



Universiteit
Leiden
The Netherlands

Getting personal: advancing personalized oncology through computational analysis of membrane proteins

Gorostiola González, M.

Citation

Gorostiola González, M. (2025, January 24). *Getting personal: advancing personalized oncology through computational analysis of membrane proteins*. Retrieved from <https://hdl.handle.net/1887/4093962>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

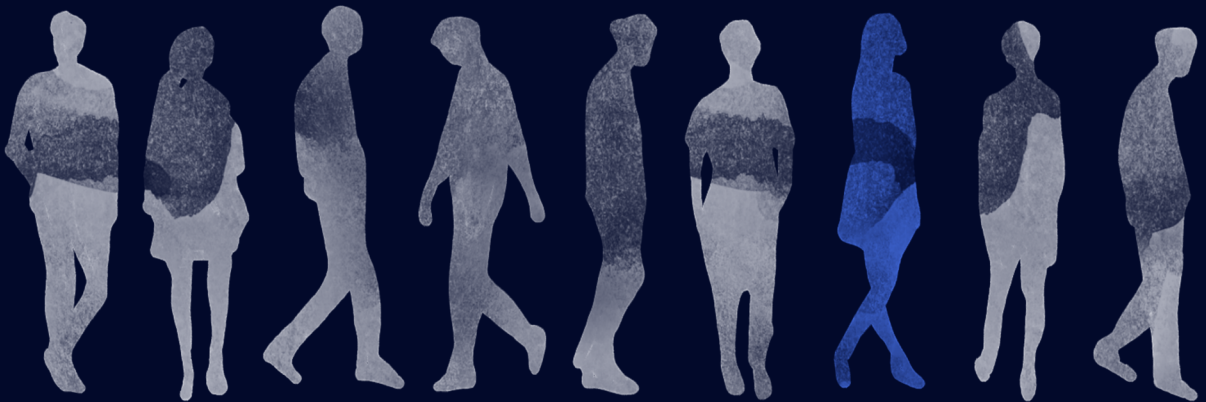
Downloaded from: <https://hdl.handle.net/1887/4093962>

Note: To cite this publication please use the final published version (if applicable).

GETT**1**NG PERS**0**NAL

Advancing personalized oncology through
computational analysis of membrane proteins

Marina Gorostiola González



Cover design and thesis layout by Marina Gorostiola González, Fernando da Silva Domingo, and Miriam González López.

This thesis was printed by Ipskamp Printing

© Marina Gorostiola González 2024

ISBN: 978-94-6473-615-1

All rights reserved. No part of this thesis may be reproduced in any form or by any means without permission of the author.

GETTING PERSONAL

Advancing personalized oncology through computational analysis of
membrane proteins

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 24 januari 2025
klokke 11.30 uur

door

Marina Gorostiola González
geboren te Medina de Pomar, Spanje
in 1994

Promotores: Prof. dr. G.J.P. van Westen
Prof. dr. L.H. Heitman
Prof. dr. A.P. IJzerman

Promotiecommissie: Prof. dr. H. Irth
Prof. dr. E.C.M. de Lange
Prof. dr. E.H.J. Danen
Prof. dr. A. Plaat
Prof. dr. A. Volkamer, Saarland University
Dr. E.L. Willighagen, Maastricht University

The research described in this thesis was performed at the Division of Drug Discovery and Safety of the Leiden Academic Centre for Drug Research (LACDR), Leiden University, The Netherlands.

To mom, dad, and anyone else who has received
the terrifying diagnosis of an undruggable tumor

About this thesis title and cover:

The primary challenge in treating cancer stems from its extensive heterogeneity, resulting in a scenario where each patient presents with a distinct disease defined by their unique genetic profile. While this complexity may seem overwhelming, computational tools offer a valuable solution for compiling individual data from a large number of patients, ultimately identifying biomarkers and targets to enable personalized diagnosis and treatment for specific subpopulations. Essentially, the tools developed in this thesis facilitate the customization of oncological treatment for each patient. In short, we are Getting Personal.

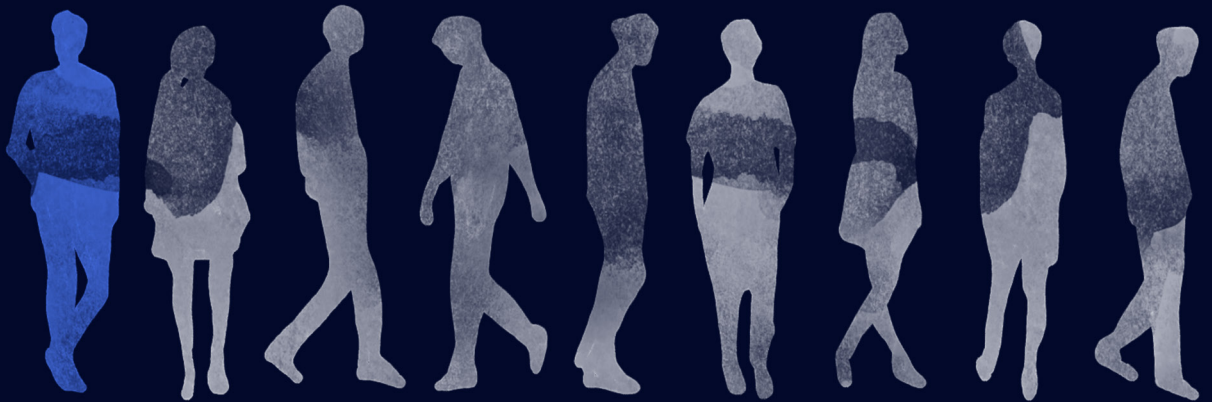
TABLE OF CONTENTS

Chapter 1	General introduction	11
Chapter 2	Oncological drug discovery: AI meets structure-based computational research	23
Chapter 3	Computational characterization of membrane proteins as anticancer targets: Current challenges and opportunities	43
Chapter 4	Excuse me, there is a mutant in my bioactivity soup! A comprehensive analysis of the genetic variability landscape of bioactivity databases and its effect on activity modeling	65
Chapter 5	Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors	123
Chapter 6	Molecular insights into disease-associated glutamate transporter (EAAT1 / SLC1A3) variants using <i>in silico</i> and <i>in vitro</i> approaches	157
Chapter 7	3DDPDs: Describing protein dynamics for proteochemometric bioactivity prediction. A case for (mutant) G protein-coupled receptors	201
Chapter 8	Connecting the dots: A patient-centric knowledge graph approach to prioritize mutants for selective anticancer targeting	241
Chapter 9	General conclusions and future perspectives	281
Appendix A	GDC SQL implementation v22.0. Quick start guide	293
Appendix B	Data and software availability	305
	Summary, List of publications, Curriculum vitae, Acknowledgements	309



Chapter 1

General introduction



Personalized oncology. Promises and challenges

Cancer research has advanced immensely in the last decades, which has materialized in novel diagnosis and treatment opportunities^{1,2}. In turn, this has translated into a decrease in cancer mortality rate despite a sustained increase in cancer incidence worldwide^{3,4}. Unfortunately, the burden of a cancer diagnosis extends beyond morbidity. Several studies have shown the high psychosocial impact of cancer on patients, caretakers, and medical professionals^{5,6}. The harshness of the treatments received, which lead to very serious acute and chronic side effects, constitutes a big factor weighting in⁶. Personalized therapies that exploit the heterogeneity of the disease have emerged as a solution, not only to improve efficacy to eradicate the tumor, but also to optimize treatment regimes, reduce side effects, and decrease the risk of relapse⁷⁻⁹.

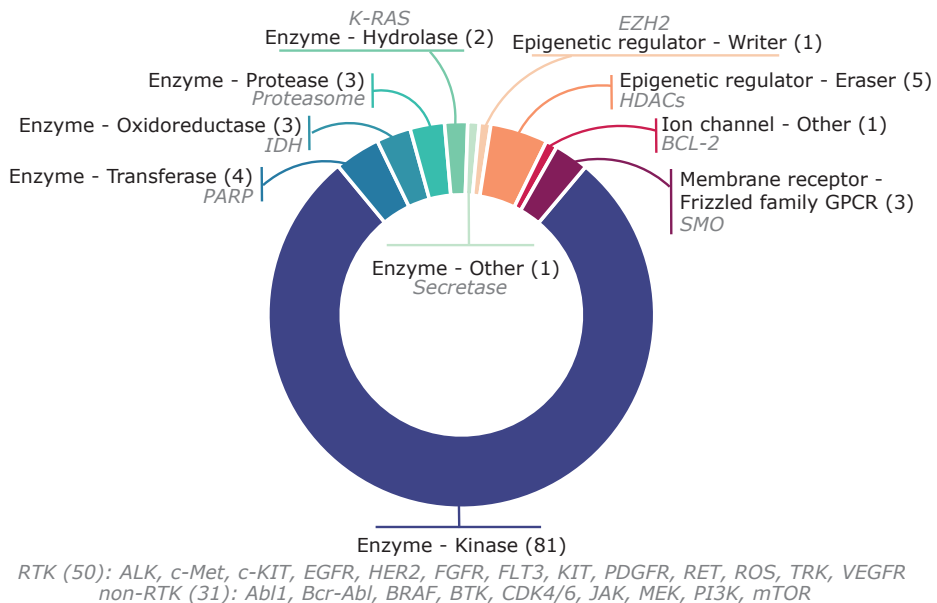


Figure 1.1. FDA-approved small molecule anticancer targeted agents from 2001 to 2023. The 104 approved drugs are distributed per target family according to the ChEMBL L1/L2 classification.

Personalized oncology comprises several therapeutical strategies that can be used when the patient meets certain specific profiling criteria⁷. This is in contrast to the “one size fits all” traditional model where general chemotherapy, radiation, or surgery treatment plans are drafted upon diagnosis of a tissue-specific tumor in a certain development stage⁸. In the personalized model, different biomarkers are used to stratify subpopulations that can benefit from specific therapies or combinations of therapies. While the location of the primary tumor and its metastases is still considered in the stratification, other biomarkers obtained via multiple “omics” analyses tend to define the therapeutic plan. These include DNA alterations such as point mutations and amplifications/deletions (genomics), but also divergences from the norm in gene and protein expression (transcriptomics, proteomics) or metabolite concentration (metabolomics)^{7,8}.

Most commonly, and throughout this thesis, I refer to targeted therapy when talking about personalized oncology, although other modalities exist such as immunotherapy, CAR-T cell therapies, or cancer vaccines⁸. Targeted therapies exploit cancer-specific traits to attack preferentially tumor tissue while avoiding healthy cells thus reducing side effects^{10–12}. This effect can be triggered by biological agents, such as monoclonal antibodies, or by small molecules, which will be the focus of this thesis¹¹. Since the approval in 2001 of the first anticancer-targeted small molecule, imatinib, 104 small molecules have been approved for anticancer treatment¹⁰. However, while the eligibility of patients for targeted therapies is increasing, it was still estimated to be less than 15% in 2020¹³.

Although substantial effort is sustained to develop new targeted therapies, the currently approved small molecules target a very limited range of proteins, of which the vast majority are kinases (**Figure 1.1**)^{10,14–16}. The associated costs to develop a new targeted drug are very elevated, and their success rate in clinical trials can be limited¹⁷. Several factors contribute to these failures, including the high incidence of therapy resistance and the use of targeted therapies only after other approaches have failed. However, the common underlying cause is still the very incomplete knowledge of cancer biology and how it is affected by inter-patient heterogeneity^{7,12,18}.

Smart prioritization of targets and small molecules via computational approaches

Computational drug discovery has emerged as a time- and cost-efficient way to prioritize targets and small molecules to pursue in therapeutics¹⁹. These methods have been integrated with molecular biology and medicinal chemistry in the early stages of the drug discovery pipeline to highlight the most promising candidates. In particular, in oncological research, these approaches can be highly beneficial in addressing the diversity of neoplastic diseases²⁰. In fact, many authors agree that the future of personalized oncology goes hand in hand with advances in the computationally driven exploration of the vast amounts of data generated^{9,12,21}.

The computational analysis of multi-omics data has proven invaluable in helping pinpoint the differences between patient subpopulations and highlight potential anticancer targets^{22–25}. Building on top of this preselection, there are three main levels where computational tools can be used to accelerate the early drug discovery pipeline in personalized oncology (**Figure 1.2**). Firstly, computational methods can further prioritize targets and alterations with predicted functional relevance^{26,27}. Secondly, further down the line towards drug discovery, the druggability of particular genetic alterations can be assessed by analyzing the structural differences that are triggered in the target of interest upon mutation²⁸. Finally, candidate drugs can be screened *in silico* to prioritize the most promising lead compounds targeting a specific target or genetic alteration with high potency and selectivity²⁹. Importantly, this multi-level prioritization can be linked to additional selection criteria to improve the success of candidate therapies by, for example, increasing the threshold to develop therapy resistance. This can be achieved by prioritizing targets in central pathways that can be targeted on key structural motifs with highly flexible molecules³⁰.

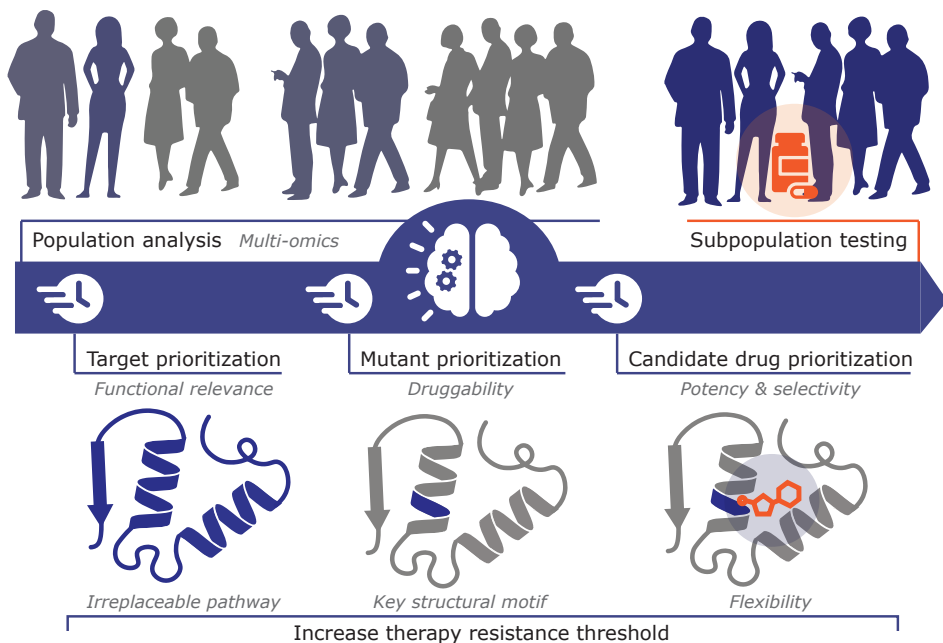


Figure 1.2. Three levels of computationally driven prioritization to accelerate personalized small molecule hit identification.

The methods used in computational drug discovery can broadly be divided into data-driven and structure-based (SB) approaches. The former class includes artificial intelligence (AI) and machine learning (ML), together with other statistical analyses. When applied to multi-omics data, data-driven tools allow us to predict cancer drivers, as well as to identify biomarkers responsible for phenotypical differences in patient subpopulations³¹. Applied to medicinal chemistry data, data-driven tools – then commonly termed ligand-based approaches – can be used to predict the characteristics of small molecules with high affinity and/or selectivity towards a target of interest. Such knowledge enables virtual screening or *de novo* generation of candidate drugs^{32,33}.

SB drug discovery, on the other hand, englobes applications dependent on the 3D structure of the target of interest and the underlying forces driving interactions between biological systems and small molecules. From the structure of a protein – experimentally determined by X-ray crystallography or Cryo-electron microscopy, modeled, or predicted with AI models such as AlphaFold³⁴, methods such as docking can be used to predict the binding mode of candidate drugs in a target of interest. Moreover, one can perform molecular dynamics (MD) simulations to explore the protein's dynamic profile. More computationally expensive methods, such as free energy perturbation (FEP), even support the calculation of binding affinities from protein-ligand complexes³³.

Standalone computational methods have been able to provide very relevant information leading to target and hit identification. However, one of the most promising outlooks following the increase in data availability and computational power is the integration of

data-driven and structural-based approaches. Particularly in oncological drug discovery, this combination can be key to tackling the complexity of the disease and provide the necessary insights to prioritize the right targets and candidate small molecules. Current methods on this front, as well as challenges and future opportunities, are explored in more detail in **Chapter 2**.

Membrane proteins as targets in personalized oncology

One of the most exciting applications of the use of computational tools in the oncological drug discovery pipeline is the possibility of expanding beyond the current clinically validated anticancer targets²¹. This opens opportunities to target novel pathways and increase patient eligibility for personalized treatments. More importantly, it facilitates the exploration of protein families that are particularly challenging to study experimentally, such as membrane proteins³⁵.

The location of membrane proteins at the cellular membrane makes them key players in the initiation of signaling cascades. In tumor cells, the aberrant initiation and propagation of signals to the cytoplasm and nucleus are directly linked to alterations in key hallmarks of cancer such as sustained cellular proliferation, evading growth suppressors, and resisting cell death^{36–38}. Moreover, thanks to their privileged location on the cellular surface, they constitute excellent biomarker and drug target candidates³⁹. The role of certain protein membrane families in cancer, particularly receptor tyrosine kinases (RTKs) has been extensively highlighted⁴⁰. In fact, almost 50% of the FDA-approved targeted anticancer small molecules target RTKs such as EGFR, ALK, or FLT3 (**Figure 1.1**). This is with good reason since these membrane receptors initiate the MAPK, JAK/STAT, and P13K/AKT/mTOR kinase cascades, which are at the center of the cancer development pathways, and are highly dysregulated in cancer patients⁴⁰.

Aside from RTKs, other membrane protein families are largely underexplored in the context of cancer, which I reviewed in **Chapter 3**. Only three non-RTK membrane proteins are the targets of anticancer drugs, namely class F G protein-coupled receptor Smoothed (SMO), ion channel B-cell lymphoma 2 (BCL-2), and enzyme γ -secretase^{10,14–16}. This disparity is also exemplified by the imbalance in the literature linking cancer to RTKs compared to the two largest membrane protein families, G protein-coupled receptors (GPCRs) and solute carriers (SLCs) (**Figure 1.3**). For reference, human receptor kinases comprise 58 genes while GPCRs and SLCs comprise around 800 and over 400 genes, respectively^{41–43}. However, new proteins are constantly annotated and these numbers could be higher as predicted based on functional and evolutionary conservation⁴⁴. GPCRs are the major signal-transducing receptors of the cell and the targets of approximately 35% of all approved drugs^{45,46}. The involvement of GPCRs in cancer has been increasingly highlighted, with patients showing hyperactivation or abnormal expression of certain receptors in the tumor tissue and the tumor microenvironment alike⁴⁷. Subsequently, GPCRs are gaining interest as anticancer targets, with some inhibitors in clinical trials particularly as immunotherapy⁴⁸. However, the underlying mechanisms of their role in cancer development need to be studied in further detail to lead to successful therapeutic strategies^{47,48}. SLCs, on the other hand, have

been historically neglected as therapeutical targets and only recently have attracted more attention from the scientific community⁴⁹. Among other substrates, SLCs transport metabolites, neurotransmitters, amino acids and ions, and their expression is dysregulated in several cancer types⁵⁰.

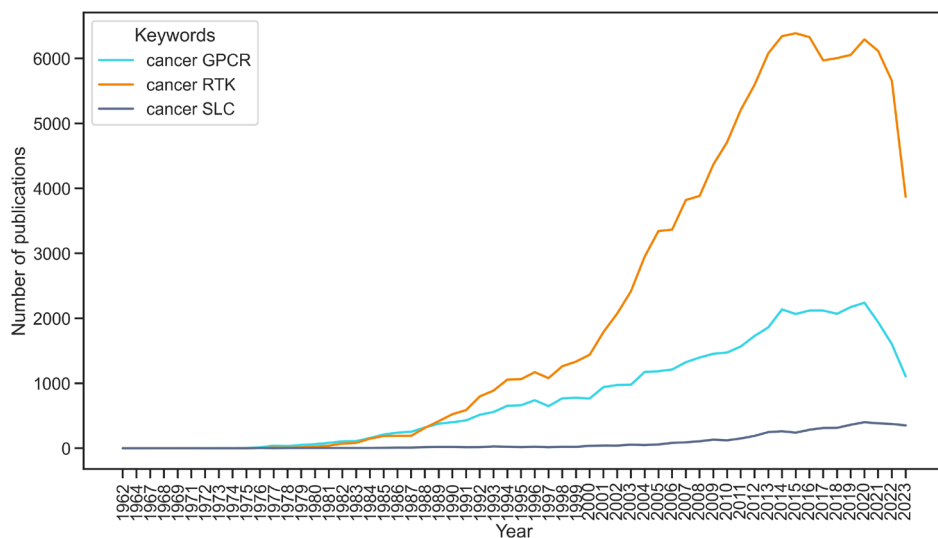


Figure 1.3. Number of publications retrieved from PubMed with the combination of keywords “cancer” and three membrane protein families: RTKs, GPCRs, and SLCs. Data was retrieved in November 2023, therefore the number of publications related to years 2020-2023 shows a drop corresponding to publication embargoes and delayed publication dates.

While the use of computational analysis of membrane proteins in the context of cancer is very promising it is, however, not exempt from challenges. The experimental difficulties linked to the study of membrane proteins result in reduced data availability, which is highly detrimental in the application of data-driven methods such as ML. Similarly, 3D structures of membrane proteins are more difficult to obtain and their conditions are more difficult to simulate, which hinders SB approaches. In **Chapter 3** I explore in detail the challenges associated with the computational analysis of membrane proteins and, in particular, GPCRs and SLCs as novel anticancer targets and the strategies available to circumvent these. Moreover, I highlight the computationally driven opportunities to improve therapeutical strategies in already established anticancer targets, namely RTKs.

Aim and outline of this thesis

This thesis aims to combine data-driven and SB computational approaches to prioritize membrane proteins as novel or improved personalized anticancer targets.

In **Chapter 2**, a selection of applications is reviewed where the integration of AI and SB methods is used to shed light on six case scenarios relevant to the oncological drug discovery pipeline. These include driver prediction, computational mutagenesis, (off)-target prediction, binding site prediction, virtual screening, and allosteric modulation analysis.

Then, in **Chapter 3**, the inherent challenges for the study of membrane proteins with computational tools as opposed to their soluble counterparts are addressed. In particular, the importance of data availability and publication bias in the context of anticancer target research is addressed. To this end, three membrane protein families with different levels of representation in the literature are exemplified: RTKs, GPCRs, and SLCs.

The topic of data availability is a constant throughout the thesis, but it is explored in detail in **Chapter 4**. Here, the available data for mutant proteins is analyzed in the most widely used public bioactivity database in computational drug discovery, ChEMBL. Subsequently, the effect this data has on bioactivity modeling is explored, thus uncovering the potential for mutant bioactivity prediction as well as the existing risk of introducing noise in wild-type modeling.

In **Chapters 5-7**, computational applications were developed aimed to accelerate the oncological drug discovery pipeline at the three levels summarized in **Figure 1.2**: target, mutant, and candidate drug prioritization. The applications in these chapters are exemplified in the three previously highlighted membrane protein families.

Chapter 5 focuses on the prioritization of GPCRs as anticancer targets based on the pan-cancer analysis of receptor somatic mutation data. This data-driven approach allowed us to identify functionally relevant highly conserved motifs as mutational hotspots in GPCRs and subsequently underline receptors with high mutation frequency in these hotspots as potential anticancer targets with functional relevance. Additionally, to support the multi-omics analyses performed in this and the following chapters, a comprehensive SQL image of the Genomic Data Commons⁵¹ data was developed to support computational analysis.

In **Chapter 6**, an SB approach was developed to analyze the effect of cancer patient-derived point mutations in SLC glutamate transporter EAAT1. A combination of docking and MD was used to analyze the impact of six cancer-related mutations on the transporter structure, function, and druggability. The results from this analysis, together with *in vitro* characterization of the mutants, provided the necessary insights to prioritize somatic mutations as potential druggable alterations.

The integration of data-driven and SB approaches was exemplified in **Chapter 7** for the prioritization of candidate drugs as (mutant) GPCR inhibitors. This approach was based on the development of MD-based protein descriptors for proteochemometric

bioactivity modeling; 3DDPDs. This combination resulted in improved predictive performance of the models while retaining high interpretability. Although the bioactivity predictive performance could not be tested on mutant GPCRs due to the lack of data availability, the 3DDPDs showed a potential to distinguish between mutants based on their dynamic profile.

Chapter 8 explores the application of holistic approaches to suggest mutated proteins as anticancer targets. This was possible to do for the membrane protein family with the most amount of data available, RTKs. A patient-centric knowledge graph was used to integrate a vast amount of kinome data, including cancer-related omics, pathways, bioactivity, and structural data. The graph enabled the analysis of the characteristics of RTK cancer mutations with the potential to be targeted selectively while suffering from the smallest therapy resistance.

Finally, in **Chapter 9**, general conclusions from the previous chapters are drawn in light of the thesis aim previously presented. The major challenges remaining are delineated, together with the future perspectives for successfully applying computational approaches to accelerate the discovery of novel personalized anticancer treatments targeting membrane proteins.

References

- Hackshaw, A., Clarke, C. A. & Hartman, A.-R. New genomic technologies for multi-cancer early detection: Rethinking the scope of cancer screening. *Cancer Cell* **40**, 109–113 (2022).
- Debela, D. T. *et al.* New approaches and procedures for cancer treatment: Current perspectives. *SAGE Open Med.* **9**, 1–10 (2021).
- Hashim, D. *et al.* The global decrease in cancer mortality: trends and disparities. *Ann. Oncol.* **27**, 926–933 (2016).
- Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *C.A. Cancer J. Clin.* **71**, 209–249 (2021).
- Wang, Y. & Feng, W. Cancer-related psychosocial challenges. *Gen. Psychiatry* **35**, e100871 (2022).
- Niedzwiedz, C. L., Knifton, L., Robb, K. A., Katikireddi, S. V. & Smith, D. J. Depression and anxiety among people living with and beyond cancer: a growing clinical and research priority. *BMC Cancer* **19**, 943 (2019).
- Hoeben, A., Joosten, E. A. J. & van den Beuken-Van Everdingen, M. H. J. Personalized medicine: Recent progress in cancer therapy. *Cancers* **13**, 1–3 (2021).
- Krzyszczak, P. *et al.* The growing role of precision and personalized medicine for cancer treatment. *Technology* **6**, 79–100 (2018).
- Lassen, U. N. *et al.* Precision oncology: a clinical and patient perspective. *Future Oncol.* **17**, 3995–4009 (2021).
- Zhong, L. *et al.* Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct. Target. Ther.* **6**, 1–48 (2021).
- Min, H.-Y. & Lee, H.-Y. Molecular targeted therapy for anticancer treatment. *Exp. Mol. Med.* **54**, 1670–1694 (2022).
- Waarts, M. R., Stonestrom, A. J., Park, Y. C. & Levine, R. L. Targeting mutations in cancer. *J. Clin. Invest.* **132**, e154943 (2022).
- Haslam, A., Kim, M. S. & Prasad, V. Updated estimates of eligibility for and response to genome-targeted oncology drugs among US cancer patients, 2006–2020. *Ann. Oncol.* **32**, 926–932 (2021).
- Mullard, A. 2021 FDA approvals. *Nat. Rev. Drug Discov.* **21**, 83–88 (2022).
- Mullard, A. 2022 FDA approvals. *Nat. Rev. Drug Discov.* **22**, 83–88 (2023).
- Mullard, A. 2023 FDA approvals. *Nat. Rev. Drug Discov.* **88**, 88–95 (2024).
- O'Dwyer, P. J. *et al.* The NCI-MATCH trial: lessons for precision oncology. *Nat. Med.* **29**, 1349–1357 (2023).
- Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2020 update. *Pharmacol. Res.* **152**, 104609 (2020).
- Hinkson, I. V., Madej, B. & Stahlberg, E. A. Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery. *Front. Pharmacol.* **11**, 770 (2020).
- Rahman, M. M. *et al.* Emerging Promise of Computational Techniques in Anti-Cancer Research: At a Glance. *Bioengineering* **9**, 335 (2022).
- Stuart, D. D. *et al.* Precision Oncology Comes of Age: Designing Best-in-Class Small Molecules by Integrating Two Decades of Advances in Chemistry, Target Biology, and Data Science. *Cancer Discov.* **13**, 2131–2149 (2023).
- Reyna, M. A. *et al.* Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **11**, 729 (2020).
- Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
- Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
- Dinalankara, W. *et al.* Digitizing omics profiles by divergence from a baseline. *Proc. Natl. Acad. Sci.* **115**, 4545–4552 (2018).
- Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385 (2018).
- Chitralla, K. N., Nagarkatti, M., Nagarkatti, P. & Yeguvapalli, S. Analysis of the TP53 deleterious single nucleotide polymorphisms impact on estrogen receptor alpha-p53 interaction: A machine learning approach. *Int. J. Mol. Sci.* **20**, 2962 (2019).
- Malhotra, S. *et al.* Understanding the impacts of missense mutations on structures and functions of human cancer-related genes: A preliminary computational analysis of the COSMIC Cancer Gene Census. *PLOS ONE* **14**, e0219935 (2019).
- Pinto, G. P., Hendrikse, N. M., Stourac, J., Damborsky, J. & Bednar, D. Virtual screening of potential anticancer drugs based on microbial products. *Semin. Cancer Biol.* **86**, 1207–1217 (2022).
- Robichaux, J. P. *et al.* Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* **597**, 732–737 (2021).
- Arjmand, B. *et al.* Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front. Genet.* **13**, 824451 (2022).
- Qureshi, R. *et al.* AI in drug discovery and its clinical relevance. *Helvion* **9**, e17575 (2023).
- Sadybekov, A. V. & Katritch, V. Computational

- approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
34. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
 35. de Jong, E. & Kocer, A. Current Methods for Identifying Plasma Membrane Proteins as Cancer Biomarkers. *Membranes* **13**, 409 (2023).
 36. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
 37. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
 38. Lazar, I. M., Karcini, A. & Haucis, J. R. S. Mapping the Cell-Membrane Proteome to the Cancer Hallmarks. Preprint at *BioRxiv* <https://doi.org/10.1101/2022.03.18.484818> (2022).
 39. Boonstra, M. C. *et al.* Selecting Targets for Tumor Imaging: An Overview of Cancer-Associated Membrane Proteins. *Biomark. Cancer* **8**, 119–133 (2016).
 40. Saraon, P. *et al.* Receptor tyrosine kinases and cancer: oncogenic mechanisms and therapeutic approaches. *Oncogene* **40**, 4079–4093 (2021).
 41. Alexander, S. P. H. *et al.* The concise guide to pharmacology 2019/20: Transporters. *Br. J. Pharmacol.* **176**, S397–S493 (2019).
 42. Alexander, S. P. H. *et al.* The concise guide to pharmacology 2019/20: Enzymes. *Br. J. Pharmacol.* **176**, S297–S396 (2019).
 43. Alexander, S. P. H. *et al.* The concise guide to pharmacology 2019/20: G protein-coupled receptors. *Br. J. Pharmacol.* **176**, S21–S141 (2019).
 44. Almén, M. S., Nordström, K. J., Fredriksson, R. & Schiöth, H. B. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.* **7**, 50 (2009).
 45. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-Function of the G-protein-Coupled Receptor Superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53**, 531–556 (2013).
 46. Sriram, K. & Insel, P. A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol. Pharmacol.* **93**, 251–258 (2018).
 47. Chaudhary, P. K. & Kim, S. An Insight into GPCR and G-Proteins as Cancer Drivers. *Cells* **10**, 3288 (2021).
 48. Usman, S., Khawer, M., Rafique, S., Naz, Z. & Saleem, K. The current status of anti-GPCR drugs against different cancers. *J. Pharm. Anal.* **10**, 517–521 (2020).
 49. César-Razquin, A. *et al.* A Call for Systematic Research on Solute Carriers. *Cell* **162**, 478–487 (2015).
 50. Lavoro, A. *et al.* In silico analysis of the solute carrier (SLC) family in cancer indicates a link among DNA methylation, metabolic adaptation, drug response, and immune reactivity. *Front. Pharmacol.* **14**, 1191262 (2023).
 51. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).

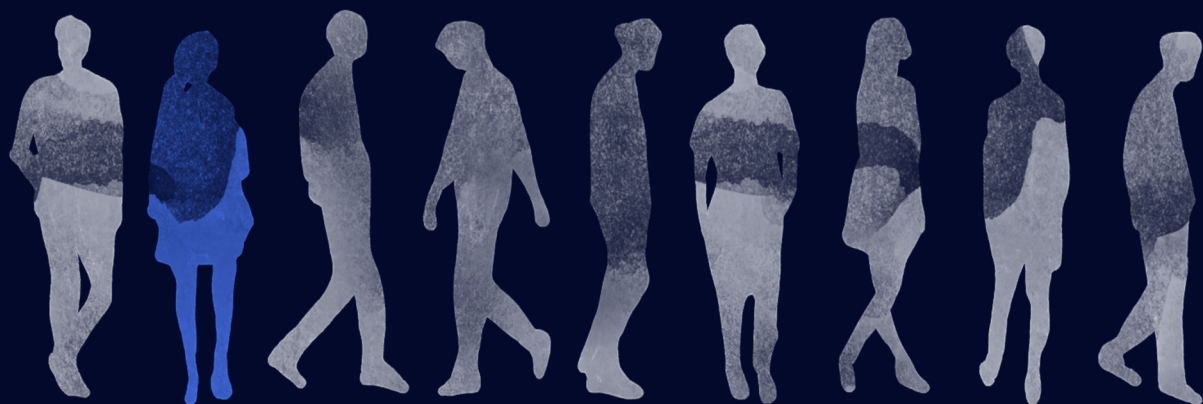


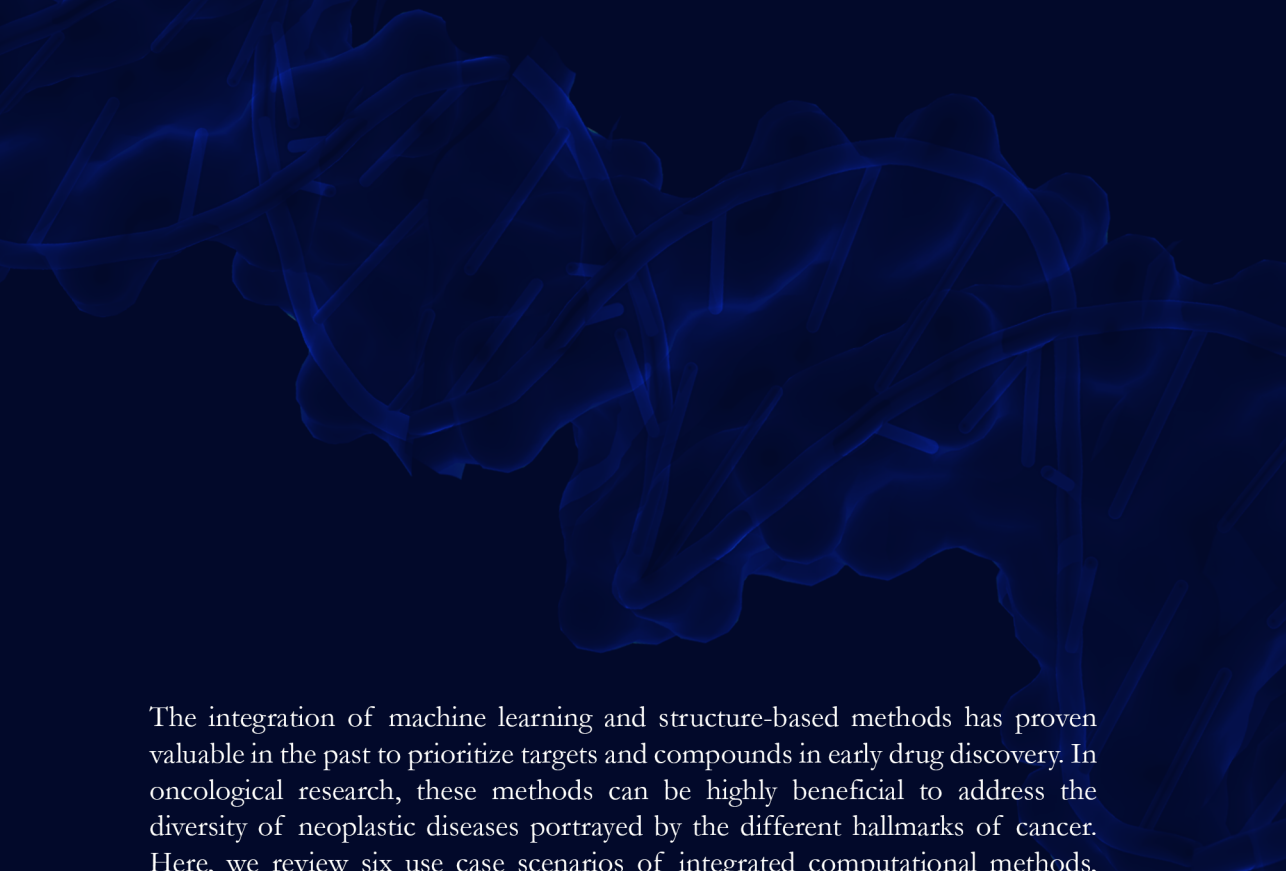
Chapter **2**

Oncological drug discovery: AI meets
structure-based computational research

Marina Gorostiola González, Antonius P.A. Janssen, Adriaan P. IJzerman,
Laura H. Heitman, Gerard J.P. van Westen

Adapted from: *Drug Discovery Today* **27**, 1661-1670 (2022)





The integration of machine learning and structure-based methods has proven valuable in the past to prioritize targets and compounds in early drug discovery. In oncological research, these methods can be highly beneficial to address the diversity of neoplastic diseases portrayed by the different hallmarks of cancer. Here, we review six use case scenarios of integrated computational methods, namely driver prediction, computational mutagenesis, (off)-target prediction, binding site prediction, virtual screening, and allosteric modulation analysis. We address the heterogeneity of integration approaches and individual methods while acknowledging their current limitations and highlighting their potential to bring drugs for oncological personalized therapies to the market faster.



Introduction

In recent years, the scientific community has seen an increased usage of computational approaches to accelerate the discovery of relevant targets and prioritize small molecules in all disease areas. These include data-driven artificial intelligence (AI) / machine learning (ML)^{1,2}, as well as structure-based (SB) methods, such as docking and molecular dynamics (MD)³. Moreover, the advances in computing power and experimental structure elucidation have made it possible to integrate these two types of methods for example to use ML-based scoring functions to rank the accuracy of docking results⁴ or use structure-derived data (e.g. interaction fingerprints or MD trajectories) as input for bioactivity prediction models^{5,6}. These advances have emerged as a joint effort of the computational drug discovery community and are generally applicable to the subfield of oncological drug discovery, which shares most of the challenges and characteristics of drug discovery in broader terms. However, it also entails its own unique traits, as represented by the complexity and diversity of neoplastic diseases summarized in the hallmarks of cancer (**Box 2.1**)^{7,8}. Understanding this diversity is an additional key aspect for the development of personalized anticancer treatments, which are increasingly being deployed in the clinical practice^{9,10}. Combined, the (computational) drug discovery field is gradually moving towards cancer-specific applications and/or demonstrating applicability in cancer-related targets.

