites on α -factor-induced cellular responses through quantitative analysis of shmoo formation, a key morphological feature associated with mating signaling in yeast. As expected, α -factor exposure in rtWT cells resulted in a significant increase in the proportion of cells forming shmoos. However, prior treatment with the tested metabolites, including LST, ZST, and CoQ6, significantly attenuated shmoo formation in rtWT cells (**Figure 4.5a, b**). These results align with previous observations using other physiological and molecular assays, reinforcing the conclusion that metabolites such as LST, ZST, and CoQ6 modulate Ste2p signaling and mitigate α -factor-induced responses.

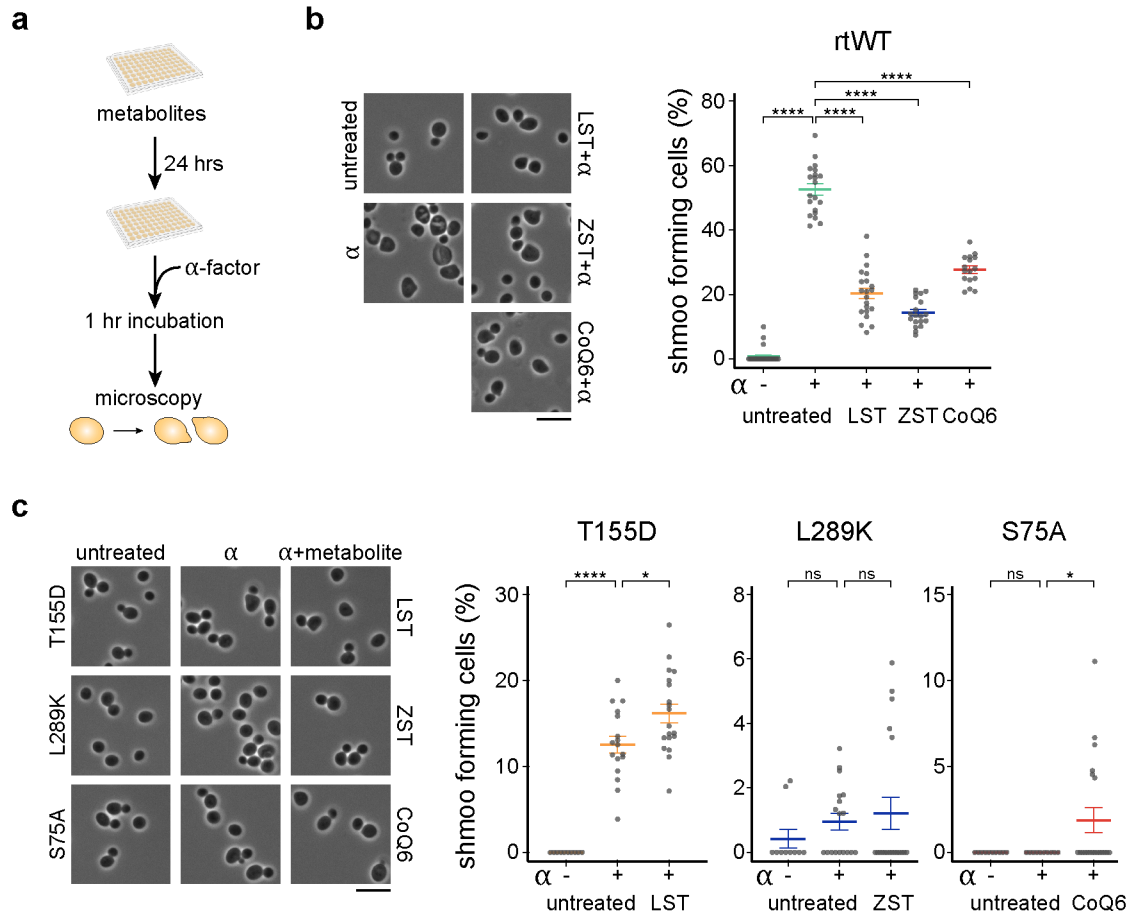


Figure 4.5 Effects of Ste2 Mutants on Shmoo Formation in Response to α -Factor.

(a) Schematic representation depicting the overall experimental workflow designed for the quantitative assessment of the shmoo formation upon α -factor treatment. **(b)** Micrographs depicting the phase contrast images of the rtWT cells, pretreated with the indicated metabolites, with and without α -factor treatment, Scale 10 μ m. Mean-Whisker plot (right) indicating the relative proportions of shmoo forming cells in the indicated conditions. **(c)** Micrographs depicting the phase contrast images of the indicated site-directed missense mutants of the STE2 gene, pretreated with the indicated metabolites, with and without α -factor treatment, Scale 10 μ m. Mean-Whisker plot (right) indicating the relative proportions of shmoo forming cells in the indicated conditions. The Student's *t*-test was used to compute the statistical significance of **b** and **c**. Asterisks indicate statistical significance, whereas ns represents non-significance. Error bars represent the standard error of the mean (SEM).

We evaluated the impact of the introduced Ste2p mutations on shmoo formation in cells pre-treated with their respective metabolites and subsequently exposed to α -factor and observed no modulations in the Ste2p signaling as inferred from the shmoo counts. Interestingly, in the S75A mutant, shmoo formation was almost entirely absent following α -factor treatment, pointing to a severe disruption in Ste2p signaling likely caused by structural perturbations introduced by the mutation (**Figure 4.5c**). These results collectively suggest that the binding of metabolites at the intracellular Ste2p-Gpa1p interface is essential

for mediating the observed rescue response to α -factor-induced programmed cell death. The lack of rescue in the mutants reinforces the importance of the targeted residues in enabling the functional modulation of Ste2p signaling by these metabolites.

4.4 Discussion

The study of G protein-coupled receptor (GPCR) signaling has long focused on its implications in a wide variety of physiological processes and its potential as a therapeutic target. Unlike the human genome, which encodes approximately 800 GPCRs, the yeast genome encodes only two GPCR systems responsible for pheromone and glucose sensing (Horn et al. 2003). Among these, the Ste2p-mediated pheromone-sensing pathway has been extensively studied, uncovering several feedback inhibition systems that primarily operate downstream of the signaling cascade (Alvaro and Thorner 2016). However, the mechanism underlying innate resistance to α -factor-induced programmed cell death (PCD) appears to function much earlier in the pathway, potentially involving a direct, physical interaction-based regulatory mechanism at the receptor level. This upstream modulation contrasts with slower transcriptomic or epigenomic feedback mechanisms and aligns with the reports of intracellular allosteric modulators regulating GPCR activity in other systems (van der Westhuizen et al. 2015; Stornaiuolo et al. 2015; Doller 2017).

The findings in this study provide strong evidence that specific endogenous metabolites, such as zymosterol (ZST), lanosterol (LST), and ubiquinone 6 (CoQ6), can act as allosteric regulators of Ste2p by targeting its intracellular interface with Gpa1p. Multiple experimental approaches, including growth kinetics, cell viability assays using propidium iodide and FUN1 staining, MAPK activation, reporter gene assay, and morphological changes associated with mating like shmoo formation assays, demonstrate the ability of these metabolites to modulate Ste2p signaling and mitigate α -factor-induced PCD. This multi-assay strategy ensures that the observed effects are not limited to a single experimental endpoint but are consistent across diverse readouts, thereby strengthening the validity of the conclusions. The selective inhibition observed in metabolite-preloaded cells suggests that these metabolites may function as endogenous allosteric regulators of Ste2p signaling, providing a novel layer of regulatory complexity within this pathway.

Additionally, site-directed mutagenesis targeting key residues predicted to interact with metabolites allowed for direct investigation of the mechanistic basis of modulation. The targeted mutations (S75A, T155D, and L289K) were designed to disrupt critical residues implicated in metabolite binding, and their analysis revealed an abrogation of the metabolite-mediated rescue phenotype. In particular, the T155D

and L289K mutations completely abolished the ability of ZST and LST to inhibit α -factor-induced PCD, MAPK activation, and shmoo formation, underscoring the functional importance of the targeted residues. These findings align with earlier studies identifying critical residues involved in Ste2p signaling (Velazhahan et al. 2021). Interestingly, the S75A mutation exhibited a more severe phenotype, characterized by near-complete loss of shmoo formation and rescue response, suggesting that this mutation may cause significant structural perturbations to Ste2p, rendering it incapable of responding to both α -factor and metabolite regulation.