Here, we review the efforts made to integrate AI/ML and SB methods in computational drug discovery that are specifically being applied or can potentially impact the field of cancer research (**Table 2.1**). The articles reviewed cover different parts of the oncology drug discovery pipeline, where we focus on six computational use case scenarios and four integration methods (**Figure 2.1**). In the following sections, we approach each of these use scenarios, namely driver prediction, computational mutagenesis, (off)-target prediction, binding site prediction, virtual screening (VS), and allosteric modulation analysis. ML-SB integration methods are classified to cover (A) the use of structural data as input for ML models, (B) ML-based scoring functions for SB applications, (C) ML as a tool to analyze MD simulations, and (D) sequential or parallel pipelines where SB and ML methods are used independently but complementarily. The biological impact in cancer research is exemplified by the link of the targets addressed in the reviewed publications to each of the ten defined hallmarks of cancer, as well as an additional eleventh “hallmark” of high relevance in oncological drug discovery, namely chemotherapy escaping capabilities (**Box 2.1**). The heterogeneity of use cases and methods (**Table 2.1**) goes hand in hand with that of molecular targets covered and illustrates the diverse potential of the combined use of AI and SB methods in oncological drug discovery.

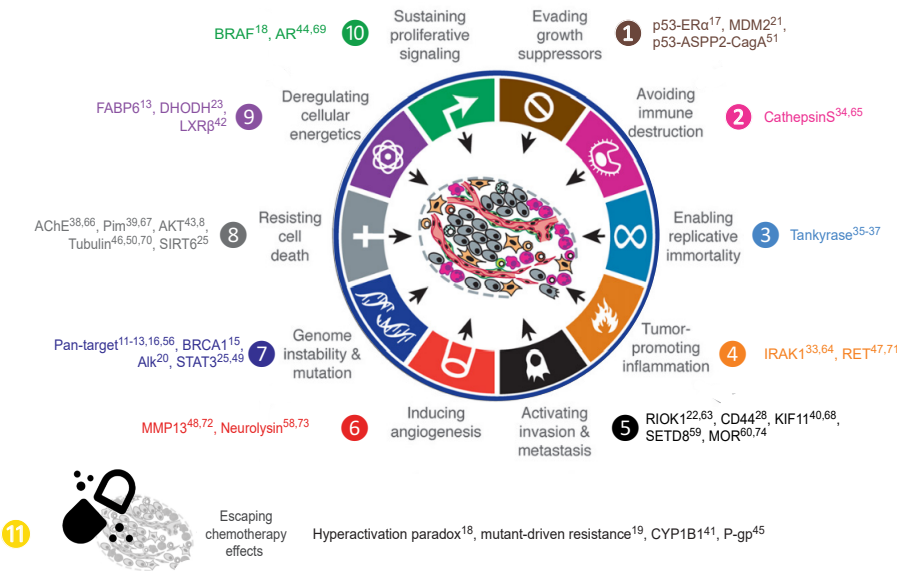
Driver prediction

One of the main use case scenarios of computational cancer research, most frequently ML-based, is the prediction of gene and mutation drivers to prioritize in anticancer therapies. These approaches are by definition pan-target and usually pan-cancer, i.e. not focused on specific targets or cancer types. They often start from multi-omics data from

cancer patients, such as the TCGA's somatic mutations^{11,12}, copy number variations¹², epigenetic¹², or RNAseq¹³ data, and their applicability depends on the availability of such data types. The work of Bailey *et al.*¹¹ provides an extensive overview of the wide array of tools available for driver prediction and, more importantly, the importance of combining different tools to maximize predictive performance. While the approach from Bailey *et al.* joined SB and ML methods in parallel, they are more frequently incorporated sequentially^{12,13}. Knijnenburg *et al.*¹² and Liñares-Blanco *et al.*¹³ created classification models (logistic regression and random forest - RF, respectively) trained on omics data to predict cancer-related outcomes such as homologous recombination deficiency and tumorigenic phenotype. In both cases, feature importance was used to prioritize genes for further SB analysis. In the case of Knijnenburg *et al.*¹², *in silico* mutagenesis studies were performed for each detected variant with a potential effect on protein stability.

Box 2.1. Targeting the hallmarks of cancer

In their description of the hallmarks of cancer, Hanahan and Weinberg (2000)⁷ defined six underlying traits that are common to tumorigenesis. In the light of new evidence, these were later complemented by two additional emerging hallmarks and two enabling characteristics⁸. These hallmarks paved the way to understanding the complexity and diversity of neoplastic diseases. Understanding this diversity is a key aspect of the development of personalized anticancer treatments. A combination of artificial intelligence (AI) and structure-based methods can be used to address cancer drug discovery research in a more holistic way, tackling all the hallmarks of cancer. In this review, we provide an overview of the biological relevance of the drug discovery targets in cancer and their relevance to the hallmarks and characteristics of cancer (numbered 1 to 10 in the box figure). An eleventh "hallmark", the ability of cancer cells to escape chemotherapy effects, is added here and is a key aspect to consider in oncology drug discovery strategies.



Some of the substitutions found were also analyzed with MD and appeared to alter protein dynamics even if they were not predicted to alter protein stability. Conversely, Liñares-Blanco *et al.*¹³ used the ML-derived information to perform a drug repurposing VS approach where FDA-approved anticancer drugs were docked into the available crystal structures of the computationally prioritized genes, such as FABP6.

It is important to note that the selection of input data, features, and outcome variables for cancer driver prediction is not homogeneous. In any case, key aspects such as the tumor microenvironment or metastasis are often neglected. Regarding cancer patient data, most of the publications use TCGA, which provides high-quality and standardized data. However, the TCGA data has been frozen since 2016, highlighting the need for updated cancer patient databases, such as the Genomic Data Commons¹⁴. Overall, the use of sequential pipelines – sometimes including experimental assays – could help account for the differential effect on tumorigenesis of the different available types of data.

Computational mutagenesis

Knowing the effect of specific point mutations on protein function and “druggability” is a key aspect for the development of personalized anticancer therapies as well as for decision-making in the clinic. *In vitro* mutagenesis studies are time- and cost-expensive, thus *in silico* computational studies are a good starting point to prioritize mutants for experimental analysis.

Most of the computational mutagenesis approaches reviewed here use structural data to train ML classifiers^{15–18}. Said structural data might originate directly from a crystal structure^{15,16}, combined with docking studies¹⁷, or MD¹⁸. The approaches developed by Masso *et al.*¹⁵ and Pandurangan *et al.*¹⁶ extract features from a geometrical representation derived from wild-type (WT) and mutant crystal structures and homology models. Those features are used in classification models to predict variant clinical significance and protein stability, respectively. Protein-protein interaction stability can also be predicted from protein-protein docking-derived features, as was done by Chitrula *et al.*¹⁷ for the p53-ER α interaction for WT and three breast cancer-related p53 polymorphisms. Moreover, computational mutagenesis studies are used to predict the effect of mutations in ligand binding dynamics. Babbitt *et al.*¹⁸ have studied the hyperactivating effect of BRAF V600E-targeting inhibitors in WT cells using MD. Here, differences in rapid dynamics in bound and unbound functional states for each amino acid were modeled in stacked classification models to detect conserved dynamic functions. They showed that the V600E mutation greatly alters dynamics, leading to lower predictive performance.

The performance of the classification models used for mutagenesis prediction varies highly depending on the amount of experimental mutagenesis data available for training and validation^{15–18}. Hence, some authors have evaluated the performance of SB methods alone compared to ML models for these tasks^{19,20}. For example, Aldeghi *et al.*¹⁹ benchmarked the performance of free energy perturbation (FEP), ML, and Monte Carlo methods to predict the change in affinity of inhibitors in Abl kinase variants. The classifier trained on a pan-target dataset was not able to generalize on the test set, but

when trained on a reduced Abl-specific dataset the performance was comparable to those of FEP and Monte Carlo methods. However, computational time was drastically reduced when using ML. Similarly, Patil *et al.*²⁰ created an MD protocol to determine the activation status of any kinase variant. This is critical information to prioritize kinase inhibitors that target the active or the inactive conformation hence preventing unwanted side effects. For that purpose, Alk kinase was selected as a case study. Here, long-term dynamics between the active and the inactive conformations were explored with metadynamics. Using results from RMSD changes and hydrogen bond occupation, a score was given for the WT and the mutant, and a final score was compared to a defined threshold. This approach outperformed a kinome-wide ML model and other common impact prediction tools, such as SIFT and Polyphen.

The here reviewed approaches in computational mutagenesis are able to capture differences in protein stability and conformation^{16,20}, protein-protein interactions¹⁷, ligand binding affinity and dynamics^{18,19}, and clinical significance¹⁵. Their applicability, however, is often limited to a particular target or mutant of interest for which there is enough data. In order to increase the impact of methods developed for members of families with highly conserved binding pockets and activation mechanisms, such as kinases (Babbitt *et al.*¹⁸, Aldeghi *et al.*¹⁹, Patil *et al.*²⁰) or G protein-coupled receptors, the training sets could be enriched with data from other members of the family. The efforts made in computational mutagenesis, therefore, could in general benefit from more extensive experimentally validated mutagenesis datasets, which should be deposited in publicly available databases following FAIR principles to favor the creation of relevant training and validation datasets.

(Off)-target prediction

Defining the (off)-target space of drugs in development is important to achieve a selective profile, but also to rationally design polypharmacological candidates, i.e. with a multi-target profile. Moreover, re-analyzing the target space of approved drugs is key to better understanding their mode of action, or to start re-purposing efforts. These endpoints are of high relevance in oncological drug discovery, where off-target effects are often responsible for grave adverse reactions. Integrated ML-SB methods have proven useful in these tasks.

The search of the target space usually starts from known information, such as ligand-protein^{21,22} or protein-protein interactions²³. Pande *et al.*²¹ set up an SB-ML integrated pipeline to identify the most likely target of natural compound resveratrol, for which the mode of action is still unknown. This study was possible due to the (recent) resolution of nine proteins in complex with the ligand. A set of forty anti-breast cancer resveratrol derivatives from the literature was used for docking, and a 3D quantitative activity-structure relationship (QSAR) CoMFA/CoMSIA PLS model was created for target-derived results from docking. Based on the performance of the models, MDM2 and QR2 were suggested as potential targets for resveratrol derivatives.

As suggested before, computational methods can also be used to rationally propose

polypharmacological approaches for novel drugs²³ or repurposing²². The implementation by Lim *et al.*²² used the original crystal structure of an approved drug as a template for a ligand binding space search in the genome. Subsequently, docking was performed and used, together with bioactivity data, as input for an ML algorithm to predict genome-wide ligand-protein interactions in a fully integrated fashion. RIOK1 was predicted, among other kinases, to be the off-target of PDE3 inhibitors such as levosimendan and proposed for drug repurposing in anticancer therapies. Conversely, Zhi *et al.*²³ used a sequential SBML pipeline to identify novel targets related to dihydroorotate dehydrogenase (DHODH) and to screen drug candidates for multiple targets in small-cell lung cancer. Firstly, protein-protein interaction information was leveraged for network pharmacology analysis. This allowed the selection of related proteins in which drugs may have a combined effect, such as UMPS, which like DHODH is involved in pyrimidine biosynthesis. Docking in both DHODH and UMPS showed eight potential multi-target compounds. These were prioritized based on predicted binding affinity towards DHODH using three multi-GNN (Graph Neural Network) regression models. The top three candidates were subjected to MD validation, where it was confirmed that they showed stable interactions with both targets.

Integrated approaches used to predict (off)-targets can have a direct impact on lead prioritization in oncological drug discovery. The application of the methodologies, however, mostly depends on the available data. Approaches such as those of Pande *et al.*²¹ and Lim *et al.*²² are relevant when true binding modes have been identified. In the case of Zhi *et al.*²³, rich interactome databases are needed as well as bioactivity data for the identified targets of interest.

Binding site prediction

Once the relevant targets have been defined, the binding sites need to be characterized for drug discovery purposes. Notably in oncological drug discovery, this task can be made more complicated with mutated binding sites or transformed protein-protein interactions. There is an extensive array of tools available for small molecule binding site prediction, as recently reviewed by Krivák and Hoksza²⁴. In their independent benchmark, they showed how some methods where SB and ML techniques were integrated showed equal or higher performance to other SB-exclusive methods. However, in their analysis, they also urged caution over the calculation of too complex features from structural data for ML analysis when using relatively small training datasets. Of particular interest in anticancer drug development is the discovery of allosteric binding sites that can be targeted selectively in cancer cells to reduce off-site adverse effects triggered by events in the orthosteric binding sites. While most of the binding site prediction methods summarized by Krivák and Hoksza²⁴ can be used to predict allosteric binding sites, these share a number of differential characteristics that have triggered the development of allosteric-specific binding site prediction tools²⁵. Some of these methods build on top of general binding site predictors with e.g. an added layer of ML classification²⁶. The application of these methods and the analysis of the effects caused by allosteric modulators will be discussed in more detail in the section *Allosteric modulation analysis*.

While the information and software needed for binding site prediction are extensively available for small molecules, the prediction of binding regions in protein-protein binding modeling is still challenging²⁷. Protein-protein interactions have been shown to be crucial in certain aspects of cancer pathogenicity⁸. In that area, integrated SB-ML approaches have proven beneficial^{28,29}. Kawaguchi *et al.*²⁸ used a Bayesian active learning-based protein-protein docking approach to predict the conformation of the dimerization interface of CD44 and the residues involved. Similarly, the approach developed by Taherzadeh *et al.*²⁹ uses ML to predict protein-peptide binding residues from protein sequence and structural data-derived features. The predicted residues from the RF classifier are used as input for a density-based clustering algorithm to define the binding region on the protein surface. The authors showed that the performance is better compared to other non-ML methods on the same dataset. In general, however, the exploratory nature of the applications in this use case scenario makes it challenging to assess the performance of the methods reviewed. To counterbalance this problem and reduce the effect of false positives, an option would be to use a consensus approach where several tools are employed and sites predicted by more than one of them are further investigated.

Largely, the feasibility of the approaches reviewed here depends on the availability of structural data. The use of homology models can be useful here, with some authors showing how their integrated ML-SB methods perform equally well in experimental structures as in homology models^{29,30}. Moreover, the recent release of AlphaFold³¹ to predict protein structures with high accuracy opens doors for the implementation of many of these methods on a genome-wide scale. The distribution of AlphaFold as open-source code has facilitated the development of related tools that will improve its biological relevance. An example is AlphaFill³², a tool that enriches AlphaFold models with ligands and co-factors. Of very high relevance in oncological drug discovery, these tools could enable the prediction of binding sites in mutants that have not been experimentally determined.

Virtual screening

The most common scenario in computational drug discovery is virtual screening (VS). Similarly to the case of computational mutagenesis, VS can be seen as a tool to prioritize compounds for experimental analysis. While VS has been extensively explored using SB and ML methods independently, their combination – both in a fully integrated or in a sequential way – allows for the use of as much data available as possible and, expectedly, more accurate results. Certainly, this use case scenario is not unique for oncological drug discovery, but the advances made in computational drug discovery in this area can very well power successful anticancer drug discovery stories.

A classic way to integrate SB and ML learning methods in VS is the use of ML-based scoring functions in docking^{33–38}. These can be directly integrated into the docking software or, more commonly, used *a posteriori* for re-scoring. Moreover, ML scoring functions are often target-specific^{33–35} but not necessarily so³⁸. One of the simplest approaches is to include docking scores as features for an ML classifier³³. Slightly more

complex, the approach developed by Yang *et al.*³⁴ starts from a similarity-based docking method to reduce the challenges presented by the large conformational space of Cathepsin S inhibitors. Subsequently, a fragmentation method is applied to the predicted poses. Furthermore, Berishvili *et al.* demonstrated the added value of including not only docking-derived features for the ML scoring function³⁵ but also MD-derived features³⁶. However, in retrospective, they showed that ML-based target-specific scoring functions were not accurate in identifying active tankyrase compounds. More complex methods, such as FEP, were needed in order to properly correlate the predicted binding affinity to the pIC_{50} values determined experimentally. Similar to other ML applications, the development of accurate ML scoring functions highly depends on the quality of the datasets available for training and validation. Adeshina *et al.*³⁸ focused on the development of a high-quality dataset (D-COID, publicly available) to train ML re-scoring functions. Importantly, they included challenging decoy complexes from the DUD-E dataset and tried to keep the dataset balanced. Also, they refrained from using docked poses in the training set.

Similar approaches might not necessarily be coined ML scoring functions, even though they also use ligand-protein interaction data as input for ML models^{39,40}. Kalali and Asadollahi-Baboli³⁹ used an approach where docking was performed as a first step to discern relevant interactions and derive ML descriptors. Using a slightly different approach, Li *et al.*⁴⁰ constructed a pharmacological space accounting for ligand, protein, and ligand-protein interaction descriptors. The latter were generated from a combined average fingerprint per protein from known binders.

In general, however, the most typical approach in VS is still the use of SB and ML methods in a sequential or parallel way^{41–50}. These often include the development of a ligand-based QSAR classification^{41–47} or regression^{48,49} model from experimental bioactivity data to prioritize compounds from a chemical database based on their predicted binding affinity. The wide array of models and databases reviewed here is collected in **Table 2.1**. Subsequently, the selected hits are filtered based on different criteria depending on the scope of the project (e.g. reverse pharmacophore mapping⁴³, ΔG calculation with MM-GBSA⁴⁴), and finally, an SB method such as docking^{41,42,44–46,49,50} and/or MD^{41–43,46,48–51} is deployed to rationalize the results of the ML model and propose compounds for *in vitro* validation. Sometimes, the SB phase is a filter on its own, with a docking-based VS^{41,46}, and occasionally it is used before the ML phase^{44,49}. Moreover, the ML model is not always built to predict binding affinity, but sometimes also anticancer activity⁵⁰, or mode of action⁴⁵. When focused on multiple on- and off-targets, sequential pipelines can also be used to prioritize polypharmacological compounds, as done for kinase inhibitors by Burggraaff *et al.*⁴⁷ Even though these VS strategies are more common in the screening of small molecules, there are also some examples from peptide VS campaigns, such as that of Junaid *et al.*⁵¹.

One of the main limitations found in VS approaches lies in the definition of relevant training and validation sets for ML. Even though databases such as ChEMBL and PubChem contain a very large amount of bioactivity data, target-specific applications still end up usually having too small datasets where generalization is difficult to achieve.

Table 2.1. Overview of reviewed literature categorized by use case scenario.

* See Box 2.1. Hallmarks of cancer to which the targets are related, as defined by Hanahan and Weinberg (Cell, 2011). Supporting references to the connection of the targets to each hallmark.
































** See Figure 2.1. Integration approach of AI and SB methods: A) Structural data as input for ML, B) ML-based scoring function, C) ML analysis of MD, and D) Sequential or parallel pipelines.

Reference	Target / Ligand dataset	Hallmark of cancer *	AI method(s)	SB method(s)	Integration approach **
Driver prediction					
Bailey <i>et al.</i> ¹¹	Pan-target / TCGA-MC3 set	7 ¹¹	Various	Various	Ⓓ
Knijnenburg <i>et al.</i> ¹²	Pan-target / TCGA-MC3 set	7 ¹²	Logistic regression classifier	FoldX, MD	Ⓓ
Liñares-Blanco <i>et al.</i> ¹³	Pan-target (FABP6) / TCGA	7 9 ¹³	RF and generalized linear classifiers	Docking	Ⓓ
Computational mutagenesis					
Masso <i>et al.</i> ¹⁵	BRCA1 / ClinVar	7 ¹⁵	RF classifier	Structure-derived features	Ⓐ
Pandurangan and Blundell ¹⁶	Pan-target / ProTherm benchmark	7 ¹⁶	ML ensemble classifier	Structure-derived features	Ⓐ
Chitralla <i>et al.</i> ¹⁷	P53-ER α / NA	1 ¹⁷	One-layer NN	Protein-protein docking	Ⓐ
Babbitt <i>et al.</i> ¹⁸	BRAF / FDA	10 11 ^{8,18}	Seven stacked classifiers	MD	Ⓒ
Aldeghi <i>et al.</i> ¹⁹	Abl / Platinum database, in-house set	11 ¹⁹	Extremely randomized regression trees	FEP	Ⓓ
Patil <i>et al.</i> ²⁰	Kinome (Alk) / UniProt, literature	7	SVM, RF, NeuralNet, LR	MD (metadynamics)	Ⓓ
(Off)-target prediction					
Pande <i>et al.</i> ²¹	Pan-target (MDM2) / Literature	1 ²¹	CoMFA/CoMSIA PLS regressor, DT, RF, KNN, MLP, SVM classifiers	Docking, MD	Ⓐ

Table 2.1 (continues)

Lim <i>et al.</i> ²²	Pan-target (RIOK1, PDE3) / ChEMBL, DrugBank, literature datasets, TCGA-CCLE	5 ⁶³	ElasticNet, SVR regressors	Ligand binding space search in genome, docking	Ⓐ
Zhi <i>et al.</i> ²³	DHODH / STRING, KEGG, ChEMBL, ZINC	9 ²³	Multi-GNN	Docking, MD	Ⓓ
Binding site prediction					
Kawaguchi <i>et al.</i> ²⁸	CD44 / NA (pre-trained)	5 ²⁸	Bayesian active learning	Protein-protein docking	Ⓑ
Taherzadeh <i>et al.</i> ²⁹	Pan-target / BioLip	1 (protein-protein binding)	RF classifier, DBSCAN	Structure-derived features	Ⓐ
Virtual screening					
Che <i>et al.</i> ³³	IRAK1 / ChEMBL, DUD-E	4 ⁶⁴	SVM classifier	Docking	Ⓑ
Yang <i>et al.</i> ³⁴	Cathepsin S / PDBbind, CSAR, GC3/4, ChEMBL	2 ⁶⁵	XGBoost regressor	Similarity-based docking	Ⓑ
Berishvili <i>et al.</i> ^{35–37}	Pan-target, Tankyrase / ZINC	3 ³⁷	DNN	Docking, MD, FEP	Ⓑ
Adeshina <i>et al.</i> ³⁸	Pan-target (AChE) / ChEMBL, DUD-E	8 ⁶⁶	XGBoost classifier	Docking	Ⓑ
Kalaki and Asadollahi-Baboli ³⁹	Pim / In-house dataset	8 ⁶⁷	PCA, PLS classifier	Docking	Ⓐ
Li <i>et al.</i> ⁴⁰	KIF11 / KEGG BRITE, DrugBank, STITCH	5 ⁶⁸	Bayesian Additive Regression Trees	Bow-pharmacological space (protein-ligand interactions)	Ⓐ
Raju <i>et al.</i> ⁴¹	CYP1B1 / ChEMBL, PubChem, literature, DUD-E, Maybridge, ChemBridge, Natural compound library	11 ⁴¹	SVM, RF, ANN classifiers	Docking, MD	Ⓓ

Table 2.1 (continues)

Chen <i>et al.</i> ⁴²	LXR β / ChEMBL, Binding DB, in-house library, GSMTL	 ⁴²	SVM, Naïve Bayes classifiers	Docking, MD	
Halder and Cordeiro ⁴³	AKT / ChEMBL, Asinex library	 ⁸	LDA, XGBoost and other classifiers	MD	
Azhagiya Singam <i>et al.</i> ⁴⁴	AR / Tox21, CompTox	 ⁶⁹	SVM classifiers	Docking	
Kadioglu and Efferth ⁴⁵	P-gp / ChEMBL	 ⁴⁵	RF classifier	Docking	
Guo <i>et al.</i> ⁴⁶	Tubulin / ChEMBL	 ⁷⁰	Naïve Bayes classifiers	Docking, MD	
Burggraaff <i>et al.</i> ⁴⁷	RET / ChEMBL, ZINC	 ⁷¹	RF classifiers	(Induced-fit) docking, metadynamics	
Chen <i>et al.</i> ⁴⁸	MMP13 / Traditional Chinese medicine database	 ⁷²	RF, gradient boosting, AdaBoost, deep learning	MD	
Chen <i>et al.</i> ⁴⁹	STAT3 / Literature set, ZINC	 ⁴⁹	Nine regressors, 3D QSAR	Docking, MD	
Guo <i>et al.</i> ⁵⁰	Tubulin / ChemDiv	 ⁷⁰	Discovery studio prediction models	Docking, MD	
Junaid <i>et al.</i> ⁵¹	p53-ASPP2-CagA / Rationally designed	 ⁵¹	ML module in MOE	MD	
Allosteric modulation analysis					
Lu <i>et al.</i> ²⁵	SIRT6, STAT3 / PDB, commercial	 ,  ²⁵	SVM	Geometric binding site predictor	
Song <i>et al.</i> ⁵⁶	Pan-target / PDB	 ⁵⁶	RF, neural networks	Structure-derived features	
Uyar <i>et al.</i> ⁵⁸	Neurolysin / PDB	 ⁷³	ElasticNet, PCA, LDA	MD	
Chen <i>et al.</i> ⁵⁹	SETD8 / cBioPortal	 ⁵⁹	Markov state model, tICA, clustering	MD	
Hu <i>et al.</i> ⁶⁰	MOR / Rationally designed	 ⁷⁴	Markov state model, tICA	MD	

This is an even more relevant bottleneck when considering cancer-related mutants, for which VS campaigns would be extremely beneficial to prioritize personalized medicine drugs. Moreover, target-specific applications present an important challenge to avoid learned biases and overfitting⁵². The inclusion of decoys in the sets (e.g. from the DUD-E dataset) is a good way to balance the presence of active and inactive compounds⁵³. In that sense, the D-COVID dataset³⁸ is a good starting point for the development of re-scoring functions, but it might require experimental expansion via collaborative work for target-specific applications.

Allosteric modulation analysis

Previously, we have mostly referred to orthosteric ligand binding when describing ligand binding, i.e. the site where the endogenous ligand or substrate binds. However, allosteric modulation has been described as a powerful tool to increase the selectivity of targeted compounds and overcome drug-resistant mutations, and it is therefore worth exploring in cancer research. Indeed, unraveling the mechanisms underlying allosteric effects can be a key step in proposing new therapeutic routes. Moreover, allosteric binding sites and modulators have been shown to exhibit differential characteristics to orthosteric counterparts⁵⁴, which calls for the development of allosteric-specific tools for most of the use case scenarios described in the sections above, as anticipated in the section *Binding site prediction*.

The work from Lu *et al.*²⁵ comprises a very complete review of the currently available SB methods for allosteric modulator discovery. Some of these methods integrate SB and ML techniques for allosteric binding site prediction²⁶, allosteric interaction scoring⁵⁵, and allosteric effect analysis of mutations⁵⁶. The authors demonstrated the applicability of these tools in oncological drug discovery with the prioritization of allosteric activators and inhibitors for anticancer (potential) targets SIRT6 and STAT3, respectively²⁵. In both cases, allosteric binding pockets were predicted and subjected to VS of commercial libraries. These computational efforts were confirmed either by experimental assays or crystallographic studies. Of direct application in oncological drug discovery is AlloDriver⁵⁶, a driver prediction tool that maps mutations from clinical cancer samples to their 3D structures, labels them as orthosteric or (potentially) allosteric, and classifies targets as driver or passenger using a combination of random forest and multi-layer neural networks. Even though periodically updated, this tool relies on the availability of annotated allosteric sites (and driver mutations), which is a common bottleneck in ML-based allostery prediction methods.

Specific to allosteric modulation analyses is the exploration of the allosteric pathways that drive the effects observed. These aspects are often better explored in a dynamic setting, given the complex conformational landscape of proteins that often is responsible for allosteric pathways^{25,57}. Hence, the efforts reviewed below use ML techniques to analyze MD trajectories and find patterns that help explain the observed effects^{58–60}. For example, the work of Uyar *et al.*⁵⁸ made possible the identification of differential dynamic patterns in apo and allosteric inhibitor-bound neurolysin structures, as well as the key residues involved. Moreover, the analysis of MD trajectories with Markov state models

using time-structure-based independent component analysis (tICA) allowed Chen *et al.*⁵⁹ and Hu *et al.*⁶⁰ to identify conformational microstates. These were then related to mutation-driven allosteric effects in catalytic activity of SEDT8, and energetic differences in Na⁺ translocation and metastable states in active and inactive MOR, respectively, which were further validated experimentally.

Even though the concept of allostery has been known for 50 years, it has only recently gained more attention in drug discovery with an exponential increase in known allosteric modulators in the last two decades²⁵. Of the 19 currently FDA-approved allosteric modulators, three are indicated as anticancer drugs⁶¹. The use of computational tools, and more specifically ML-based methods, still suffers from the lack of experimentally determined allosteric interactions and mechanisms. In the near future, we expect this area of research to play a more important role in oncological drug discovery in combination with experimental validation as it holds promise to bring more selective anticancer drugs to the market.

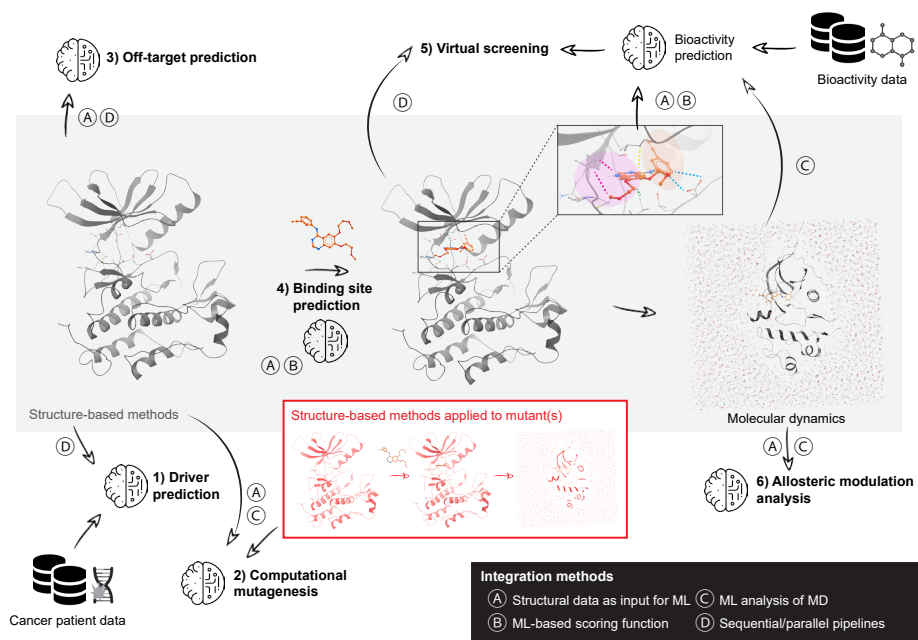


Figure 2.1. Use case scenarios of integrated structure-based (SB) and machine learning (ML) methods in oncological drug discovery and the integration methods employed. In this review we address six use case scenarios, namely 1) driver prediction, 2) computational mutagenesis, 3) (off)-target prediction, 4) binding site prediction, 5) virtual screening, and 6) allosteric modulation analysis. Integration approaches that achieve a full integration include those where (A) structural data derived from SB methods is used as input for ML models, with emphasis on the predicted output; (B) docking poses are analyzed with ML-based scoring functions; and (C) output trajectories from molecular dynamics (MD) simulations are analyzed with ML. However, it is still more common to combine SB and ML methods without full integration, with the implementation done in a sequential or parallel way (D) where ML acts as a pre-filter for the SB phase, or vice versa.

Conclusions

Integrated ML-SB methods are useful to investigate different aspects of oncological drug discovery. These methods apply to a variety of use case scenarios that can be cancer-specific or general for computational drug discovery with potential application in oncological research. There is no rule of thumb for the selection of approaches because these largely depend on the scope of the study. However, some ML-SB integration methods are primarily leveraged in specific use case scenarios, for example, ML-based scoring functions in VS or the use of ML to analyze MD simulations in allosteric modulation analyses. VS use cases are still the most common ones, but integrated methods are also gaining relevance in fields such as driver prediction and computational mutagenesis, where the use of structural data has proven to be a significant complement to omics data. Despite their broad domain of applicability, the approaches reviewed here still present certain limitations worth discussing. In general, data availability and computational requirements present common bottlenecks that need to be assessed on a project-specific basis. Moreover, it has been shown that sometimes less expensive approaches outperform more complex ones in the same tasks. Future research will probably extend more into the use of more complex algorithms currently underrepresented, such as DNNs, to be able to capture all relevant information from structural data. Finally, a common drawback in computational drug discovery that can be observed in the articles reviewed here is the lack of experimental validation. These aspects trigger some open questions on the use of integrated computational methods in oncological drug research, which we address in **Box 2.2**. However, the approaches presented here are considered a good way to prioritize targets and small molecules in the field, and their combination with experimental validation will likely be a key factor in bringing drugs for oncological personalized therapies faster to the market. During the revision of our manuscript, a proposal for a further conceptual extension of the hallmarks of cancer was published⁶². This exemplifies the fast pace at which oncological research advances and the need to constantly revisit the biological relevance of the methods applied in oncological drug discovery.

Box 2.2. Open questions on present and future directions

The articles reviewed here exemplify the added value of integrated AI-SB methods in oncological drug discovery. However, some questions worth exploring in the future arise from their interpretation, which we outline below.

- Structural data availability is a common bottleneck. How beneficial is its inclusion in pan-target analyses when it results in a reduced target space? Will approaches like AlphaFold³¹ be able to solve this issue?
- Currently, the analysis of trajectories from MD with ML is rather restricted to cases with small datasets (i.e. allosteric modulation analyses). However, we expect that with increasing amounts of data and computing power this approach will become more relevant in big-scale virtual screening.
- Is it pertinent to continue expanding the research into integrated approaches without conducting exhaustive benchmarking against classical individual methods?
- Are there enough resources devoted to enlarging and standardize publicly available datasets for computational oncological drug discovery? Will these expand into aspects often neglected, such as tumor microenvironment?
- We hypothesize the rise of allosteric modulation analyses to bring more selective drugs to the market. Will we also see a boom in publicly available allosteric structural and experimental data for machine learning applications?
- Is the potential added value of more complex approaches worth the likely resulting increase in computing power/time and data storage needs? Will this aspect limit the use of deep learning approaches in the near future?
- A common drawback in computational drug discovery is the lack of experimental validation. We strongly advise an increase of collaborative work leading both to validated tools and larger datasets available for ML training.

References

1. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem Rev.* **18**, 119 (2019)
2. Azuaje, F. Artificial intelligence for precision oncology: beyond patient stratification. *npj Precis Oncol.* **3**, 6 (2019)
3. Duran-Frigola, M., Mosca, R. & Aloy, P. Structural systems pharmacology: The role of 3D structures in next-generation drug development. *Chem Biol.* **20**, 674-684 (2013)
4. Li, H., Sze, K.H., Lu, G. & Ballester P.J. Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip Rev Comput Mol Sci.* **10**, 1-20 (2020)
5. Batool, M., Ahmad, B. & Choi, S. A structure-based drug discovery paradigm. *Int J Mol Sci.* **20**, 2783 (2019)
6. Sydow, D., Burggraaff L., Szengel A., *et al.* Advances and Challenges in Computational Target Prediction. *J Chem Inf Model.* **59**, 1728-1742 (2019)
7. Hanahan, D. & Weinberg, R.A. The Hallmarks of Cancer. *Cell.* **100**, 57-70 (2000)
8. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell.* **144**, 646-674 (2011)
9. Krzyszczyk, P., Acevedo A., Davidoff, E.J., *et al.* The growing role of precision and personalized medicine for cancer treatment. *Technology.* **6**, 79-100 (2018)
10. Wu, F., Zhou, Y., Li, L., *et al.* Computational Approaches in Preclinical Studies on Drug Discovery and Development. *Front Chem.* **8**, 726 (2020)
11. Bailey, M.H., Tokheim, C., Porta-Pardo, E., *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell.* **173**, 371-385 (2018)
12. Knijnenburg, T.A., Wang, L., Zimmermann, M.T., *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239-254 (2018)
13. Liñares-Blanco, J., Munteanu, C.R., Pazos, A. & Fernandez-Lozano C. Molecular docking and machine learning analysis of Abemaciclib in colon cancer. *BMC Mol Cell Biol.* **21**, 1-18 (2020)
14. Jensen, M.A., Ferretti, V., Grossman, R.L. & Staudt, L.M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood.* **130**, 453-459 (2017)
15. Masso, M., Bansal, A., Bansal, A. & Henderson, A. Structure-based functional analysis of BRCA1 RING domain variants: Concordance of computational mutagenesis, experimental assay, and clinical data. *Biophys Chem.* **266**, 106442 (2020)
16. Pandurangan, A.P. & Blundell, T.L. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Sci.* **29**, 247-257 (2020)
17. Chitrula, K.N., Nagarkatti, M., Nagarkatti, P. & Yeguvapalli, S. Analysis of the TP53 deleterious single nucleotide polymorphisms impact on estrogen receptor alpha-p53 interaction: A machine learning approach. *Int J Mol Sci.* **20**, 2962 (2019)
18. Babbitt, G.A., Lynch, M.L., McCoy, M., Fokoue, E.P. & Hudson A.O. Function and evolution of B-Raf loop dynamics relevant to cancer recurrence under drug inhibition. *J Biomol Struct Dyn.* **40**, 468-483 (2022)
19. Aldeghi, M., Gapsys, V. & De Groot, B.L. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS Cent Sci.* **5**, 1468-1474 (2019)
20. Patil, K., Jordan, E.J., Park, J.H., *et al.* Computational studies of anaplastic lymphoma kinase mutations reveal common mechanisms of oncogenic activation. *Proc Natl Acad Sci U S A.* **118**, e2019132118 (2021)
21. Pande, A., Manchanda, M., Bhat, H.R., Bairy, P.S., Kumar, N. & Gahtori, P. Molecular insights into a mechanism of resveratrol action using hybrid computational docking/CoMFA and machine learning approach. *J Biomol Struct Dyn.* **40**, 8286-8300 (2022)
22. Lim, H., He D., Qiu, Y., Krawczuk, P., Sun, X. & Xie, L. Rational discovery of dual-indication multi-target pde/kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology. *PLoS Comput Biol.* **15**, 1-21 (2019)
23. Zhi, H.Y., Zhao, L., Lee, C.C. & Chen C.Y.C. A novel graph neural network methodology to investigate dihydroorotate dehydrogenase inhibitors in small cell lung cancer. *Biomolecules.* **11**, 477 (2021)
24. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform.* **10**, 1-12 (2018)
25. Lu, S., He, X., Ni, D. & Zhang J. Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *J Med Chem.* **62**, 6405-6421 (2019)
26. Huang, W., Lu, S., Huang, Z., *et al.* Allosite: A method for predicting allosteric sites. *Bioinformatics.* **29**, 2357-2357 (2013)
27. Vakser, I.A. Challenges in protein docking. *Curr Opin Struct Biol.* **64**, 160-165 (2020)
28. Kawaguchi, M., Dashzeveg, N., Cao, Y., *et al.*

- Extracellular Domains I and II of cell-surface glycoprotein CD44 mediate its trans-homophilic dimerization and tumor cluster aggregation. *J Biol Chem.* **295**, 2640-2649 (2020)
29. Taherzadeh G., Zhou, Y., Liew, A.W.C. & Yang, Y. Structure-based prediction of protein-peptide binding regions using Random Forest. *Bioinformatics.* **34**, 477-484 (2018)
 30. Li, L., Khanna, M., Jo, I., *et al.* Target-specific support vector machine scoring in structure-based virtual screening: Computational validation, in vitro testing in kinases, and effects on lung cancer cell proliferation. *J Chem Inf Model.* **51**, 755-759 (2011)
 31. Jumper, J., Evans, R., Pritzel, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature.* **596**, 583-589 (2021)
 32. Hekkelman, M.L., Vries, I. de, Joosten, R.P. & Perrakis, A. AlphaFill: enriching the AlphaFold models with ligands and co-factors. *Nat. Methods.* **20**, 205-2013 (2023)
 33. Che, J., Feng, R., Gao, J., *et al.* Evaluation of Artificial Intelligence in Participating Structure-Based Virtual Screening for Identifying Novel Interleukin-1 Receptor Associated Kinase-1 Inhibitors. *Front Oncol.* **10**, 1-12 (2020)
 34. Yang, Y., Lu, J., Yang, C. & Zhang, Y. Exploring fragment-based target-specific ranking protocol with machine learning on cathepsin S. *J Comput Aided Mol Des.* **33**, 1095-1105 (2019)
 35. Berishvili, V.P., Voronkov, A.E., Radchenko, E.V. & Palyulin, V.A. Machine Learning Classification Models to Improve the Docking-based Screening: A Case of PI3K-Tankyrase Inhibitors. *Mol Inform.* **37**, 1-10 (2018)
 36. Berishvili, V.P., Perkin, V.O., Voronkov, A.E., *et al.* Time-Domain Analysis of Molecular Dynamics Trajectories Using Deep Neural Networks: Application to Activity Ranking of Tankyrase Inhibitors. *J Chem Inf Model.* **59**, 3519-3532 (2019)
 37. Berishvili, V.P., Kuimo, A.N., Voronkov, A.E., *et al.* Discovery of novel tankyrase inhibitors through molecular docking-based virtual screening and molecular dynamics simulation studies. *Molecules.* **25**, 1-15 (2020)
 38. Adeshina, Y., Deeds, E. & Karanickolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *PNAS.* **117**, 18477-18488 (2020)
 39. Kalaki, Z. & Asadollahi-Baboli, M. Molecular docking-based classification and systematic QSAR analysis of indoles as Pim kinase inhibitors. *SAR QSAR Environ Res.* **31**, 399-419 (2020)
 40. Li, L., Koh, C.C., Reker, D., *et al.* Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees. *Sci Rep.* **9**, 7703 (2019)
 41. Raju, B., Verma, H., Narendra, G., Sapra, B. & Silakari, O. Multiple machine learning, molecular docking, and ADMET screening approach for identification of selective inhibitors of CYP1B1. *J Biomol Struct Dyn.* **40**, 7975-7990 (2022)
 42. Chen, H., Chen, Z., Zhang, Z., *et al.* Discovery of new LXR β agonists as glioblastoma inhibitors. *Eur J Med Chem.* **194**, 112240 (2020)
 43. Halder, A.K. & Cordeiro, M.N.D.S. Akt inhibitors: The road ahead to computational modeling-guided discovery. *Int J Mol Sci.* **22**, 3944 (2021)
 44. Azhagiya Singam, E.R., Tachachartvanich, P., Fourches, D., *et al.* Structure-based virtual screening of perfluoroalkyl and polyfluoroalkyl substances (PFASs) as endocrine disruptors of androgen receptor activity using molecular docking and machine learning. *Environ Res.* **190**, 109920 (2020)
 45. Kadioglu, O. & Efferth, T. A Machine Learning-Based Prediction Platform for P-Glycoprotein Modulators and Its Validation by Molecular Docking. *Cells.* **8**, 1286 (2019)
 46. Guo, Q., Zhang, H., Deng, Y., *et al.* Ligand- and structural-based discovery of potential small molecules that target the colchicine site of tubulin for cancer treatment. *Eur J Med Chem.* **196**, 112328 (2020)
 47. Burggraaff, L., Lenselink, E.B., Jespers, W., *et al.* Successive statistical and structure-based modeling to identify chemically novel kinase inhibitors. *J Chem Inf Model.* **60**, 4286-4295 (2020)
 48. Chen, J.Q., Chen, H.Y., Dai, W.J. & Lv, Q.J., Chen CYC. Artificial Intelligence Approach to Find Lead Compounds for Treating Tumors. *J Phys Chem Lett.* **10**, 4382-4400 (2019)
 49. Chen, X., Chen, H.Y., Chen, Z.D., Gong, J.N. & Chen, C.Y.C. A novel artificial intelligence protocol for finding potential inhibitors of acute myeloid leukemia. *J Mater Chem B.* **8**, 2063-2081 (2020)
 50. Guo, Q., Luo, Y., Zhai, S., *et al.* Discovery, biological evaluation, structure-activity relationships and mechanism of action of pyrazolo[3,4-b]pyridin-6-one derivatives as a new class of anticancer agents. *Org Biomol Chem.* **17**, 6201-6214 (2019)
 51. Junaid, M., Shah, M., Khan, A., *et al.* Structural-dynamic insights into the H. pylori cytotoxin-associated gene A (CagA) and its abrogation to interact with the tumor suppressor protein ASP22 using decoy peptides. *J Biomol Struct Dyn.* **37**, 4035-4050 (2019)
 52. Sieg, J., Flachsenberg, F. & Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J Chem Inf Model.* **59**, 947-961 (2019)
 53. Allen, B.K., Mehta, S., Ember, S.W.J., Schonbrunn, E., Ayad, N. & Schürer, S.C. Large-Scale Computational Screening Identifies First in Class Multitarget Inhibitor of EGFR Kinase and BRD4. *Sci Rep.* **5**, 1-16 (2015)

54. van Westen, G.J.P., Gaulton, A. & Overington, J.P. Chemical, Target, and Bioactive Properties of Allosteric Modulation. *PLoS Comput Biol.* **10**, e1003559 (2014)
55. Li, S., Shen, Q., Su, M., *et al.* Alloscore: A method for predicting allosteric ligand-protein interactions. *Bioinformatics.* **32**, 1547-1576 (2016)
56. Song, K., Li, Q., Gao, W., *et al.* AlloDriver: A method for the identification and analysis of cancer driver targets. *Nucleic Acids Res.* **47**, W315-W321 (2019)
57. Nussinov, R., Tsai, C.J. & Jang, H. Dynamic protein allosteric regulation and disease. *Adv Exp Med Biol.* **1163**, 25-43 (2019)
58. Uyar, A., Karamyan, V.T. & Dickson, A. Long-Range Changes in Neurolysin Dynamics Upon Inhibitor Binding. *J Chem Theory Comput.* **14**, 444-452 (2018)
59. Chen, S., Wiewiora, R.P., Meng, F., *et al.* The dynamic conformational landscape of the protein methyltransferase setd8. *Elife.* **8**, e45403 (2019)
60. Hu, X., Wang, Y., Hunkele, A., Provasi, D., Pasternak, G.W. & Filizola, M. Kinetic and thermodynamic insights into sodium ion translocation through the μ -opioid receptor from molecular dynamics and machine learning analysis. *PLoS Comput Biol.* **15**, 1-19 (2019)
61. Amamuddy, O.S., Veldman, W., Manyumwa, C., *et al.* Integrated computational approaches and tools for allosteric drug discovery. *Int J Mol Sci.* **21**, 847 (2020)
62. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31-46 (2022)
63. Weinberg, F., Reischmann, N., Fauth, L., *et al.* The Atypical Kinase R1OK1 Promotes Tumor Growth and Invasive Behavior. *EBioMedicine.* **20**, 79-97 (2017)
64. Wee, Z.N., Yatim, S.M.J.M., Kohlbauer, V.K., *et al.* IRAK1 is a therapeutic target that drives breast cancer metastasis and resistance to paclitaxel. *Nat Commun.* **6**, 8746 (2015)
65. Fuchs, N., Meta, M., Schuppan, D., Nuhn, L. & Schirmeister, T. Novel Opportunities for Cathepsin S Inhibitors in Cancer Immunotherapy by Nanocarrier-Mediated Delivery. *Cells.* **9**, 1-17 (2020)
66. Xi, H.J., Wu, R.P., Liu, J.J., Zhang, L.J. & Li, Z.S. Role of acetylcholinesterase in lung cancer. *Thorac Cancer.* **6**, 390-398 (2015)
67. Keane, N.A., Reidy, M., Natoni, A., Raab, M.S. & O'Dwyer, M. Targeting the Pim kinases in multiple myeloma. *Blood Cancer J.* **5**, e325 (2015)
68. Zhou, J., Chen, W.R., Yang, L.C., *et al.* KIF11 functions as an oncogene and is associated with poor outcomes from breast cancer. *Cancer Res Treat.* **51**, 1207-1221 (2019)
69. Shafi, A.A., Yen, A.E. & Weigel, N.L. Androgen receptors in hormone-dependent and castration-resistant prostate cancer. *Pharmacol Ther.* **140**, 223-238 (2013)
70. Dolhyi, V., Avierin, D., Hojouj, M. & Bondarenko, I. Tubulin Role in Cancer Development and Treatment. *Asploro J Biomed Clin Case Reports.* **2**, 15-22 (2019)
71. Castellone, M.D. & Melillo, R.M. RET-mediated modulation of tumor microenvironment and immune response in multiple endocrine neoplasia type 2 (MEN2). *Endocr Relat Cancer.* **25**, T105-T119 (2018)
72. Kudo, Y., Iizuka, S., Yoshida, M., *et al.* Matrix metalloproteinase-13 (MMP-13) directly and indirectly promotes tumor angiogenesis. *J Biol Chem.* **287**, 38716-38728 (2012)
73. Karamyan, V.T. The role of peptidase neurolysin in neuroprotection and neural repair after stroke. *Neural Regen Res.* **16**, 21-25 (2021)
74. Chen, D.T., Pan, J.H., Chen, Y.H., *et al.* The mu-opioid receptor is a molecular marker for poor prognosis in hepatocellular carcinoma and represents a potential therapeutic target. *Br J Anaesth.* **122**, e157-e167 (2019)

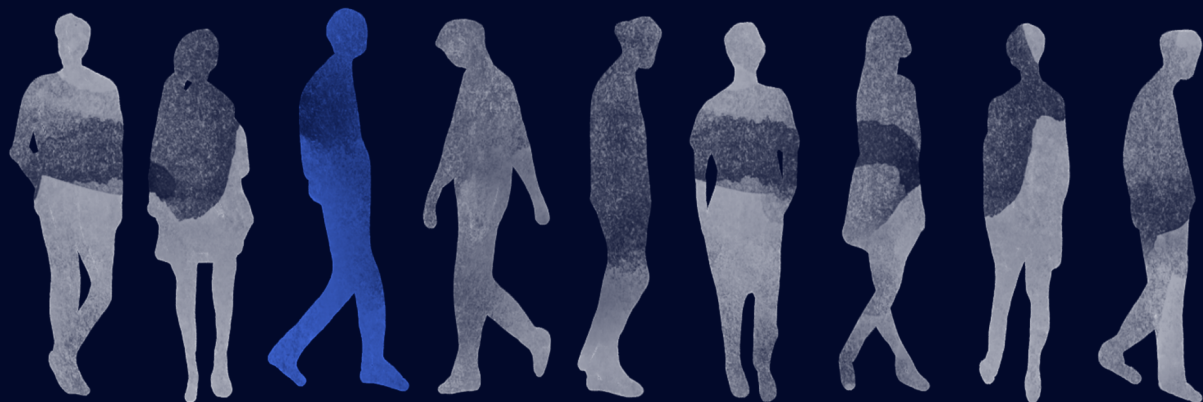



Chapter **3**

Computational characterization of membrane proteins as anticancer targets: Current challenges and opportunities

Marina Gorostiola González, Pepijn Rakers, Willem Jespers, Adriaan P. IJzerman, Laura H. Heitman, Gerard J.P. van Westen

Adapted from: *International Journal of Molecular Sciences* **25**, 3698 (2024)



An abstract, glowing blue ribbon structure, resembling a protein or a complex network, dominates the upper half of the page. It has a dynamic, flowing appearance with many loops and branches.

Cancer remains a leading cause of mortality worldwide and calls for novel therapeutic targets. Membrane proteins are key players in various cancer types but present unique challenges compared to soluble proteins. The advent of computational drug discovery tools offers a promising approach to address these challenges, allowing for the prioritization of “wet lab” experiments. In this review, we explore the applications of computational approaches in membrane protein oncological characterization, particularly focusing on three prominent membrane protein families: receptor tyrosine kinases (RTKs), G protein-coupled receptors (GPCRs), and solute carrier proteins (SLCs). We chose these families due to their varying levels of understanding and research data availability, which leads to distinct challenges and opportunities for computational analysis. We discuss the utilization of multi-omics data, machine learning, and structure-based methods to investigate aberrant protein functionalities associated with cancer progression within each family. Moreover, we highlight the importance of considering the broader cellular context and, in particular cross-talk between proteins. Despite existing challenges, computational tools hold promise in dissecting membrane protein dysregulation in cancer. With advancing computational capabilities and data resources, these tools are poised to play a pivotal role in identifying and prioritizing membrane proteins as personalized anticancer targets.



Introduction

Cancer remains one of the main causes of death worldwide, being responsible for nearly 10 million deaths each year¹. There is therefore continuous need for novel biomarkers and disease-modifying targets. These biomarkers can be leveraged for diagnosis and progression tracking, while the identified targets can be the focus of effective and safe drug development efforts. The etiology of this multifaceted disease often involves aberrant functionality in specific proteins, resulting in increased cellular proliferation and a decrease in standard checkpoints². Notably, membrane proteins have emerged as central players in the development of the most prevalent cancer types^{3–5}. Unfortunately, their study presents additional challenges compared to their soluble counterparts, as has been extensively reviewed^{5–7}. On the bright side, the rise of computational methodologies applied to drug discovery in the past decades has provided researchers with a new set of tools to study these protein classes^{4,8}. These computational pipelines allow scientists to accelerate and streamline the identification of challenging targets and novel hits by prioritizing experiments and reducing the “wet” experimental burden^{9,10}.

Computational methods have applications at multiple levels of the oncological drug discovery pipeline, as highlighted in **Chapter 2**¹¹. Machine learning (ML) and other statistical models can be used to analyze a wealth of multi-omics data and pinpoint the driver mutations in proteins that may be responsible for the onset of a tumor¹². These approaches can also be used in the *in silico* characterization of the effect of point mutations on protein stability, function, and pharmacology¹³. Cancer-related mutations can also be analyzed structurally (as done in **Chapter 6**¹⁴), and ML models such as AlphaFold make it possible even when structural data is not available¹⁵. Furthermore, pharmacophore or quantitative structure-activity relationship (QSAR) models can be used to find chemical structures that either inhibit or activate selectively the target of interest^{16,17}. Once a satisfactory selection of candidate molecules is found, structure-based approaches such as molecular docking or molecular dynamics (MD) simulations can help further refine the favorable protein-drug, or in the case of biological drugs, protein-protein interactions¹⁸. Such detailed knowledge, which many times can only be obtained with computational approaches, is key to enabling personalized oncological treatments¹⁹.

Computational drug discovery pipelines can ease some experimental challenges in membrane protein research, but they face their own issues, mainly stemming from limited data availability due to experimental difficulties^{5,8,10}. Different membrane protein families have varying degrees of experimental investigation, particularly in the context of cancer, which together with differences in structural and functional characteristics leads to family-specific challenges. This review will primarily focus on three membrane protein families: receptor tyrosine kinases (RTKs), G protein-coupled receptors (GPCRs), and solute carrier proteins (SLCs), which exhibit differing levels of general understanding and their connection to cancer^{20–22}. RTKs have been extensively studied, particularly in the context of cancer, with a wealth of research available²³. GPCRs, on the other hand, have received significant attention in drug discovery, but their connection to cancer has only recently become the subject of investigation²⁰. In contrast, SLCs have been relatively understudied in general²⁴. These trends explain the amount of literature linking each

of these protein families with cancer, computational drug discovery, and both (Figure 3.1) despite GPCRs and SLCs being the two largest families of membrane proteins.

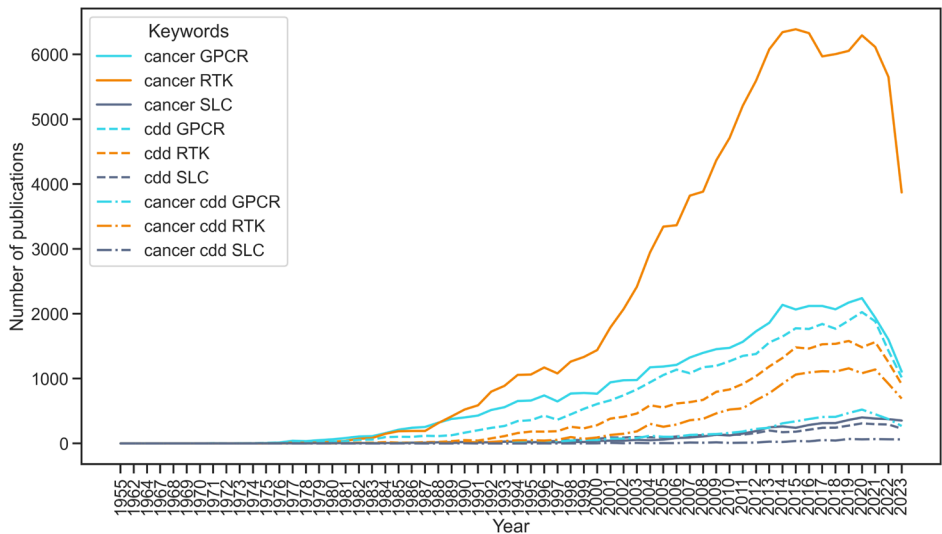


Figure 3.1. Number of publications in PubMed linking three membrane protein families – Receptor tyrosine kinases (RTK), G protein-coupled receptors (GPCR), and Solute carriers (SLC) – to cancer, computational drug discovery (CDD), and the combination of both. The search was computed using the Python package paperscraper²⁵. The list of keywords for each term was: cancer – “cancer”; CDD – “computational”, “computational drug discovery”, “artificial intelligence”, “deep learning”, “machine learning”, “expert systems”, “QSAR”, “PCM”, “molecular dynamics”, “docking”, “molecular modeling”, “FEP”; RTK – “RTK”, “receptor tyrosine kinase”; GPCR – “GPCR”, “G protein-coupled receptor”; SLC – “SLC”, “solute carrier”. Data was retrieved in November 2023, therefore the number of publications related to years 2020-2023 shows a drop corresponding to publication embargoes and delayed publication dates.

In the upcoming sections, we first expand on the key experimental and computational challenges particular to the study of membrane proteins. Then, we outline the primary structural and functional characteristics of RTKs, GPCRs, and SLCs, focusing particularly on alterations that are associated with the progression of cancer. Subsequently, we explore the use of multi-omics, ML, and structure-based methods to investigate these anomalies for each protein family, highlighting their inherent challenges. Finally, we place these membrane protein families within the broader landscape of cellular biology, focusing in particular on inter-family crosstalk. To conclude, we delineate potential avenues to further improve the computational characterization of membrane proteins as anticancer targets.

Key experimental and computational challenges in the study of membrane proteins

Membrane proteins are embedded into the cell membrane yielding an extracellular part, a transmembrane (TM) part, and an intracellular part (Figure 3.2). It is known that the

interactions between the lipids in the cell membrane and the membrane protein are crucial for the protein to acquire the right structure and to function properly²⁶. This means that removing the membrane proteins from the membrane lipids, necessary for many experimental studies such as protein structural determination, brings additional challenges compared to cytosolic proteins^{27,28}. To alleviate a part of this problem, researchers have come up with intricate membrane mimetics that try to imitate the natural environment of the membrane protein instead of using detergents, such as nanodiscs, lipid cubic phase, and styrene malic acid lipid particles²⁸. Another problem is the often low expression level of membrane proteins compared to cytosolic proteins. As such, prokaryotes such as *Escherichia coli* may be used to overexpress a certain membrane protein, but this often leads to aggregation, lack of proper posttranslational modifications, and misfolding of the membrane protein in the process²⁷.

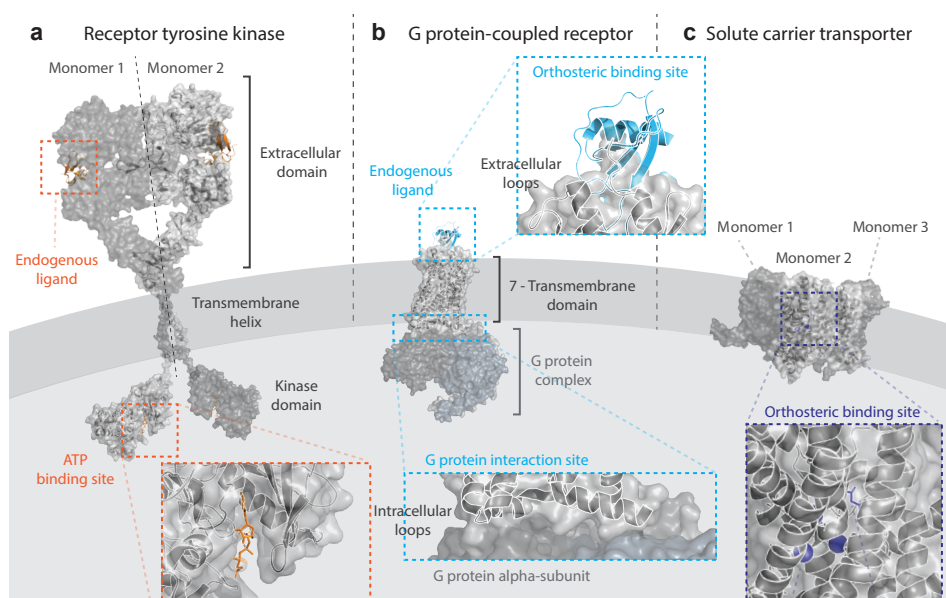


Figure 3.2. Structural models of three membrane protein family members and their main interacting partners. **a)** Receptor tyrosine kinases, represented by a dimer model of the epidermal growth factor – EGFR. The model was constructed using the following determined structures from the RCSB Protein Data Bank (PDB): 3NJF for the dimeric extracellular domain in complex with endogenous ligand EGF; 2M20 for the dimeric transmembrane helix; and 2GS6 for the monomeric kinase domain in complex with ATP. **b)** G protein-coupled receptors, represented by a model of the chemokine receptor CCR2 bound to its endogenous ligand CCL2 and the G_i protein complex. The PDB code used was 7XA3. **c)** Solute carrier transporters, represented by a trimeric model of the glutamate transporter SLC1A3/EAAT1 in complex with its endogenous substrate L-Aspartate and in coordination with three sodium ions needed for transport. The PDB code was 7AWM. The models were built using Pymol²⁹.

Crystallization itself is more difficult for membrane proteins, particularly due to protein instability outside of the membrane. This means that additional measures are needed, such as the addition of stabilizing molecules, or the construction of protein orthologues, which in turn reduces the throughput of structural characterization²⁸. While X-ray crystallization has been the classical method to determine the structure of membrane

proteins, cryo-EM has been on the rise, allowing near-native structural determination with resolutions that have vastly increased in the past years³⁰. The aforementioned challenges do not apply equally to the protein classes that are reviewed here, as demonstrated by their structural coverage. At the time of writing this review, over 2,000 structures were available for 41 out of 58 human RTKs, although in most cases they only represent the soluble kinase domain³¹. Moreover, 187 out of the 826 GPCRs in the human genome have been structurally determined with a total of 1,160 receptor structures in different conformations³². However, due to their dynamic nature, under 5% of human SLC protein structures are available – although the number is increasing rapidly thanks to cryo-EM structures³³. Computational tools, and in particular AlphaFold, are increasingly expanding the availability of good quality predicted protein structures, although this task is less accurate for protein families with fewer structures available for training, which is the case for membrane proteins³⁴.

Besides finding the right structure of membrane proteins, it requires additional computing power to be able to simulate the behavior of membrane proteins in their natural environment, with detailed simulations requiring large computing clusters to be achieved³⁵. The supplemental amount of interactions that come with the oligomerization of membrane proteins adds to this computing power restraint³⁶. Choices have to be made about how the membrane is represented and set up, which lipids are used, and how many atoms are used in the simulation, to name a few issues. Trade-offs have to be made between the amount of detail in the simulation and the time scale at which the simulation takes place³⁵. The ever-increasing computing power and efficiency of simulation algorithms help in overcoming these obstacles and, in the last decades, the number of applications has been on the rise, as we review in the following sections.

Receptor tyrosine kinases

RTKs are characterized by a single TM helix, with an extracellular region that recognizes a ligand and an intracellular tyrosine kinase domain (**Figure 3.2a**)²². This global structure of RTKs is highly evolutionary conserved, both within the human genome and between different species³⁷. The main activation of RTKs is through dimerization or oligomerization after binding a ligand. Often these ligands themselves have a dimeric nature to assist in the activation process³⁷. After dimerization, the RTKs are able to auto-phosphorylate each other's kinase domains. The phosphorylated kinase domain then allows proteins that contain an SH2 domain to bind. The intracellular kinase domain can phosphorylate downstream kinases, which can induce a range of different cellular effects such as cellular differentiation, growth, and proliferation²². Aberrant activation of some of these pathways plays a key role in multiple cancers. One example is the Ras/MAPK pathway, which contains the extracellular-signal-regulated ERK5 and p38 kinases downstream. ERK5 is known to play a role in tumor invasion, while p38 is able to regulate the activity of the transcription factor p53, a central protein that is found to be mutated in over half of all tumors^{38,39}. Another RTK signaling pathway that often plays a role in cancer is the PI3K/Akt/mTOR pathway. PI3K is phosphorylated by RTKs, after which PI3K phosphorylates Akt. When Akt is activated, it is able to

phosphorylate transcription factors such as mTOR which can increase cell survival and cellular proliferation³⁸.

The RTK family is the most strongly associated with cancer of the three membrane protein families that are reviewed here. Famous examples of aberrant RTK signaling in cancer are epidermal growth factor receptors HER2 in breast cancer⁴⁰ and EGFR in multiple cancers^{41,42}. For a more detailed overview of the role of RTKs in cancer, the review by Du et al. is recommended²². Many inhibitors of RTKs are currently used in clinics to treat several oncological diseases. As of January 1st, 2024, 43 small molecule RTK inhibitors have been approved for anticancer indications, besides 37 inhibitors targeting other kinase families⁴³. Apart from small-molecule drugs, biologics are also used in clinics to combat aberrant RTK signaling. A famous example is the HER2 inhibitor trastuzumab, which is effective in the treatment of HER2+ breast cancer⁴⁴.

The main alterations in RTK structure and function leading to cancer development result in increased kinase activity (**Figure 3.3a**). Alterations in the kinase domain can increase the stability of the RTK dimer, leading to constitutive activation independent of the ligand²². This can be the result of point mutations that are then considered to be drivers of oncogenicity. Indeed in the case of EGFR, 90% of the mutations in lung cancer are found in the genetic regions that contain the kinase domain⁴⁵. Several mutant driver prediction tools, which rely on different ML tools, are available to forecast the pathogenic effect of these mutations, although their level of agreement is limited. Interestingly, it is higher in RTKs compared to other kinase families, possibly due to the wealth of available training data⁴⁶. Structurally, the effect of mutations in the kinase domain is easier to study, since the intracellular kinase domain of RTKs can be determined experimentally as a soluble protein. For example, structure-based approaches have been able to shed light on the mechanisms of constitutive activation triggered by the D816V mutation in the kinase domain of c-Kit⁴⁷. However, recent studies have highlighted the importance of mutations in the extracellular and TM domains of RTKs^{48,49}. Despite the different characteristics of the oncogenic mutations across domains, driver prediction ML models have been able to predict with equal success the oncogenicity of mutations in the extracellular and kinase domains⁵⁰. Moreover, structural studies have allowed us to understand that these mutations in non-kinase domains trigger constitutive ligand-free activation via covalent extracellular or TM dimerization^{50,51}.

Constitutive activation can also be the effect of chromosomal rearrangements leading to fusion proteins, which are very common in RTKs. Targeting these fusion proteins in anticancer therapies is very promising because they are not present in healthy cells. Fusion genes can be detected from sequencing data, although the characterization of their oncogenic and druggability potential, and therefore their clinical relevance, is not trivial⁵². Computational analyses of genomic, transcriptomics, and drug sensitivity data can be used to prioritize oncogenic⁵³ and actionable RTK gene fusions⁵⁴. From validated genetic fusions, ML models can also be constructed to further improve detection, which at the moment still has a problem of high false positive rates. For example, the method developed by Hafstað et al. showed that including an ML-based filter on top of RNA-seq-based fusion detecting algorithms improved the true positive detection

rate, although domain-specific information needed to be provided to the model⁵⁵. Deep learning models have also been developed to predict oncogenicity starting from the fusion protein sequence without providing any oncogenic domain-specific features⁵⁶.

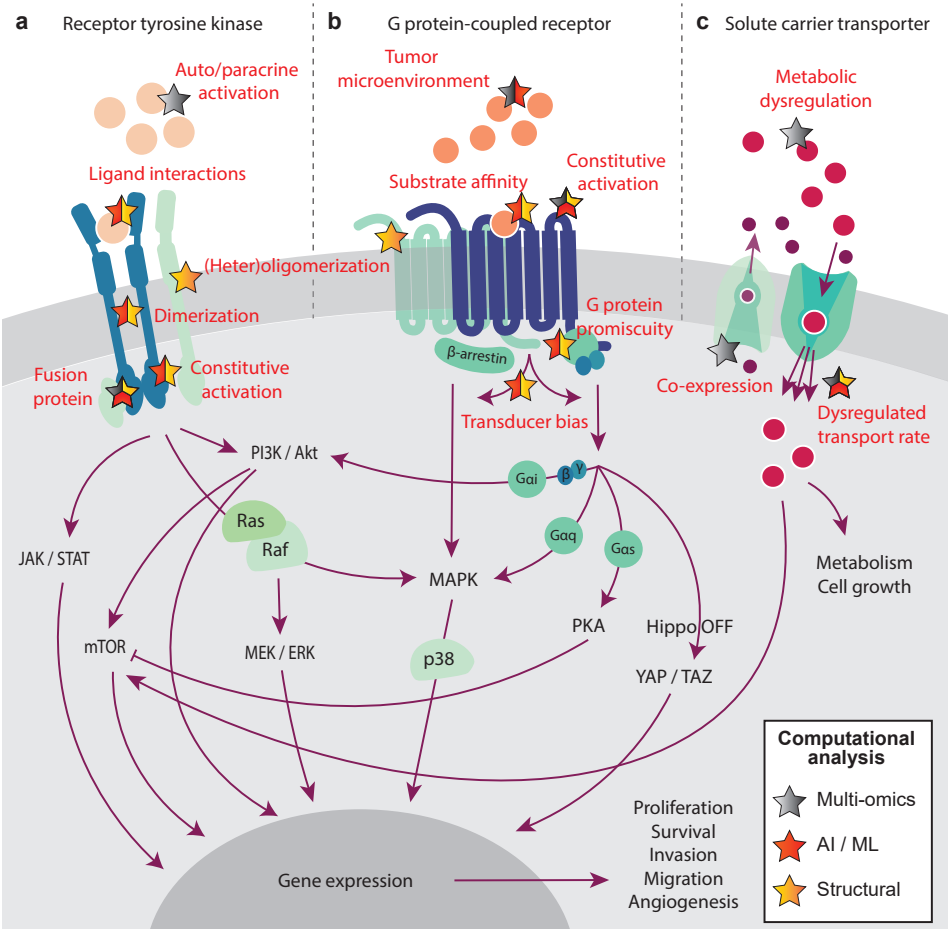


Figure 3.3. Functional and structural alterations in membrane proteins leading to cancer progression that can be characterized by one or several computational methods, including multi-omics analyses (grey star), artificial intelligence or machine learning algorithms (AI/ML, red star), and structure-based approaches (yellow star). In particular, three membrane protein families are explored: **a)** Receptor tyrosine kinases - RTK, activated by endogenous ligands represented by light orange spheres **b)** G protein-coupled receptors - GPCR activated by endogenous ligands represented by orange spheres, and **c)** Solute carriers - SLC that transport substrates represented by pink and purple spheres.

Given the historical focus on kinase domains, the study of the effect of cancer alterations in substrate affinity on the extracellular domain of the RTK is not very extended. In fact, there are very few structures of full-length RTKs containing the TM and extracellular domains⁵⁷. However, these can be very useful in designing monoclonal antibodies targeting the ligand-binding region⁵⁸. Structural characterization has also enabled

the determination of the mechanism behind the inhibitory synergy of Pertuzumab and Trastuzumab, showing with cryo-EM structures that they do induce cooperative binding⁵⁹. ML models based on structural signatures have been leveraged to improve the design of mAbs^{60,61}. The structural analysis of the kinase domain, on the other hand, has been very useful to identify and prioritize small molecules targeting this domain, and to explain the reasons for resistance⁶². The wealth of experimental data for kinase inhibitors has made it possible to use ML models to screen not only potency for wild-type⁶³ and mutant RTKs⁶⁴ but also clinical responses associated with gene expression signatures⁶⁵. Beyond small molecule screening, ML models have also been employed to generate de novo RTK inhibitors by combining 2D and 3D features of known kinase inhibitors⁶⁶. While most drugs are initially designed to bind to a specific target, some drugs bind to multiple kinases, which can be predicted through poly-pharmacology ML models⁶⁷. The combination of ML and structure-based methods was also leveraged to identify potent small molecules that block the dimerization of the kinase domain¹⁸. What is more, basing these models on structural features has enabled the prediction of drug response toward specific cancer-related RTK mutants⁶⁸.

Finally, pathogenic mutations in RTKs can induce aberrant dimerization or oligomerization that leads to increased signaling. These alterations can be studied through structural analyses. For example, MD simulations helped identify ephrin type-A receptor EphA4 melanoma mutation L920F in the C-terminus as the destabilizing factor leading to receptor trimerization instead of dimerization⁶⁹. Similarly, MD simulations showed that oncogenic mutation V536E in the TM domain of platelet-derived growth factor receptor PDGFRA is responsible for stabilizing a tetrameric conformation responsible for constitutive activation. Further dimerization alterations leading to cancer, such as heterodimerizations of EGFR with other RTKs, have been studied with MD, for which only the kinase domain is needed⁷⁰. Beyond receptor-specific abnormalities, increased RTK signaling in cancer can also be triggered by autocrine and paracrine activation. This is the result of ligand overexpression, which in turn can be studied by multi-omics computational approaches⁷¹.

G protein-coupled receptors

Ever since the initial characterization of the rhodopsin structure by Schertler et al.⁷², a GPCR structure is recognized by its seven TM α -helices, collectively the 7TM domain (**Figure 3.2b**). Additionally, they include an N-terminus, three extracellular loops (ECL), three intracellular loops (ICL), and a C-terminus⁷³. Of the three ECLs, ECL2 is usually the longest loop and the most structurally diverse between different GPCRs. An exception is the highly conserved disulfide bond between ECL2 and the TM3 α -helix⁷⁴. Whereas ECL2 is of paramount importance to change the conformation of the GPCR after ligand binding, contacts with the ligand binding to the orthosteric pocket usually happen with ECL1 or ECL3⁷⁴. The orthosteric binding pocket is in the extracellular side of the 7TM domain and usually comprises TM3, TM5, TM6, and TM7⁷⁵. Additionally, many allosteric pockets have been described for GPCRs⁷⁶. Upon activation of the GPCR by a ligand, the structure suffers a conformational rearrangement that primarily involves

TM5 and TM6^{75,77}. This conformational change often induces the heterotrimeric G protein that is bound to the ICL2 and ICL3 of the GPCR to exchange its bound GDP for GTP, activating the G protein in the process⁷⁵. Alterations in the GPCR structure due to mutations can lead to e.g., constitutive activation of the receptor, where it remains in an active state in the absence of the (endogenous) agonist⁷⁴.

GPCRs are the most commonly targeted proteins by drugs, with estimations showing that 35% of all developed drugs have a GPCR as their target⁷⁸. Multiple GPCRs have been extensively associated with cancer, for example, the thyrotropin receptor in thyroid adenomas⁷⁹, estrogen receptors GPER1 and GPR30 in breast cancer^{80,81}, and gonadotropin-releasing hormone receptor GnRH in prostate cancer. In fact, hormonal therapy targeting GnRH is used in the clinic to combat prostate cancer⁸². The smoothened receptor SMO, which is part of the Hedgehog pathway is also currently targeted by small molecule antagonists in the treatment of basal cell carcinomas⁸³. Furthermore, there is a range of GPCR antagonists that have been or are currently tested in (pre)clinical trials such as Ki16198 LPA receptor inhibitor for pancreatic cancer⁸⁴ and astrasentan endothelin receptor inhibitor in prostate cancer⁸⁵. The astrasentan trial however, like multiple other GPCR antagonists that were developed, had to be ended due to unwanted side-effects occurring in patients⁸⁵. Many clinical candidates target GPCRs involved in the tumor microenvironment (TME), such as chemokine receptor CXCR4, which are promising targets in immunotherapy^{20,83}. For an extensive overview of the current state of GPCR targeting drugs in oncology, the reviews by Arang and Gutkind and Usman *et al.* are recommended^{83,86}.

Computational analyses have proven relevant in identifying the role of GPCRs in the TME and their potential role in immunotherapy. For example, the computational analysis of multi-omics data helped pinpoint chemokine receptor axes relevant to particular cancer types and, more importantly, the epigenetic mechanisms responsible for their overexpression (**Figure 3.3b**)⁸⁷. Beyond cancer types, GPCR expression signatures extracted with ML models have also been shown to allow head and neck cancer patient stratification into subtypes leading to differential sensitivity to immunotherapy⁸⁸. A similar approach enabled the classification of melanoma patients based on survival and response to immunotherapy based on combined GPCR-TME multi-omics data⁸⁹. These applications have a high potential to define GPCRs as immune biomarkers to help in cancer treatment and patient stratification.