Despite the strengths of this study, there are limitations that warrant further investigation to fully elucidate the mechanisms underlying the observed effects. One primary shortcoming is the focus on endpoint assays to infer signaling activity. While these assays are reliable indicators of downstream outcomes, they do not capture the signaling events occurring at intermediate stages of Ste2p signaling. Ste2p signaling involves multiple steps, from receptor activation and G-protein coupling to MAPK activation and transcriptional responses. For a more detailed analysis, techniques like Förster Resonance Energy Transfer (FRET) or Bioluminescence Resonance Energy Transfer (BRET) can be employed to monitor real-time receptor conformational changes and G-protein activation dynamics (Schihada et al. 2018; Salahpour et al. 2012; Picard et al. 2018). Another limitation is the lack of direct evidence demonstrating binding between the tested metabolites and Ste2p. Techniques such as surface plasmon resonance (SPR) (Navratilova, Besnard, and Hopkins 2011), isothermal titration calorimetry (ITC) (Xiangyu Liu et al. 2019), or microscale thermophoresis (MST) (Seidel et al. 2013) can be used to validate these interactions. Structural approaches like cryo-electron microscopy (cryo-EM) or X-ray crystallography would further elucidate the molecular basis of metabolite binding and its impact on receptor conformation. Another unexplored aspect is whether the observed mechanism is conserved across different yeast species or higher eukaryotes. Investigating this mechanism in other yeast species, such as *Candida albicans* or *Schizosaccharomyces pombe*, could provide insights into the evolutionary conservation of this regulatory process. Comparative studies using orthologs of Ste2p from these species would help determine whether the identified residues and metabolites play similar roles in modulating signaling. Additionally, the study relies heavily on exogenous elevation of metabolite levels, which, while informative, may not fully reflect physiological conditions. Future studies could focus on quantifying endogenous metabolite concentrations during different cellular states, such as stress, nutrient deprivation or mating, to assess their physiological relevance. Metabolomic profiling combined with genetic perturbations could provide more insights into the physiological relevance of these metabolites.

Despite these limitations, the study opens new avenues to uncover the untapped potential of intracellular metabolites in regulating complex signaling networks. Intracellular metabolites offer the potential to

regulate GPCR activity with high specificity, targeting distinct allosteric sites that are less conserved across receptor subtypes. This specificity minimizes off-target effects, which are a major concern in orthosteric drug development due to the high sequence conservation of ligand-binding pockets. By targeting allosteric interactions at GPCR interfaces, novel therapeutic strategies could be developed to fine-tune receptor activity with high specificity and minimal side effects. This study provides compelling evidence that intracellular metabolites can modulate GPCR signaling through direct interactions at the receptor-G α interface. Unlike conventional GPCR regulation, which predominantly focuses on extracellular ligands binding to orthosteric sites, the findings demonstrate that endogenous intracellular metabolites can act as allosteric modulators, directly altering receptor activity from within the cell. While limitations such as the lack of direct binding evidence remain, the integrative approach of this study lays a strong foundation for future investigations into GPCR regulation by endogenous metabolites. Exploring these mechanisms further could not only enhance our understanding of GPCR signaling but also open new avenues for therapeutic targeting of allosteric sites in diverse biological systems.

Chapter 5 - Conclusion and Summary

Metabolites, long regarded as mere intermediates in metabolic pathways, play multifaceted roles in cellular function, signaling, and pathophysiology. This research deciphers these roles, unraveling their potential to act as internal threats, signaling modulators, and regulators of cellular processes. By integrating computational frameworks, advanced AI methodologies, and experimental validation, this work provides critical insights into the diverse functional landscape of metabolites and their broader implications for cellular biology and disease.

In Chapter 2, the focus was on the nature of metabolites as potential internal threats. Metabokiller, a computational framework developed for carcinogenicity prediction, revealed how intrinsic biochemical properties of metabolites can transform them into internal threats. This tool identified endogenous metabolites with carcinogenic potential validated experimentally by assessing genotoxicity, mutagenesis, and malignant transformation. The findings underscore that metabolites such as 4-nitrocatechol and 3,4-dihydroxyphenylacetic acid can disrupt genomic stability and drive tumorigenesis. These results not only enhance our understanding of metabolite-mediated carcinogenesis but also demonstrate the utility of integrative computational and experimental approaches for predictive toxicology.

Building on this, Chapter 3 expanded the exploration of metabolites for their roles in GPCR-mediated signaling through OdoriFy and EvOlf, AI-driven platforms for decoding ligand-receptor interactions. OdoriFy was designed to decode odorant-receptor interactions in human olfactory systems, providing a detailed understanding of the molecular basis of olfaction. Beyond olfactory signaling, the multimodal deep-learning framework EvOlf was developed to predict ligand interactions across the GPCR superfamily. The comprehensive datasets and predictive accuracy of these frameworks will facilitate the identification of novel receptor-ligand interactions. This research emphasizes the evolutionary and structural diversity of GPCRs and their centrality in metabolite-mediated regulation, laying the foundation for therapeutic applications in drug discovery and receptor-targeted interventions.

Chapter 4 shifted focus to the regulatory potential of metabolites in GPCR signaling using the *Saccharomyces cerevisiae* Ste2p pheromone receptor as a model system. Metabolites such as zymosterol, lanosterol, and ubiquinone-6 were identified as endogenous allosteric regulators of the Ste2p-Gpa1p interface, suppressing pheromone-induced programmed cell death and altering mating responses. Multi-assay validations confirmed their role in modulating GPCR signaling, with site-directed mutagenesis pinpointing critical residues (S75A, T155D, and L289K) that mediate this regulation. These

findings provide compelling evidence of a novel mechanism by which intracellular metabolites exert precise control over GPCR activity, offering an additional layer of complexity in cellular signaling.

Collectively, these chapters demonstrate the diverse and impactful roles of metabolites in cellular biology. While these findings represent significant advances, they also highlight the vast potential for future research. The physiological relevance of metabolite-mediated signaling under various environmental and cellular contexts remains to be fully explored. Techniques such as metabolomics, cryo-electron microscopy, and *in vivo* imaging could further illuminate the interactions between metabolites and their molecular targets. Additionally, extending the predictive computational frameworks like Metabokiller and EvOlf to include other disease-related metabolites or receptor families would broaden their applicability and uncover new regulatory paradigms and therapeutic opportunities.

By bridging computational and experimental approaches, this work establishes a foundation for understanding the nuanced roles of metabolites as internal threats, signaling regulators, and therapeutic targets. This research not only enhances our understanding of metabolite functionality but also provides tools and methodologies to decode their roles in complex biological systems. The integration of computational predictions with experimental validations ensures that these findings are both robust and actionable, marking a significant step forward in the study of metabolites as dynamic and multifaceted cellular entities. The implications of this work extend beyond fundamental biology, offering translational potential in areas such as precision medicine, cancer biology, and drug discovery.

References

- Abaffy, Tatjana, James R. Bain, Michael J. Muehlbauer, Ivan Spasojevic, Shweta Lodha, Elisa Bruguera, Sara K. O’Neal, So Young Kim, and Hiroaki Matsunami. 2018. “A Testosterone Metabolite 19-Hydroxyandrostenedione Induces Neuroendocrine Trans-Differentiation of Prostate Cancer Cells via an Ectopic Olfactory Receptor.” *Frontiers in Oncology* 8 (May):162.
- 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.
- Ahmad, Walid, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. “ChemBERTa-2: Towards Chemical Foundation Models.” *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/2209.01712>.
- Ahmed, Marawan, Horia Jalily Hasani, Subha Kalyaanamoorthy, and Khaled Barakat. 2021. “GPCR_LigandClassify.py; a Rigorous Machine Learning Classifier for GPCR Targeting Compounds.” *Scientific Reports* 11 (1): 9510.
- Ahuja, Gaurav, Deniz Bartsch, Wenjie Yao, Simon Geissen, Stefan Frank, Aitor Aguirre, Nicole Russ, et al. 2019. “Loss of Genomic Integrity Induced by Lysosphingolipid Imbalance Drives Ageing in the Heart.” *EMBO Reports* 20 (4). <https://doi.org/10.15252/embr.201847407>.
- Aloni, Ronny, Tsviya Olender, and Doron Lancet. 2006. “Ancient Genomic Architecture for Mammalian Olfactory Receptor Clusters.” *Genome Biology* 7 (10): R88.
- 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.
- Anand, Preetha, Ajaikumar B. Kunnumakkara, Chitra Sundaram, Kuzhuvelil B. Harikumar, Sheeja T. Tharakan, Oiki S. Lai, Bokyung Sung, and Bharat B. Aggarwal. 2008. “Cancer Is a Preventable Disease That Requires Major Lifestyle Changes.” *Pharmaceutical Research* 25 (9): 2097–2116.
- Armitage, Emily G., and Coral Barbas. 2014. “Metabolomics in Cancer Biomarker Discovery: Current Trends and Future Perspectives.” *Journal of Pharmaceutical and Biomedical Analysis* 87 (January):1–11.
- Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” *PloS One* 10 (7): e0130140.
- Ballante, Flavio, Axel Rudling, Alexey Zeifman, Andreas Lutten, Duy Duc Vo, John J. Irwin, Jan Kihlberg, Jose Brea, Maria Isabel Loza, and Jens Carlsson. 2020. “Docking Finds GPCR Ligands in