On the tumor side, computational tools can help study the constitutive activation of GPCRs, which may lead to the onset of cancer by inducing downstream cellular pathways^{79,90}. An example is the frizzled receptors, which indirectly activate the Wnt pathway, a pathway that is strongly linked to the progression of cancer^{91,92}. Genomic analyses have been able to identify oncogenic mutational drivers among GPCR genes, particularly SMO⁹³. However, most GPCR mutants do not share the characteristics of classical drivers, such as a high prevalence. With a much lower mutation prevalence than in RTKs, identifying GPCR drivers needed the integration of multi-omics data⁹⁴, or the characterization of multi-gene oncdrivers⁹⁵, for which computational tools have been crucial. GPCRs of interest in cancer have also been pinpointed based solely on dysregulated

expression in different cancer types, for which there does not seem to be a pan-cancer common profile⁹⁶. Given the challenges of predicting the oncogenic status of GPCR mutations, many authors have opted to study the structural impact of cancer-related mutations on the receptor's stability and activation mechanism. This can be useful to pinpoint novel potential biomarkers, such as the olfactory receptor OR2T7 destabilizing mutation D125V in glioblastoma⁹⁷, or to gain further insights into the activating and binding mechanisms of established anticancer targets, such as CXCR4⁹⁸ or SMO⁹⁹, to improve the development of targeted therapies.

The combination of structure-based and ML tools has also made it possible to get an insight into the different mechanisms behind GPCR involvement in cancer. The principal pathways leading to aberrant GPCR signaling in cancer concern G protein promiscuity and biased signaling⁸⁵. Canonically, every GPCR preferentially activates one of the main four subtypes of G protein α subunits – G_{α_i} , G_{α_q} , G_{α_s} , and $G_{\alpha_{12/13}}$. However, some G proteins are more important than others in the development of cancer, which explains why certain cancer-related GPCR mutations lead to G protein promiscuity^{85,86}. ML models that predict the probability of different GPCR variants binding to the different G_{α} protein subunits have been developed, where the GPCR embeddings were generated from the receptor's sequence¹⁰⁰. This method also allowed the assessment of bias towards different signaling partners beyond G proteins, by considering also β -arrestins as interacting partners, which could have very important implications in the development of biased ligands that trigger preferentially one signaling pathway. However, structurally there does not seem to be a clear conformation basis for transducer biases¹⁰¹, which would introduce an important risk of side effects to these therapies.

Similarly to RTKs, the formation of heterodimers has been shown to trigger aberrant GPCR signaling in cancer¹⁰². The structural analysis of the homo- and heterodimerization patterns and stability therefore introduces novel avenues for treatment. However, in the case of GPCRs, the lipidic environment seems to be extremely determinant in the formation of GPCR oligomers¹⁰³, which introduces additional experimental and computational constraints in the mechanistic analyses¹⁰⁴. Surpassing the technical challenges, however, can help gain insights into cancer-related mutants leading to distinct di/oligomerization patterns that in turn result in biased signaling¹⁰⁵.

Despite the challenges to computationally assess GPCR oncogenic mechanisms due to the limited availability of training data, the wealth of data collected in non-oncological GPCR drug discovery campaigns is a very good starting point for the discovery of anticancer therapies targeting GPCRs. There are several examples in this area, such as structure-based virtual screenings of novel small molecules targeting free fatty acid receptor FFAR4 for colorectal cancer¹⁰⁶, or adhesion receptor ADGRF5 for breast cancer^{107,108}; or ligand-based screenings of small molecules targeting oxoeicosanoid receptor OXER1 that signal specifically through G_{α_i} and/or $G_{\beta\gamma}$ for prostate cancer¹⁰⁹. Beyond providing a wealth of data for novel hit identification, approved GPCR therapies can be considered to be repurposed for oncological applications. To this end, the analysis of omics data can assist in identifying GPCRs with approved drugs that play an important role in cancer survival, such as dopamine receptor 2 in osteosarcoma¹¹⁰. Structure-based

and ML applications are also common in drug repurposing and can equally be leveraged for cancer applications. However, in the case of GPCRs, there are many risks of off-target effects and unintended implications in the cancer phenotype¹¹¹, thus omics-aware approaches are preferred.

Solute carriers

In contrast to the GPCR and RTK families, which both have recognizable basic structures, the SLCs family consists of very diverse proteins. A structure that many SLCs do have in common consists of 10 to 14 TM α -helices (**Figure 3.2c**)¹¹². However, due to the difficulties of obtaining SLC protein structures, they are mostly classified on the basis of their known sequence. SLCs are normally considered of the same family when they have an overlap in sequence of at least 20%¹¹². ML approaches have been relevant in classifying SLCs into families¹¹³. While SLC families are structurally diverse, they remain highly evolutionary conserved within Bilaterian species, with glucose transporters being conserved within all eukaryotes¹¹⁴. SLCs do not only have a high sequential and structural variety, but their transport mechanisms are also very diverse both conformationally and dynamically, which poses a big strain for structure-based methods²⁴.

SLCs transport differing molecules through the cell membrane, such as ions, lipids, and carbohydrates¹¹⁵. As Warburg *et al.* noticed back in 1927, the metabolism in tumor cells differs from that in non-tumor cells¹¹⁶. This change in metabolism is achieved through, among other things, changes in the expression of SLCs in the cell¹¹⁷. A well-known change in SLC expression in cancer is the upregulation of the glucose transporters to meet the increased demand for glucose in tumor cells¹¹⁸. However, no drugs are currently targeting SLCs in an oncological setting. Liu *et al.* performed a preclinical study in which a glucose transporter GLUT1 inhibitor was able to inhibit cancer cell growth, but this compound was not pursued any further¹¹⁹. The monocarboxylate transporter MCT1 inhibitor AZD3965 was tested in a phase 1 clinical trial on patients suffering from advanced stages of cancer, but this drug did not enter phase 2 clinical trials¹²⁰.

Multi-omics analyses of SLCs in cancer have been crucial in detecting aberrant mechanisms leading to dysregulated transport rates (**Figure 3.3c**). Most commonly, abnormal expression of SLCs beyond glucose transporters has been associated with increased transport of metabolites and building blocks necessary for cancer development, which has been used to identify SLCs as prognostic biomarkers in pan-cancer^{121,122} and cancer-specific studies^{123–125}. In this context, ML has also helped further discriminate between the genes with the biggest effect within the cancer signatures¹²⁶. What is more, these transcriptomic analyses have been able to identify SLC co-expression patterns that effectively influence cancer development and that can be used as more precise biomarkers than unique SLC signatures^{127,128}. Moreover, additional omics data types can help further classify tumor subtypes and make sense of the mechanisms leading to cancer development, for example by linking expression profiles to genomic^{128,129} or metabolomic data^{130,131}. The latter provides an additional advantage since metabolic dysregulation is a good candidate for faster biomarker detection in liquid biopsies¹³².

ML and structure-based approaches have also been able to elucidate the role of point mutations triggering changes in transport function, although not frequently in cancer. Polymorphisms in many SLC families are related to several non-oncological diseases, such as cystinuria or ataxia, as well as drug sensitivity, and have mostly been studied in this context¹³³. Several analyses have demonstrated the effect of point mutations in SLC structural and functional changes, as well as the potential risk posed by rare uncharacterized mutations^{133–136}. These methods can be further explored in the context of cancer-related mutations, which have also been shown to affect transport efficiency and conformational dynamics in **Chapter 6**¹⁴. Structural changes can further be exploited to design and virtually screen SLC targeting compounds, as demonstrated for organic anion transporting polypeptide – OATPs¹³⁷. ML-based virtual screening is also possible, but a relative lack of bioactivity data for SLCs is still a big drawback compared to other protein families more broadly explored, such as RTKs and GPCRs^{137,138}.

Crosstalk between membrane proteins

The fact that proteins do not exist in isolation is one of the most difficult aspects to tackle in cancer research. In turn, crosstalk between proteins can lead to compensation mechanisms, synergistic effects, and therapy resistance. Protein interplay has been extensively characterized for different members of the same family, for example triggering synergy by co-expression in SLCs¹²⁸, or the activation of compensatory networks by RTKs as a mechanism of drug resistance¹³⁹. However, the crosstalk can also happen with members from other membrane protein families (**Figure 3.4**), which in turn opens opportunities for novel therapeutic avenues that can also be explored computationally.

Crosstalk between GPCRs and other membrane proteins can lead to oncogenic events. For example, the insulin receptor is known to interact with multiple GPCRs to commence the mTOR pathway¹⁴⁰. Multiple transactivations between GPCRs and EGFR that induce oncogenic pathways are described, such as GPR30 and EGFR to activate the oncogenic MAPK and PI3K/Akt pathways or the protease-activated receptor 1 and EGFR in breast cancer^{141,142}. Moreover, activated RTKs in cancer have been shown to activate GPCR signaling pathways via direct interaction with G proteins¹⁴³. Computationally, structure-based approaches can be used to gain insights into the mechanisms leading to aberrant signaling¹⁴³. Where RTKs and GPCRs are often concerned with activating downstream proteins to exert an effect, the prime task of SLCs is to transport molecules through the cell membrane, meaning SLCs themselves cannot activate oncogenic pathways. There are however interactions between RTKs and GPCRs with SLCs that aid tumor cells. For example, EGFR is known to be able to stabilize the Glucose transporter SGLT1 in tumor cells to increase cell survival¹⁴⁴.

Further crosstalk with SLCs is characterized by shared substrates. This is the case for many GPCR ligands, whereby the expression of SLCs serves as a regulatory mechanism for GPCR ligand availability¹⁴⁵. An example is monocarboxylate transporter MCT1, which is able to efflux succinate from the cell. This succinate is then able to bind the succinate receptor SUCNR1 (a GPCR), inducing a proinflammatory response¹⁴⁵. ML models can identify novel and approved small molecules with shared GPCR and SLC

targets¹⁴⁶ that can be exploited for drug repurposing or poly-pharmacology approaches in cancer when linked to multi-omics cancer analyses.

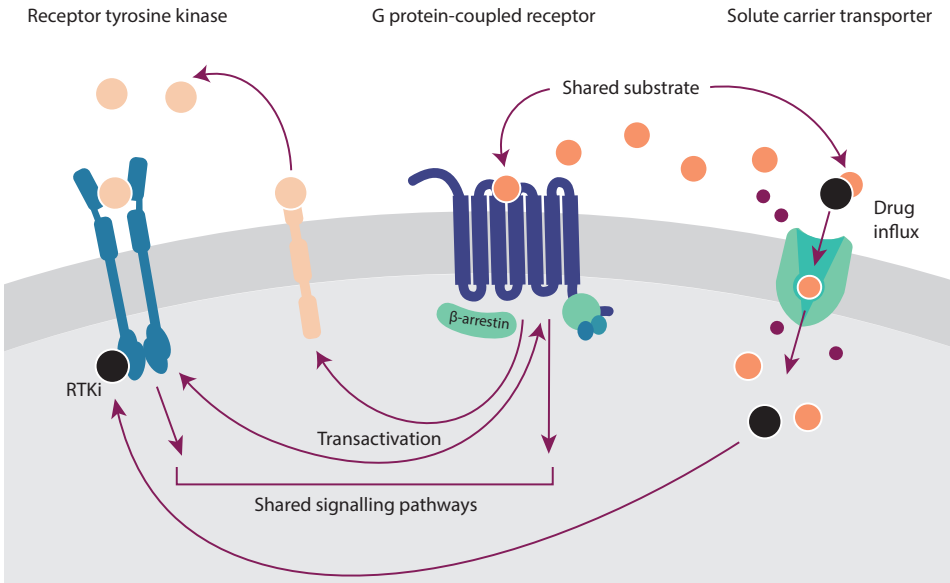


Figure 3.4. Crosstalk between three membrane protein families – Receptor tyrosine kinases (RTK), G protein-coupled receptors (GPCR), and Solute carriers (SLC) – in cancer. The expression of RTK endogenous ligands, represented by light orange spheres, can be induced by GPCRs. GPCR ligands, represented by orange spheres, can be also the substrates of SLCs. SLCs also transport RTK inhibitors (RTKi) into the cell.

Of particular relevance in cancer treatment is the transport of many anticancer drugs, including RTK inhibitors such as sunitinib, via SLCs. Thus, alterations in SLCs are a prominent cause of therapy resistance, which can be explored via multi-omics and drug sensitivity analyses^{147,148}. Of note, alterations in SLCs - together with other genes - can be responsible not only for resistance to targeted therapies but also for first-line chemotherapy¹⁴⁹. As a result of these alterations in membrane transporters, not only drug sensitivity is affected, but also prognosis¹⁵⁰, which can help stratify populations for treatment selection.

Conclusions

Membrane proteins are very promising anticancer targets, but their study is hindered by experimental challenges. Computational tools used in drug discovery pipelines can help overcome some of these challenges, although they are not free of their own obstacles. In particular, the main bottleneck is the lack of experimental data to train ML algorithms or to apply and validate structure-based approaches on. Not all membrane protein families, however, suffer equally from these issues. Historically relevant families such as RTKs have a vast wealth of experimental cancer data and many approved anticancer small molecules, which provide an excellent starting point for ML applications.

Moreover, the kinase domain can be experimentally determined and simulated as a soluble kinase, decreasing the threshold for structure-based approaches. This, however, means that the TM and extracellular domains of RTKs are rather unexplored computationally, even though targeting these domains could be key to avoiding off-target effects. Pharmacologically relevant families underexplored in cancer research, such as GPCRs, lack cancer-related data but they compensate for that in non-oncological data. In fact, many computational approaches have been used to study and bring GPCR-targeting molecules to the market. These tools and knowledge can be easily repurposed for oncological applications, although their relevance for this particular applicability domain should be backed up by multi-omics analyses. Finally, in membrane protein families where the lack of experimental data is very prominent, such as SLCs, family-wide tools should be explored that leverage data from other membrane protein families. These can facilitate the prediction of the effect of mutations in TM domains in particular^{151,152}, or assess the relevance of soluble counterparts of membrane proteins in experimental and computational approaches¹⁵³. Regardless of the wealth of data available for each membrane protein family, they can all benefit from additional computational approaches that consider a holistic view of the tumor and its environment. Some examples include the prediction of the effect of mutations in gene expression¹⁵⁴, or the occurrence of mutant signatures as latent drivers¹⁵⁵, which could be further explored to prioritize personalized cancer therapies¹⁵⁶. Moreover, the extrapolation of methods beyond their conventional use cases, for example, the application of ML algorithms to analyze structural complexes, can help circumvent some of the classical bottlenecks and assist in the design of novel therapies¹⁵⁷. In conclusion, computational tools can help analyze the relevance and mechanisms behind membrane protein dysregulation in cancer and will be crucial tools for prioritizing anticancer targets and improved therapies with increasing amounts of data and computational power.

References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).
2. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
3. Kampen, K. R. Membrane Proteins: The Key Players of a Cancer Cell. *J Membrane Biol* **242**, 69–74 (2011).
4. Lin, C.-Y. *et al.* Membrane protein-regulated networks across human cancers. *Nat Commun* **10**, 3131 (2019).
5. de Jong, E. & Kocer, A. Current Methods for Identifying Plasma Membrane Proteins as Cancer Biomarkers. *Membranes* **13**, 409 (2023).
6. Sojo, V., Dessimoz, C., Pomiankowski, A. & Lane, N. Membrane Proteins Are Dramatically Less Conserved than Water-Soluble Proteins across the Tree of Life. *Molecular Biology and Evolution* **33**, 2874–2884 (2016).
7. Hedin, L. E., Illergård, K. & Elofsson, A. An Introduction to Membrane Proteins. *J. Proteome Res.* **10**, 3324–3331 (2011).
8. Sowlati-Hashjin, S., Gandhi, A. & Garton, M. Dawn of a New Era for Membrane Protein Design. *BioDesign Research* **2022**, 9791435 (2022).
9. Rahman, M. M. *et al.* Emerging Promise of Computational Techniques in Anti-Cancer Research: At a Glance. *Bioengineering* **9**, 335 (2022).
10. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
11. Gorostiola González, M., Janssen, A. P. A., IJzerman, A. P., Heitman, L. H. & van Westen, G. J. P. Oncological drug discovery: AI meets structure-based computational research. *Drug Discovery Today* **27**, 1661–1670 (2022).
12. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385 (2018).
13. Rodrigues, C. H. *et al.* DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research* **46**, W350–W355 (2018).
14. Gorostiola González, M. *et al.* Molecular insights into disease-associated glutamate transporter (EAAT1 / SLC1A3) variants using in silico and in vitro approaches. *Frontiers in Molecular Biosciences* **10**, 3389 (2023).
15. Zheng, L. *et al.* MoDAFold: a strategy for predicting the structure of missense mutant protein based on AlphaFold2 and molecular dynamics. *Briefings in Bioinformatics* **25**, bbae006 (2024).
16. Burggraaff, L. *et al.* Successive statistical and structure-based modeling to identify chemically novel kinase inhibitors. *Journal of Chemical Information and Modeling* **60**, 4283–4295 (2020).
17. Weng, C.-W. *et al.* Pharmacophore-based virtual screening for the identification of the novel Src inhibitor SJG-136 against lung cancer cell growth and motility. *Am J Cancer Res* **10**, 1668–1690 (2020).
18. Mohanan, A., Melge, A. R. & Mohan, C. G. Predicting the Molecular Mechanism of EGFR Domain II Dimer Binding Interface by Machine Learning to Identify Potent Small Molecule Inhibitor for Treatment of Cancer. *Journal of Pharmaceutical Sciences* **110**, 727–737 (2020).
19. Sakellaropoulos, T. *et al.* A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Reports* **29**, 3367–3373.e4 (2019).
20. Wu, V. *et al.* Illuminating the Onco-GPCRome: Novel G protein-coupled receptor-driven oncoendocrine networks and targets for cancer immunotherapy. *Journal of Biological Chemistry* **294**, 11062–11086 (2019).
21. Lavoro, A. *et al.* In silico analysis of the solute carrier (SLC) family in cancer indicates a link among DNA methylation, metabolic adaptation, drug response, and immune reactivity. *Front. Pharmacol.* **14**, 1191262 (2023).
22. Du, Z. & Lovly, C. M. Mechanisms of receptor tyrosine kinase activation in cancer. *Molecular Cancer* **17**, 58 (2018).
23. Saraon, P. *et al.* Receptor tyrosine kinases and cancer: oncogenic mechanisms and therapeutic approaches. *Oncogene* **40**, 4079–4093 (2021).
24. Wang, W., Gallo, L., Jadhav, A., Hawkins, R. & Parker, C. G. The Druggability of Solute Carriers. *Journal of Medicinal Chemistry* **63**, 3834–3867 (2020).
25. Born, J. & Manica, M. Trends in Deep Learning for Property-driven Drug Design. *Curr Med Chem* **28**, 7862–7886 (2021).
26. Lee, A. G. How lipids affect the activities of integral membrane proteins. *Biochim Biophys Acta* **1666**, 62–87 (2004).
27. Kermani, A. A. A guide to membrane protein X-ray crystallography. *FEBS J* **288**, 5788–5804 (2021).
28. Errasti-Murugarren, E., Bartoccioni, P. & Palacín, M. Membrane Protein Stabilization Strategies for Structural and Functional Studies. *Membranes (Basel)* **11**, 155 (2021).
29. The PyMOL Molecular Graphics System, Version 2.5.2 Schrödinger, LLC.
30. Piper, S. J., Johnson, R. M., Wootten, D. & Sexton, P. M. Membranes under the Magnetic Lens: A Dive into the Diverse World of Membrane Protein

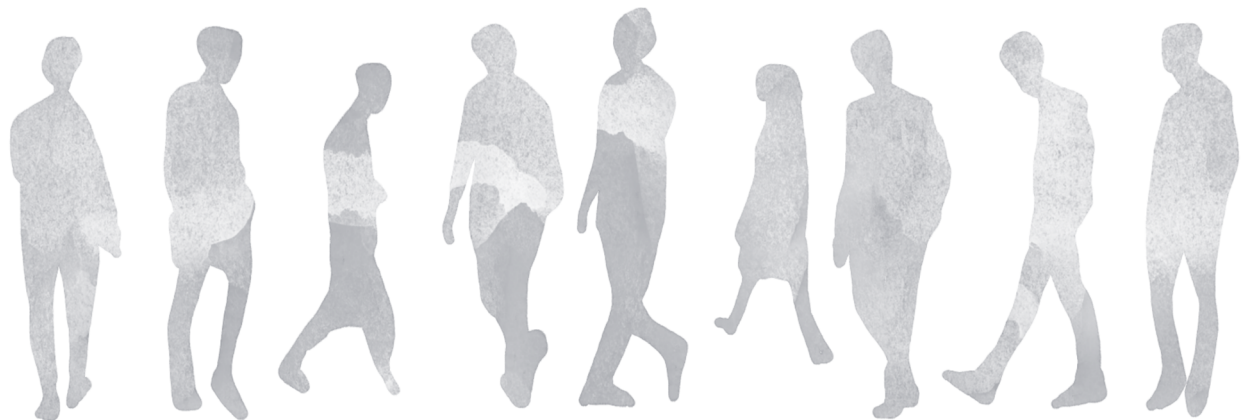
- Structures Using Cryo-EM. *Chem. Rev.* **122**, 13989–14017 (2022).
31. Kooistra, A. J. *et al.* KLIFS: a structural kinase-ligand interaction database. *Nucleic Acids Res* **44**, 365–371 (2015).
32. Pándy-Szekeres, G. *et al.* GPCRdb in 2018: Adding GPCR structure models and ligands. *Nucleic Acids Research* **46**, D440–D446 (2018).
33. Schlessinger, A., Zatorski, N., Hutchinson, K. & Colas, C. Targeting SLC transporters: small molecules as modulators and therapeutic opportunities. *Trends in Biochemical Sciences* **48**, 801–814 (2023).
34. Jambrich, M. A., Tusnady, G. E. & Dobson, L. How AlphaFold2 shaped the structural coverage of the human transmembrane proteome. *Sci Rep* **13**, 20283 (2023).
35. Goossens, K. & De Winter, H. Molecular Dynamics Simulations of Membrane Proteins: An Overview. *J. Chem. Inf. Model.* **58**, 2193–2202 (2018).
36. Škerle, J. *et al.* Membrane Protein Dimerization in Cell-Derived Lipid Membranes Measured by FRET with MC Simulations. *Biophys J* **118**, 1861–1875 (2020).
37. Lemmon, M. A. & Schlessinger, J. Cell signaling by receptor-tyrosine kinases. *Cell* **141**, 1117–1134 (2010).
38. Regad, T. Targeting RTK signaling pathways in cancer. *Cancers* **7**, 1758–1784 (2015).
39. Hoyos, D., Greenbaum, B. & Levine, A. J. The genotypes and phenotypes of missense mutations in the proline domain of the p53 protein. *Cell Death Differ* **29**, 938–945 (2022).
40. Browne, B. C., O'Brien, N., Duffy, M. J., Crown, J. & O'Donovan, N. HER-2 signaling and inhibition in breast cancer. *Curr Cancer Drug Targets* **9**, 419–438 (2009).
41. Wang, Z. *et al.* Mechanistic insights into the activation of oncogenic forms of EGF receptor. *Nat Struct Mol Biol* **18**, 1388–1393 (2011).
42. Lopez-Gines, C. *et al.* New pattern of EGFR amplification in glioblastoma and the relationship of gene copy number with gene expression profile. *Mod Pathol* **23**, 856–865 (2010).
43. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2024 update. *Pharmacol Res* **200**, 107059 (2024).
44. Kreutzfeldt, J., Rozeboom, B., Dey, N. & De, P. The trastuzumab era: current and upcoming targeted HER2+ breast cancer therapies. *Am J Cancer Res* **10**, 1045–1067 (2020).
45. Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nature Reviews Cancer* **7**, 169–181 (2007).
46. Akula, S. *et al.* Large-scale pathogenicity prediction analysis of cancer-associated kinase mutations reveals variability in sensitivity and specificity of computational methods. *Cancer Medicine* **12**, 17468–17474 (2023).
47. Raghav, P. K., Singh, A. K. & Gangenahalli, G. A change in structural integrity of c-Kit mutant D816V causes constitutive signaling. *Mutat Res* **808**, 28–38 (2018).
48. Nair, S. *et al.* Novel EGFR ectodomain mutations associated with ligand-independent activation and cetuximab resistance in head and neck cancer. *PLoS One* **15**, e0229077 (2020).
49. Cleary, J. M. *et al.* FGFR2 Extracellular Domain In-Frame Deletions Are Therapeutically Targetable Genomic Alterations That Function as Oncogenic Drivers in Cholangiocarcinoma. *Cancer Discov* **11**, 2488–2505 (2021).
50. Ishiyama, N. *et al.* Computational and Functional Analyses of HER2 Mutations Reveal Allosteric Activation Mechanisms and Altered Pharmacologic Effects. *Cancer Research* **83**, 1531–1542 (2023).
51. Wagner, A. *et al.* Identification of Activating Mutations in the Transmembrane and Extracellular Domains of EGFR. *Biochemistry* **61**, 2049–2062 (2022).
52. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research* **44**, 4487–4503 (2016).
53. Hernandez, A. *et al.* In silico validation of RNA-Seq results can identify gene fusions with oncogenic potential in glioblastoma. *Sci Rep* **12**, 14439 (2022).
54. Li, J. *et al.* A functional genomic approach to actionable gene fusions for precision oncology. *Science Advances* **8**, eabm2382 (2022).
55. Hafstad, V. *et al.* Improved detection of clinically relevant fusion transcripts in cancer by machine learning classification. *BMC Genomics* **24**, 783 (2023).
56. Lovino, M., Urgese, G., Macii, E., Di Cataldo, S. & Ficarra, E. A Deep Learning Approach to the Screening of Oncogenic Gene Fusions in Humans. *Int J Mol Sci* **20**, 1645 (2019).
57. Diwanji, D., Thaker, T. & Jura, N. More than the Sum of the Parts: Towards Full-Length Receptor Tyrosine Kinase Structures. *IUBMB Life* **71**, 706–720 (2019).
58. Cruz, V. L. *et al.* Binding Affinity of Trastuzumab and Pertuzumab Monoclonal Antibodies to Extracellular HER2 Domain. *International Journal of Molecular Sciences* **24**, 12031 (2023).
59. Hao, Y., Yu, X., Bai, Y., McBride, H. J. & Huang, X. Cryo-EM Structure of HER2-trastuzumab-pertuzumab complex. *PLOS ONE* **14**, e0216095 (2019).
60. Shانهsazzadeh, A. *et al.* Unlocking de novo

- antibody design with generative artificial intelligence. Preprint at *BioRxiv* <https://doi.org/10.1101/2023.01.08.523187> (2023).
61. Balakrishnan, N., Baskar, G., Balaji, S., Kullappan, M. & Krishna Mohan, S. Machine learning modeling to identify affinity improved biobetter anticancer drug trastuzumab and the insight of molecular recognition of trastuzumab towards its antigen HER2. *Journal of Biomolecular Structure and Dynamics* **40**, 11638–11652 (2022).
 62. Majumdar, S., Di Palma, F., Spyarakis, F., Decherchi, S. & Cavalli, A. Molecular Dynamics and Machine Learning Give Insights on the Flexibility–Activity Relationships in Tyrosine Kinome. *J. Chem. Inf. Model.* **63**, 4814–4826 (2023).
 63. Reid, T.-E., Fortunak, J. M., Wutoh, A. & Wang, X. S. Cheminformatic-based Drug Discovery of Human Tyrosine Kinase Inhibitors. *Curr Top Med Chem* **16**, 1452–1462 (2016).
 64. Weng, C.-W. *et al.* Hybrid Pharmacophore- and Structure-Based Virtual Screening Pipeline to Identify Novel EGFR Inhibitors That Suppress Non-Small Cell Lung Cancer Cell Growth. *Int J Mol Sci* **23**, 3487 (2022).
 65. Ferrato, M. H. *et al.* Machine learning classifier approaches for predicting response to RTK-type-III inhibitors demonstrate high accuracy using transcriptomic signatures and ex vivo data. *Bioinform Adv* **3**, vbad034 (2023).
 66. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* **37**, 1038–1040 (2019).
 67. Ma, X. H. *et al.* Virtual screening of selective multitarget kinase inhibitors by combinatorial support vector machines. *Mol Pharm* **7**, 1545–1560 (2010).
 68. Robichaux, J. P. *et al.* Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* **597**, 732–737 (2021).
 69. Light, T. P. *et al.* A cancer mutation promotes EphA4 oligomerization and signaling by altering the conformation of the SAM domain. *Journal of Biological Chemistry* **297**, 100876 (2021).
 70. Zhu, M., Wang, D. D. & Yan, H. Genotype-determined EGFR-RTK heterodimerization and its effects on drug resistance in lung Cancer treatment revealed by molecular dynamics simulations. *BMC Mol Cell Biol* **22**, 34 (2021).
 71. Lam, I., Pickering, C. M. & Mac Gabhann, F. Context-Dependent Regulation of Receptor Tyrosine Kinases: Insights from Systems Biology Approaches. *Wiley Interdiscip Rev Syst Biol Med* **11**, e1437 (2019).
 72. Schertler, G. F. X., Villa, C. & Henderson, R. Projection structure of rhodopsin. *Nature* **362**, 770–772 (1993).
 73. Rosenbaum, D. M., Rasmussen, S. G. F. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459**, 356–363 (2009).
 74. Wheatley, M. *et al.* Lifting the lid on GPCRs: the role of extracellular loops. *Br J Pharmacol* **165**, 1688–1703 (2012).
 75. Weis, W. I. & Kobilka, B. K. The Molecular Basis of G Protein–Coupled Receptor Activation. *Annual Review of Biochemistry* **87**, 897–919 (2018).
 76. Hedderich, J. B. *et al.* The pocketome of G-protein-coupled receptors reveals previously untargated allosteric sites. *Nat Commun* **13**, 2567 (2022).
 77. Schwartz, T. W., Frimurer, T. M., Holst, B., Rosenkilde, M. M. & Elling, C. E. Molecular Mechanism of 7tm Receptor Activation—A Global Toggle Switch Model. *Annual Review of Pharmacology and Toxicology* **46**, 481–519 (2006).
 78. Sriram, K. & Insel, P. A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol Pharmacol* **93**, 251–258 (2018).
 79. Parma, J. *et al.* Somatic mutations in the thyrotropin receptor gene cause hyperfunctioning thyroid adenomas. *Nature* **365**, 649–651 (1993).
 80. Yang, H., Wang, C., Liao, H. & Wang, Q. Activation of GPER by E2 promotes proliferation, invasion and migration of breast cancer cells by regulating the miR-124/CD151 pathway. *Oncology Letters* **21**, 432 (2021).
 81. Steiman, J., Peralta, E. A., Louis, S. & Kamel, O. Biology of the estrogen receptor, GPR30, in triple negative breast cancer. *The American Journal of Surgery* **206**, 698–703 (2013).
 82. Cook, T. & Sheridan, W. P. Development of GnRH antagonists for prostate cancer: new approaches to treatment. *Oncologist* **5**, 162–168 (2000).
 83. Usman, S., Khawer, M., Rafique, S., Naz, Z. & Saleem, K. The current status of anti-GPCR drugs against different cancers. *J Pharm Anal* **10**, 517–521 (2020).
 84. Komachi, M. *et al.* Orally active lysophosphatidic acid receptor antagonist attenuates pancreatic cancer invasion and metastasis in vivo. *Cancer Sci* **103**, 1099–1104 (2012).
 85. Chaudhary, P. K. & Kim, S. An Insight into GPCR and G-Proteins as Cancer Drivers. *Cells* **10**, 3288 (2021).
 86. Arang, N. & Gutkind, J. S. G Protein-Coupled receptors and heterotrimeric G proteins as cancer drivers. *FEBS Lett* **594**, 4201–4232 (2020).
 87. Qualliotine, J. R. *et al.* A Network Landscape of HPVOPC Reveals Methylation Alterations as Significant Drivers of Gene Expression via an Immune-Mediated GPCR Signal. *Cancers (Basel)* **15**, 4379 (2023).
 88. Huang, C., Zhu, F., Zhang, H., Wang, N. & Huang, Q. Identification of S1PR4 as an immune modulator for favorable prognosis in HNSCC

- through machine learning. *iScience* **26**, 107693 (2023).
89. Shen, K. *et al.* Prediction of survival and immunotherapy response by the combined classifier of G protein-coupled receptors and tumor microenvironment in melanoma. *Eur J Med Res* **28**, 352 (2023).
 90. Vecchio, E. A. *et al.* Ligand-Independent Adenosine A2B Receptor Constitutive Activity as a Promoter of Prostate Cancer Cell Proliferation. *J Pharmacol Exp Ther* **357**, 36–44 (2016).
 91. Lappano, R. & Maggiolini, M. GPCRs and cancer. *Acta Pharmacologica Sinica* **33**, 351–362 (2012).
 92. Clevers, H. Wnt/beta-catenin signaling in development and disease. *Cell* **127**, 469–480 (2006).
 93. Sherman, M. A. *et al.* Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nat Biotechnol* **40**, 1634–1643 (2022).
 94. Li, S. *et al.* Multi-omics integration analysis of GPCRs in pan-cancer to uncover inter-omics relationships and potential driver genes. *Comput Biol Med* **161**, 106988 (2023).
 95. Raimondi, F. *et al.* Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm. *Oncogene* **38**, 6491–6506 (2019).
 96. Suteau, V. *et al.* Identification of Dysregulated Expression of G Protein Coupled Receptors in Endocrine Tumors by Bioinformatics Analysis: Potential Drug Targets? *Cells* **11**, 703 (2022).
 97. Sharp, A. K. *et al.* Biophysical insights into OR217: Investigation of a potential prognostic marker for glioblastoma. *Biophysical Journal* **121**, 3706–3718 (2022).
 98. Rebolledo-Bustillo, M. *et al.* Structural Basis of the Binding Mode of the Antineoplastic Compound Motixafortide (BL-8040) in the CXCR4 Chemokine Receptor. *Int J Mol Sci* **24**, 4393 (2023).
 99. Kumari, S., Mitra, A. & Bulusu, G. Structural dynamics of Smoothened (SMO) in the ciliary membrane and its interaction with membrane lipids. *Biochim Biophys Acta Biomembr* **1864**, 183946 (2022).
 100. Matic, M. *et al.* PRECOGx: exploring GPCR signaling mechanisms with deep protein representations. *Nucleic Acids Research* **50**, W598–W610 (2022).
 101. Seyedabadi, M., Gharghabi, M., Gurevich, E. V. & Gurevich, V. V. Structural basis of GPCR coupling to distinct signal transducers: implications for biased signaling. *Trends Biochem Sci* **47**, 570–581 (2022).
 102. Coke, C. J. *et al.* Simultaneous Activation of Induced Heterodimerization between CXCR4 Chemokine Receptor and Cannabinoid Receptor 2 (CB2) Reveals a Mechanism for Regulation of Tumor Progression. *J Biol Chem* **291**, 9991–10005 (2016).
 103. Gahbauer, S., Pluhackova, K. & Böckmann, R. A. Closely related, yet unique: Distinct homo- and heterodimerization patterns of G protein coupled chemokine receptors and their fine-tuning by cholesterol. *PLOS Computational Biology* **14**, e1006062 (2018).
 104. Di Marino, D., Conflitti, P., Motta, S. & Limongelli, V. Structural basis of dimerization of chemokine receptors CCR5 and CXCR4. *Nat Commun* **14**, 6439 (2023).
 105. Paradis, J. S. *et al.* Computationally designed GPCR quaternary structures bias signaling pathway activation. *Nat Commun* **13**, 6826 (2022).
 106. Pal, A., Curtin, J. F. & Kinsella, G. K. In silico and in vitro screening for potential anticancer candidates targeting GPR120. *Bioorg Med Chem Lett* **31**, 127672 (2021).
 107. Muthiah, I., Rajendran, K. & Dhanaraj, P. In silico molecular docking and physicochemical property studies on effective phytochemicals targeting GPR116 for breast cancer treatment. *Mol Cell Biochem* **476**, 883–896 (2021).
 108. Muthiah, I., Rajendran, K., Dhanaraj, P. & Vallinayagam, S. In silico structure prediction, molecular docking and dynamic simulation studies on G Protein-Coupled Receptor 116: a novel insight into breast cancer therapy. *Journal of Biomolecular Structure and Dynamics* **39**, 4807–4815 (2021).
 109. Panagiotopoulos, A. A., Konstantinou, E., Pirintzos, S. A., Castanas, E. & Kampa, M. Mining the ZINC database of natural products for specific, testosterone-like, OXER1 antagonists. *Steroids* **199**, 109309 (2023).
 110. Tan, M. *et al.* Prediction and Identification of GPCRs Targeting for Drug Repurposing in Osteosarcoma. *Frontiers in Oncology* **12**, 828849 (2022).
 111. Cornwell, A. C. & Feigin, M. E. Unintended Effects of GPCR-Targeted Drugs on the Cancer Phenotype. *Trends in Pharmacological Sciences* **41**, 1006–1022 (2020).
 112. Schlessinger, A. *et al.* Comparison of human solute carriers. *Protein Sci* **19**, 412–428 (2010).
 113. Meixner, E. *et al.* A substrate-based ontology for human solute carriers. *Mol Syst Biol* **16**, e9652 (2020).
 114. Höglund, P. J., Nordström, K. J. V., Schiöth, H. B. & Fredriksson, R. The solute carrier families have a remarkably long evolutionary history with the majority of the human families present before divergence of Bilaterian species. *Mol Biol Evol* **28**, 1531–1541 (2011).
 115. Colas, C., Ung, P. M. U. & Schlessinger, A. SLC transporters: Structure, function, and drug discovery. *MedChemComm* **7**, 1069–1081 (2016).