- Dark Chemical Matter.” *Journal of Medicinal Chemistry* 63 (2): 613–20.
- Banerjee, Priyanka, Andreas O. Eckert, Anna K. Schrey, and Robert Preissner. 2018. “ProTox-II: A Webserver for the Prediction of Toxicity of Chemicals.” *Nucleic Acids Research* 46 (W1): W257–63.
- Bardwell, Lee. 2005. “A Walk-through of the Yeast Mating Pheromone Response Pathway.” *Peptides* 26 (2): 339–50.
- Bar-Even, Arren, Elad Noor, Avi Flamholz, Joerg M. Buescher, and Ron Milo. 2011. “Hydrophobicity and Charge Shape Cellular Metabolite Concentrations.” *PLoS Computational Biology* 7 (10): e1002166.
- Benigni, R., A. Giuliani, R. Franke, and A. Gruska. 2000. “Quantitative Structure-Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines.” *Chemical Reviews* 100 (10): 3697–3714.
- Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42.
- 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.
- Bilsland, Alan E., Angelo Pugliese, Yu Liu, John Revie, Sharon Burns, Carol McCormick, Claire J. Cairney, et al. 2015. “Identification of a Selective G1-Phase Benzimidazolone Inhibitor by a Senescence-Targeted Virtual Screen Using Artificial Neural Networks.” *Neoplasia (New York, N.Y.)* 17 (9): 704–15.
- Block, Geoffrey A., Kevin J. Martin, Angel L. M. de Francisco, Stewart A. Turner, Morrell M. Avram, Michael G. Suranyi, Gavril Hercz, et al. 2004. “Cinacalcet for Secondary Hyperparathyroidism in Patients Receiving Hemodialysis.” *The New England Journal of Medicine* 350 (15): 1516–25.
- 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.
- Bokoch, Michael P., Yaozhong Zou, Søren G. F. Rasmussen, Corey W. Liu, Rie Nygaard, Daniel M. Rosenbaum, Juan José Fung, et al. 2010. “Ligand-Specific Regulation of the Extracellular Surface of a G-Protein-Coupled Receptor.” *Nature* 463 (7277): 108–12.
- Bonidia, Robson P., Douglas S. Domingues, Danilo S. Sanches, and André C. P. L. F. de Carvalho. 2022. “MathFeature: Feature Extraction Package for DNA, RNA and Protein Sequences Based on Mathematical Descriptors.” *Briefings in Bioinformatics* 23 (1): bbab434.
- Bourcier, Todd, Tim McGovern, Lidiya Stavitskaya, Naomi Kruhlak, and David Jacobson-Kram. 2015. “Improving Prediction of Carcinogenicity to Reduce, Refine, and Replace the Use of Experimental Animals.” *Journal of the American Association for Laboratory Animal Science: JAALAS* 54 (2):

163–69.

- Braakman, Rogier, and Eric Smith. 2012. “The Emergence and Early Evolution of Biological Carbon-Fixation.” *PLoS Computational Biology* 8 (4): e1002455.
- Brahma, Rahul, Sunghyun Moon, Jae-Min Shin, and Kwang-Hwi Cho. 2025. “AiGPro: A Multi-Tasks Model for Profiling of GPCRs for Agonist and Antagonist.” *Journal of Cheminformatics* 17 (1): 12.
- Brower, K. R., and Rollie Schafer. 1975. “The Recognition of Chemical Types by Odor. The Effect of Steric Hindrance at the Functional Group.” *Journal of Chemical Education* 52 (8): 538.
- 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.
- Budczies, Jan, Berit M. Pfitzner, Balazs Györfy, Klaus-Jürgen Winzer, Cornelia Radke, Manfred Dietel, Oliver Fiehn, and Carsten Denkert. 2015. “Glutamate Enrichment as New Diagnostic Opportunity in Breast Cancer: Glutamate and Glutamine in Breast Cancer.” *International Journal of Cancer: Journal International Du Cancer* 136 (7): 1619–28.
- Budin, Itay, Noam Prwyys, Na Zhang, and Jack W. Szostak. 2014. “Chain-Length Heterogeneity Allows for the Assembly of Fatty Acid Vesicles in Dilute Solutions.” *Biophysical Journal* 107 (7): 1582–90.
- Buschdid, C., M. O. Magnasco, L. B. Vosshall, and A. Keller. 2014. “Humans Can Discriminate More than 1 Trillion Olfactory Stimuli.” *Science (New York, N.Y.)* 343 (6177): 1370–72.
- Buschdid, 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.
- Busse, Daniela, Philipp Kudella, Nana-Maria Grüning, Günter Gisselmann, Sonja Ständer, Thomas Luger, Frank Jacobsen, et al. 2014. “A Synthetic Sandalwood Odorant Induces Wound-Healing Processes in Human Keratinocytes via the Olfactory Receptor OR2AT4.” *The Journal of Investigative Dermatology* 134 (11): 2823–32.
- Butterworth, Byron E., Lesa L. Aylward, and Sean M. Hays. 2007. “A Mechanism-Based Cancer Risk Assessment for 1,4-Dichlorobenzene.” *Regulatory Toxicology and Pharmacology: RTP* 49 (2): 138–48.
- 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.
- Caballero-Vidal, Gabriela, Cédric Bouysset, Hubert Grunig, Sébastien Fiorucci, Nicolas Montagné, Jérôme Golebiowski, and Emmanuelle Jacquin-Joly. 2020. “Machine Learning Decodes Chemical Features to Identify Novel Agonists of a Moth Odorant Receptor.” *Scientific Reports* 10 (1): 1655.
- Cao, Yiqun, Anna Charisi, Li-Chang Cheng, Tao Jiang, and Thomas Girke. 2008. “ChemmineR: A

- Compound Mining Framework for R.” *Bioinformatics (Oxford, England)* 24 (15): 1733–34.
- 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.
- Carthew, Richard W. 2021. “Gene Regulation and Cellular Metabolism: An Essential Partnership.” *Trends in Genetics: TIG* 37 (4): 389–400.
- Casadó, Vicent, and Verònica Casadó-Anguera. 2023. “What Are the Current Trends in G Protein-Coupled Receptor Targeted Drug Discovery?” *Expert Opinion on Drug Discovery* 18 (8): 815–20.
- Castleman, P., G. Szwabowski, D. Bowman, J. Cole, A. L. Parrill, and D. L. Baker. 2022. “Ligand-Based G Protein Coupled Receptor Pharmacophore Modeling: Assessing the Role of Ligand Function in Model Development.” *Journal of Molecular Graphics & Modelling* 111 (108107): 108107.
- Chan, Wallace K. B., Hongjiu Zhang, Jianyi Yang, Jeffrey R. Brender, Junguk Hur, Arzucan Özgür, and Yang Zhang. 2015. “GLASS: A Comprehensive Database for Experimentally Validated GPCR-Ligand Associations.” *Bioinformatics (Oxford, England)* 31 (18): 3035–42.
- Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2011. “SMOTE: Synthetic Minority over-Sampling Technique.” *arXiv [cs.AI]*. arXiv. <http://arxiv.org/abs/1106.1813>.
- Chéret, Jérémy, Marta Bertolini, Leslie Ponce, Janin Lehmann, Teresa Tsai, Majid Alam, Hanns Hatt, and Ralf Paus. 2018. “Olfactory Receptor OR2AT4 Regulates Human Hair Growth.” *Nature Communications* 9 (1): 3624.
- Chithrananda, Seyone, Gabe Grand, and Bharath Ramsundar. 2020. “ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction.” *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/2010.09885>.
- Christopoulos, Arthur. 2002. “Allosteric Binding Sites on Cell-Surface Receptors: Novel Targets for Drug Discovery.” *Nature Reviews. Drug Discovery* 1 (3): 198–210.
- Clish, Clary B. 2015. “Metabolomics: An Emerging but Powerful Tool for Precision Medicine.” *Cold Spring Harbor Molecular Case Studies* 1 (1): a000588.
- Cogliano, Vincent James, Robert A. Baan, Kurt Straif, Yann Grosse, Marie Béatrice Secretan, Fatiha El Ghissassi, and Paul Kleihues. 2004. “The Science and Practice of Carcinogen Identification and Evaluation.” *Environmental Health Perspectives* 112 (13): 1269–74.
- Cong, Xiaojing, Wenwen Ren, Jody Pacalon, Rui Xu, Lun Xu, Xuwen Li, Claire A. de March, et al. 2022. “Large-Scale G Protein-Coupled Olfactory Receptor-Ligand Pairing.” *ACS Central Science* 8 (3): 379–87.
- Couvineau, Alain, Rosa P. Gomariz, and Yossan-Var Tan. 2020. “Editorial: GPCR in Inflammatory and

- Cancer Diseases.” *Frontiers in Endocrinology* 11 (September):588157.
- Damavandi, Mohammad Sadegh, Hasan Shojaei, and Bahram Nasr Esfahani. 2023. “The Anticancer and Antibacterial Potential of Bioactive Secondary Metabolites Derived From Bacterial Endophytes in Association with *Artemisia Absinthium*.” *Scientific Reports* 13 (1): 18473.
- Dando, Ilaria, Elisa Dalla Pozza, Giulia Ambrosini, Margalida Torrens-Mas, Giovanna Butera, Nidula Mullappilly, Raffaella Pacchiana, Marta Palmieri, and Massimo Donadelli. 2019. “Oncometabolites in Cancer Aggressiveness and Tumour Repopulation.” *Biological Reviews of the Cambridge Philosophical Society* 94 (4): 1530–46.
- Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics (Oxford, England)* 27 (15): 2156–58.
- Dang, Lenny, David W. White, Stefan Gross, Bryson D. Bennett, Mark A. Bittinger, Edward M. Driggers, Valeria R. Fantin, et al. 2010. “Cancer-Associated IDH1 Mutations Produce 2-Hydroxyglutarate.” *Nature* 465 (7300): 966.
- Dearden, John C. 2017. “The history and development of quantitative structure-activity relationships (QSARs).” *International journal of quantitative structure-property relationships* 2 (2): 36–46.
- Dennis, James W., and C. Fred Brewer. 2013. “Density-Dependent Lectin-Glycan Interactions as a Paradigm for Conditional Regulation by Posttranslational Modifications.” *Molecular & Cellular Proteomics: MCP* 12 (4): 913–20.
- Doller, Dario. 2017. “Endogenous Allosteric Modulators of G Protein-Coupled Receptors: Implications in Drug Design.” In *2017 Medicinal Chemistry Reviews*, 343–60. Medicinal Chemistry Division of the American Chemical Society.
- Doyle, Timothy M., Kali Janes, Zhoumou Chen, Peter M. Grace, Emanuela Esposito, Salvatore Cuzzocrea, Tally M. Largent-Milnes, et al. 2020. “Activation of Sphingosine-1-Phosphate Receptor Subtype 1 in the Central Nervous System Contributes to Morphine-Induced Hyperalgesia and Antinociceptive Tolerance in Rodents.” *Pain* 161 (9): 2107–18.
- Duncan, Elaine M., Luca Vita, Bethany Dibnah, and Brian D. Hudson. 2023. “Metabolite-Sensing GPCRs Controlling Interactions between Adipose Tissue and Inflammation.” *Frontiers in Endocrinology* 14 (July):1197102.
- Ebenhöh, O., and R. Heinrich. 2001. “Evolutionary Optimization of Metabolic Pathways. Theoretical Reconstruction of the Stoichiometry of ATP and NADH Producing Systems.” *Bulletin of Mathematical Biology* 63 (1): 21–55.
- Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, et al. 2022. “ProtTrans: Toward Understanding the Language of Life through Self-Supervised

- Learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (10): 7112–27.
- Fan, Teresa W. M., Andrew N. Lane, Richard M. Higashi, Mohamed A. Farag, Hong Gao, Michael Bousamra, and Donald M. Miller. 2009. “Altered Regulation of Metabolic Pathways in Human Lung Cancer Discerned by (13)C Stable Isotope-Resolved Metabolomics (SIRM).” *Molecular Cancer* 8 (1): 41.
- Farland, William H., Angela Lynch, Neeraja K. Erraguntla, and Lynn H. Pottenger. 2019. “Improving Risk Assessment Approaches for Chemicals with Both Endogenous and Exogenous Exposures.” *Regulatory Toxicology and Pharmacology: RTP* 103 (April):210–15.
- Feldmesser, Ester, Tsviya Olender, Miriam Khen, Itai Yanai, Ron Ophir, and Doron Lancet. 2006. “Widespread Ectopic Expression of Olfactory Receptor Genes.” *BMC Genomics* 7 (1): 121.
- Fiehn, Oliver, Dinesh K. Barupal, and Tobias Kind. 2011. “Extending Biochemical Databases by Metabolomic Surveys.” *The Journal of Biological Chemistry* 286 (34): 30244.
- Fjodorova, Natalja, Marjan Vračko, Marjan Tušar, Aneta Jezierska, Marjana Novič, Ralph Kühne, and Gerrit Schüürmann. 2010. “Quantitative and Qualitative Models for Carcinogenicity Prediction for Non-Congeneric Chemicals Using CP ANN Method for Regulatory Uses.” *Molecular Diversity* 14 (3): 581–94.
- Förster, Jochen, Iman Famili, Patrick Fu, Bernhard Ø. Palsson, and Jens Nielsen. 2003. “Genome-Scale Reconstruction of the *Saccharomyces Cerevisiae* Metabolic Network.” *Genome Research* 13 (2): 244–53.
- Franke, R., A. Gruska, A. Giuliani, and R. Benigni. 2001. “Prediction of Rodent Carcinogenicity of Aromatic Amines: A Quantitative Structure-Activity Relationships Model.” *Carcinogenesis* 22 (9): 1561–71.
- Fridmanis, Davids, Robert Fredriksson, Ivo Kapa, Helgi B. Schiöth, and Janis Klovins. 2007. “Formation of New Genes Explains Lower Intron Density in Mammalian Rhodopsin G Protein-Coupled Receptors.” *Molecular Phylogenetics and Evolution* 43 (3): 864–80.
- Fu, Accalia, Juan Carlos Alvarez-Perez, Daina Avizonis, Tatsuya Kin, Scott B. Ficarro, Dong Wook Choi, Esra Karakose, et al. 2020. “Glucose-Dependent Partitioning of Arginine to the Urea Cycle Protects β -Cells from Inflammation.” *Nature Metabolism* 2 (5): 432–46.
- Fukushima, Shoji, Anna Kinoshita, Rawiwan Puatanachokchai, Masahiko Kushida, Hideki Wanibuchi, and Keiichirou Morimura. 2005. “Hormesis and Dose-Response-Mediated Mechanisms in Carcinogenesis: Evidence for a Threshold in Carcinogenicity of Non-Genotoxic Carcinogens.” *Carcinogenesis* 26 (11): 1835–45.
- Garrick, R. 2010. “A Randomized, Double-Blind, Placebo-Controlled Study to Assess the Efficacy and Safety of Cinacalcet HCl in Participants with CKD Not Receiving Dialysis.” *The Year Book of*

- Medicine* 2010 (January):245–48.
- Gelis, Lian, Nikolina Jovancevic, Sophie Veitinger, Bhubaneswar Mandal, Hans-Dieter Arndt, Eva M. Neuhaus, and Hanns Hatt. 2016. “Functional Characterization of the Odorant Receptor 51E2 in Human Melanocytes.” *The Journal of Biological Chemistry* 291 (34): 17772–86.
- Gerkin, Richard C., and Jason B. Castro. 2015. “The Number of Olfactory Stimuli That Humans Can Discriminate Is Still Unknown.” *eLife* 4 (July):e08127.
- Gibson, Daniel G., John I. Glass, Carole Lartigue, Vladimir N. Noskov, Ray-Yuan Chuang, Mikkel A. Algire, Gwynedd A. Benders, et al. 2010. “Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome.” *Science* 329 (5987): 52–56.
- Gibson, Daniel G., Lei Young, Ray-Yuan Chuang, J. Craig Venter, Clyde A. Hutchison 3rd, and Hamilton O. Smith. 2009. “Enzymatic Assembly of DNA Molecules up to Several Hundred Kilobases.” *Nature Methods* 6 (5): 343–45.
- Giera, Martin, Oscar Yanes, and Gary Siuzdak. 2022. “Metabolite Discovery: Biochemistry’s Scientific Driver.” *Cell Metabolism* 34 (1): 21–34.
- Gimpl, Gerald. 2016. “Interaction of G Protein Coupled Receptors and Cholesterol.” *Chemistry and Physics of Lipids* 199 (September):61–73.
- Goupil, E., S. A. Laporte, and T. E. Hébert. 2012. “Functional Selectivity in GPCR Signaling: Understanding the Full Spectrum of Receptor Conformations.” *Mini Reviews in Medicinal Chemistry* 12 (9): 817–30.
- Guedes, Joana P., Vitória Baptista, Cátia Santos-Pereira, Maria João Sousa, Stéphen Manon, Susana R. Chaves, and Manuela Côrte-Real. 2022. “Acetic Acid Triggers Cytochrome c Release in Yeast Heterologously Expressing Human Bax.” *Apoptosis: An International Journal on Programmed Cell Death* 27 (5-6): 368–81.
- Guijas, Carlos, J. Rafael Montenegro-Burke, Benedikt Warth, Mary E. Spilker, and Gary Siuzdak. 2018. “Metabolomics Activity Screening for Identifying Metabolites That Modulate Phenotype.” *Nature Biotechnology* 36 (4): 316–20.
- Guimerà, Roger, and Luís A. Nunes Amaral. 2005. “Functional Cartography of Complex Metabolic Networks.” *Nature* 433 (7028): 895–900.
- 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, Ria, Aayushi Mittal, Vishesh Agrawal, Sushant Gupta, Krishan Gupta, Rishi Raj Jain, Prakriti Garg, et al. 2021. “OdoriFy: A Conglomerate of Artificial Intelligence-Driven Prediction Engines for