116. Warburg, O., Wind, F. & Negelein, E. The metabolism of tumors in the body. *J Gen Physiol* **8**, 519–530 (1927).
117. El-Gebali, S., Bentz, S., Hediger, M. A. & Anderle, P. Solute carriers (SLCs) in cancer. *Molecular Aspects of Medicine* **34**, 719–734 (2013).
118. Scafoglio, C. *et al.* Functional expression of sodium-glucose transporters in cancer. *Proc Natl Acad Sci U S A* **112**, E4111–4119 (2015).
119. Liu, Y. *et al.* A small-molecule inhibitor of glucose transporter 1 downregulates glycolysis, induces cell-cycle arrest, and inhibits cancer cell growth in vitro and in vivo. *Mol Cancer Ther* **11**, 1672–1682 (2012).
120. Lin, L., Yee, S. W., Kim, R. B. & Giacomini, K. M. SLC transporters as therapeutic targets: Emerging opportunities. *Nature Reviews Drug Discovery* **14**, 543–560 (2015).
121. Xie, M. *et al.* Systematic pan-cancer analysis identifies SLC35C1 as an immunological and prognostic biomarker. *Sci Rep* **13**, 5331 (2023).
122. Li, J. *et al.* A pan-cancer analysis revealed the role of the SLC16 family in cancer. *Channels (Austin)* **15**, 528–540 (2021).
123. Zhu, J. *et al.* Identification of a Six-Gene SLC Family Signature With Prognostic Value in Patients With Lung Adenocarcinoma. *Frontiers in Cell and Developmental Biology* **9**, 803198 (2021).
124. Zhao, X., Jin, L., Liu, Y., Liu, Z. & Liu, Q. Bioinformatic analysis of the role of solute carrier-glutamine transporters in breast cancer. *Annals of Translational Medicine* **10**, 777–777 (2022).
125. Sun, T., Bi, F., Liu, Z. & Yang, Q. SLC7A2 serves as a potential biomarker and therapeutic target for ovarian cancer. *Aging* **12**, 13281–13296 (2020).
126. Samaržija, I., Trošelj, K. G. & Konjevoda, P. Prognostic Significance of Amino Acid Metabolism-Related Genes in Prostate Cancer Retrieved by Machine Learning. *Cancers* **15**, 1309 (2023).
127. Zhang, Y. *et al.* Co-expression pattern of SLC transporter genes associated with the immune landscape and clinical outcomes in gastric cancer. *Journal of Cellular and Molecular Medicine* **27**, 4181–4194 (2023).
128. Zhou, R. *et al.* Integrative analysis of co-expression pattern of solute carrier transporters reveals molecular subtypes associated with tumor microenvironment hallmarks and clinical outcomes in colon cancer. *Heliyon* **10**, e22775 (2024).
129. Zhang, P. *et al.* SLC31A1 Identifying a Novel Biomarker with Potential Prognostic and Immunotherapeutic Potential in Pan-Cancer. *Biomedicines* **11**, 2884 (2023).
130. Danzi, F. *et al.* To metabolomics and beyond: a technological portfolio to investigate cancer metabolism. *Sig Transduct Target Ther* **8**, 137 (2023).
131. Poplawski, P. *et al.* Coordinated reprogramming of renal cancer transcriptome, metabolome and secretome associates with immune tumor infiltration. *Cancer Cell International* **23**, 2 (2023).
132. Wang, W. *et al.* Cancer metabolites: promising biomarkers for cancer liquid biopsy. *Biomarker Research* **11**, 66 (2023).
133. Schaller, L. & Lauschke, V. M. The genetic landscape of the human solute carrier (SLC) transporter superfamily. *Human Genetics* **138**, 1359–1377 (2019).
134. Koleske, M. L. *et al.* Functional genomics of OCTN2 variants informs protein-specific variant effect predictor for Carnitine Transporter Deficiency. *Proceedings of the National Academy of Sciences* **119**, e2210247119 (2022).
135. Pasquadibisceglie, A., Quadrotta, V. & Polticelli, F. In Silico Analysis of the Structural Dynamics and Substrate Recognition Determinants of the Human Mitochondrial Carnitine/Acylcarnitine SLC25A20 Transporter. *International Journal of Molecular Sciences* **24**, 3946 (2023).
136. Wu, Q. *et al.* Ataxia-linked SLC1A3 mutations alter EAAT1 chloride channel activity and glial regulation of CNS function. *Journal of Clinical Investigation* **132**, e154891 (2022).
137. Tuerkova, A. *et al.* Identifying Novel Inhibitors for Hepatic Organic Anion Transporting Polypeptides by Machine Learning-Based Virtual Screening. *J. Chem. Inf. Model.* **62**, 6323–6335 (2022).
138. Burggraaff, L. *et al.* Identification of novel small molecule inhibitors for solute carrier SGLT1 using proteochemometric modeling. *Journal of Cheminformatics* **11**, 15 (2019).
139. Xu, A. M. & Huang, P. H. Receptor tyrosine kinase coactivation networks in cancer. *Cancer Res* **70**, 3857–3860 (2010).
140. Kisfalvi, K., Rey, O., Young, S. H., Sinnett-Smith, J. & Rozengurt, E. Insulin potentiates Ca²⁺ signaling and phosphatidylinositol 4,5-bisphosphate hydrolysis induced by Gq protein-coupled receptor agonists through an mTOR-dependent pathway. *Endocrinology* **148**, 3246–3257 (2007).
141. Prossnitz, E. R. & Maggiolini, M. Mechanisms of estrogen signaling and gene expression via GPR30. *Mol Cell Endocrinol* **308**, 32–38 (2009).
142. Arora, P., Cuevas, B. D., Russo, A., Johnson, G. L. & Trejo, J. Persistent transactivation of EGFR and ErbB2/HER2 by protease-activated receptor-1 promotes breast carcinoma cell invasion. *Oncogene* **27**, 4434–4445 (2008).
143. Kalogiropoulos, N. A. *et al.* Receptor tyrosine kinases activate heterotrimeric G proteins via phosphorylation within the interdomain cleft of Gαi. *Proceedings of the National Academy of Sciences* **117**, 28763–28774 (2020).
144. Weihua, Z. *et al.* Survival of cancer cells is

- maintained by EGFR independent of its kinase activity. *Cancer Cell* **13**, 385–393 (2008).
145. Sijben, H. J., Superti-Furga, G., IJzerman, A. P. & Heitman, L. H. Targeting solute carriers to modulate receptor-ligand interactions. *Trends Pharmacol Sci* **43**, 358–361 (2022).
 146. Oh, J., Ceong, H.-T., Na, D. & Park, C. A machine learning model for classifying G-protein-coupled receptors as agonists or antagonists. *BMC Bioinformatics* **23**, 346 (2022).
 147. van de Geer, W. S. *et al.* Identifying somatic changes in drug transporters using whole genome and transcriptome sequencing data of advanced tumors. *Biomedicine & Pharmacotherapy* **159**, 114210 (2023).
 148. Qu, Y.-Y., Guo, R.-Y., Luo, M.-L. & Zhou, Q. Pan-Cancer Analysis of the Solute Carrier Family 39 Genes in Relation to Oncogenic, Immune Infiltrating, and Therapeutic Targets. *Frontiers in Genetics* **12**, 757582 (2021).
 149. Buttarelli, M. *et al.* Identification of a novel gene signature predicting response to first-line chemotherapy in BRCA wild-type high-grade serous ovarian cancer patients. *Journal of Experimental & Clinical Cancer Research* **41**, 50 (2022).
 150. Alves, R. *et al.* Genetic Variants of ABC and SLC Transporter Genes and Chronic Myeloid Leukaemia: Impact on Susceptibility and Prognosis. *International Journal of Molecular Sciences* **23**, 9815 (2022).
 151. Pires, D. E. V., Rodrigues, C. H. M. & Ascher, D. B. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Research* **48**, W147–W153 (2020).
 152. Ge, F. *et al.* MutTMPredictor: Robust and accurate cascade XGBoost classifier for prediction of mutations in transmembrane proteins. *Comput Struct Biotechnol J* **19**, 6400–6416 (2021).
 153. Ma, L. *et al.* CrMP-Sol database: classification, bioinformatic analyses and comparison of cancer-related membrane proteins and their water-soluble variant designs. *BMC Bioinformatics* **24**, 360 (2023).
 154. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196–1203 (2021).
 155. Yavuz, B. R., Tsai, C.-J., Nussinov, R. & Tuncbag, N. Pan-cancer clinical impact of latent drivers from double mutations. *Commun Biol* **6**, 202 (2023).
 156. Mateo, L. *et al.* Personalized cancer therapy prioritization based on driver alteration co-occurrence patterns. *Genome Medicine* **12**, 78 (2020).
 157. Qi, X., Zhao, Y., Qi, Z., Hou, S. & Chen, J. Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges. *Molecules* **29**, 903 (2024).



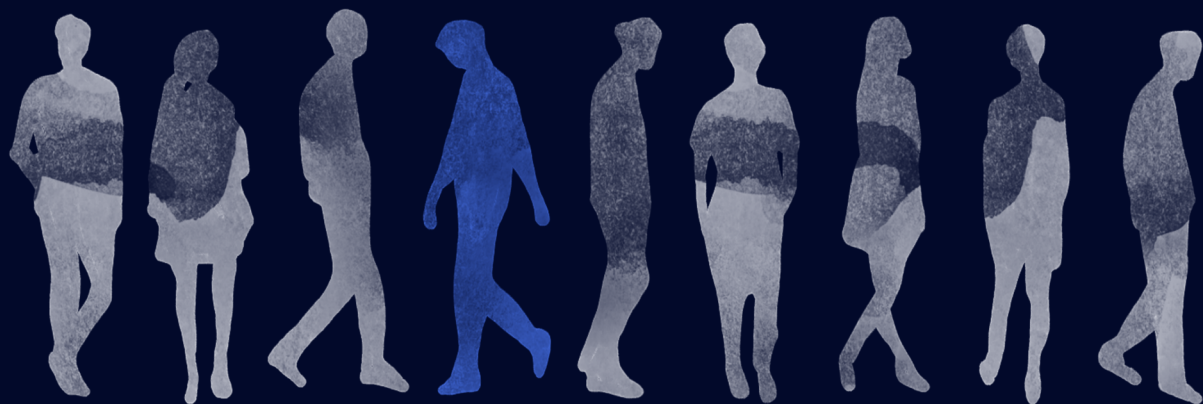
Chapter 4

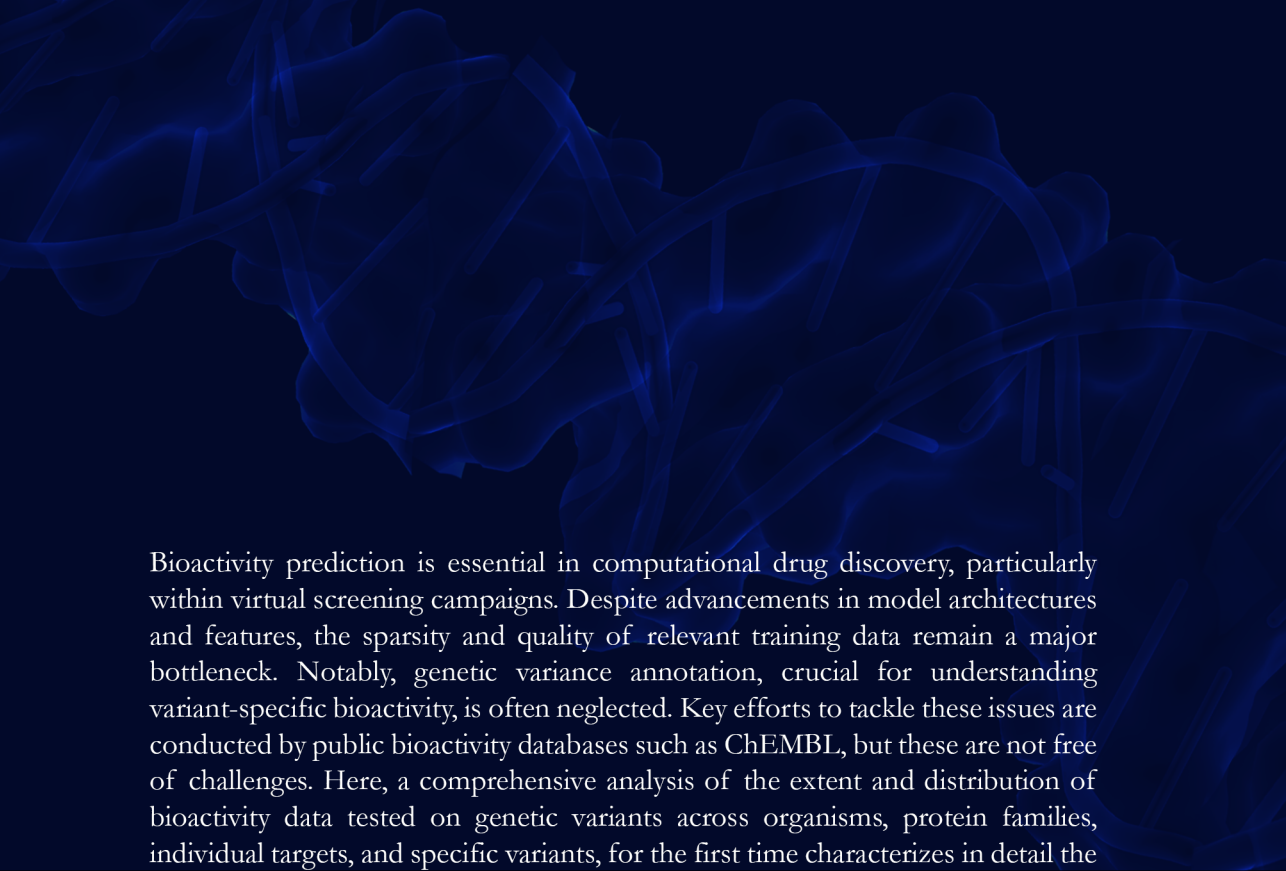
Excuse me, there is a mutant in my bioactivity soup!
A comprehensive analysis of the genetic variability
landscape of bioactivity databases and its effect
on activity modeling

Marina Gorostiola González[†], Olivier J.M. Béquignon[†], Emma J. Manners, Barbara
Zdrazil, Andrew R. Leach, Adriaan P. IJzerman, Laura H. Heitman,
Gerard J.P. van Westen

Adapted from: *ChemRxiv* doi:10.26434/chemrxiv-2024-kxlgm (2024)

[†]These authors contributed equally



An abstract graphic in shades of blue, resembling a complex molecular structure or a network of interconnected lines, occupies the upper half of the page. It features various loops, branches, and dense clusters of lines, giving it a three-dimensional, organic feel.

Bioactivity prediction is essential in computational drug discovery, particularly within virtual screening campaigns. Despite advancements in model architectures and features, the sparsity and quality of relevant training data remain a major bottleneck. Notably, genetic variance annotation, crucial for understanding variant-specific bioactivity, is often neglected. Key efforts to tackle these issues are conducted by public bioactivity databases such as ChEMBL, but these are not free of challenges. Here, a comprehensive analysis of the extent and distribution of bioactivity data tested on genetic variants across organisms, protein families, individual targets, and specific variants, for the first time characterizes in detail the genetic variability landscape in the ChEMBL database and sheds light on the range and consequences of protein amino acid substitutions in bioactivity data distribution and modeling. Furthermore, an extensive set of analysis resources (Python package and notebooks) and a variant-annotated bioactivity dataset are made available to help replicate the analyses described here for any protein of interest and make informed decisions regarding the quality of data for modeling. Finally, the potential to extract variants and subsets of the chemical space with desirable inter-variant bioactivity profiles is demonstrated for data-rich proteins. This approach contributes to more reliable bioactivity modeling, aids noise reduction, and informs decision-making in computational drug discovery.



Introduction

Bioactivity prediction is one of the key techniques in the computational drug discovery pipeline, mostly applied in virtual screening campaigns^{1,2}. Quantitative structure-activity relationship (QSAR) modeling has been around for a long time and can be used to predict ligand bioactivity for a target of interest based on the compound's chemical structural characteristics³. Over time other bioactivity prediction strategies have emerged that include information other than chemistry-derived features⁴⁻⁸. An example is proteochemometric (PCM) modeling, where the protein characteristics are considered in addition to ligand molecular structure, allowing for bioactivity predictions on several targets simultaneously⁸⁻¹⁰.

Every year an increasing number of articles showcase improvements in machine learning and artificial intelligence (AI/ML) bioactivity modeling in the form of novel model architectures or chemical and protein descriptors, among other innovations¹¹⁻¹⁶. Still, previous literature shows that one of the main bottlenecks in bioactivity prediction is the amount and quality of the available data for model training and testing^{17,18}. Several databases, such as ChEMBL and PubChem, aim to compile as much data as possible by extracting it from the literature or accepting deposited datasets, which on its own can introduce errors^{19,20}. Certain annotations like assay cell type, tissue, or genetic variants are not present in all articles or are described differently. In turn, this can result in inconsistencies in information content that affect the quality and comprehensiveness of the data^{21,22}.

Variant annotation in particular is one of the key aspects that should be considered when analyzing bioactivity data²³. The same compound can have a very different bioactivity on different genetic variants of the same protein²⁴⁻²⁷. In fact, some compounds are explicitly designed to have differential bioactivity across variants to, for example, reduce side effects by avoiding targeting the wild-type (WT) protein in anticancer therapies, or to target escape variants in antibiotics or antivirals^{28,29}. However, variant annotation tends to be overlooked in bioactivity databases where, in many cases, it is not present or lacks validation. Moreover, even when variants are annotated - as is the case in the ChEMBL database - they are often ignored when constructing a bioactivity dataset, which only recently has been explicitly described as a potential source of noise^{30,31}. The advantage of modeling variant-annotated data has been demonstrated in variant-rich organisms, such as HIV³², and the implications in human proteins could be similarly important.

Here, we thoroughly evaluate the risks and opportunities presented by variant annotation in bioactivity databases by extensively characterizing variant-annotated bioactivity data in the ChEMBL 31 database. Through an assessment of annotation fidelity, the non-triviality of this task is highlighted, and adjustments are proposed to improve the ChEMBL variant annotation pipeline for future releases. A revised bioactivity dataset with protein amino acid substitution annotations is derived from this work and enriched with curated data from literature³³ (Christmann subset) previously curated as part of the Papyrus dataset³⁴. The additional data is aggregated in this work with the ChEMBL annotated data following the pipeline with rigorous data curation and filtering, and

standardization of molecular structures that were applied to obtain the Papyrus dataset. Furthermore, we investigate the distribution of variant-annotated bioactivity data points in the combined dataset across organisms, protein families, individual targets, and specific variants; and evaluate the effect of variants in bioactivity distribution and modeling. These findings not only contribute to advancing our understanding of the effects of amino acid substitutions in bioactivity but also provide invaluable insights for refining bioactivity data curation practices, particularly concerning variants, for enhanced predictive modeling purposes. Our work also highlights the importance of reporting comprehensively the full sequences of proteins used in bioassays and bioactivity measurements, in both the literature and when depositing data directly into databases.

Results

Variant annotation in bioactivity databases is far from trivial

Genetic variants are currently annotated in the ChEMBL database by manually extracting this information from the original articles for data originating from the scientific literature. Since ChEMBL 22 this information has been mapped to protein targets (alongside their UniProt accessions) and made available in a structured format via the *variant_sequences* table. In this work, an orthogonal approach has been used to evaluate the fidelity and comprehensiveness of these annotations and to include as many variants as possible for the analysis of bioactivities against proteins carrying amino acid substitutions (**Figure 4.1**, steps 1-7). This approach is expert knowledge-agnostic and embodies an automatic pipeline based exclusively on data previously extracted from the database. Its first step consisted of the automatic extraction of amino acid substitution patterns from the assay descriptions of unique pairs of assays and protein targets, and their subsequent validation against the WT protein sequence (**Figure 4.1**, step 2). The extracted substitutions were then compared to the ChEMBL variant annotations in a feedback loop in which mismatches were semi-automatically classified and used to rescue or revert annotations (**Figure 4.1**, step 3). Finally, variant targets were annotated based on this feedback and mapped to ChEMBL bioactivity data. The final variant-enhanced bioactivity dataset (VEBD) was constructed by keeping exclusively bioactivity data for proteins with at least one variant annotated and was lastly enriched with variant-annotated bioactivity data from the Christmann dataset.

Regular expressions were used to extract amino acid substitution patterns from assay descriptions, starting from 376,233 assay-protein target pairs in the ChEMBL 31 database with data suitable for regression modeling. Assay descriptions are extracted and curated in ChEMBL from the primary literature sources in a combined manual and semi-automated pipeline. Of note, genetic alterations other than amino acid substitutions were deemed out of the scope for the initial stage of this project. As exemplified in **Figure 4.1** (step 2) for the assay-target pair CHEMBL832660 - P47900, these expression patterns could extract true substitutions, such as *Y306F*, but also incorrect patterns from the assay description, like *P2Y*. This first step yielded potential substitutions in 52,922 assay-target pairs. Therefore, exceptions were defined from other fields related

to the assay and the target protein, in particular cell type, target preferred name, and target synonyms. This helped to refine the pipeline by rejecting extracted patterns such as *P2Y* that map to a part of the name of the assay target (purinoceptor *P2Y1* in this case, UniProt accession *P47900*) and does not refer to a proline to tyrosine substitution. Indeed, 34,676 assay-target pair substitutions raised at least one exception flag. Of note, these exceptions are less of an issue in the original ChEMBL variant annotation pipeline, since some manual curation is performed. The substitution patterns that had not been flagged as exceptions were validated in the next step by checking the existence of the WT amino acid at the specified position in the target sequence. For example, in the case of the aforementioned *Y306F* substitution pattern, *P2Y1* has indeed a tyrosine residue at position 306 of its sequence, hence this extracted substitution was validated. At this point, several additional exceptions were introduced by extracting patterns that were likely to be falsely validated, such as *M1*, as substitutions are unlikely to appear at the first position of the sequence, yet they would be given a false valid status as the starting codon AUG codes for methionine. This resulted in 8,455 assay-target pairs with WT sequence-validated extracted substitutions.

Next, the extracted and validated substitutions were compared to the originally annotated ChEMBL variants for all assay-target pairs (**Figure 4.1**, step 3, **Supplementary Figure 4.1**). This step, which we refer to as the annotation feedback loop, was included for three reasons, namely 1) to pinpoint highlights and pitfalls, 2) to suggest improvements to the ChEMBL variant annotation pipeline, and 3) to include additional ChEMBL variants and collect the most complete dataset with variant annotated data in the scope of this project. Additionally, it served as a reminder of the non-triviality of the variant annotation process. Given its complexity, the feedback loop is now under review and remains subject to revision. The updated results will be incorporated in a revised version, therefore it is advisable to approach the following preliminary results with caution. Out of the 8,455 assay-target pairs with extracted substitutions, 7,622 (90%) had an identical annotation in ChEMBL. The remaining 833 were missing in ChEMBL, either completely (651) or because they had been flagged as “Undefined mutation” (182). Mismatching variants were further classified to determine their suitability for the VEBD (**Supplementary Table 4.1**, **Supplementary Figure 4.1**). Assays assessing more than one target were rejected for this analysis, as well as assays testing targets with variation corresponding to alterations or genotypes with ambiguous definitions. If a multiple substituted protein was only partially validated; the annotations were rejected. If a validated amino acid substitution was combined with an insertion/deletion/truncation then the substitution was included in this analysis. Finally, non-substitution patterns that had been incorrectly validated against the WT sequence were identified as potential novel exceptions for improving the pipeline. Subsequently, these 833 entries were manually classified into 648 true positives that represent potential novel annotations missed by ChEMBL, and 185 false positives that arise from substitution extraction errors and will be used to refine the current pipeline. The true positive group was included in the final VEBD. Of note, among these were assay-target pairs with either completely novel extracted substitutions or rescues from previously undefined variants that were not fully annotated but were deemed inside the scope of this project. For example, we deemed within scope, variants with co-occurring amino acid substitutions and deletions/duplications, flagged

by ChEMBL as undefined variants and “rescued” for this project.

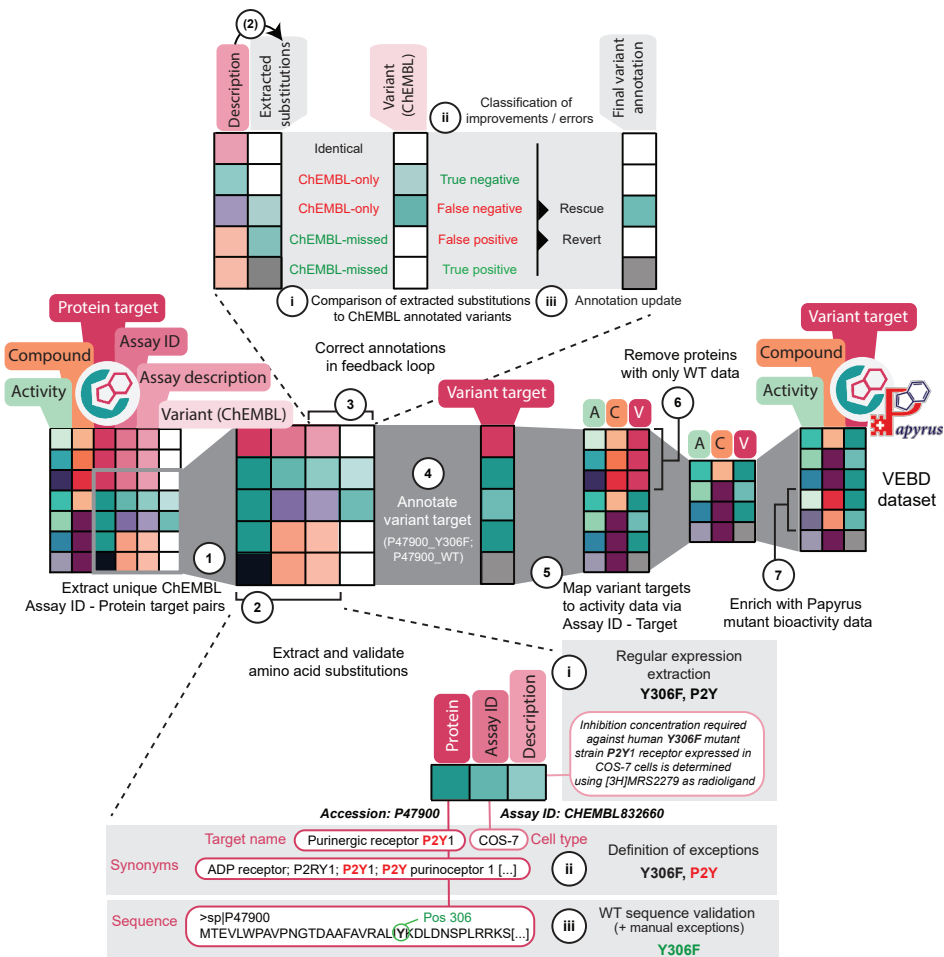


Figure 4.1. Pipeline to construct the variant-enhanced bioactivity dataset (VEBD) from ChEMBL and Papyrus data. (1) Unique assay-target pairs with bioactivity data are extracted from ChEMBL 31. (2) Regular expressions are used to extract amino acid substitution patterns, which are validated by introducing exceptions and mapping them to wild-type (WT) sequences. (3) Extracted substitutions are compared to ChEMBL annotated variants, and the classification of mismatches is used to determine the final annotations. More details of this step are available in Supplementary Figure 4.1. (4) A variant target identifier is defined based on the final variant annotations. (5) The variant target identifier is mapped back to the ChEMBL bioactivity dataset. (6) Proteins with only WT data are filtered out. (7) The bioactivity dataset is standardized and curated similarly to, and enriched with variant data from the Papyrus dataset.

Apart from the 8,455 assay-target pairs with extracted substitutions, 1,600 pairs were found to be annotated only in ChEMBL and not identified by the current variant annotation pipeline. These ChEMBL-only annotated pairs were further evaluated in light of the underlying reasons that led to their exclusion from the current variant annotation pipeline (Supplementary Table 4.2, Supplementary Figure 4.1). ChEMBL

substitutions missed by the regular expression, such as those with unconventional definitions, were incorporated into this analysis unless their initial annotation was “undefined” or a deletion. Extracted substitutions failing validation against the WT sequence were categorized into three groups: 1) If the extracted substitutions matched those in ChEMBL in all aspects except the residue number, the original substitutions were considered a sequence number shift exception and included. 2) If the extracted substitutions fully matched the original ChEMBL annotation but were not valid according to the WT sequence, they were either a) excluded (i.e. if the associated target was a protein family) or b) classified as ambiguous due to a sequence mismatch. 3) Finally, if the extracted substitutions did not align with the original annotation, they were deemed ambiguous due to substitution mismatch or omission and are under review. This analysis led to the classification of ChEMBL-only variants into true negatives (686 misclassified ChEMBL annotations), false negatives (798 ChEMBL expert annotations), and ambiguous (416 ChEMBL-only annotations). True negatives were excluded from the final dataset, while false negatives were rescued from ChEMBL and included. Pairs in the ambiguous group were flagged as undefined variants and included in the final dataset. After the annotation feedback loop, 9,229 assay-target pairs (774 additional assays) were annotated with variants. These were annotated with a variant target identifier as done in the Papyrus dataset by adding the amino acid substitutions as a suffix to the UniProt accession code of the protein. Similarly, bioactivity data points tested on WT proteins were identified by the suffix “WT” after the accession code. Note that the final number of annotated pairs relies on the feedback loop, which is currently under revision; thus, the ultimate count is subject to change in an updated version.

To construct the VEBD, the variant target identifiers were mapped to ChEMBL bioactivity data based on assay-target pairs. Duplicated data from several assays for the same variant target were joined into one single point by dropping data with questioned validity, considered low-quality, and calculating the mean pchembl value or the most common activity flag. This resulted in 1,870,748 compound-target pairs across 6,777 targets, of which 25,259 contained variant targets - 736 with undefined variants - and the rest were WT. The ChEMBL 31 annotated set was merged with the fraction of the Papyrus dataset version 5.5 originating from the Christmann subset, keeping only targets with at least one variant defined for the follow-up analysis of variant-annotated bioactivity. The final combined VEBD for bioactivity analysis contained 455,839 compound-target pairs across 335 proteins, of which 25,086 data points represented data on variant proteins. Of these, 22,992 compound-target pairs originated from ChEMBL 31, 672 from the Papyrus Christmann subset, which were not present in ChEMBL, and 1,422 from both sources. In the following sections, we explore in detail the VEBD.

Variants are heterogeneously represented in bioactivity datasets across protein families

The first observation from the review of the VEBD was that bioactivity data points were not homogeneously distributed across protein families. Proteins were assigned to their corresponding protein families using the levels L1-L5 in the protein family classification

table in ChEMBL. Out of the 455,839 bioactivity data points in the VEBD, more than half were in enzymes (266,328), followed by membrane receptors (96,037), and then the remaining protein families (Figure 4.2a, Supplementary Table 4.3). The percentage of variant-tested bioactivity data with respect to the total amount of bioactivity data – hereby referred to as variant bioactivity percentage – was highest for secreted proteins (10.8%) followed by enzymes (7.8%), but in both cases, it was in the same order of magnitude as the variant bioactivity percentage for the whole dataset (5.5%).

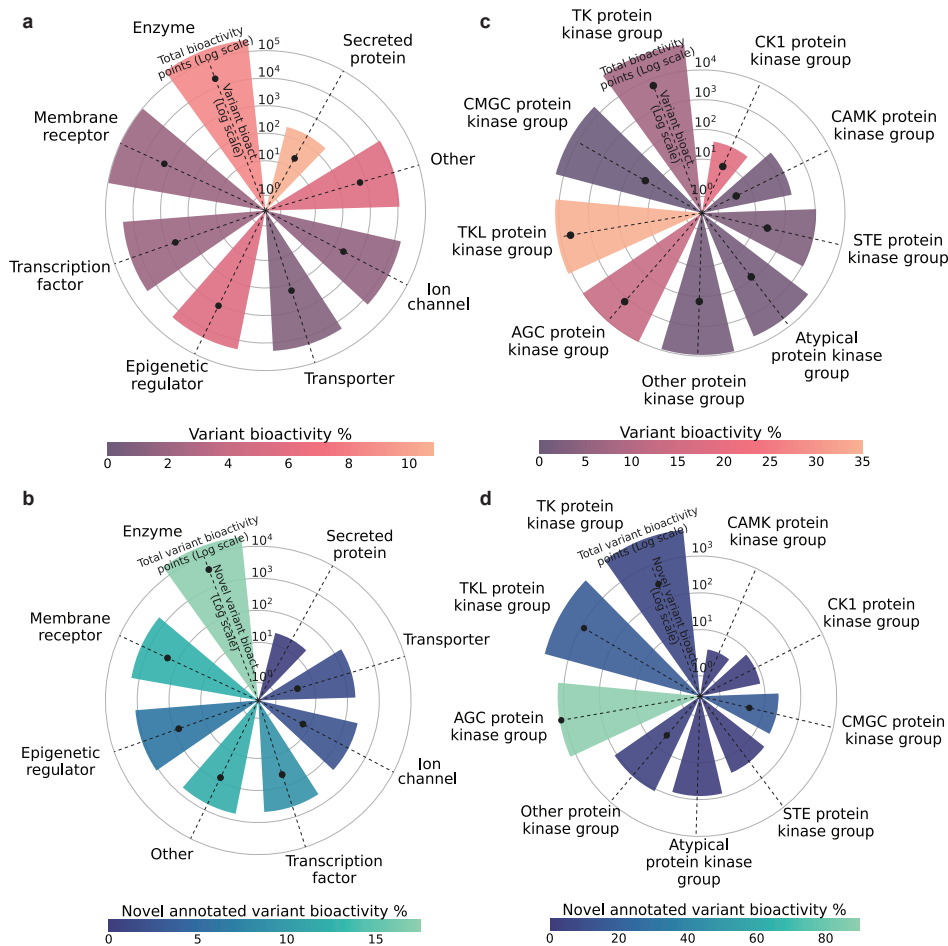


Figure 4.2. Distribution of variant bioactivity data across protein families in targets with at least one annotated variant. **a)** Bioactivity data in the VEBD for all protein families (L1 classification). **b)** Comparison of originally ChEMBL-annotated and novel variant data for all protein families (L1 classification). **c)** Bioactivity data in the VEBD for subfamilies of the Kinase enzymes family (L4 classification for L2 = Kinase). **d)** Comparison of originally ChEMBL-annotated and novel variant data for subfamilies of the Kinase enzymes family (L4 classification for L2 = Kinase). Bar heights represent the number of total bioactivity points (in a,c) or total variant bioactivity points (in b,d) on a logarithmic scale. The height of the black dots along the dashed lines represents the number of variant bioactivity points (in a,c) or novel annotated variant bioactivity points (in b,d). The color gradient represents the percentage of variant bioactivity data with respect to total bioactivity data (in a,c) or the percentage of novel annotated variant data with respect to total novel annotated variant data (in b,d).

Of note, the secreted proteins family included only one protein while the enzymes family included 195. Compared to the highest classification level of protein families, the variant load drastically differed between protein subfamilies. For example, the variant bioactivity percentage across subfamilies of the kinase enzyme family ranged from 0.1% for the CMGC protein kinase group to 35% for the TKL protein kinase group (**Figure 4.2c, Supplementary Table 4.4**).

Similar trends were observed while focusing only on ChEMBL-exclusive data and exploring the differences between the original and the current variant annotation pipelines. The highest amount of bioactivity data points with potential novel variant annotations corresponded to enzymatic targets (3,631), followed by membrane receptors (218). However, at the highest protein classification level, the percentage of potentially novel annotated bioactivity data to the totality of the variant-annotated data significantly differed across protein families, ranging from 0% in secreted proteins to 17.5% in enzymes (**Figure 4.2b, Supplementary Table 4.5**). Again, this effect was exacerbated across kinase subfamilies. Here, in four subfamilies (i.e. atypical, STE, CK1, and CAMK protein kinase groups) the totality of the variant bioactivity data had previously been annotated in ChEMBL, resulting in a novel annotated variant bioactivity percentage of 0%, while in the AGC protein kinase group, 89.7% of the variant data was introduced by the current variant annotation pipeline (**Figure 4.2d, Supplementary Table 4.6**). Similarly, in the kinase subfamily with the highest amount of variant data (i.e. TK protein kinase group), 5.1% of the variant data had not been previously annotated in ChEMBL.

The distribution of data in the VEBD across individual proteins was similarly unbalanced. Of the 335 proteins included in the annotated dataset, eight viral and bacterial proteins and one human protein did not include any WT data. However, only three of these (Hepatitis C viral NS3 protease Q0ZMF1 and polyprotein K7XJL6, and Human immunodeficiency virus 1 – HIV-1 – reverse transcriptase Q9WKE8) had more than 30 bioactivity data points. From the remaining 326 proteins, the vast majority (315) had simultaneously less than 20 variants and less than 10,000 bioactivity data (**Figure 4.3, Supplementary Table 4.7**). Only three human proteins (aldehyde dehydrogenase AL1A1 - P00352, phosphatidylinositol kinase PK3CA - P42336, and epidermal growth factor receptor EGFR - P00533) had more than 10,000 bioactivity data points, of which only one (EGFR) had a variant bioactivity percentage over 2%, specifically 18.36%. Moreover, eight different proteins had more than 20 annotated variants, including WT (**Figure 4.3a**). Some of these variants were single amino acid substitutions, while other variants accumulated several substitutions (**Supplementary Table 4.8**). The two most tested proteins among these eight with high genetic variance were viral proteins from HIV-1, namely polyprotein RNase H - reverse transcriptase (RNaseH-RT, Q72547) and polyprotein Q72874. The other six were mammalian membrane proteins, some of which may have been subjected to experimental mutagenesis programs: five class A G protein-coupled receptors (GPCRs) – the human gonadotropin-releasing hormone receptor GNRHR (P30968), the rat muscarinic receptor ACM3 (P08483), the human chemokine receptor CXCR4 (P61073), the rat opioid receptor OPRK (P34975), and P2Y1 (P47900) – and one solute carrier transporter – human betaine transporter S6A12 (P48065). The protein with the largest number of annotated variants was GNRHR, with

70 variants other than the WT. Among the eight proteins with high genetic variance, the variant bioactivity percentages ranged between 1.72% and 71.83%.

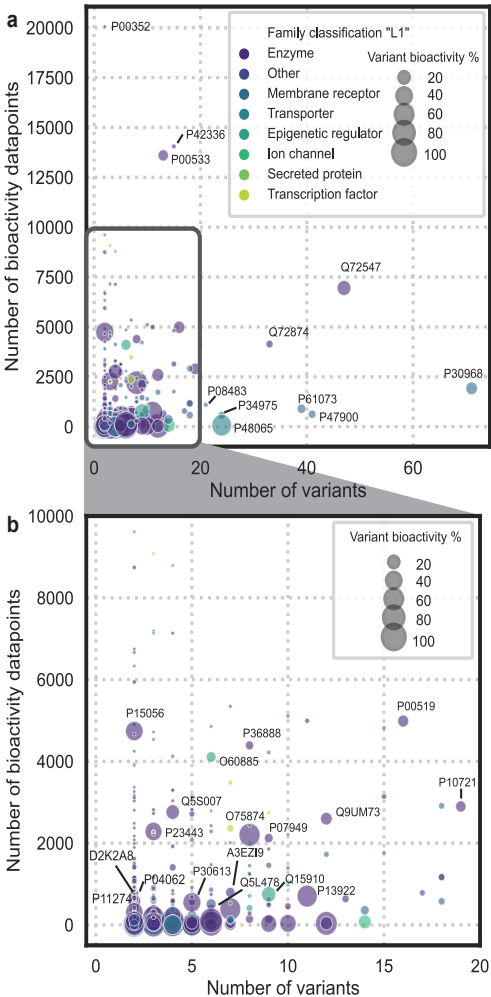


Figure 4.3. Variant annotation load per protein in terms of the number of variants and bioactivity data, as well as variant bioactivity percentage. **a)** Overall. Labelled are proteins with more than 10,000 data and/or more than 20 annotated variants, including WT. **b)** Proteins with less than 10,000 bioactivity data and less than 20 variants. Labelled are proteins with a variant bioactivity percentage higher than 10% and more than 500 data.

namely IDHC (seven variants apart from WT), BRAF (one variant), and RPKS6B1 (two variants), and variant bioactivity percentages of 86.29%, 60.27%, and 55.21%.