- Olfactory Decoding.” *The Journal of Biological Chemistry* 297 (2): 100956.
- Gusenleitner, Daniel, Scott S. Auerbach, Tisha Melia, Harold F. Gómez, David H. Sherr, and Stefano Monti. 2014. “Genomic Models of Short-Term Exposure Accurately Predict Long-Term Chemical Carcinogenicity and Identify Putative Mechanisms of Action.” *PloS One* 9 (7): e102579.
- Haas, Robert, Aleksej Zelezniak, Jacopo Iacovacci, Stephan Kamrad, Stjohn Townsend, and Markus Ralser. 2017. “Designing and Interpreting ‘Multi-Omic’ Experiments That May Change Our Understanding of Biology.” *Current Opinion in Systems Biology* 6 (December):37–45.
- Hanson, Michael A., Vadim Cherezov, Mark T. Griffith, Christopher B. Roth, Veli-Pekka Jaakola, Ellen Y. T. Chien, Jeffrey Velasquez, Peter Kuhn, and Raymond C. Stevens. 2008. “A Specific Cholesterol Binding Site Is Established by the 2.8 Å Structure of the Human β 2-Adrenergic Receptor.” *Structure (London, England: 1993)* 16 (6): 897–905.
- Hart, Gerald W., Chad Slawson, Genaro Ramirez-Correa, and Olof Lagerlof. 2011. “Cross Talk between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease.” *Annual Review of Biochemistry* 80 (1): 825–58.
- Hartwell, L. H. 1980. “Mutants of *Saccharomyces Cerevisiae* Unresponsive to Cell Division Control by Polypeptide Mating Hormone.” *The Journal of Cell Biology* 85 (3): 811–22.
- Hayden, Sara, and Emma C. Teeling. 2014. “The Molecular Biology of Vertebrate Olfaction.” *Anatomical Record (Hoboken, N.J.: 2007)* 297 (11): 2216–26.
- Hernández, Lya G., Harry van Steeg, Mirjam Luijten, and Jan van Benthem. 2009. “Mechanisms of Non-Genotoxic Carcinogens and Importance of a Weight of Evidence Approach.” *Mutation Research* 682 (2-3): 94–109.
- Hoeijmakers, Jan H. J. 2009. “DNA Damage, Aging, and Cancer.” *The New England Journal of Medicine* 361 (15): 1475–85.
- Hojo, M., T. Morimoto, M. Maluccio, T. Asano, K. Morimoto, M. Lagman, T. Shimbo, and M. Suthanthiran. 1999. “Cyclosporine Induces Cancer Progression by a Cell-Autonomous Mechanism.” *Nature* 397 (6719): 530–34.
- Holmes, Elaine, Ruey Leng Loo, Jeremiah Stamler, Magda Bictash, Ivan K. S. Yap, Queenie Chan, Tim Ebbels, et al. 2008. “Human Metabolic Phenotype Diversity and Its Association with Diet and Blood Pressure.” *Nature* 453 (7193): 396–400.
- 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.
- Husted, Anna Sofie, Mette Trauelsen, Olga Rudenko, Siv A. Hjorth, and Thue W. Schwartz. 2017. “GPCR-Mediated Signaling of Metabolites.” *Cell Metabolism* 25 (4): 777–96.

- Jabeen, Amara, Claire A. de March, Hiroaki Matsunami, and Shoba Ranganathan. 2021. "Machine Learning Assisted Approach for Finding Novel High Activity Agonists of Human Ectopic Olfactory Receptors." *International Journal of Molecular Sciences* 22 (21): 11546.
- Jaeger, Sabrina, Simone Fulle, and Samo Turk. 2018. "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition." *Journal of Chemical Information and Modeling* 58 (1): 27–35.
- Jafurulla, Md, G. Aditya Kumar, Bhagyashree D. Rao, and Amitabha Chattopadhyay. 2019. "A Critical Analysis of Molecular Mechanisms Underlying Membrane Cholesterol Sensitivity of GPCRs." *Advances in Experimental Medicine and Biology* 1115:21–52.
- Jang, Jae Wan, Nam-Chul Cho, Sun-Joon Min, Yong Seo Cho, Ki Duk Park, Seon Hee Seo, Kyoung Tai No, and Ae Nim Pae. 2016. "Novel Scaffold Identification of mGlu1 Receptor Negative Allosteric Modulators Using a Hierarchical Virtual Screening Approach." *Chemical Biology & Drug Design* 87 (2): 239–56.
- Ježek, Petr. 2020. "2-Hydroxyglutarate in Cancer Cells." *Antioxidants & Redox Signaling* 33 (13): 903–26.
- Jiménez-Luna, José, Francesca Grisoni, and Gisbert Schneider. 2020. "Drug Discovery with Explainable Artificial Intelligence." *Nature Machine Intelligence* 2 (10): 573–84.
- Johnson, Brett A., Haleh Farahbod, Sepideh Saber, and Michael Leon. 2005. "Effects of Functional Group Position on Spatial Representations of Aliphatic Odorants in the Rat Olfactory Bulb." *The Journal of Comparative Neurology* 483 (2): 192–204.
- Johnson, Caroline H., and Frank J. Gonzalez. 2012. "Challenges and Opportunities of Metabolomics." *Journal of Cellular Physiology* 227 (8): 2975–81.
- Johnson, Caroline H., Julijana Ivanisevic, and Gary Siuzdak. 2016. "Metabolomics: Beyond Biomarkers and towards Mechanisms." *Nature Reviews. Molecular Cell Biology* 17 (7): 451–59.
- Joyce, Andrew R., and Bernhard Ø. Palsson. 2006. "The Model Organism as a System: Integrating 'Omics' Data Sets." *Nature Reviews. Molecular Cell Biology* 7 (3): 198–210.
- Kaddurah-Daouk, Rima, Bruce S. Kristal, and Richard M. Weinshilboum. 2008. "Metabolomics: A Global Biochemical Approach to Drug Response and Disease." *Annual Review of Pharmacology and Toxicology* 48 (1): 653–83.
- Kanehisa, M., and S. Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30.
- Karpov, Pavel, Guillaume Godin, and Igor V. Tetko. 2020. "Transformer-CNN: Swiss Knife for QSAR Modeling and Interpretation." *Journal of Cheminformatics* 12 (1): 17.
- Kell, Douglas B., and Stephen G. Oliver. 2016. "The Metabolome 18 Years on: A Concept Comes of Age." *Metabolomics: Official Journal of the Metabolomic Society* 12 (9): 148.

- Knerr, Stefanie, and Dieter Schrenk. 2006. "Carcinogenicity of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin in Experimental Models." *Molecular Nutrition & Food Research* 50 (10): 897–907.
- Kokhlikyan, Narine, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, et al. 2020. "Captum: A Unified and Generic Model Interpretability Library for PyTorch." *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/2009.07896>.
- Kolakowski, L. F., Jr. 1994. "GCRDb: A G-Protein-Coupled Receptor Database." *Receptors & Channels* 2 (1): 1–7.
- Kumar, Surendra, Mahesh K. Teli, and Mi-Hyun Kim. 2024. "GPCR-IPL Score: Multilevel Featurization of GPCR-Ligand Interaction Patterns and Prediction of Ligand Functions from Selectivity to Biased Activation." *Briefings in Bioinformatics* 25 (2): bbae105.
- Kursa, M., A. Jankowski, and W. Rudnicki. 2010. "Boruta - A System for Feature Selection." *Fundamenta Informaticae* 101 (4): 271–85.
- Lalis, Maxence, Matej Hladiš, Samar Abi Khalil, Loïc Briand, Sébastien Fiorucci, and Jérémie Topin. 2024. "M2OR: A Database of Olfactory Receptor-Odorant Pairs for Understanding the Molecular Mechanisms of Olfaction." *Nucleic Acids Research* 52 (D1): D1370–79.
- 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.
- Lewis, D. F., C. Ioannides, and D. V. Parke. 1998. "Cytochromes P450 and Species Differences in Xenobiotic Metabolism and Activation of Carcinogen." *Environmental Health Perspectives* 106 (10): 633–41.
- Liao, Yang, Gordon K. Smyth, and Wei Shi. 2019. "The R Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads." *Nucleic Acids Research* 47 (8): e47.
- Liehr, J. G. 2000. "Is Estradiol a Genotoxic Mutagenic Carcinogen?" *Endocrine Reviews* 21 (1): 40–54.
- Liska, Orsolya, Gábor Boross, Charles Rocabert, Balázs Szappanos, Roland Tengölics, and Balázs Papp. 2023. "Principles of Metabolome Conservation in Animals." *Proceedings of the National Academy of Sciences of the United States of America* 120 (35): e2302147120.
- Liu, Shan, Wen-Jia Kang, Anna Abrimian, Jin Xu, Luca Cartegni, Susruta Majumdar, Patrick Hesketh, Alex Bekker, and Ying-Xian Pan. 2021. "Alternative Pre-mRNA Splicing of the Mu Opioid Receptor Gene, OPRM1: Insight into Complex Mu Opioid Actions." *Biomolecules* 11 (10): 1525.
- Liu, Xiangyu, Ali Masoudi, Alem W. Kahsai, Li-Yin Huang, Biswaranjan Pani, Dean P. Staus, Paul J. Shim, et al. 2019. "Mechanism of β 2AR Regulation by an Intracellular Positive Allosteric Modulator." *Science (New York, N.Y.)* 364 (6447): 1283–87.
- Liu, Xinyi, Xubo Su, Fei Wang, Zhimin Huang, Qi Wang, Zhen Li, Ruina Zhang, et al. 2011.