From the 315 proteins that had simultaneously less than 20 variants and less than 10,000 bioactivity data, only 100 displayed a variant bioactivity percentage equal to or greater than 10% (**Figure 4.3b**), and only 10 of these had more than 1,000 bioactivity data points. For reference, we consider 1,000 data points as an arbitrary threshold to enable bioactivity modeling. Constraining the variant bioactivity percentage to 20% resulted in only 62 proteins out of which only six had more than 1,000 bioactivity data points; most of these contained clinically relevant mutations. The five proteins with the largest amount of bioactivity data were all tyrosine, tyrosine-like, or AGC kinases, namely ABL1 (P00519), BRAF (P15056), leucine-rich repeat kinase LRRK2 (Q5S007), ALK (Q9UM73), and ribosomal protein kinase RPKS6B1 (P23443) in descending order of bioactivity data points and in line with the distributions per protein family (**Figure 4.2a,c**). The sixth protein was the oxidoreductase isocitrate dehydrogenase IDHC (O75874).

Save for the exceptions mentioned above, generally higher variant bioactivity percentages correlated with lower total bioactivity data, regardless of the number of variants annotated (**Supplementary Figure 4.2d**). From the total of 335 proteins in the dataset, only 32 showed as much or more bioactivity data for variants than for WT (i.e. 50% variant bioactivity percentage or higher), and out of these, only three had more than 1,000 bioactivity data points,

In general, annotated proteins with more than 1,000 data points had a small number of variants, and most of their data was tested on the WT protein (**Supplementary Figure 4.2d**). However, the data-rich protein targets highlighted in this section emphasized the potential relevance of hidden variant data in bioactivity modeling and were therefore the focus for the rest of the analysis. In particular, we defined a set of 13 data-rich proteins (**Table 4.1**) with the highest variant bioactivity percentages (i.e. equal or above 10%) that had simultaneously sufficient data for bioactivity modeling (i.e. equal or above 1,000 bioactivity data points) and that were subsequently analyzed in more detail in the following sections.

Amino acid substitution types represented in bioactivity datasets align with organism mutation rates

The type of amino acid substitutions represented in bioactivity datasets was also not homogeneously represented and may reflect the community's interest in protein variant sampling. As such, the majority of the reported variants were amino acid substitutions to alanine (**Figure 4.4a**), as part of the commonplace alanine scanning strategies to determine key structural and functional residues. Indeed, as expected, the alanine enrichment was not maintained in the number of bioactivity data points (**Figure 4.4b**). Instead, biologically relevant variants such as cancer-related BRAF V600E and EGFR T790M and L858R were responsible for the largest density of bioactivity data around particular amino acid substitutions. For example, the amino acid substitution with the largest amount of associated bioactivity data was valine to glutamic acid, with 2,864 bioactivity data points, out of which 99.7% corresponded to the BRAF V600E variant.

In line with the amount of data in ChEMBL per organism (**Supplementary Table 4.9**), the most frequently tested variants were in human proteins (BRAF, IDHC, RPKS6B1, EGFR). Indeed, out of the variant annotated bioactivity data, 90.56% corresponded to *Homo sapiens*. Viral and bacterial variants were also represented, however with only

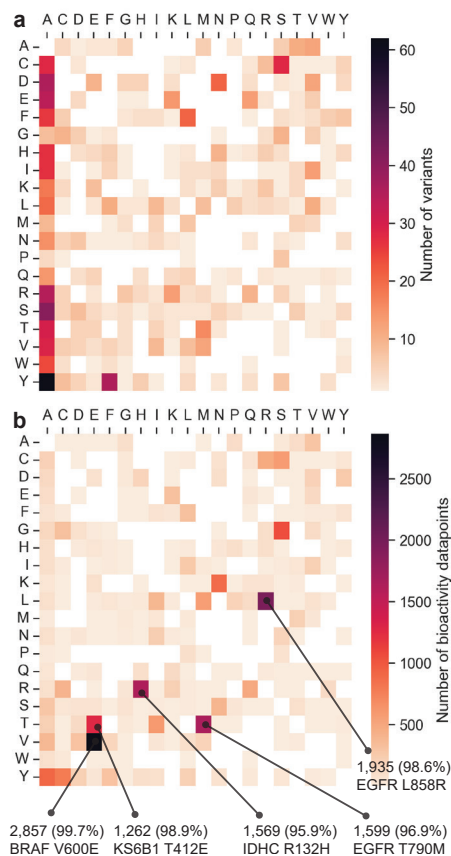


Figure 4.4. Amino acid substitutions reported in bioactivity databases. **a)** Unique variants reported per amino acid substitution. **b)** Number of bioactivity data points per amino acid substitution. Highlighted, is the substitution with the highest representation for the top five amino acid substitutions. In variants with multiple substitutions reported, each variant was accounted for individually.

4.82% and 0.70% of the bioactivity data. The remaining bioactivity data corresponded to 13 non-human Eukaryotic organisms of interest in preclinical studies, such as *Rattus norvegicus* or *Mus musculus*, among others. The type of amino acid substitutions reported in bacterial variants were similar to human variants (**Figure 4.5a,b**). These featured many disruptive amino acid substitutions (91.53% in bacteria and 89.67% in humans), either affecting the size or polarity of the original amino acid or, in most cases, both. To further characterize the disruptive potential of each amino acid substitution, we calculated the Epstein coefficient of difference³⁵, which is higher for more disruptive changes. In line with the previous observations, the Epstein coefficient of difference for most of the variants was higher than 0.4 (50.00% of the bacterial and 55.30% of the human variants), thus indicating changes in amino acid properties that would likely affect the protein's function. On the other hand, viral variants featured a larger proportion of conservative amino acid substitutions (17.81%, **Figure 4.5c**). This observation was also backed up by a lower proportion of amino acid substitutions with an Epstein coefficient of difference higher than 0.4 (41.21%), even when the size or polarity was affected. From a biological perspective, organisms with a higher mutation rate, such as viruses, are indeed prone to accumulate fewer damaging substitutions than organisms with a lower mutation rate subjected to more checkpoints, such as humans.

Among the 14 viruses and 16 bacteria for which 217 and 115 variants were tested, respectively, two organisms concentrated the majority of the data available (**Supplementary Table 4.9**). HIV-1 accumulated 54.8% of the viral variants and 70.6% of the viral bioactivity data in just five proteins. Similarly, *Escherichia coli* concentrated 20.9% of the bacterial variants and 42.0% of the bacterial bioactivity data tested in eight proteins. A closer look into the nature of the substitutions reported in these organisms offered some interesting insights when compared to EGFR as a proxy for a human protein with disease-relevant variants. In line with the general observation across human proteins, the nine single substitutions reported for EGFR were few but of high relevance, with only one conservative substitution and Epstein coefficients of difference around (three) or higher than (five) 0.4 (**Figure 4.5d**). Based on the 77 crystal structures available, all reported EGFR substituted amino acids were located from 8Å to almost 25Å of the center of geometry (centroid) of the protein ligands. Of note, the two most tested substitutions (resistance substitution T790M and activating substitution L858R) showed very high coefficients of difference but different locations with respect to the binding pocket (0.80 and 9.77Å, and 1.01 and 16.60Å, respectively). These two substituted residues are in the binding pocket of EGFR and correspond, respectively, to the gatekeeper residue and the back cleft. In contrast, HIV-1 RNaseH-RT harbored 31 single substitutions, of which 64.52% had an Epstein coefficient of difference lower than 0.4 (**Figure 4.5f**). Of note, these substitutions were concentrated around the non-nucleoside reverse transcriptase inhibitor (NNRTI) binding site, with distances to the ligand centroid mostly below 15Å. The only *E. coli* proteins with structural data, acetylglucosamine deacetylase LPXC (P0A725), and dihydrofolate reductase DYR (P0ABQ4), showed six substitutions affecting either size or polarity (**Figure 4.5e**), and were located around 15Å of the ligand centroid. The type of amino acid substitution, as well as the distance from the substituted residue to the ligand binding site, could affect the bioactivity of certain small molecules towards different variants. From a biological point of view, enriched

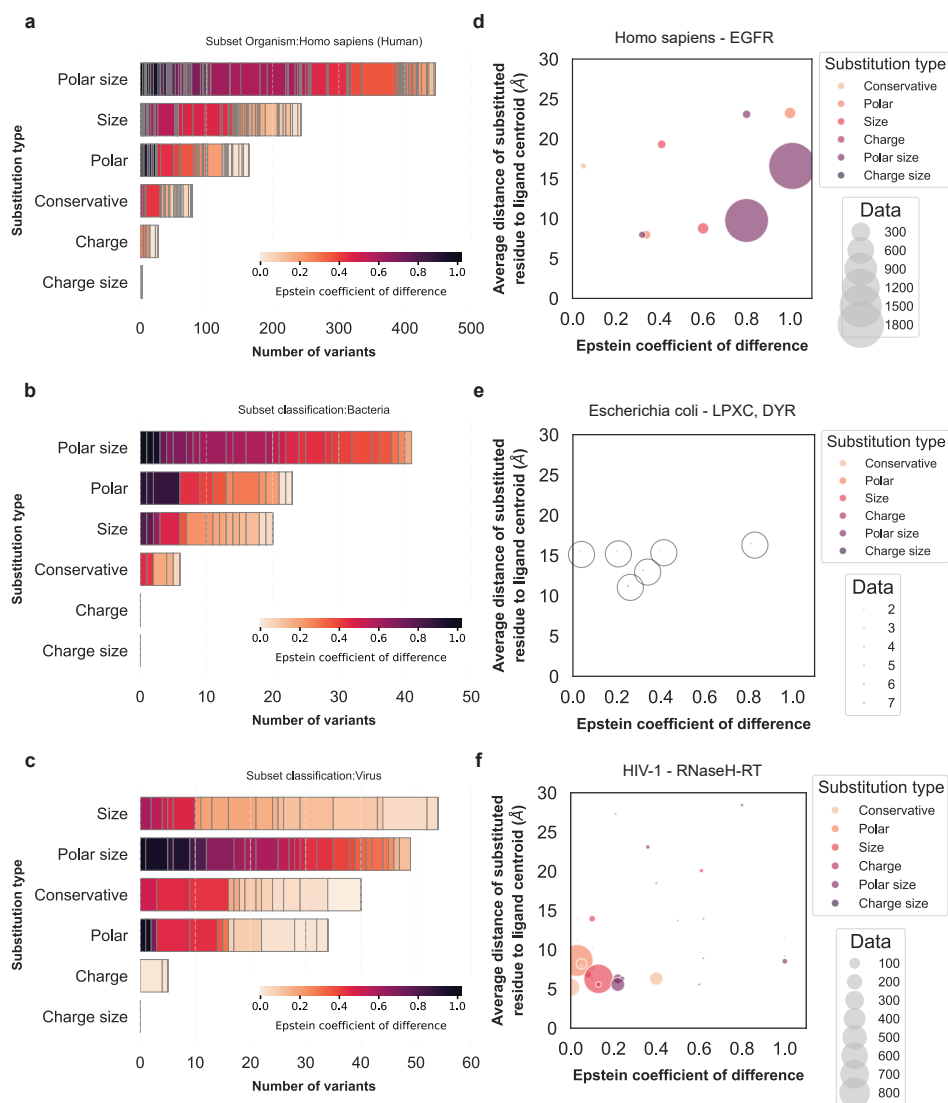


Figure 4.5. Types of amino acid substitutions in bioactivity databases across taxonomic categories: *Homo sapiens* (a,d), Bacteria (b,e), and Viruses (c,f). **a-c)** Number of variants according to their amino acid change, divided into six categories related to the effect in the amino acid polarity and size and colored by the Epstein coefficient of difference of the corresponding amino acid substitution. **d-f)** Correlation between amino acid change relevance (Epstein coefficient of difference, x-axis), distance to ligand (average distance of substituted residue to ligand center of geometry or centroid, y-axis), and sampling frequency (number of bioactivity data points, bubble size) in variants of **d)** *Homo sapiens* EGFR (P00533), **e)** *Escherichia coli* LPXC (P0A725) and DYR (P0ABQ4), and **f)** *Human immunodeficiency virus 1* (HIV-1) polyprotein RNase H - reverse transcriptase (RNaseH-RT, Q72547). Note that although Q72547 is the code for RNaseH-RT, the substitutions were concentrated in the RT domain, with only three substitutions in the RNaseH domain. In variants with multiple substitutions reported, each variant was accounted for individually.

human variants are likely to be disease-related whereas variants in pathogenic organisms are more likely linked to drug resistance. The extent of such an effect and its potential relevance in bioactivity modeling was analyzed in the following sections.

Genetic variants affect bioactivity at different levels

Heterogeneity was found in annotated variants not only regarding the type and location of amino acid substitutions but also the number and structure of small molecules tested across them, as well as their relative bioactivity compared to WT. These observations reflected the interest in therapeutically targeting disease-relevant variants. In previous sections, it was shown that the majority of proteins have a small amount of variant bioactivity data compared to WT, in particular in proteins with sufficient data for modeling (**Figure 4.3**). Even in the proteins with the highest variant bioactivity percentages (i.e. equal to or above 10%) that had sufficient data for bioactivity modeling (i.e. equal to or above 1,000 data), data density across variants was rather uneven. Out of the 13 data-rich proteins satisfying these conditions, WT was the most populated variant in all cases except for BRAF (P15056) V600E, IDHC (O75874) R132H, and RPKS6B1 (P23443) T412E (**Supplementary Table 4.10**), with the two first mutations corresponding to clinically relevant variants in cancer. BRAF and RPKS6B1 were also the only proteins, together with LRRK2 (Q5S007), where the most populated variant-annotated target had less than twice the amount of data of the second most populated variant, namely 1.52, 1.21, and 1.96 times. The rest of the proteins ranged from 4.73 (ALK, Q9UM73) to 104.64 (GNRHR, P30968) times more data in the most populated variant-annotated target – generally WT – compared to the second. The proteins with the largest relative data density differences between the first and second variants were those with the largest number of variants annotated (**Supplementary Figure 4.3a**). In these cases, the existence of many variants compensated for their data scarcity and still amounted to a relevant variant bioactivity percentage, above 10%. However, for all 13 data-rich proteins, only up to three variants – generally the most established clinically relevant – contained more than 500 data points, with some of the remaining variants dropping to as little data as one data point (**Supplementary Figure 4.3b**). These numbers corroborated the high data sparsity and hinted at the potential challenges to accurately reflect the differences in bioactivity caused by variants.

Two scenarios were contemplated to reduce the effect of chemical data sparsity across variants. The first one simulated an ideal scenario where all compounds would have been tested on all variants. For this purpose, fully dense common subsets were computed for targets with sufficient data, where only those compounds tested across all available variants were kept. Given the number of variants with extremely low data density, this task was not trivial. In fact, approximately two-thirds of the 335 targets in the VEBD did not have a single compound that had been tested on all reported variants. For the other third consisting of 114 targets, the fully dense common subset represented a small portion of the target's set, with only 18 targets exceeding 10% and the maximum representation being 50%. Moreover, the size of their fully dense common subsets was very small, with only four targets surpassing 35 compounds tested across all their annotated variants

(**Supplementary Figure 4.4**). However, the computation of fully dense common subsets proved to be relevant to achieve fair comparisons. In many cases, like for breakpoint cluster region protein BCR (P11274) and JAK2 kinase (O60674), the modeling protein set was highly biased towards WT bioactivity, making the fully dense common subset valuable for comparison (**Supplementary Figure 4.4b,c,f,g**). Given these results, a strategy was developed to compute non-fully dense common subsets - referred to as common subsets - for the previously mentioned two-thirds of proteins for which a fully dense common subset was not available. Common subsets generated for compounds tested in at least two variants with a variant coverage of at least 20% identified 115 targets for which a fully dense common subset was not possible. Overall, using these parameters to compute the common subsets resulted in very diverse subsets covering 229 targets with an average common subset of 35 ± 121 unique compounds and 5 ± 6 variants. This was a clear improvement in terms of subset size from the original 114 fully dense common subsets, which had an average of 10 ± 33 unique compounds and 4 ± 7 variants. Additional measures were taken in very sparse targets by allowing the previous filters to be computed based on pairwise molecular similarity. This allowed us to include compounds only tested in one variant if a highly similar compound (e.g. Tanimoto similarity ≥ 0.80) had been tested in a different variant. The similarity option with the previously defined parameters allowed rescuing an additional four targets but did not improve the existing subset sizes, given the stringent 80% similarity threshold. The obtained similarity-expanded common subsets maintained the bioactivity distribution per variant of the VEBD, and all reached a higher balance and reduced sparsity as intended (**Supplementary Table 4.11**).

The generation of common subsets with varying parameters made it possible to analyze complete panels of compounds across variants. The versatility of such analysis on different protein families was exemplified for targets previously highlighted based on bioactivity data density and variant bioactivity percentage, namely EGFR (**Figure 4.6**, **Supplementary Figures 4.5,4.6**), HIV-1 RNaseH-RT, IDHC, and bromodomain-containing protein BRD4 - O60885 (**Supplementary Figures 4.7-4.9**, respectively). For EGFR, this analysis allows the user to follow some of the most biologically relevant activating – L858R, G719C/S, A750P, P753S, L861Q – and resistance – T790M – substitutions and the different generations of EGFR inhibitors (EGFRi) developed to achieve selective bioactivity profiles (as a reference commonly used in drug discovery we will consider a potency difference over 30-fold against specific variants of interest, which translates to a pchembl value difference over 1.5). The bioactivity analysis set for EGFR was generated from a common subset with compounds tested on at least three EGFR variants and variants covering at least 10% of the compounds. The analysis subset contained 22 compounds tested on nine out of the 14 annotated EGFR variants with clear differences in bioactivity (**Figure 4.6**, see **Supplementary Figure 4.5** for compound ID mapping). Out of these 22 compounds, 10 were approved drugs – EGFRi but also pan-kinase and other inhibitors – and the rest were either preclinical or clinical candidates (**Supplementary Figure 4.5,4.6**). The first two generations of EGFRi were represented in this analysis. First-generation EGFRi are reversible compounds developed to target activating mutations, in particular substitution L858R. Second-generation EGFRi are irreversible compounds aiming at a similar selectivity profile. Three compounds

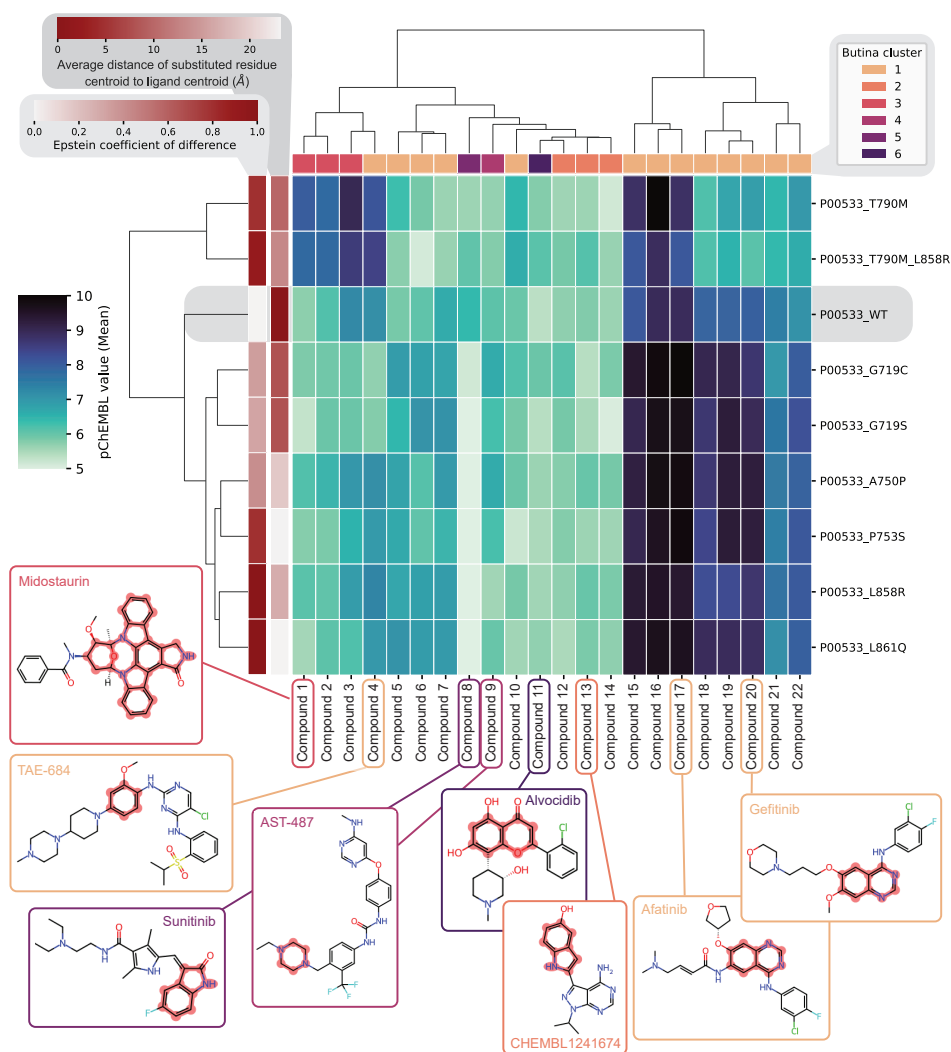


Figure 4.6. Full-panel bioactivity analysis of the effect of EGFR (P00533) variants. Bioactivity is represented in the heatmap as the pchembl value of different compounds, on the x-axis, tested on several variants, on the y-axis. See Supplementary Figure 4.5 for the mapping of compound numbers to their connectivity ID, preferred name, and approval status. Compounds and variants were clustered by their overall bioactivity profile. Compounds are further represented by their corresponding Butina clusters upon clustering of the subset with a cutoff of 0.7. Compounds that are representatives of particular clusters or bioactivity profiles are highlighted and their 2D structures are displayed with the preferred molecule name (ChEMBL). The rest of the molecules can be found in Supplementary Figure 4.6. The biggest ring system in each molecule is highlighted in red for reference as a less stringent proxy for the maximum common substructure to visually distinguish molecules with similar scaffolds. Variants are further represented by the distance from the substituted residue to the centroid of the ligand in the structure of the protein and by the Epstein coefficient of difference calculated for the amino acid substitution. In variants with multiple substitutions, average distance and coefficient of difference are reported.

(15-17), including second-generation EGFRi afatinib, showed consistently high pchembl values over 8.07, while seven (8-14) showed consistently low activity across variants with maximum pchembl value of 6.66. Moreover, four compounds (1-4) showed very high activity – pchembl value between 7.80 and 8.99 – against the two variants containing the resistance substitution T790M compared to the rest of the variants, including WT – where the maximum pchembl value was 7.34. These two variants, single substituted T790M and double substituted T790M/L858R, also exhibited the most different overall bioactivity patterns, as expected given their biological relevance. Indeed, five first-generation EGFRi (18-22) exhibited lower activity against the two T790M-containing variants (pchembl values between 6.09-7.00, compared to 7.01-9.33), as this resistance substitution is known to appear as a response to treatment with first- and second-generation EGFRi. Despite high activity overall, afatinib exclusively showed a decrease in bioactivity for the double mutant L858R/T790M. In terms of the location with respect to the ligand binding site, T790 is one of the closest substituted residues, below 10 Å from the ligand centroid, and effectively in the binding site of EGFR. Additionally, the threonine to methionine amino acid change is highly disruptive with an Epstein coefficient of difference over 0.80. The rest of the variants behaved more similarly to the WT, with two major compound clusters with low (pchembl values between 5.00 and 7.34) and high activity (between 7.01 and 10.00), respectively. From these, WT was the odd one with the least marked differences between the two groups of compounds, as seen in the hierarchical clustering per variant (**Figure 4.6**). This was expected, as most EGFRi were developed to be variant-selective and reduce the side effects of anticancer therapies. The single substituted variant L858R behaved very differently from the double substituted T790M/L858R variant, in line with the different biological roles of these substitutions. Although the substitution to arginine is highly disruptive, L858 is further away from the ligand than T790. The Butina clustering performed on the 22 compounds showed that similar compounds exhibit similar effects across variants, as observed for clusters 2-6, and in line with the sequential development of EGFRi generations. Clusters 2-6 were populated by compounds with clear similarities, resulting in a diverse cluster 1 (**Supplementary Figure 4.6**) showing multiple patterns across variants but mostly containing first- and second-generation EGFRi. An interesting example was compound 4, which is structurally very different from the compounds in cluster 3 (compounds 1-3) yet exhibited the same bioactivity pattern. As such, this analysis can aid in the exploration of compounds with variant-selective profiles beyond the most well-known chemical groups. For other proteins, it can be a tool to rationalize the chemical modifications needed to develop drugs targeting specific resistance substitutions (**Supplementary Figure 4.7**); an instrument for extracting starting scaffolds with specific selectivity profiles (**Supplementary Figure 4.8**); or to distinguish between compounds with different binding modes (**Supplementary Figure 4.9**).

The different effects observed for different chemical clusters in common subsets could also be expanded to bigger yet sparser subsets. This allowed us to analyze the overall effect of variants on different subsets of the chemical space tested for one protein. While this analysis is possible for the whole protein subset, in targets with a clear bias towards WT testing, selecting subsets of compounds tested on at least two variants was still preferred to increase the significance of comparisons across variants. Particularly for

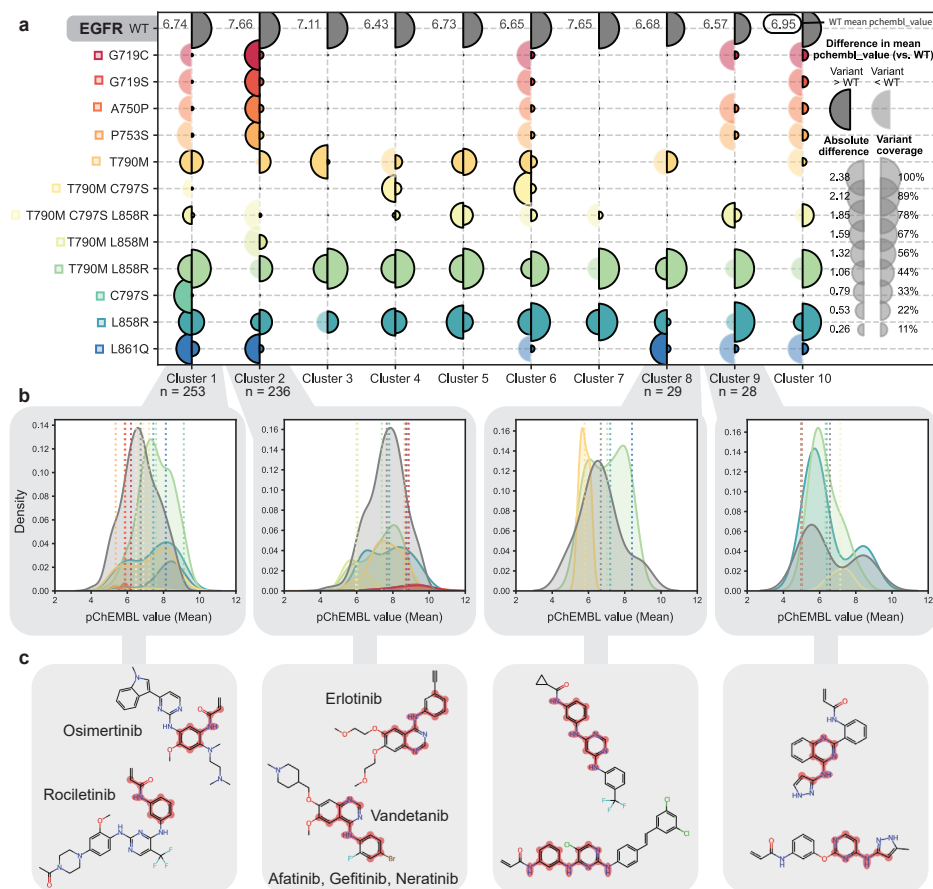


Figure 4.7. EGFR (P00533) bioactivity variability across variants compared to WT for compounds in the 10 most populated Butina Clusters upon clustering compounds tested on at least two variants with a clustering threshold of 0.5. **a**) Differences between mean *pchembl* value in WT, displayed at the first row as calculated for the compounds in each cluster, and the mean *pchembl* value in each of the variants for the compounds in the same clusters. The left bubbles represent the result of subtracting the variant mean from the WT mean. The bubble size represents the absolute value of this difference (error). Opaque left bubbles represent a positive error (i.e. the mean calculated for the variant is higher than for WT), and translucent left bubbles represent a negative error (i.e. the mean calculated for the variant is lower than for WT). Right bubble sizes represent the variant coverage, in other words, the percentage of compounds in each cluster that was tested on a specific variant. **b**) Distribution density of *pchembl* values across compounds in each cluster. Different colors represent the different variants where compounds of the cluster were tested, according to the color code of panel a. Dashed lines represent the mean *pchembl* value, which was used to calculate the differences in panel a. **c**) Two compound examples per cluster with the atoms corresponding to the maximum common substructure of all the compounds in the cluster highlighted in red. When available, approved compounds or preclinical candidates are displayed.

EGFR, the set of 1,219 compounds tested on at least two variants was clustered using the Butina algorithm³⁶ with a threshold of 0.5 resulting in 118 clusters. Clear differences in bioactivity across variants were observed among the top 10 biggest clusters (**Figure 4.7**). Chemistry-related changes in bioactivity distribution were already somewhat

apparent on the WT level (**Figure 4.7a,b**), with mean pchembl values between 6.43 and 7.66 from slightly divergent distributions. The compounds in the two most populated clusters ($n=253$ and $n=236$, respectively) were tested across 11 and 10 out of the 12 variants, respectively, with various rates of variant coverage (**Figure 4.7a**). These two clusters included approved first (cluster 2), second (cluster 2), and third generation (cluster 1) EGFRis, as well as pan-kinase inhibitors (cluster 2). Third-generation EGFRis were not present in **Figure 4.6** and were developed to selectively target the L858R/T790M double substitution. Furthermore, the average differences in bioactivity compared to WT across variants were virtually the opposite between the two clusters, in line with the known selectivity profiles of different generations of EGFRi. For example, compounds tested on rare variants G719C, G719S, A750P, and P753S all showed lower activity than compounds tested on the WT in cluster 1 (0.54, 0.85, and 0.88 points below WT – 6.74) but higher in cluster 2 (1.21, 1.11, and 1.06 points above WT – 7.66).

The opposite effect was observed for compounds tested on the double substituted T790M/L858R variant, which had a mean pchembl value 0.85 points higher than compounds tested on the WT in cluster 1 (7.59 vs. 6.74) and 0.27 points lower than compounds tested on the WT in cluster 2 (7.39 vs. 7.66). Of note, the bioactivity distributions across compounds tested in each variant were highly diverse (**Figure 4.7b**), thus relevant in addition to the point mean differences. Together, this type of analysis pinpoints chemical patterns (as highlighted in **Figure 4.7c** for the maximum common substructures of compounds in each cluster) driving differences in bioactivity across variants. Similarly to EGFR, this analysis can help expand the results observed in the full-panel bioactivity analysis for other proteins as exemplified for HIV-1 RNaseH-RT, IDHC, and BRD4 (**Supplementary Figures 4.10–4.12**, respectively). In an explorative fashion, results derived from this analysis can be the starting point of drug design campaigns satisfying certain activity characteristics. Alternatively, in virtual screening campaigns, they can be relevant for decision-making to reduce noise in models or increase the modeling performance by constructing variant-aware models, as explored in the following section.

Variant awareness improves modeling performance

The effects of variant bioactivity data on the performance of machine learning modeling were investigated by comparing results obtained from three scenarios. The first scenario corresponds to modeling in a variant-agnostic situation, wherein all bioactivity data measurements are (mistakenly) assumed to derive from assays carried out on WT proteins only (*QSAR-All*). The two other scenarios correspond to modeling in a variant-aware situation, wherein data points assayed on variant targets are either kept in (*PCM-All*) or filtered out of the training set (*QSAR-WT*).

First, modeling performance was evaluated based on random split cross-validation on the VEBD in its entirety, splitting out each protein in turn, to assess the overall effect of introducing variant-aware strategies. As expected, on average the performance of models decreased with a scarcer number of bioactivity data points (**Table 4.1, Supplementary Figure 4.8a and 4.8c, and Supplementary Table 4.12**), characterized by the average

Table 4.1. Modeling performance of variant-annotated proteins following three modeling strategies: PCM explicitly modeling variants (PCM-All), QSAR with all protein data without considering variants (QSAR-All), and QSAR removing variant data (QSAR-WT). The performance of PCM and QSAR models depends on the number of data points and the variant bioactivity percentage. Performance is reported for the entire training set, focused protein families, and data-rich proteins (more than 1,000 data points with at least 10% variant bioactivity percentage) for a random split 5-fold cross-validation strategy as the average Pearson correlation coefficient for each group or protein and, between brackets, as the average per group or protein of the standard deviation of Pearson r between cross-validation folds for each protein. The best average Pearson r is reported in bold for each row. Pearson r of PCM and/or QSAR-WT models significantly differing from QSAR-All models are starred. Pearson r of PCM or QSAR-WT models significantly differing from all other models (i.e. QSAR-WT and QSAR-All, and QSAR-All and PCM-All respectively) are underlined. Statistical results are detailed in Supplementary Table 4.17.

	Average Pearson correlation coefficient (average standard deviations)			Number data points	Variant bioactivity (%)
	PCM-All	QSAR-All	QSAR-WT		
All	0.653 (0.117)*	0.634 (0.116)	0.654 (0.121)*	453,660	5.5
5 to 100 data	0.396 (0.322)	0.352 (0.323)	0.363 (0.378)	3,257	29.1
100 to 500 data	0.704 (0.085)	0.690 (0.083)	0.691 (0.094)	19,694	10.0
500 to 2,000 data	0.746 (0.038)	0.737 (0.039)	0.747 (0.041)*	84,426	4.5
2,000 to 20,000 data	0.769 (0.018)*	0.763 (0.017)	0.764 (0.017)	346,283	5.2
Family A GPCRs	0.731 (0.046)	0.735 (0.035)	0.752 (0.037)	93,454	1.8
Ion Channels	0.620 (0.142)	0.613 (0.134)	0.646 (0.168)	16,635	1.5
Nuclear Receptors	0.704 (0.047)	0.690 (0.036)	0.714 (0.034)	14,344	2.5
Protein Kinases	<u>0.716 (0.068)*</u>	0.701 (0.068)	0.700 (0.080)*	133,396	9.1
P00533 (EGFR)	0.822 (0.009)*	0.802 (0.008)	0.809 (0.004)	13,601	18.4
Q72547 (HIV-1 RNaseH- RT)	0.809 (0.013)*	0.764 (0.005)	0.776 (0.012)	6,953	34.0
P00519 (ABL1)	0.867 (0.008)	0.850 (0.019)	0.857 (0.012)	4,985	22.3
P15056 (BRAF)	0.847 (0.012)	0.834 (0.013)	0.858 (0.014)	4,740	60.3
P36888 (FLT3)	0.813 (0.022)	0.812 (0.016)	0.798 (0.018)	4,390	11.8
O60885 (BRD4)	0.856 (0.007)*	0.714 (0.038)	0.858 (0.013)	4,106	17.1
P10721 (KIT)	0.748 (0.028)*	0.708 (0.010)	0.716 (0.015)	2,897	19.4
Q5S007 (LRRK2)	0.853 (0.017)	0.851 (0.013)	0.827 (0.009)	2,760	34.0
Q9UM73 (ALK)	0.854 (0.017)	0.829 (0.011)	0.837 (0.021)	2,598	24.9
P23443 (RPKS6B1)	0.854 (0.005)	0.853 (0.012)	0.682 (0.042)*	2,286	55.2
O75874 (IDHC)	0.804 (0.014)	0.759 (0.031)	0.775 (0.045)	2,203	86.3
P07949 (RET)	0.778 (0.027)*	0.752 (0.033)	0.718 (0.020)	2,123	13.2
P30968 (GNRHR)	0.758 (0.047)	0.724 (0.030)	0.720 (0.045)	1,921	23.7

cross-validated Pearson correlation coefficient (Pearson's r) below 0.40 when modeling proteins with 5 to 100 data points, around 0.70 with 100 to 500 data points, around 0.75 with 500 to 200 data points, and above 0.76 with more than 2000 data points, respectively. In any case, variant-aware models showed increased performance, with all *QSAR-WT* models showing an increased correlation with experimental values compared to *QSAR-All* models. Data balance between the data points obtained on WT and the ones on variant targets had an impact on the significance of the differences in performance observed (**Table 4.1** and **Supplementary Figure 4.13b** and **4.13d**). This was demonstrated in protein families with substantial experimental data by the significantly increased performance of the *PCM-All* model (0.716) for protein kinases (p -value= 4.1×10^{-5}), with 9.1% of variant bioactivity percentage, compared to that of *QSAR-All* and *QSAR-WT* models (0.700 and 0.701, respectively). In contrast, no significant difference was observed for family A GPCRs, ion channels, and nuclear receptors, which all had a lower data balance (between 1.5 and 2.5% variant bioactivity percentage), and for which PCM was not the best strategy. Indeed, all points relating to protein kinases in **Figure 4.8a zoom-in** were very close to or below the identity line, and most data-rich kinases showed a significant performance increase when using PCM. These included EGFR (P00533), ABL1 (P00519), LRKK2 (Q5S007), ALK (Q9UM73), RPKS6B1 (P23443) and proto-oncogene RET (P07949) kinases, all having more than 2,000 associated data points with at least 10% variant bioactivity percentage, with correlation coefficients between 0.75 and 0.85 (**Table 4.1**, **Figure 4.8a**). Interestingly, of data-rich proteins, only BRAF (P15056) showed a decreased performance when including data points of variants, with a Pearson's r of 0.847 for *PCM-All* and 0.858 for *QSAR-WT*. This could be the result of the very large amount of data points associated with variants (60.3%) and due to the distinctively divergent but overlapping trends in the distributions of bioactivities between WT and variants (**Supplementary Figure 4.4**). These results highlight the importance of variant awareness in bioactivity modeling but do not provide a solid basis for general recommendations on the variant-aware strategy that should be used.

Next, the capacity of models to predict the bioactivity of compounds on unseen variants was investigated. To this end, Leave-One-Variant-Out (LOVO) cross-validation was carried out. This confirmed the trend previously observed of the ability of PCM models to interpolate in the protein feature space, especially for richer sets of proteins (more than 2000 data points) with an average Pearson's r of 0.325 compared to 0.311 for other proteins (**Supplementary Table 4.13**). To decrease the sparsity of datasets, similarity-expanded common subsets were derived to focus on a subset of molecules and their analogs tested across a subset of variants. The latter drastically decreased the applicability domains of models (**Supplementary Table 4.14**) and affected the performance of most models (**Figure 4.8b** and **Supplementary Table 4.15**) but improved the performance of models when used in combination with LOVO cross-validation (**Figure 4.8c** and **Supplementary Table 4.16**) for most proteins. Nonetheless, the general trend showed no clear difference between *QSAR-All* and *PCM-All* models derived from LOVO cross-validation from the common subsets (**Figure 4.8d**), suggesting that the extrapolation to new variants using PCM is similar to random prediction. These results show the complexity of accurately predicting bioactivity for individual variants. Moreover, they highlight the impact of data sparsity on model performance and how the

limited size of current datasets restricts extrapolation in the protein feature space when focusing on analog molecules.