- “ODORactor: A Web Server for Deciphering Olfactory Coding.” *Bioinformatics (Oxford, England)* 27 (16): 2302–3.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.
- Lowe, Daniel M., Peter T. Corbett, Peter Murray-Rust, and Robert C. Glen. 2011. “Chemical Name to Structure: OPSIN, an Open Source Solution.” *Journal of Chemical Information and Modeling* 51 (3): 739–53.
- Luch, Andreas. 2005. “Nature and Nurture - Lessons from Chemical Carcinogenesis.” *Nature Reviews. Cancer* 5 (2): 113–25.
- Lutz, W. K., and T. Fekete. 1996. “Endogenous and Exogenous Factors in Carcinogenesis: Limits to Cancer Prevention.” *International Archives of Occupational and Environmental Health* 68 (2): 120–25.
- 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.
- Manna, Moutusi, Miia Niemelä, Joonas Tynkkynen, Matti Javanainen, Waldemar Kulig, Daniel J. Müller, Tomasz Rog, and Ilpo Vattulainen. 2016. “Mechanism of Allosteric Regulation of β 2-Adrenergic Receptor by Cholesterol.” *eLife* 5 (November). <https://doi.org/10.7554/eLife.18432>.
- Mansouri, Kamel, and Richard S. Judson. 2016. “In Silico Study of in Vitro GPCR Assays by QSAR Modeling.” *Methods in Molecular Biology (Clifton, N.J.)* 1425:361–81.
- Martins, Vítor M., Tânia R. Fernandes, Diana Lopes, Catarina B. Afonso, Maria R. M. Domingues, Manuela Côrte-Real, and Maria J. Sousa. 2019. “Contacts in Death: The Role of the ER-Mitochondria Axis in Acetic Acid-Induced Apoptosis in Yeast.” *Journal of Molecular Biology* 431 (2): 273–88.
- 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, Annika Simon, Dieter Häussinger, Verena Keitel, Günter Gisselmann, Heike Conrad, and Hanns Hatt. 2015. “Monoterpene (-)-Citronellal Affects Hepatocarcinoma Cell Signaling via an Olfactory Receptor.” *Archives of Biochemistry and Biophysics* 566 (January):100–109.
- Maunz, Andreas, Martin Gütlein, Micha Rautenberg, David Vorgrimmler, Denis Gebele, and Christoph Helma. 2013. “Lazar: A Modular Predictive Toxicology Framework.” *Frontiers in Pharmacology* 4 (April):38.
- Meister, Markus. 2015. “On the Dimensionality of Odor Space.” *eLife* 4 (July):e07865.
- Melhem, Hassan, Berna Kaya, C. Korcan Ayata, Petr Hruz, and Jan Hendrik Niess. 2019.

- “Metabolite-Sensing G Protein-Coupled Receptors Connect the Diet-Microbiota-Metabolites Axis to Inflammatory Bowel Disease.” *Cells (Basel, Switzerland)* 8 (5): 450.
- Mirnaghi, Fatemeh S., and Amy A. Caudy. 2014. “Challenges of Analyzing Different Classes of Metabolites by a Single Analytical Method.” *Bioanalysis* 6 (24): 3393–3416.
- Mohanty, Sanjay Kumar, Aayushi Mittal, Aakash Gaur, Namra, Subhadeep Duari, Saveena Solanki, Anmol Kumar Sharma, et al. 2023. “Endogenous Intracellular Metabolites Allosterically Modulate GPCR-G α Interface.” *Bioinformatics*. bioRxiv. <https://www.biorxiv.org/content/10.1101/2023.02.14.528587v3>.
- Mollo, Ernesto, Angelo Fontana, Vassilios Roussis, Gianluca Polese, Pietro Amodeo, and Michael T. Ghiselin. 2014. “Sensing Marine Biomolecules: Smell, Taste, and the Evolutionary Transition from Aquatic to Terrestrial Life.” *Frontiers in Chemistry* 2 (October):92.
- Morales, Aliuska Helguera, Miguel Angel Cabrera Pérez, Robert D. Combes, and Maykel Pérez González. 2006. “Quantitative Structure Activity Relationship for the Computational Prediction of Nitrocompounds Carcinogenicity.” *Toxicology* 220 (1): 51–62.
- Moriwaki, Hirotomo, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. 2018. “Mordred: A Molecular Descriptor Calculator.” *Journal of Cheminformatics* 10 (1): 4.
- 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.
- Narayanan, Annamalai, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. “Graph2vec: Learning Distributed Representations of Graphs.” *arXiv [cs.AI]*. arXiv. <http://arxiv.org/abs/1707.05005>.
- National Research Council (US) Committee on Comparative Toxicity of Naturally Occurring Carcinogens. 1996. *Methods for Evaluating Potential Carcinogens and Anticarcinogens*. National Academies Press (US).
- Navratilova, Iva, Jérémy Besnard, and Andrew L. Hopkins. 2011. “Screening for GPCR Ligands Using Surface Plasmon Resonance.” *ACS Medicinal Chemistry Letters* 2 (7): 549–54.
- Neuhaus, Eva M., Weiyi Zhang, Lian Gelis, Ying Deng, Joachim Noldus, and Hanns Hatt. 2009. “Activation of an Olfactory Receptor Inhibits Proliferation of Prostate Cancer Cells.” *The Journal of Biological Chemistry* 284 (24): 16218–25.
- Newgard, Christopher B. 2017. “Metabolomics and Metabolic Diseases: Where Do We Stand?” *Cell Metabolism* 25 (1): 43–56.
- Noor, Elad, Eran Eden, Ron Milo, and Uri Alon. 2010. “Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy.” *Molecular Cell* 39 (5): 809–20.

- 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.
- Oh, Jooseong, Hyi-Thaek Ceong, Dokyun Na, and Chungoo Park. 2022. "A Machine Learning Model for Classifying G-Protein-Coupled Receptors as Agonists or Antagonists." *BMC Bioinformatics* 23 (Suppl 9): 346.
- Oliveira de Souza, Camila, Xuenan Sun, and Dayoung Oh. 2021. "Metabolic Functions of G Protein-Coupled Receptors and β -Arrestin-Mediated Signaling Pathways in the Pathophysiology of Type 2 Diabetes and Obesity." *Frontiers in Endocrinology* 12 (August):715877.
- Ollitrault, Guillaume, Rayane Achebouché, Antoine Dreux, Samuel Murail, Karine Audouze, Anne Tromelin, and Olivier Taboureau. 2024. "Pred-O3, a Web Server to Predict Molecules, Olfactory Receptors and Odor Relationships." *Nucleic Acids Research* 52 (W1): W507–12.
- Orecchioni, Marco, Hiroaki Matsunami, and Klaus Ley. 2022. "Olfactory Receptors in Macrophages and Inflammation." *Frontiers in Immunology* 13 (October):1029244.
- Pan, Shanshan, Luan Yin, Jie Liu, Jie Tong, Zichuan Wang, Jiahui Zhao, Xuesong Liu, et al. 2024. "Metabolomics-Driven Approaches for Identifying Therapeutic Targets in Drug Discovery." *MedComm* 5 (11): e792.
- Pardella, Elisa, Luigi Ippolito, Elisa Giannoni, and Paola Chiarugi. 2022. "Nutritional and Metabolic Signalling through GPCRs." *FEBS Letters* 596 (18): 2364–81.
- Passi, Anurag, Juan D. Tibocha-Bonilla, Manish Kumar, Diego Tec-Campos, Karsten Zengler, and Cristal Zuniga. 2021. "Genome-Scale Metabolic Modeling Enables in-Depth Understanding of Big Data." *Metabolites* 12 (1): 14.
- Patti, Gary J., Oscar Yanes, and Gary Siuzdak. 2012. "Innovation: Metabolomics: The Apogee of the Omics Trilogy." *Nature Reviews. Molecular Cell Biology* 13 (4): 263–69.
- Penev, Petar I., Claudia Alvarez-Carreño, Eric Smith, Anton S. Petrov, and Loren Dean Williams. 2021. "TwinCons: Conservation Score for Uncovering Deep Sequence Similarity and Divergence." *PLoS Computational Biology* 17 (10): e1009541.
- Peregrin-Alvarez, José Manuel, Sophia Tsoka, and Christos A. Ouzounis. 2003. "The Phylogenetic Extent of Metabolic Enzymes and Pathways." *Genome Research* 13 (3): 422–27.
- Peregrín-Alvarez, José M., Chris Sanford, and John Parkinson. 2009. "The Conservation and Evolutionary Modularity of Metabolism." *Genome Biology* 10 (6): R63.
- Picard, Louis-Philippe, Anne Marie Schönege, Martin J. Lohse, and Michel Bouvier. 2018.