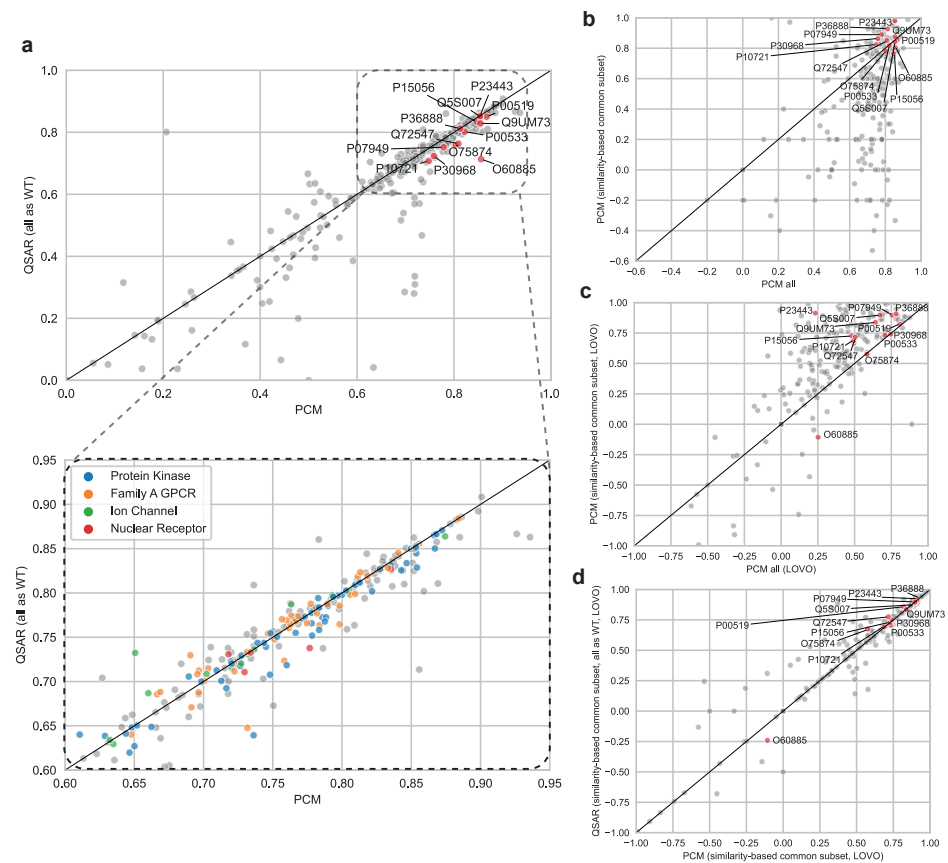


Figure 4.8. a) Comparison of the performance (random split cross-validated Pearson correlation coefficient average) of models in a variant-aware (PCM-All) and variant-agnostic (QSAR-All) setting considering the newly annotated VEBD. Zoom-in coloured by of protein families. **b,c)** Comparison of performances between the PCM models obtained from the entire VEBD set and the similarity-based common subset using random split cross-validation (b), or using Leave-One-Variant-Out (LOVO) cross-validation (c). **d)** Comparison between PCM and QSAR models derived from the similarity-based common subset. Labelled points correspond to data-rich proteins (see Table 4.1).

The general trends highlighted above were consistent across the data-rich proteins, although few of them had a significant performance improvement when using variant-aware models (Table 4.1, Figure 4.8). On a protein-specific level, this effect can be traced back to data sparsity and imbalance across variants and subsets of the chemical space (Figure 4.7 and Supplementary Figure 4.10 for EGFR and HIV-1 RNaseH-T, respectively). In fact, tackling these issues by reducing the applicability domain with a similarity-expanded common subset resulted in equivalent or improved PCM performance in random split cross-validation compared to complete sets for these proteins, with a clear advantage over the variant-agnostic model (Supplementary Table 4.13, Figure 4.8b). Moreover, the analysis of the bioactivity patterns can help explain

discrepancies from the general modeling trend. For example, among data-rich proteins, BRD4 (O60885) displayed the biggest increase in performance when using variant-aware models in random split cross-validation (**Figure 4.8a**). Following the general trend, we expected a good extrapolation to novel variants for this protein, which was not the case (**Figure 4.8c**). The examination of the substituted residue distance to the ligand's centroid on the bioactivity cluster map for BRD4 (**Supplementary Figure 4.9**) highlighted that the two most represented variants, Y97A and Y390A respectively, are each part of different protein domains, bromodomains 1 and 2 respectively, corresponding to different binding sites, and had therefore opposite effects on bioactivity for the subset of compounds examined. This was confirmed in the protein's structure and explained the lack of generalization power of the model, which might be improved by splitting the chemical space into domain-specific binders. Still looking at the data-rich proteins, IDHC (O75874) showed poor extrapolation, which could be traced back to the very similar bioactivity profiles across the tested variants, all of them occurring in the clinically relevant R132 residue (**Supplementary Figures 4.8,4.11**). Based on this information, model performance could be improved by pooling all variant data or designing protein descriptors able to capture the subtle differences in one residue. These results stress the importance of informed decision-making via the analysis of bioactivity trends to design relevant training sets and strategies for variant-aware modeling.

Discussion

Bioactivity modeling is one of the cornerstones of computational drug discovery. Despite the most recent advances in modeling techniques and capacities, data quality and quantity remain a major bottleneck, particularly for those working in the public sector without access to large proprietary or commercial datasets. As a consequence, large, curated, and open bioactivity databases such as the ChEMBL database or the Papyrus dataset constitute key resources for the community. Despite the many benefits that the expert extraction and curation processes for these databases provide, the user still needs to navigate the often-complex database structures and make informed decisions to select and curate data for the modeling task at hand. This does of course also reflect the fact that developing, running, and processing the data from bioactivity assays is a complex scientific endeavor. Careful selection of several fields in these databases, such as activity comments and assay types can have a big impact on the quality of the modeling data. Here, the effect of a commonly overlooked field in bioactivity databases, amino acid substitutions constituting protein variants, was extensively analyzed. The genetic variability landscape in the ChEMBL database has been explored in detail here for the first time, including the annotation strategy, the extent of variant data at different levels, the effect on bioactivity distributions, and finally the effect on bioactivity modeling. The dataset and results from this are made available to facilitate modeling with consideration of genetic variants. Moreover, a full analysis Python package is made available to promote variant analysis in proteins of interest to the user and thus help make informed decisions about data selection and curation for modeling.

A variant annotation strategy parallel to that of ChEMBL was developed that extracted

82.65% of the original variant annotations from the assay descriptions, which reinforced the confidence in the original ChEMBL variant annotation pipeline (which delivers these annotations by manual extraction of protein variant information from original papers). A clear advantage in the ChEMBL pipeline was the access to expert knowledge to rescue variants otherwise missed by a regular expression match. For example, sequence number shifts and non-canonical amino acid substitution definitions were identified among these expert rescues. However, mis-annotations reported by ChEMBL were also identified, for example, derived from mistakenly linking assays to protein families rather than single proteins. The current annotation strategy also retrieved several substitutions that had not been previously reported in ChEMBL 31. Nevertheless, these results need to be considered cautiously since they are based on fields previously extracted by ChEMBL rather than the original source in the literature and might miss important aspects of the experimental set-up. Importantly, this approach also relies on accurate reporting of tested variants in the scientific literature in order for their subsequent capture in bioactivity databases. Collaborative work such as reported here is key to improving the ChEMBL database^{37,38} for the wider community; for future releases of ChEMBL, we will aim to improve and enhance our reporting of variant data based upon the findings in this paper. Although several drug and protein databases contain variant data, the effect of drugs on specific variants is very sparse and conflicting^{39,40}. An expert-curated dataset derived from our analysis could therefore serve as a user-friendly central repository for variant bioactivity data regularly retrieved from ChEMBL and additional sources. As a result of this collaboration, a revised version of this work will be released, integrating the alterations recommended through the feedback loop (see ChEMBL comments in **Supplementary Table 4.1** and **4.2**, revision ongoing).

The variant landscape in ChEMBL 31 and additional Papyrus sources is, as expected, a reflection of the clinical relevance and interest of the community in particular organisms, protein families, targets, and individual variants. Unsurprisingly, human proteins concentrated the bulk of the variant data, but several mammalian orthologs and human pathogens were also identified. Of note, curated drug resistance databases for significant pathogens such as HIV⁴¹, tuberculosis⁴², and other antibiotic-resistant bacteria⁴³ are available independently of bioactivity databases and should be queried separately. Apart from being more complete, these databases have a more domain-focused curation process e.g. strain annotation in microorganisms. Although different organisms show significant differences in the amounts of data available, the amino acid substitution trends align with nature-observed patterns. Indeed, organisms with smaller genome sizes and higher mutation rates, such as viruses and to a lesser extent bacteria, accumulated larger amounts of non-disrupting substitutions compared to human proteins^{44,45}.

Among human protein families, enzymes, in particular kinases, amassed the most variant data, though not always proportionate to the overall data volume. While these numbers do not correspond to evolutionary mutation rates⁴⁶, they are certainly correlated to the high interest in protein kinase variants in cancer research⁴⁷. Indeed, the targets that simultaneously displayed high variant bioactivity percentages and large amounts of data overall were predominantly cancer-related kinases with clinically relevant somatic substitutions such as EGFR⁴⁸, ABL1⁴⁹, BRAF⁵⁰, and ALK⁵¹. Nonetheless, in this category

were also cancer-related kinases with no reported disease-related somatic substitutions like RPKS6B1⁵², where experimental mutations are common, or kinases responsible for other pathologies, such as LRRK2 in Parkinson's⁵³. Of note, the individual variants reported for specific targets also reflect the interest within the scientific community and do not necessarily include all reported and clinically relevant variants⁵⁴. Other than clinically relevant variants, experimentally important variants were found, such as activating substitutions in downstream cascades⁵⁵, or alanine scanning panels for functional⁵⁶ or thermostabilizing assessment⁵⁷ in GPCRs. Far from negligible, such panels can be repurposed for model training, consequently reducing the need for experimental assays⁵⁸.

The Python package and notebooks that accompany this work have been carefully designed to allow complete reproducibility of the annotation and variant landscape analysis. However, their primary purpose is to empower readers to self-assess variant effects on protein bioactivity. As shown here for the clinically relevant kinase EGFR, among other data-rich targets, these analyses can identify clusters of chemical space with varying effects on bioactivity, specific protein structural traits causing differing bioactivity patterns, and compounds with desirable selectivity profiles. These results not only are in line with the literature and enabled the analysis of activating and resistance-inducing substitutions, but also extended beyond the most widely-recognized variants and chemical classes⁵⁹. In turn, they can be used as hypothesis generators in drug design⁶⁰ as well as recommendation systems to include or remove certain chemical clusters⁶¹ or variants from a prospective modeling or virtual screening task⁶². Indeed, for a target like EGFR with a high variant bioactivity percentage and differential bioactivity profiles across variants and chemical groups, our bioactivity modeling results indicated a decrease in predictive performance when variants were not accounted for, generalizing the effects previously observed when modeling cyclooxygenases 1 and 2⁶³. Both removing variant data from the QSAR model and explicitly modeling each variant in a PCM model increased performance in random split cross-validation, likely by reducing the negative effect of noise^{64,65}. Similar results were observed for other proteins with a high variant bioactivity percentage despite large inter-target variability. Nevertheless, non-optimized protein sequence descriptors were used in this work. Furthermore, the average length of protein sequences varies greatly - for instance considering the 566 amino acids of HIV-1 RNaseH-RT and the 2549 amino acids of the human mammalian target of rapamycin (MTOR) - and could influence the sensitivity significantly and hence the ability of PCM to detect signal from the averaged representation used herein. To remedy these challenges, the use of alignment-dependent or autocorrelation descriptors could be explored^{8,66}. Moreover, as previously mentioned, some mutants are disease-causing and are often the drug target. For these cases, in which molecules are optimized away from the WT, the baseline for the *QSAR-WT* could be substituted with the disease-causing mutant. The modeling results presented here for all proteins containing variant data can be used for decision-making regarding additional data curation or the selection of modeling tasks for individual proteins. As a rule of thumb, targets with small datasets and/or high variant bioactivity percentages are the most susceptible to the presence of variants. These should be thoroughly examined before modeling and, if needed, additional measures should be implemented to tackle the drawbacks in the dataset⁶⁷.

Beyond bioactivity modeling with a focus on the WT protein, the dataset and results presented here can be exploited in variant bioactivity prediction with some precautions. First, variant data is still too sparse for large-scale modeling of new variants, as represented by the low performance of PCM models with LOVO validation. However, small-scale campaigns following data balancing strategies showed promising results and should be considered in light of each particular project's scope⁶⁸. Second, in this work only amino acid substitutions were considered, however, other aberrations such as deletions, insertions, amplifications, or copy number variations are known to be clinically relevant and affect both protein function and pharmacology^{48,69}. A protocol should therefore be devised to also map these variations in bioactivity databases accurately. Third, the biological context of the variants studied – activating vs. resistance substitutions, as an example – is correlated with the effect in bioactivity, and should be considered in database annotation and extrapolated to modeling. Fourth, new clinical variants are constantly identified and have limited data in bioactivity databases compared to established variants⁷⁰. This does not mean that these variants are less important, and thus more appropriate channels for variant tracking should be consulted simultaneously to assess clinical relevance. Finally, the data and results presented here should not be restricted to bioactivity modeling for virtual screening, and thus the exploration of other modeling tasks considering protein variants is highly encouraged including (and not restricted to) selectivity modeling⁷¹, drug design by fragment merging⁷², or pharmacophore modeling⁶².

Conclusions

The genetic variability landscape of ChEMBL, the most widely used public bioactivity database in computational drug discovery, was comprehensively analyzed for the first time. Key advantages resulting from years of expert knowledge gathering in ChEMBL's variant annotation pipeline were identified through parallel annotation. Additionally, mis-annotations requiring future correction were found. Recommendations for pipeline enhancement were provided, alongside a proposal for simplified annotation of target variants for bioactivity modeling, which are made available in a modeling dataset. The amount and distribution of variant data across protein organisms, families, individual proteins, and variants were extensively described. Furthermore, a Python package and notebooks were developed to assess variant effects on bioactivity distributions and modeling performance. The potential of these analysis tools to extract variants and promising chemical candidates was demonstrated, particularly for data-rich proteins. Particularly, informed decisions for noise reduction in bioactivity models and modeling variant bioactivity can be facilitated using our approach.

Materials and Methods

Bioactivity data sources

Bioactivity data was collected from ChEMBL (version 31) and the Papyrus dataset (version 5.5). The Papyrus dataset contains highly curated data from ChEMBL version 31,

ExCAPE-DB, and other individual datasets. Protein targets in the Papyrus dataset are identified either by *accession* (i.e. UniProt accession code) or *target_id*. The latter is constructed from the accession and the amino acid substitutions present in the variant analyzed, with *accession_WT* for wild-type (WT) proteins. In its current version, the Papyrus dataset does not reflect variants described in ChEMBL.

ChEMBL data was collected using the ChEMBL Python client (**Supplementary Figure 4.14a**, full query available on the associated GitHub repository, see **Appendix B**). The data queried included activities (i.e. *pchembl_value* and *activity_comment*), assay descriptions, molecular structures (i.e. SMILES – *canonical_smiles*), protein identifiers and sequences, and ChEMBL-annotated variants (i.e. *mutation* in the *variant_sequences* table).

After assay-based amino acid substitution annotation (see *Amino acid substitution annotation* section and **Figure 4.1**), ChEMBL assay-target pairs were given Papyrus-like identifiers based on the validated substitutions. Target variants were henceforward identified by *target_id*. Subsequently, individual ChEMBL activity points were mapped to annotated variant targets (*target_id*) based on their *assay_id* and *accession*. Duplicated activity data (*target_id*-*compound* *chembl_id* pairs) from several assays were joined into one single point by dropping low-quality data and calculating the mean *pchembl* value or most common activity label (**Supplementary Figure 4.14b**). The *data_validity* field was used to drop low-quality data (author confirmed error), as done in the Papyrus dataset.³⁴ The *activity_comment* field was also used to define active and inactive binary labels when *pchembl_value* was not available.

Before variant bioactivity analysis, the Papyrus and ChEMBL datasets were integrated. Firstly, only the Papyrus entries originating from the Christmann subset were considered, filtering out de facto any Papyrus data point with ChEMBL as a source, avoiding duplicates. ChEMBL compounds were given Papyrus-like identifiers (*connectivity*). Then, the average *pchembl_value* was calculated for unique *target_id*-*connectivity* pairs. For data points with no *pchembl_value*, the most common activity label was kept. Finally, the VEBD for analysis was constrained to only targets with at least one variant annotated other than the WT.

Amino acid substitution annotation

ChEMBL amino acid substitutions were extracted from assay *descriptions* for unique assay-target (i.e. *assay_id*-*accession*) pairs following a three-step approach (**Figure 4.1**).

- i) First, regular expressions were used to extract from the assay description amino acid substitution patterns. This is, either a one-letter amino acid code followed by an unlimited number of digits and another one-letter code, or a three-letter amino acid code followed by digits and another three-letter code. Subsequently, three-letter codes were transformed into one-letter codes.
- ii) Second, exceptions were defined from assay-associated metadata and filtered out. These exceptions included assay cell types, target names, and target gene

names and synonyms. At this level, an option was included to manually define exceptions from a JSON file for specific assays. Here, most “M1” and “D2” instances were filtered out as they could easily get a false positive validation status in step iii. The complete JSON file used for manual exception definition is included in the associated GitHub repository (see **Appendix B**).

- iii) Third, the remaining substitutions were validated by mapping the first amino acid of the substitution pattern to the WT sequence. If the mapping was successful, the substitutions were included for further analysis.

The resulting annotated assay-target pairs from the first round of annotation were introduced in an annotation feedback loop where they were compared to the original ChEMBL-annotated variants (**Supplementary Figure 4.1**). Annotations missed by ChEMBL were manually checked to assess their validity and classified accordingly into different categories of true and false positives. True positives included likely correct new annotations and likely correct rescue instances of “UNDEFINED MUTATION” labels in ChEMBL. New annotations and rescues with deletions were also categorized as true positives given the scope of this work. ChEMBL-only annotations were parsed and categorized into different categories of true and false negatives. True negatives included misclassified annotations due to the mis-linking of single protein assays to protein families. Missed deletions were also categorized as true negatives in light of this work’s scope. False negatives included instances where expert knowledge was required. These were, for example, variants for which the amino acid substitution extracted matched but the sequence position was different due to sequence number shifts. Another example was constituted by completely missed substitutions because they did not correspond to the canonical regular expression. On the verge between true and false negatives were other ambiguous sequence number and amino acid substitution mismatches that did not correspond to the categories defined before. Without further manual curation, these could correspond either to potential ChEMBL miss-annotations or missed correct annotations requiring expert knowledge. In a second round of annotation following the annotation feedback loop, the defined false positives were excluded from the annotated variants and reverted to WT. Similarly, false negatives were rescued by using the ChEMBL-annotated variants. The ambiguous cases were annotated as undefined variants given the lower confidence. The assay-target annotations from the second round were further linked to ChEMBL activity data to annotate variant targets (see section *Bioactivity data sources*).

Family and taxonomic distribution analysis

Protein family annotations were retrieved from ChEMBL version 31 by querying levels L1-L5 from the SQL table *protein_family_classification* for all unique UniProt accession codes. Proteins in the VEBD were mapped to their corresponding family levels based on their accession code. Non-defined levels were labeled as “Other”. On levels L1 and L2, small-sized families were grouped into larger families as follows. L1 tags “Auxiliary”, “Unclassified”, “Structural”, and “Surface” were grouped into “Other”. L2 tags “Primary active”, “Ligase”, “Isomerase”, and “Writer” were grouped into “Other”. Additionally, all G protein-coupled receptor L2 tags were grouped into a single L2 family, “GPCR”.

Subsequently, the total number of bioactivity data points as well as the number of variant bioactivity data points in the VEBD were calculated across families for each level. From these, the variant bioactivity percentage per family was calculated by dividing the amount of variant data by the amount of total data and multiplying the result by 100. Similarly, the novel variant bioactivity annotation percentage was calculated exclusively in ChEMBL data by dividing the number of bioactivity data points in potentially novel annotated variants (i.e. not previously defined in the ChEMBL “mutation” variable) by the total number of variant bioactivity data and multiplying the result by 100.

Organism names and HGNC gene symbols were mapped on accession codes from the Papyrus version 05.5 protein table. Moreover, the proteins’ taxonomy was retrieved and mapped for all unique UniProt accession codes using the UniProt API via the UniProtMapper package. The two *Escherichia coli* strains present in the dataset were aggregated under one single *Escherichia coli* organism. The number of variants and bioactivity data points were subsequently calculated at different taxonomy levels.

Statistical analysis per protein and variant

The amount and distribution of variant bioactivity data across individual proteins and variants were analyzed in detail. For each protein, the number of variants and bioactivity data points were calculated, as well as the variant bioactivity percentage compared to the totality of the protein’s data. Within proteins, variants were ordered from most to least populated in terms of bioactivity data. The relative amount of data in the most populated compared to each of the following variants was calculated by dividing the amount of data in the first variant by the amount of data in the variant of interest.

Amino acid substitution type analysis

Amino acid substitution types were extracted from the variants. For variants with multiple substitutions, all the substitutions were considered individually. Three substitution-type definitions were implemented:

- i) Categorical: Six substitution-type categories were defined based on the type of amino acid substitution regarding side chain size and polarity. “Conservative” for amino acid substitutions where the size and polarity remained similar. “Size” when size changed but polarity remained the same. “Polar” and “Charge” when the size remained similar but either the polarity or the actual charge, respectively, changed. And “Polar size” and “Charge size” as a combination of the aforementioned size and polarity changes. To define the changes, amino acids were grouped into four polarity groups and three size groups. Polarity groups included non-polar (alanine, glycine, isoleucine, leucine, proline, valine, methionine, phenylalanine), polar neutral (asparagine, glutamine, serine, threonine, tyrosine, cysteine, tryptophan), polar acidic (glutamic acid, aspartic acid), and polar basic (arginine, histidine, lysine). Size groups were defined based on the relative side

chain size previously defined by Epstein³⁵ and included bulky (tryptophan, tyrosine, arginine, phenylalanine), intermediate (histidine, glutamic acid, glutamine, lysine, methionine, asparagine, leucine, isoleucine, proline), and small (cysteine, threonine, valine, alanine, glycine).

- ii) Continuous and non-directional (Grantham's distance): A value from 5 (most similar, leucine-isoleucine) to 215 (most dissimilar, cysteine-tryptophan) was assigned to each amino acid substitution mapping it to Grantham's distance matrix. This distance depends on three properties: composition, polarity, and molecular volume; and is independent of the directionality of the change (e.g. leucine > isoleucine is the same as isoleucine > leucine).
- iii) Continuous and directional (Epstein's coefficient of difference): A value from 0 (most similar) to 1 (most different) was assigned to each amino acid substitution mapping it to Epstein's coefficient of difference matrix. This coefficient depends on the polarity and size of the replaced amino acids and takes into account directionality (e.g. leucine > tyrosine is 0.28 and tyrosine > leucine is 0.22).

The number of variants and bioactivity data was subsequently calculated per substitution type for different subsets of proteins. For variants with multiple substitutions, each substitution was considered, and therefore accounted for, separately.

Amino acid substitution location analysis

Amino acid substitutions in a protein were defined by their location within the protein with respect to its binding pocket. To this end, each protein was mapped by its UniProt accession code to the available PDB structures with a co-crystallized ligand, which were downloaded as PDB files. Next, for each structure, the structure's first chain with the crystallized ligand was extracted and, for that chain, the ligand's coordinates in the PDB file were retrieved. Based on these coordinates, the ligand's center of geometry (centroid) was calculated. Similarly, the centroid of each residue in the chain was also calculated. Finally, the distance between the ligand's centroid and each residue's centroid was computed, and the average distance was calculated for each residue across all PDB structures available for a protein. The average distance between the substituted residues' centroid and the ligand's centroid was subsequently used as a metric to differentiate variants based on the location of the substituted residue in the protein. Of note, the average distance between centroids will by definition be larger than the shortest distance to the ligand, which is generally considered when using distances of 5 Å to define the binding pocket. This metric was constructed to be as ligand-agnostic as possible, which in turn leads to non-generalizable distance ranges and should therefore be considered carefully (as an example two ligands with different sizes and binding modes leading to different distances to key residues in EGFR are presented in **Supplementary Figure 4.15**). In variants with multiple substitutions, each substitution was considered separately. For the analysis of HIV-1 RNaseH-RT (Q72547), only the first of two retrieved PDB codes (2JLE and 3HYF) was used to annotate substitutions located in the reverse transcriptase domain (**Supplementary Figure 4.16**).

Common subset design

The analysis of variant bioactivity data was done on common subsets of small molecules to ensure fair and accurate comparisons between distributions (**Supplementary Figure 4.17**). When possible, fully dense common subsets were computed, where all compounds of the subset had been tested on all annotated variants. More typically, non-fully dense common subsets - referred to as common subsets - were defined for each *accession* by first keeping molecules that meet a threshold of being tested on a minimum number of variants. For further analysis, this minimum variant threshold was set to at least two variants. Secondly, variant coverage was calculated as the percentage of molecules in the subset that were tested on a specific variant. Subsequently, variants above a certain coverage threshold were kept for analysis. Ideally, variant coverage would be set to 100% but, due to high data sparsity, it was set to 20% for analysis.

To increase the density of the common subset, a strategy was introduced where similarity-based filters were used for calculating the minimum variant and the variant coverage thresholds. To obtain these similarity-expanded common subsets, we first computed pair-wise Tanimoto similarities for all molecules in our dataset. Then, we assigned to each molecule a similarity group containing all molecules with a Tanimoto similarity above a certain threshold (0.80). Next, we computed common subset thresholds considering not only true activity points but also activity points in the similarity groups. This is, for threshold calculation a non-existing activity point of molecule X in variant A was counted as existing if compound Y, similar to X, was tested in variant A.

Common subsets were also computed to enable full-panel bioactivity analysis of proteins without a true fully dense common subset. For example, for EGFR (P00533), a bioactivity analysis subset was derived from a common subset computed with a minimum variant threshold of three and a variant coverage of 10%. For HIV-1 RNaseH-RT (Q72574), from a common subset for variants with a compound coverage greater than 3%. For IDHC (O75874), from a common subset for compounds tested on at least two variants and variants with a compound coverage greater than 20%. Finally, for epigenetic regulator BRD4 (O60885), from a common subset for variants with a compound coverage greater than 2%.

The differences between the bioactivity distributions across different types of common subsets were analyzed by calculating the Wasserstein distance between distributions of the *pchembl_value_Mean* variable separately for the WT and all variants combined.

Molecular clustering and visualization

Small molecules in a subset of compounds were clustered using the Butina algorithm to represent their structural similarity across the subset. Starting from compounds represented by canonical SMILES, molecular objects were generated using RDKit. Subsequently, RDKit Daylight-like topological fingerprints were generated and the Tanimoto distance matrix was calculated based on these. Finally, the Butina cluster algorithm was applied

to the similarity matrix with a varying cutoff for each subset to minimize the number of single-element clusters. Clusters generated to analyze variant bioactivity distributions in **Figure 4.7** were computed for subsets including all compounds tested on at least two variants and a Butina cluster cutoff of 0.5. Clusters generated to analyze the full-panel bioactivity differences of compounds in the EGFR (P00533; **Figure 4.6**), BRD4 (O60885), and IDHC (O75874) bioactivity analysis subsets were computed for said bioactivity analysis subsets with a Butina cluster cutoff of 0.7. For HIV-1 RNaseH-RT (Q72574), the cluster cutoff was set to 0.5.

To visualize the molecules in a subset of compounds, 2D molecular representations were computed with RDKit. Molecular substructures of interest were matched and highlighted in red. These included either the largest ring system in the molecule or the atoms corresponding to the maximum common substructure of all the compounds in a given cluster.

Variant bioactivity distribution analysis

The distribution of bioactivity values across variants per protein was analyzed for three different types of subsets: i) modeling, ii) common, and iii) Butina clusters. These subsets were computed to capture differences in bioactivity across variants covering, respectively, i) all compounds tested on a given protein, ii) a common subset of compounds tested across variants, and iii) different areas of the chemical space tested on a given protein. Common subsets were computed as defined in the section *Common subset design*. In all cases, univariate pchembl value distributions were plotted using kernel density estimations in Seaborn for each variant present in the protein subset.

To give an idea of the data sparsity across variants in the different subsets, variant coverage was calculated and reported as defined in the section *Common subset design*. To summarize the bioactivity distribution information, the mean and standard deviation pchembl value for each variant was calculated. Moreover, the difference in mean pchembl value with respect to the WT was calculated for each variant by subtracting the variant's mean pchembl value from the WT's mean pchembl value.

Modeling of bioactivities

Three sets were considered for modeling with machine learning. The first set consisted of the original set of bioactivity values obtained for both WT and variant proteins. The second set consisted of data points relating to the WT protein sequences only. Finally, the third set consisted of the similarity-derived common subsets.

All three sets were independently modeled with a quantitative structure-activity relationship (QSAR) model for each accession without any protein sequence-derived feature and with a proteochemometrics (PCM) model for all accessions altogether with sequence features. Protein sequences containing other than the 20 natural amino acids were not considered for modeling with PCM. The collected negative logarithmically

scaled bioactivities values were modeled using the XGBoost (version 1.7.5) implementation of gradient-boosted regression trees⁷³. Molecules were represented with the 777 physicochemical and topological Mold2 molecular descriptors⁷⁴. Unaligned protein sequences were described with ProDEC⁷⁵ by splitting them into 50 equal parts and averaging the first three principal components (PCs) of Sandberg *et al.*'s amino acid descriptors over each part and over the entire sequence for each of the three PCs, resulting in 153 features (50 parts x 3 PCs + 3 averages PCs)^{11,76,77}. Models were 5-fold cross-validated using a random split with a random seed set to 1234 and using a leave-one-out strategy applied for each sequence variant (LOVO). Accessions with less than five data points were disregarded for QSAR modeling and data points related to only one variant were not considered for PCM modeling. Applicability domains were derived using MLChemAD (version 1.2.0) with isolation forests by fitting the training subsets and evaluating them on the Enamine Hit Locator Library (downloaded on 24/01/2024), emulating a typical real-world virtual screening. Finally, the performances of cross-validated models were statistically evaluated between *PCM-All*, *QSAR-WT*, and *QSAR-All* models using Friedman's test for repeated samples using Scipy (version 1.11.2). Significant differences (p-value<0.05) were further investigated using pairwise uncorrected post-hoc Conover-Friedman tests (p-value<0.05) using scikit_posthocs (version 0.8.0).

References

1. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **12**, 2694–2718 (2016).
2. Hessler, G. & Baringhaus, K.-H. Artificial Intelligence in Drug Design. *Molecules* **23**, 2520 (2018).
3. Mager, P. P. Theoretical approaches to drug design and biological activity: critical comments to the use of mathematical methods applied to univariate and multivariate quantitative structure-activity relationships (QSAR). *Med Res Rev* **2**, 93–121 (1982).
4. Matsuzaka, Y. & Uesawa, Y. Ensemble Learning, Deep Learning-Based and Molecular Descriptor-Based Quantitative Structure-Activity Relationships. *Molecules* **28**, 2410 (2023).
5. Trapotsi, M.-A. *et al.* Comparison of Chemical Structure and Cell Morphology Information for Multitask Bioactivity Predictions. *J Chem Inf Model* **61**, 1444–1456 (2021).
6. Norinder, U., Spjuth, O. & Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J Chem Inf Model* **60**, 2830–2837 (2020).
7. Li, Y. *et al.* Introducing block design in graph neural networks for molecular properties prediction. *Chem Eng J* **414**, 128817 (2021).
8. Bongers, B. J. *et al.* Proteochemometric Modeling Identifies Chemically Diverse Norepinephrine Transporter Inhibitors. *J Chem Inf Model* **63**, 1745–1755 (2023).
9. Bongers, B. J., IJzerman, A. P. & Van Westen, G. J. P. Proteochemometrics – recent developments in bioactivity and selectivity modeling. *Drug Discov. Today Technol.* **32**, 89–98 (2019).
10. Kim, P. T., Winter, R. & Clevert, D.-A. Unsupervised Representation Learning for Proteochemometric Modeling. *Int J Mol Sci* **22**, 12882 (2021).
11. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminformatics* **9**, 45 (2017).
12. Zakharov, A. V. *et al.* Novel Consensus Architecture to Improve Performance of Large-Scale Multitask Deep Learning QSAR Models. *J. Chem. Inf. Model.* **59**, 4613–4624 (2019).
13. Cortes-Ciriano, I. *et al.* Proteochemometric modeling in a Bayesian framework. *J. Cheminformatics* **6**, 35 (2014).
14. Wang, D. D., Xie, H. & Yan, H. Proteochemometrics interaction fingerprints of protein-ligand complexes predict binding affinity. *Bioinformatics* **37**, 2570–2579 (2021).
15. Sokouti, B. & Hamzeh-Mivehroud, M. 6D-QSAR for predicting biological activity of human aldose reductase inhibitors using quasar receptor surface modeling. *BMC Chem Biol* **17**, 63 (2023).
16. Atas Guvenilir, H. & Doğan, T. How to approach machine learning-based prediction of drug/compound-target interactions. *J Cheminform* **15**, 16 (2023).
17. Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* **9**, 5441–5451 (2018).
18. Zhao, L., Wang, W., Sedykh, A. & Zhu, H. Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *ACS Omega* **2**, 2805–2812 (2017).
19. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
20. Tükkäinen, P., Bellis, L., Light, Y. & Franke, L. Estimating error rates in bioactivity databases. *J Chem Inf Model* **53**, 2499–2505 (2013).
21. Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of mixed IC₅₀ data - a statistical analysis. *PLoS One* **8**, e61007 (2013).
22. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The experimental uncertainty of heterogeneous public K_i data. *J Med Chem* **55**, 5165–5173 (2012).
23. Geng, C., Vangone, A. & Bonvin, A. M. J. J. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Eng Sel* **29**, 291–299 (2016).
24. Feng, C. *et al.* Cancer-Associated Mutations of the Adenosine A2A Receptor Have Diverse Influences on Ligand Binding and Receptor Functions. *Molecules* **27**, 4676 (2022).
25. den Hollander, L. S. *et al.* Impact of cancer-associated mutations in CC chemokine receptor 2 on receptor function and antagonism. *Biochem. Pharmacol.* **208**, 115399 (2023).
26. Hu, Y. *et al.* Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir. Preprint at *BioRxiv* <https://doi.org/10.1101/2022.06.28.497978> (2022).
27. Du, Y. *et al.* Evolution of Multiple Domains of the HIV-1 Envelope Glycoprotein during Coreceptor Switch with CCR5 Antagonist Therapy. *Microbiol Spectr* **10**, e0072522 (2022).
28. Yver, A. Osimertinib (AZD9291)-a science-driven, collaborative approach to rapid drug design and development. *Ann Oncol* **27**, 1165–1170 (2016).
29. Musharrafieh, R., Ma, C. & Wang, J. Discovery of M2 channel blockers targeting the drug-resistant double mutants M2-S31N/L26I and M2-S31N/V27A from the influenza A viruses. *Eur J Pharm Sci*

- 141, 105124 (2020).
30. Landrum, G. A. & Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).
31. Leeson, P. D. *et al.* Target-Based Evaluation of “Drug-Like” Properties and Ligand Efficiencies. *J. Med. Chem.* **64**, 7210–7230 (2021).
32. Van Westen, G., Hendriks, A., Wegner, J. K., Ijzerman, A. P. & Van Vlijmen, H. W. T. Significantly Improved HIV Inhibitor Efficacy Prediction Employing Proteochemometric Models Generated From Antivirogram Data. *PLoS Comput Biol* **9**, 1002899 (2013).
33. Christmann-Franck, S. *et al.* Unprecedentedly Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound-Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **56**, 1654–1675 (2016).
34. Béquignon, O. J. M. *et al.* Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J. Cheminformatics* **15**, 3 (2023).
35. Epstein, C. J. Non-randomness of Ammonoacid Changes in the Evolution of Homologous Proteins. *Nature* **215**, 355–359 (1967).
36. Butina, D. Unsupervised Data Base Clustering Based on Daylight’s Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).
37. Hunter, F. M. I. *et al.* Drug Safety Data Curation and Modeling in ChEMBL: Boxed Warnings and Withdrawn Drugs. *Chem. Res. Toxicol.* **34**, 385–395 (2021).
38. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics* **12**, 51 (2020).
39. Starlinger, J. *et al.* Variant information systems for precision oncology. *BMC Med. Inform. Decis. Mak.* **18**, 107 (2018).
40. Masoudi-Sobhanzadeh, Y., Omid, Y., Amanlou, M. & Masoudi-Nejad, A. Drug databases and their contributions to drug repurposing. *Genomics* **112**, 1087–1095 (2020).
41. Shafer, R. W. Rationale and Uses of a Public HIV Drug-Resistance Database. *J. Infect. Dis.* **194**, S51–S58 (2006).
42. Sandgren, A. *et al.* Tuberculosis Drug Resistance Mutation Database. *PLoS Med.* **6**, e1000002 (2009).
43. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistance prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
44. Duffy, S. Why are RNA virus mutation rates so damn high? *PLoS Biol.* **16**, e3000003 (2018).
45. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
46. Lopez-Bigas, N., De, S. & Teichmann, S. A. Functional protein divergence in the evolution of Homo sapiens. *Genome Biol.* **9**, R33 (2008).
47. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
48. Liu, H., Zhang, B. & Sun, Z. Spectrum of EGFR aberrations and potential clinical implications: insights from integrative pan-cancer analysis. *Cancer Commun.* **40**, 43–59 (2020).
49. Testoni, E. *et al.* Somatic mutated ABL1 is an actionable and essential NSCLC survival gene. *EMBO Mol. Med.* **8**, 105–116 (2016).
50. Śmiech, M., Leszczyński, P., Kono, H., Wardell, C. & Taniguchi, H. Emerging BRAF Mutations in Cancer Progression and Their Possible Effects on Transcriptional Networks. *Genes* **11**, 1342 (2020).
51. Bresler, S. C. *et al.* ALK mutations confer differential oncogenic activation and sensitivity to ALK inhibition therapy in neuroblastoma. *Cancer Cell* **26**, 682 (2014).
52. Artemenko, M., Zhong, S. S. W., To, S. K. Y. & Wong, A. S. T. p70 S6 kinase as a therapeutic target in cancers: More than just an mTOR effector. *Cancer Lett.* **535**, 215593 (2022).
53. Rocha, E. M., Keeney, M. T., Maio, R. D., Miranda, B. R. D. & Greenamyre, J. T. LRRK2 and idiopathic Parkinson’s disease. *Trends Neurosci.* **45**, 224–236 (2022).
54. Bracht, J. W. P. *et al.* BRAF Mutations Classes I, II, and III in NSCLC Patients Included in the SLIP Trial: The Need for a New Pre-Clinical Treatment Rationale. *Cancers* **11**, 1381 (2019).
55. Sunami, T. *et al.* Structural Basis of Human p70 Ribosomal S6 Kinase-1 Regulation by Activation Loop Phosphorylation. *J. Biol. Chem.* **285**, 4587–4594 (2010).
56. Yang, L.-K. & Tao, Y.-X. Alanine Scanning Mutagenesis of the DRYxxI Motif and Intracellular Loop 2 of Human Melanocortin-4 Receptor. *Int. J. Mol. Sci.* **21**, 7611 (2020).
57. Tate, C. G. & Schertler, G. F. Engineering G protein-coupled receptors to facilitate their structure determination. *Curr. Opin. Struct. Biol.* **19**, 386–395 (2009).
58. Muk, S. *et al.* Machine Learning for Prioritization of Thermostabilizing Mutations for G-Protein Coupled Receptors. *Biophys. J.* **117**, 2228–2239 (2019).
59. Gilmer, T. M. *et al.* Impact of Common Epidermal Growth Factor Receptor and HER2 Variants on Receptor Activity and Inhibition by Lapatinib. *Cancer Res.* **68**, 571–579 (2008).
60. Singh, H., Singh, S., Singla, D., Agarwal, S. M.