- “Bioluminescence Resonance Energy Transfer-Based Biosensors Allow Monitoring of Ligand- and Transducer-Mediated GPCR Conformational Changes.” *Communications Biology* 1 (1): 106.
- Pinu, Farhana R., Seyed Ali Goldansaz, and Jacob Jaine. 2019. “Translational Metabolomics: Current Challenges and Future Opportunities.” *Metabolites* 9 (6): 108.
- Rappaport, Stephen M. 2018. “Redefining Environmental Exposure for Disease Etiology.” *Npj Systems Biology and Applications* 4 (1): 30.
- Rattray, Nicholas J. W., Nicole C. Deziel, Joshua D. Wallach, Sajid A. Khan, Vasilis Vasiliou, John P. A. Ioannidis, and Caroline H. Johnson. 2018. “Beyond Genomics: Understanding Exposotypes through Metabolomics.” *Human Genomics* 12 (1): 4.
- Reed, Randall R. 2004. “After the Holy Grail: Establishing a Molecular Basis for Mammalian Olfaction.” *Cell* 116 (2): 329–36.
- 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.” In *Molecular Pharmacology*. IntechOpen.
- Reznik, Ed, Augustin Luna, Bülent Arman Aksoy, Eric Minwei Liu, Konnor La, Irina Ostrovnaya, Chad J. Creighton, A. Ari Hakimi, and Chris Sander. 2018. “A Landscape of Metabolic Variation across Tumor Types.” *Cell Systems* 6 (3): 301–13.e3.
- Rifaioğlu, Ahmet Sureyya, Esra Nalbat, Volkan Atalay, Maria Jesus Martin, Rengul Cetin-Atalay, and Tunca Doğan. 2020. “DEEPScreen: High Performance Drug-Target Interaction Prediction with Convolutional Neural Networks Using 2-D Structural Compound Representations.” *Chemical Science (Royal Society of Chemistry: 2010)* 11 (9): 2531–57.
- Rinschen, Markus M., Julijana Ivanisevic, Martin Giera, and Gary Siuzdak. 2019. “Identification of Bioactive Metabolites Using Activity Metabolomics.” *Nature Reviews. Molecular Cell Biology* 20 (6): 353–67.
- Rios, Carl, Ivone Gomes, and Lakshmi A. Devi. 2006. “Mu Opioid and CB1 Cannabinoid Receptor Interactions: Reciprocal Inhibition of Receptor Signaling and Neuritogenesis: Opioid-Cannabinoid Interactions.” *British Journal of Pharmacology* 148 (4): 387–95.
- Rodriguez-Antona, C., and M. Ingelman-Sundberg. 2006. “Cytochrome P450 Pharmacogenetics and Cancer.” *Oncogene* 25 (11): 1679–91.
- Rogers, David, and Mathew Hahn. 2010. “Extended-Connectivity Fingerprints.” *Journal of Chemical Information and Modeling* 50 (5): 742–54.
- Ruiz, P., O. Faroon, C. J. Moudgal, H. Hansen, C. T. De Rosa, and M. Mumtaz. 2008. “Prediction of the Health Effects of Polychlorinated Biphenyls (PCBs) and Their Metabolites Using Quantitative Structure-Activity Relationship (QSAR).” *Toxicology Letters* 181 (1): 53–65.

- Saad, Mona M. G., Samir A. M. Abdelgaleil, and Yoshihito Shiono. 2021. "Antibacterial and Herbicidal Properties of Secondary Metabolites from Fungi." *Natural Product Research* 35 (23): 5446–51.
- Salahpour, Ali, Stefano Espinoza, Bernard Masri, Vincent Lam, Larry S. Barak, and Raul R. Gainetdinov. 2012. "BRET Biosensors to Study GPCR Biology, Pharmacology, and Signal Transduction." *Frontiers in Endocrinology* 3 (August):105.
- Sanz, Guenhaël, Isabelle Leray, Aurélie Dewaele, Julien Sobilo, Stéphanie Lerondel, Stéphan Bouet, Denise Grébert, Régine Monnerie, Edith Pajot-Augy, and Lluís M. Mir. 2014. "Promotion of Cancer Cell Invasiveness and Metastasis Emergence Caused by Olfactory Receptor Stimulation." *PloS One* 9 (1): e85110.
- Schihada, Hannes, Sylvie Vandenabeele, Ulrike Zabel, Monika Frank, Martin J. Lohse, and Isabella Maiellaro. 2018. "A Universal Bioluminescence Resonance Energy Transfer Sensor Design Enables High-Sensitivity Screening of GPCR Activation Dynamics." *Communications Biology* 1 (1): 105.
- Schmidt, Fredrik H. 2014. "A New Way to Understand Chemical Carcinogenesis and Cancer Prevention." *Research and Reports in Medicinal Chemistry* 4 (January):23.
- Schreiber, Stuart L. 2005. "Small Molecules: The Missing Link in the Central Dogma." *Nature Chemical Biology* 1 (2): 64–66.
- Schyman, Patric, Ruifeng Liu, Valmik Desai, and Anders Wallqvist. 2017. "VNN Web Server for ADMET Predictions." *Frontiers in Pharmacology* 8 (December):889.
- Seidel, Susanne A. I., Patricia M. Dijkman, Wendy A. Lea, Geert van den Bogaart, Moran Jerabek-Willemsen, Ana Lazic, Jeremiah S. Joseph, et al. 2013. "Microscale Thermophoresis Quantifies Biomolecular Interactions under Previously Challenging Conditions." *Methods (San Diego, Calif.)* 59 (3): 301–15.
- Seo, Sangmin, Jonghwan Choi, Soon Kil Ahn, Kil Won Kim, Jaekwang Kim, Jaehyuck Choi, Jinho Kim, and Jaegyeon Ahn. 2018. "Prediction of GPCR-Ligand Binding Using Machine Learning Algorithms." *Computational and Mathematical Methods in Medicine* 2018 (January):6565241.
- Severin, F. F., and A. A. Hyman. 2002. "Pheromone Induces Programmed Cell Death in *S. Cerevisiae*." *Current Biology: CB* 12 (7): R233–35.
- Sharma, Anju, Rajnish Kumar, Rahul Semwal, Imlimaong Aier, Pankaj Tyagi, and Pritish K. Varadwaj. 2022. "DeepOlf: Deep Neural Network Based Architecture for Predicting Odorants and Their Interacting Olfactory Receptors." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19 (1): 418–28.
- Shellhammer, James P., Amy E. Pomeroy, Yang Li, Lorena Dujmusic, Timothy C. Elston, Nan Hao, and Henrik G. Dohlman. 2019. "Quantitative Analysis of the Yeast Pheromone Pathway." *Yeast (Chichester, England)* 36 (8): 495–518.

- Siramshetty, Vishal B., Janette Nickel, Christian Omieczynski, Bjoern-Oliver Gohlke, Malgorzata N. Drwal, and Robert Preissner. 2016. "WITHDRAWN--a Resource for Withdrawn and Discontinued Drugs." *Nucleic Acids Research* 44 (D1): D1080–86.
- Smith, Martyn T., Kathryn Z. Guyton, Catherine F. Gibbons, Jason M. Fritz, Christopher J. Portier, Ivan Rusyn, David M. DeMarini, et al. 2016. "Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis." *Environmental Health Perspectives* 124 (6): 713–21.
- 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.
- Stein, Reed M., Hye Jin Kang, John D. McCorvy, Grant C. Glatfelter, Anthony J. Jones, Tao Che, Samuel Slocum, et al. 2020. "Virtual Discovery of Melatonin Receptor Ligands to Modulate Circadian Rhythms." *Nature* 579 (7800): 609–14.
- 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 (1): 15453.
- Strassheim, Derek, Timothy Sullivan, David C. Irwin, Evgenia Gerasimovskaya, Tim Lahm, Dwight J. Klemm, Edward C. Dempsey, Kurt R. Stenmark, and Vijaya Karoor. 2021. "Metabolite G-Protein Coupled Receptors in Cardio-Metabolic Diseases." *Cells (Basel, Switzerland)* 10 (12): 3347.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." Edited by Doina Precup and Yee Whye Teh. *International Conference on Machine Learning* 70 (March):3319–28.
- Swenberg, James A., Kun Lu, Benjamin C. Moeller, Lina Gao, Patricia B. Upton, Jun Nakamura, and Thomas B. Starr. 2011. "Endogenous versus Exogenous DNA Adducts: Their Role in Carcinogenesis, Epidemiology, and Risk Assessment." *Toxicological Sciences: An Official Journal of the Society of Toxicology* 120 Suppl 1 (Supplement 1): S130–45.
- Tang, Jane. 2011. "Microbial Metabolomics." *Current Genomics* 12 (6): 391–403.
- Tan, Jian K., Craig McKenzie, Eliana Mariño, Laurence Macia, and Charles R. Mackay. 2017. "Metabolite-Sensing G Protein-Coupled Receptors-Facilitators of Diet-Related Immune Regulation." *Annual Review of Immunology* 35 (April):371–402.
- Teng, Xinchun, and J. Marie Hardwick. 2009. "Reliable Method for Detection of Programmed Cell Death in Yeast." *Methods in Molecular Biology (Clifton, N.J.)* 559:335–42.
- Thal, David M., Alisa Glukhova, Patrick M. Sexton, and Arthur Christopoulos. 2018. "Structural Insights

- into G-Protein-Coupled Receptor Allostery.” *Nature* 559 (7712): 45–53.
- Timasheva, Yanina, Kaido Lepik, Orsolya Liska, Balázs Papp, and Zoltán Kutalik. 2024. “Widespread Natural Selection on Metabolite Levels in Humans.” *Genome Research* 34 (8): 1121–29.
- Tingle, Benjamin I., Khanh G. Tang, Mar Castanon, John J. Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii S. Moroz, and John J. Irwin. 2023. “ZINC-22—A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery.” *Journal of Chemical Information and Modeling* 63 (4): 1166–76.
- Tsai, Teresa, Sophie Veitinger, Irina Peek, Daniela Busse, Josephine Eckardt, Dilyana Vladimirova, Nikolina Jovancevic, et al. 2017. “Two Olfactory Receptors-OR2A4/7 and OR51B5-Differentially Affect Epidermal Proliferation and Differentiation.” *Experimental Dermatology* 26 (1): 58–65.
- Van Norman, Gail A. 2019. “Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is It Time to Rethink Our Current Approach?” *JACC. Basic to Translational Science* 4 (7): 845–54.
- Van Zandycke, Sylvie M., Olivier Simal, Sara Gualdoni, and Katherine A. Smart. 2003. “Determination of Yeast Viability Using Fluorophores.” *Journal of the American Society of Brewing Chemists. American Society of Brewing Chemists* 61 (1): 15–22.
- Vaser, Robert, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C. Ng. 2016. “SIFT Missense Predictions for Genomes.” *Nature Protocols* 11 (1): 1–9.
- 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.
- Velloso, João Paulo L., David B. Ascher, and Douglas E. V. Pires. 2021. “pdCSM-GPCR: Predicting Potent GPCR Ligands with Graph-Based Signatures.” *Bioinformatics Advances* 1 (1): vbab031.
- Wang, Mukuo, Shujing Hou, Yu Wei, Dongmei Li, and Jianping Lin. 2021. “Discovery of Novel Dual Adenosine A1/A2A Receptor Antagonists Using Deep Learning, Pharmacophore Modeling and Molecular Docking.” *PLoS Computational Biology* 17 (3): e1008821.
- Warszycki, Dawid, Łukasz Struski, Marek Śmieja, Rafał Kafel, and Rafał Kurczab. 2021. “Pharmacoprint: A Combination of a Pharmacophore Fingerprint and Artificial Intelligence as a Tool for Computer-Aided Drug Design.” *Journal of Chemical Information and Modeling* 61 (10): 5054–65.
- Watson, J. D., and F. H. Crick. 1953. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38.
- Weber, Lea, Désirée Maßberg, Christian Becker, Janine Altmüller, Burkhard Ubrig, Gabriele Bonatz, Gerhard Wölk, et al. 2018. “Olfactory Receptors as Biomarkers in Human Breast Carcinoma Tissues.” *Frontiers in Oncology* 8 (February):33.

- Weng, J-K, and J. P. Noel. 2012. "The Remarkable Pliability and Promiscuity of Specialized Metabolism." *Cold Spring Harbor Symposia on Quantitative Biology* 77 (0): 309–20.
- 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.
- Winkler, Wade C., and Ronald R. Breaker. 2003. "Genetic Control by Metabolite-Binding Riboswitches." *Chembiochem: A European Journal of Chemical Biology* 4 (10): 1024–32.
- Wishart, David S. 2016. "Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine." *Nature Reviews. Drug Discovery* 15 (7): 473–84.
- Wishart, David S. 2019. "Metabolomics for Investigating Physiological and Pathophysiological Processes." *Physiological Reviews* 99 (4): 1819–75.
- Wishart, David S., Anchi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, et al. 2022. "HMDB 5.0: The Human Metabolome Database for 2022." *Nucleic Acids Research* 50 (D1): D622–31.
- Wishart, David S., Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, et al. 2007. "HMDB: The Human Metabolome Database." *Nucleic Acids Research* 35 (Database issue): D521–26.
- Xiao, Nan, Dong-Sheng Cao, Min-Feng Zhu, and Qing-Song Xu. 2015. "protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences." *Bioinformatics (Oxford, England)* 31 (11): 1857–59.
- Yadav, Prakarsh, Parisa Mollaei, Zhonglin Cao, Yuyang Wang, and Amir Barati Farimani. 2022. "Prediction of GPCR Activity Using Machine Learning." *Computational and Structural Biotechnology Journal* 20 (May):2564–73.
- Yamane, Haruki, and Takashi Ishida. 2023. "Helix Encoder: A Compound-Protein Interaction Prediction Model Specifically Designed for Class A GPCRs." *Frontiers in Bioinformatics* 3 (May):1193025.
- 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): 7.
- Yang, Ming, Tomoyoshi Soga, and Patrick J. Pollard. 2013. "Oncometabolites: Linking Altered Metabolism with Cancer." *The Journal of Clinical Investigation* 123 (9): 3652–58.
- Zeng, Zhen, Chunxiang Ma, Kexin Chen, Mingshan Jiang, Reshma Vasu, Rui Liu, Yinglan Zhao, and Hu Zhang. 2021. "Roles of G Protein-Coupled Receptors (GPCRs) in Gastrointestinal Cancers: Focus on Sphingosine 1-Shosphate Receptors, Angiotensin II Receptors, and Estrogen-Related GPCRs." *Cells (Basel, Switzerland)* 10 (11): 2988.

- Zhang, Ai-Hua, Zhi-Ming Ma, Ling Kong, Hong-Lei Gao, Hui Sun, Xiang-Qian Wang, Jing-Bo Yu, Ying Han, Guang-Li Yan, and Xi-Jun Wang. 2020. "High-Throughput Lipidomics Analysis to Discover Lipid Biomarkers and Profiles as Potential Targets for Evaluating Efficacy of Kai-Xin-San against APP/PS1 Transgenic Mice Based on UPLC-Q/TOF-MS." *Biomedical Chromatography: BMC* 34 (2): e4724.
- Zhang, Ai-Hua, Zhi-Ming Ma, Hui Sun, Ying Zhang, Jian-Hua Liu, Fang-Fang Wu, and Xi-Jun Wang. 2019. "High-Throughput Metabolomics Evaluate the Efficacy of Total Lignans from *Acanthopanax Senticosus* Stem against Ovariectomized Osteoporosis Rat." *Frontiers in Pharmacology* 10 (May):553.
- Zhang, Haiping, Hongjie Fan, Jixia Wang, Tao Hou, Konda Mani Saravanan, Wei Xia, Hei Wun Kan, et al. 2024. "Revolutionizing GPCR-Ligand Predictions: DeepGPCR with Experimental Validation for High-Precision Drug Discovery." *Briefings in Bioinformatics* 25 (4): bbae281.
- 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, Li, Haixin Ai, Wen Chen, Zimo Yin, Huan Hu, Junfeng Zhu, Jian Zhao, Qi Zhao, and Hongsheng Liu. 2017. "CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods." *Scientific Reports* 7 (1): 2118.
- Zhang, Mingyang, Ting Chen, Xun Lu, Xiaobing Lan, Ziqiang Chen, and Shaoyong Lu. 2024. "G Protein-Coupled Receptors (GPCRs): Advances in Structures, Mechanisms, and Drug Discovery." *Signal Transduction and Targeted Therapy* 9 (1): 88.
- 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 (Clifton, N.J.)* 1324:191–203.
- Zhou, Zhigang, Qianhuan Dai, and Tong Gu. 2003. "A QSAR Model of PAHs Carcinogenesis Based on Thermodynamic Stabilities of Biactive Sites." *ChemInform* 34 (21). <https://doi.org/10.1002/chin.200321214>.