- & Raghava, G. P. S. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct* **10**, 10 (2015).
61. Burggraaff, L. *et al.* Annotation of allosteric compounds to enhance bioactivity modeling for class A GPCRs. *J. Chem. Inf. Model.* **60**, 4664–4672 (2020).
 62. Dera, A. A. *et al.* Identification of Potent Inhibitors Targeting EGFR and HER3 for Effective Treatment of Chemoresistance in Non-Small Cell Lung Cancer. *Molecules* **28**, 4850 (2023).
 63. Cortes-Ciriano, I., Murrell, D. S., Van Westen, G. J., Bender, A. & Malliavin, T. E. Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. *J. Cheminformatics* **7**, 1 (2015).
 64. Hasan, R. & Chu, C. Noise in Datasets: What Are the Impacts on Classification Performance? *Proc. 11th Int. Conf. Pattern Recognit. Appl. Methods* (2022) doi:10.5220/0010782200003122.
 65. Kumar, V., De, P., Ojha, P. K., Saha, A. & Roy, K. A Multi-layered Variable Selection Strategy for QSAR Modeling of Butyrylcholinesterase Inhibitors. *Curr. Top. Med. Chem.* **20**, 1601–1627 (2020).
 66. Burggraaff, L. *et al.* Successive statistical and structure-based modeling to identify chemically novel kinase inhibitors. *J. Chem. Inf. Model.* **60**, 4283–4295 (2020).
 67. Caiafa, C. F., Sun, Z., Tanaka, T., Marti-Puig, P. & Solé-Casals, J. Machine Learning Methods with Noisy, Incomplete or Small Datasets. *Appl. Sci.* **11**, 4132 (2021).
 68. Aldeghi, M., Gapsys, V. & De Groot, B. L. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS Cent. Sci.* **5**, 1468–1474 (2019).
 69. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
 70. An, L. *et al.* Defining the sensitivity landscape of EGFR variants to tyrosine kinase inhibitors. *Transl. Res.* **255**, 14–25 (2023).
 71. Burggraaff, L., Van Vlijmen, H. W. T., Ijzerman, A. P. & Van Westen, G. J. P. Quantitative prediction of selectivity between the A1 and A2A adenosine receptors. *J. Cheminformatics* **12**, 1–16 (2020).
 72. Andrianov, G. V., Gabriel Ong, W. J., Serebriiskii, I. & Karanickolas, J. Efficient Hit-to-Lead Searching of Kinase Inhibitor Chemical Space via Computational Fragment Merging. *J. Chem. Inf. Model.* **61**, 5967–5987 (2021).
 73. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *KDD '16* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
 74. Hong, H. *et al.* Mold 2 , Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. doi:10.1021/ci800038f.
 75. Béquignon, O. J. M. OlivierBeq/ProDEC: Version 1.0.2. Zenodo <https://doi.org/10.5281/ZENODO.7007058> (2022).
 76. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets. *J. Cheminformatics* **5**, 41 (2013).
 77. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets. *J. Cheminformatics* **5**, 42 (2013).

Supplementary Information

Supplementary Tables 4.1, 4.2, 4.7, 4.8, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15, 4.16, and 4.17 are not included in this thesis due to spatial constraints. Please, check the Supplementary Information available in the data repository for this chapter linked in **Appendix B**.

Supplementary Table 4.1. Analysis of ChEMBL-missed substitutions. ChEMBL Assay ID - target (accession) pairs for which a novel annotation was derived from our amino acid substitution extraction and validation pipeline based on the assay description. This list was further manually classified into True Positive and False Positive labels. False Positives are additionally given a reason for their labeling to help explain the caveats of the extraction and validation pipeline. These reasons are further grouped into a “reason group” that is represented in Supplementary Figure 4.1. ChEMBL collaborators further analyzed these annotations and provided a comment that will be used in the future to improve the variant annotation pipeline.

Supplementary Table 4.2. Analysis of ChEMBL-only annotations. ChEMBL Assay ID - target (accession) pairs with annotated variants in ChEMBL that did not match the annotation that was derived - or was missing altogether - from our amino acid substitution extraction and validation pipeline based on the assay description. This list was further automatically classified based on rejection flags into True Negative and False Negative labels. Rejection flags are mapped to sub-labels that explain the reason for the mismatch and that are represented in Supplementary Figure 4.1. ChEMBL collaborators further analyzed these annotations and provided a comment that will be used in the future to improve the variant annotation pipeline.

Supplementary Table 4.3. Distribution of variant bioactivity data across protein families in targets with at least one annotated variant. ChEMBL family classification level L1.

L1 classification	Variant activity data	All data	Variant bioactivity %
Enzyme	20,759	266,328	7.80
Membrane receptor	1,730	96,037	1.80
Epigenetic regulator	1,105	21,244	5.20
Other	590	11,432	5.20
Transcription factor	458	23,975	1.90
Ion channel	245	17,036	1.40
Transporter	176	19,575	0.90
Secreted protein	23	212	10.80
Total	25,086	455,839	5.50

Supplementary Table 4.4. Comparison of novel and originally annotated variant data in ChEMBL for all protein families (L1 classification).

L1 classification	Novel variant data	All variant data	Novel annotated variant bioactivity %
Enzyme	3,631	20,779	17.50
Membrane receptor	218	1,758	12.40
Epigenetic regulator	70	1,174	6.00
Other	75	626	12.00
Transcription factor	42	472	8.90
Ion channel	6	250	2.40
Transporter	3	177	1.70
Secreted protein	-	23	0.00
Total	4,045	25,259	16.00

Supplementary Table 4.5. Distribution of variant bioactivity data across protein kinase subfamilies in targets with at least one annotated variant. ChEMBL family classification level L4, with L2=Kinase.

L4 classification	Variant data	All data	Variant bioactivity %
TK protein kinase group	5,925	76,095	7.80
CMGC protein kinase group	24	18,749	0.10
TKL protein kinase group	4,644	13,284	35.00
AGC protein kinase group	1,385	10,570	13.10
Other protein kinase group	146	9,161	1.60
Atypical protein kinase group	80	4,888	1.60
STE protein kinase group	28	1,073	2.60
CAMK protein kinase group	3	186	1.60
CK1 protein kinase group	8	42	19.00
Total	12,243	134,048	9.10

Supplementary Table 4.6. Comparison of novel and originally annotated data in ChEMBL for subfamilies of the Kinase enzymes family (L4 classification for L2 = Kinase).

L4 classification	Novel variant data	All variant data	Novel annotated variant bioactivity %
TK protein kinase group	280	5,473	5.10
TKL protein kinase group	825	4,695	17.60
AGC protein kinase group	1,302	1,452	89.70
Other protein kinase group	4	116	3.40
Atypical protein kinase group	-	80	0.00
STE protein kinase group	-	28	0.00
CMGC protein kinase group	4	24	16.70
CK1 protein kinase group	-	8	0.00
CAMK protein kinase group	-	3	0.00
Total	2,415	11,879	20.30

Supplementary Table 4.7. Annotated data statistics per UniProt accession code. Proteins are sorted in descending order of bioactivity data points in the dataset. L1-L5 ChEMBL classification reported.

Supplementary Table 4.8. Statistics of variant-annotated targets with respect to the number of variants and amino acid substitutions per variant. Proteins are sorted from largest to smallest number of variants. This value includes WT. The number of single amino acid substitutions per variant equals “-1” for variants with undefined substitutions and “0” for the WT.

Supplementary Table 4.9. Distribution of organisms, variants, and variant bioactivity data across taxonomic domains in targets with at least one annotated variant. In grey, are the statistics for the most highly represented organism in each domain. *Three proteins were not annotated taxonomically. **Percentage of total bioactivity data was calculated with respect to the total number of bioactivity data points, including the three proteins without taxonomic annotation (455,839).

Domain	Organisms	Proteins	Variants (incl. WT)	Bioactivity data	% of total bioactivity data **
Virus	14	28	217	21,972	4.82
	<i>Human immunodeficiency virus 1</i>	5	119	15,512	3.40
Archaea	1	1	2	2	0.00
Bacteria	16	28	115	3,203	0.70
	<i>Escherichia coli</i>	8	24	1,345	0.30
Eukaryota	14	275	1,410	429,998	94.33
	<i>Homo sapiens</i>	235	1,225	412,797	90.56
Total	45	332*	1,744	455,175*	99.85

Supplementary Table 4.10. Distribution of data across the top three most populated variants for proteins with over 10% variant bioactivity percentage and over 1,000 data.

Supplementary Table 4.11. Comparison between the bioactivity distributions of the variant-enhanced bioactivity dataset (VEBD) set and common subset (at least two variants, variant coverage 20% and similarity 80%) of each protein with a common subset, in decreasing size of the common subset. Dataset size represents the number of bioactivity data points in the dataset. The variant data percentage was calculated by dividing non-WT bioactivity data by the total dataset size multiplied by 100. Dataset sparsity was calculated as the dataset size divided by the potential full matrix size, calculated as the number of unique variants multiplied by the number of unique compounds. Wasserstein distance equal to or greater than 1.5 is highlighted in red to represent changes in the distribution. A variant data percentage equal to or greater than 50% is highlighted in green to represent a higher data balance. Dataset sparsity equal to or smaller than 0.5 is highlighted in green to represent lower data sparsity.

Supplementary Table 4.12. The performance of PCM and QSAR models depends on the number of data points and the variant bioactivity percentage. Performance is reported as the average Pearson correlation coefficient for protein and, between brackets, as the average of the standard deviation of Pearson *r* per protein between cross-validation folds. The best average Pearson *r* is reported in bold for each row. Pearson *r* of PCM and/or QSAR-WT models significantly differing from QSAR-All models are starred. Pearson *r* of PCM or QSAR-WT models significantly differing from all other models (i.e. QSAR-WT and QSAR-All, and QSAR-All and PCM-All respectively) are underlined.

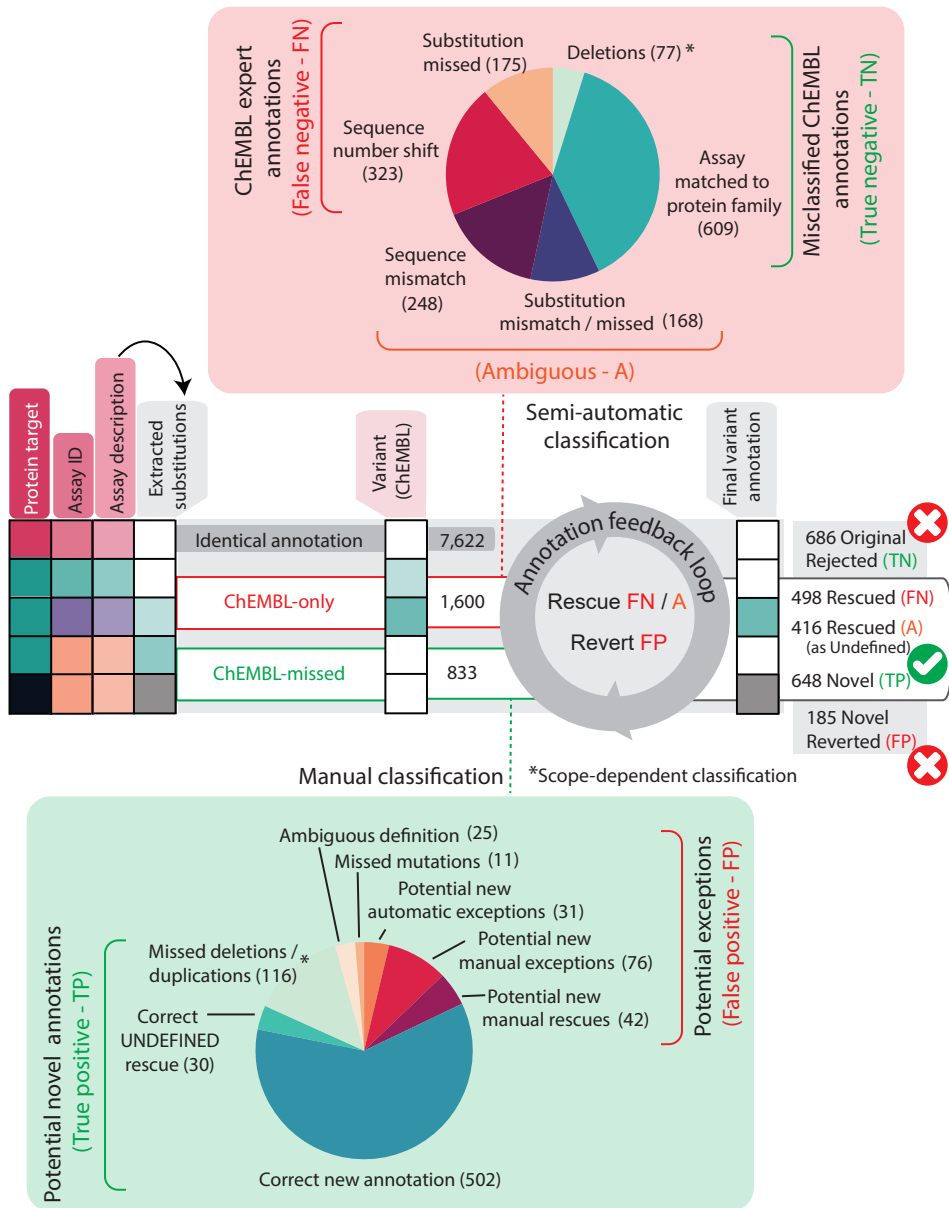
Supplementary Table 4.13. Performance of PCM and QSAR models obtained through Leave-One-Variant-Out (LOVO) cross-validation on the entire training set, proteins with specific numbers of data points, focused protein families, and data-rich proteins (more than 1,000 data points with at least 10% measured on variants). Performance is reported as the average Pearson correlation coefficient for each group or protein and, between brackets, as the average per group or protein of the standard deviation of Pearson *r* between cross-validation folds for each protein.

Supplementary Table 4.14. Molecular applicability domains of models evaluated through the Enamine Hit Locator Library as the fraction of the library's molecules close to molecules of the training set, using an isolation forest algorithm.

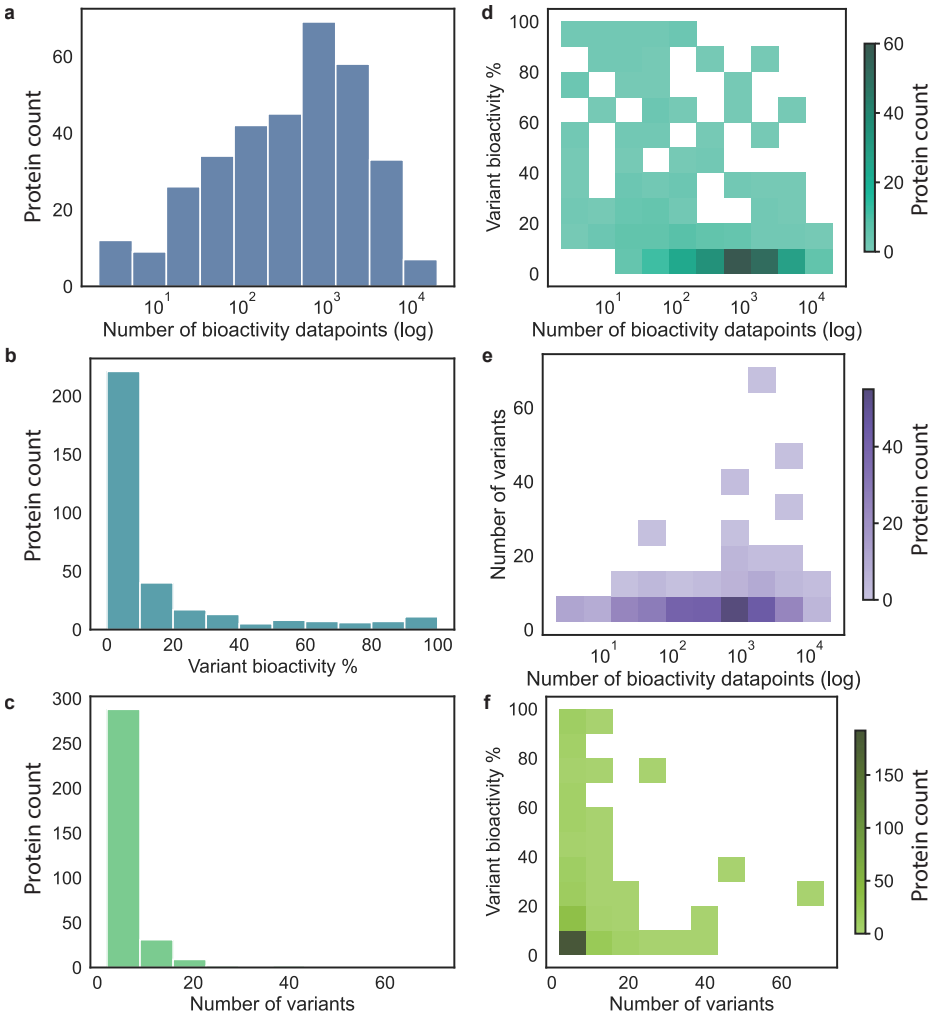
Supplementary Table 4.15. Performance of PCM and QSAR models obtained through cross-validation on the entire similarity-based common subset, on focused protein families, and data-rich proteins (more than 1,000 data points with at least 10% variant bioactivity percentage). Performance is reported as the average Pearson correlation coefficient for each group or protein and, between brackets, as the average per group or protein of the standard deviation of Pearson *r* between cross-validation folds for each protein.

Supplementary Table 4.16. Performance of PCM and QSAR models obtained through LOVO cross-validation on the entire similarity-based common subset, on focused protein families, and data-rich proteins (more than 1,000 data points with at least 10% variant bioactivity percentage). Performance is reported as the average Pearson correlation coefficient for each group or protein and, between brackets, as the average per group or protein of the standard deviation of Pearson *r* between cross-validation folds for each protein.

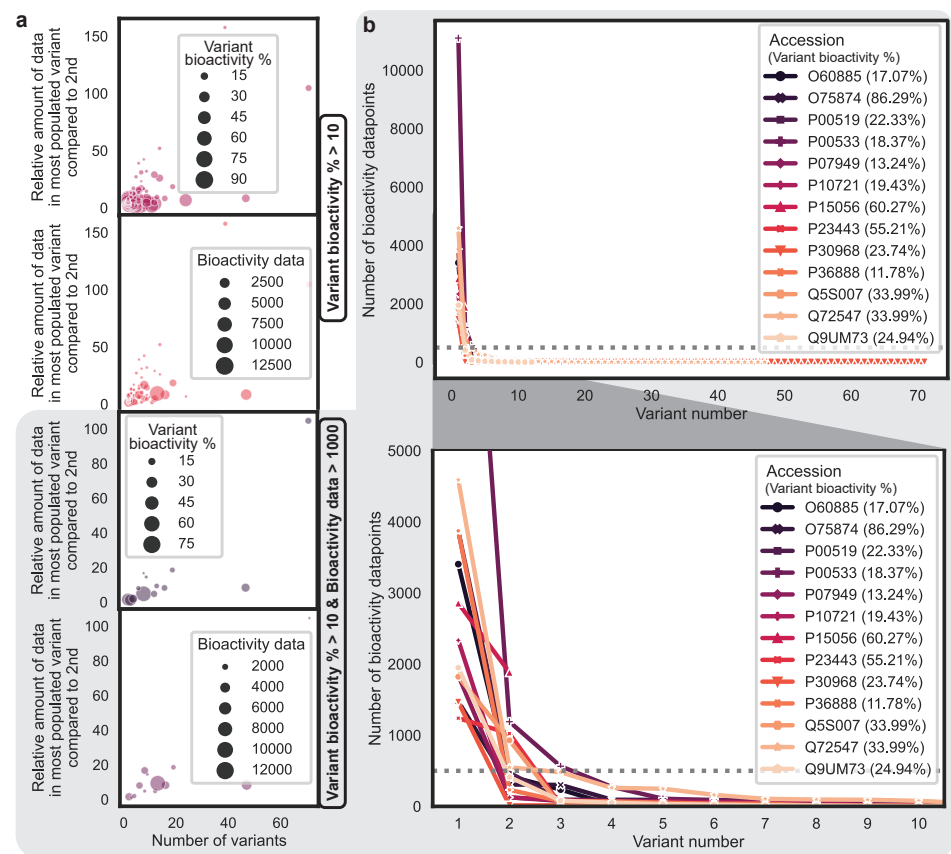
Supplementary Table 4.17. Statistical analysis of the difference in performance between PCM models and QSAR models both considering all data points as WT or only WT.



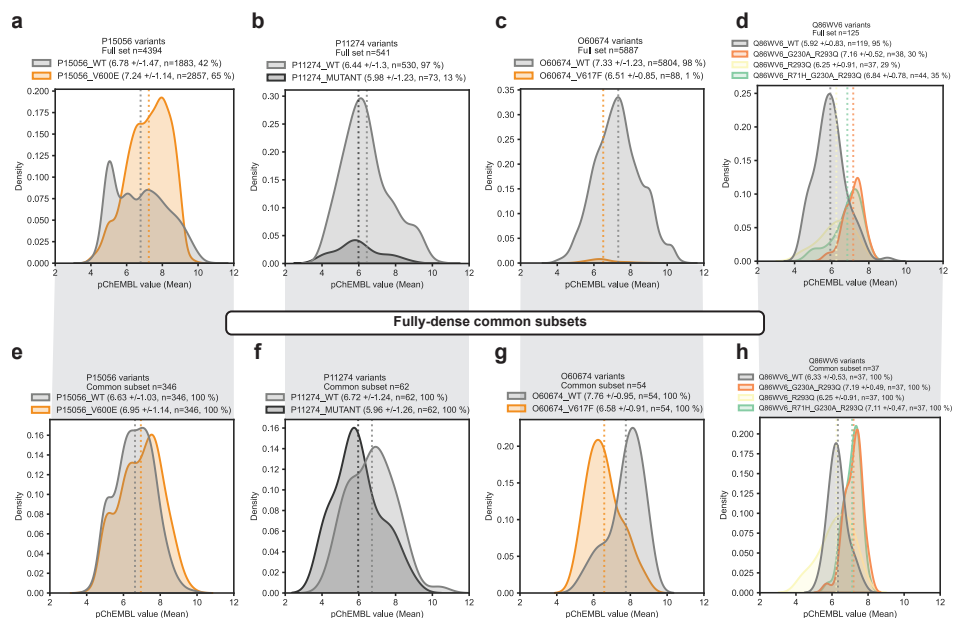
Supplementary Figure 4.1. Annotation feedback loop for unique ChEMBL assay-target pairs. Amino acid substitutions annotated and validated following step 2 of the pipeline shown in Figure 4.1 were compared to the original ChEMBL-annotated variants. ChEMBL-missed substitutions were manually checked to assess their validity, and classified accordingly into different categories of true and false positives. ChEMBL-only annotations were parsed and categorized into different categories of true and false negatives based on the nature of the mismatch. A third group of ambiguous ChEMBL-only variants was also flagged. The flags derived from the annotation feedback loop were used to rescue false negatives from the ChEMBL-only annotations and to revert false positive ChEMBL-missed annotations. This resulted in the final annotations used to construct the VEBD.



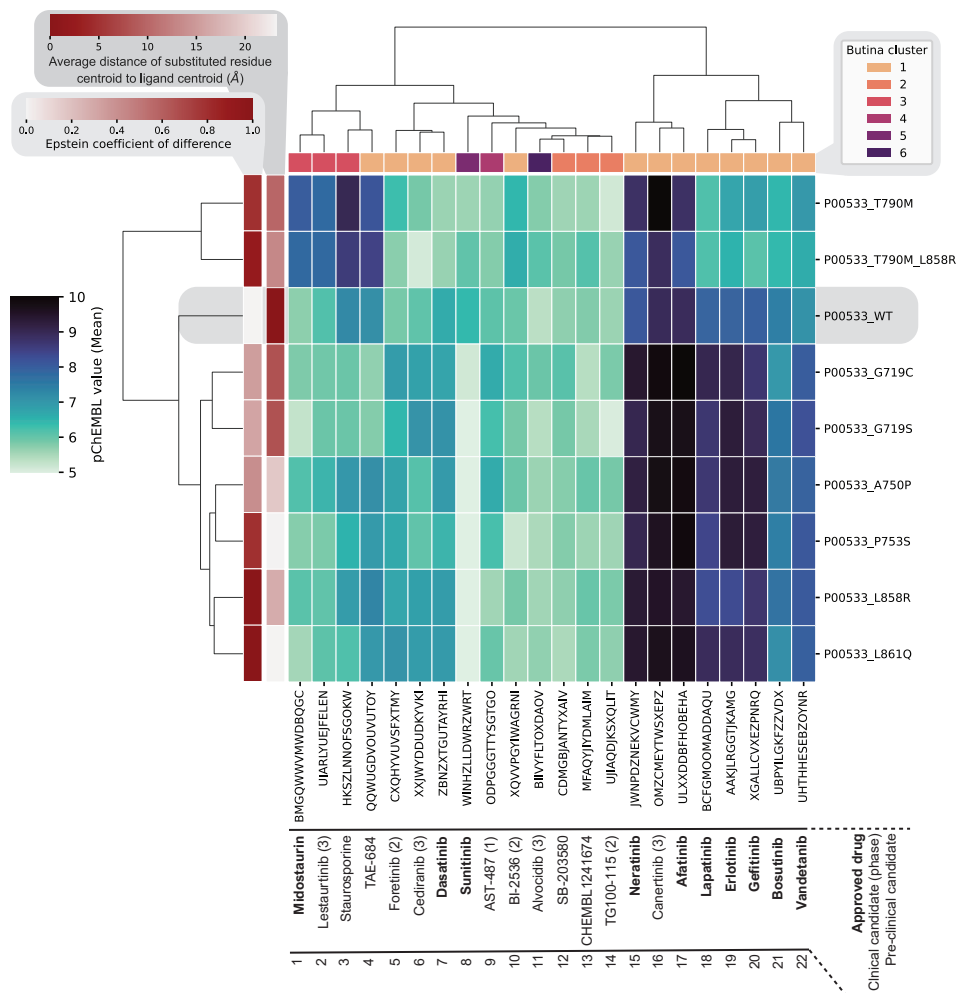
Supplementary Figure 4.2. Number of annotated proteins according to one variable: **a)** amount of bioactivity data (log scaled), **b)** variant bioactivity percentage, **c)** number of annotated variants, including wild-type (WT); or according to two variables: **d)** amount of bioactivity data and variant bioactivity percentage, **e)** amount of bioactivity data and number of annotated variants, **f)** number of annotated variants and variant bioactivity percentage.



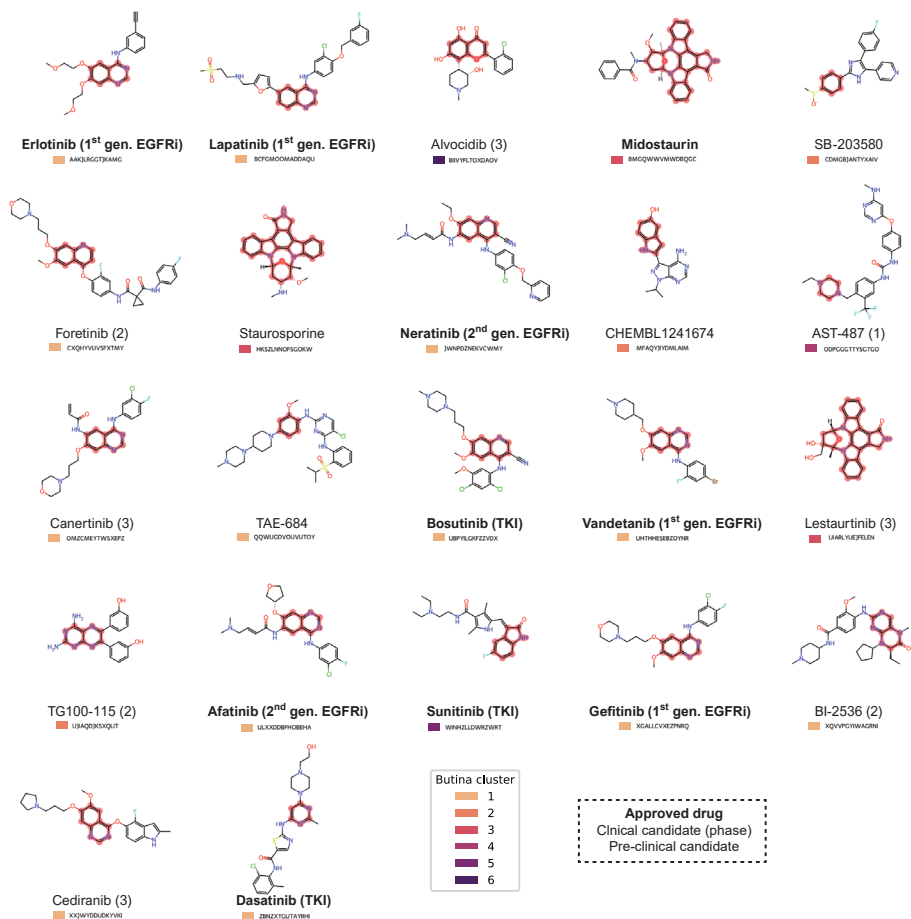
Supplementary Figure 4.3. Bioactivity data density across variants for data-rich proteins with a variant bioactivity percentage equal to or higher than 10%. **a)** Correlation between the number of annotated variants and the relative amount of data in the most populated variant compared to the second most populated variant. Bubble size represents either the variant bioactivity percentage or the total amount of bioactivity data for the protein. The two bottom panels are subsets of the two top panels, where only proteins with more than 1,000 bioactivity data are plotted. **b)** Number of bioactivity data per variant in order of decreasing amount of data for the 13 proteins with a variant bioactivity percentage equal to or higher than 10 and an amount of bioactivity data bigger than 1,000. The bottom panel is a zoom into the first panel. The dashed grey line represents 500 bioactivity data points, for reference.



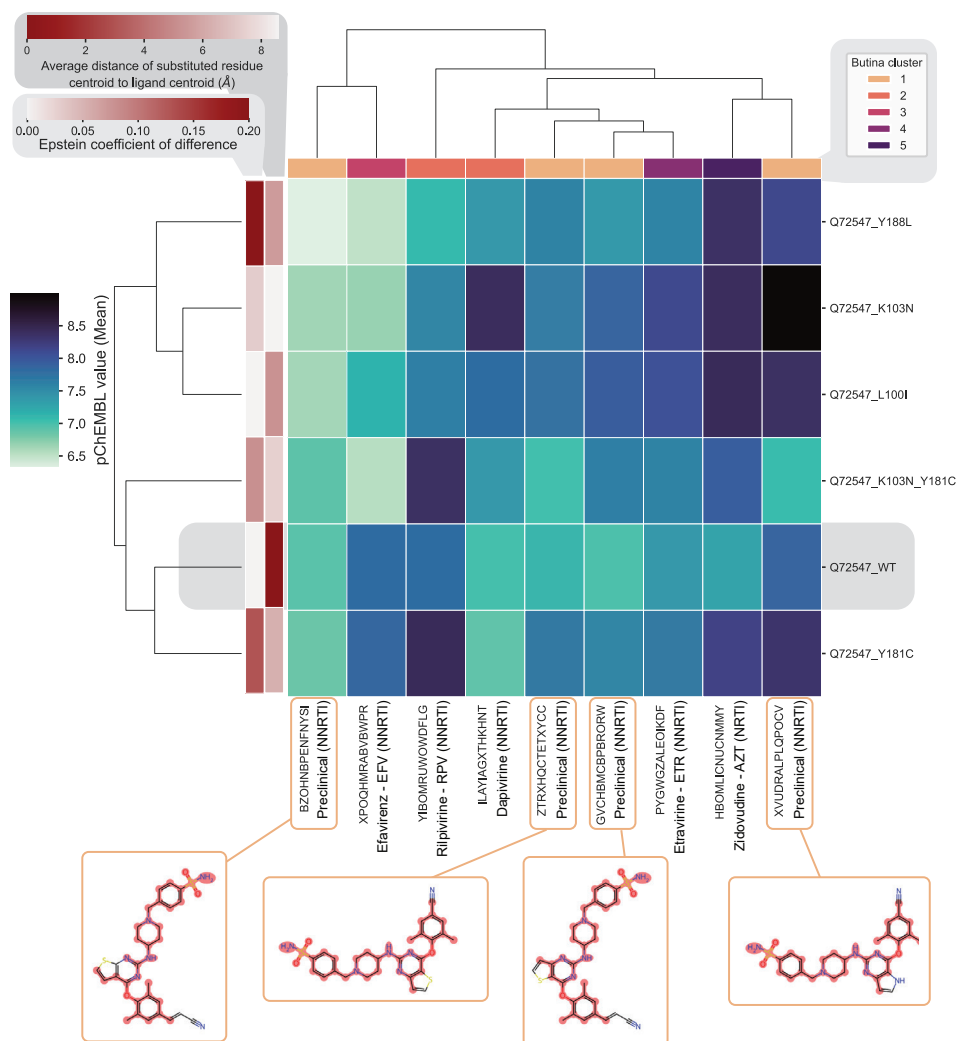
Supplementary Figure 4.4. Bioactivity distribution across variants on VEBD and fully dense common subsets. Displayed are a selection of four proteins with the biggest fully dense common subsets, i.e. BRAF - P15056 (a,e), BCR - P11274 (b,f), JAK2 - O60674 (c,g), and STING - Q86WV6 (d,h). The “MUTANT” variant label in (b,f) corresponds to undefined variants in the ambiguous ChEMBL-only group defined in the annotation feedback loop. The top row (a-d) is the distribution of the VEBD of compounds tested on the protein. The bottom row (e-h) is the distribution in the fully dense common subset of compounds tested on all annotated variants. Dashed vertical lines represent the average of the pchembl value variant distribution. This average and its corresponding standard deviation are also collected in the legend for each variant between brackets, followed by the size of the subset on which it was calculated and the percentage that this subset represents among the totality of the compounds tested on the protein.



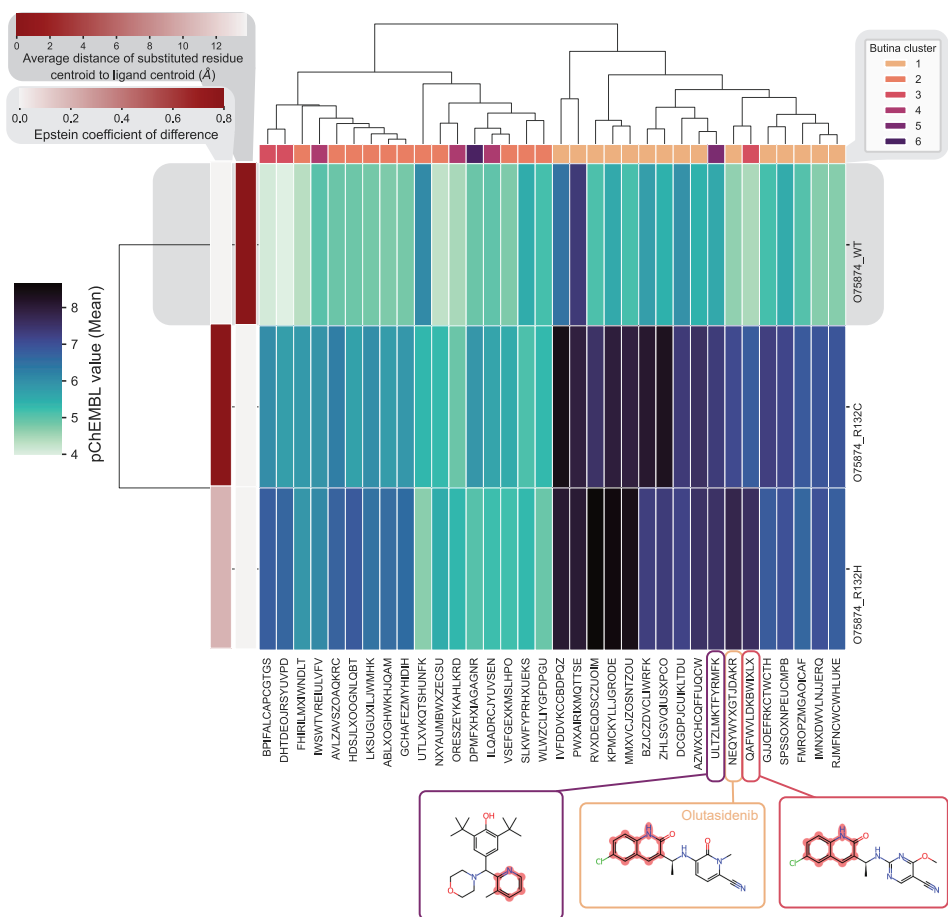
Supplementary Figure 4.5. Full-panel bioactivity analysis of the effect of EGFR (P00533) variants. The bioactivity analysis subset was computed from a common subset for compounds tested on at least three variants and variants with a compound coverage greater than 10%. Bioactivity is represented in the heatmap as the pchembl value of different compounds, on the x-axis, tested on several variants, on the y-axis. Compounds are annotated by their connectivity and preferred name, which is linked to their approval status. Compounds and variants were clustered by their overall bioactivity profile. Compounds are further represented by their corresponding Butina clusters upon clustering of the subset with a cut-off of 0.7. Variants are further represented by the distance from the substituted residue to the centroid of the ligand in the structure of the protein and by the Epstein coefficient of difference calculated for the amino acid substitution. In variants with multiple substitutions reported, the average distance and Epstein coefficient of difference are reported.



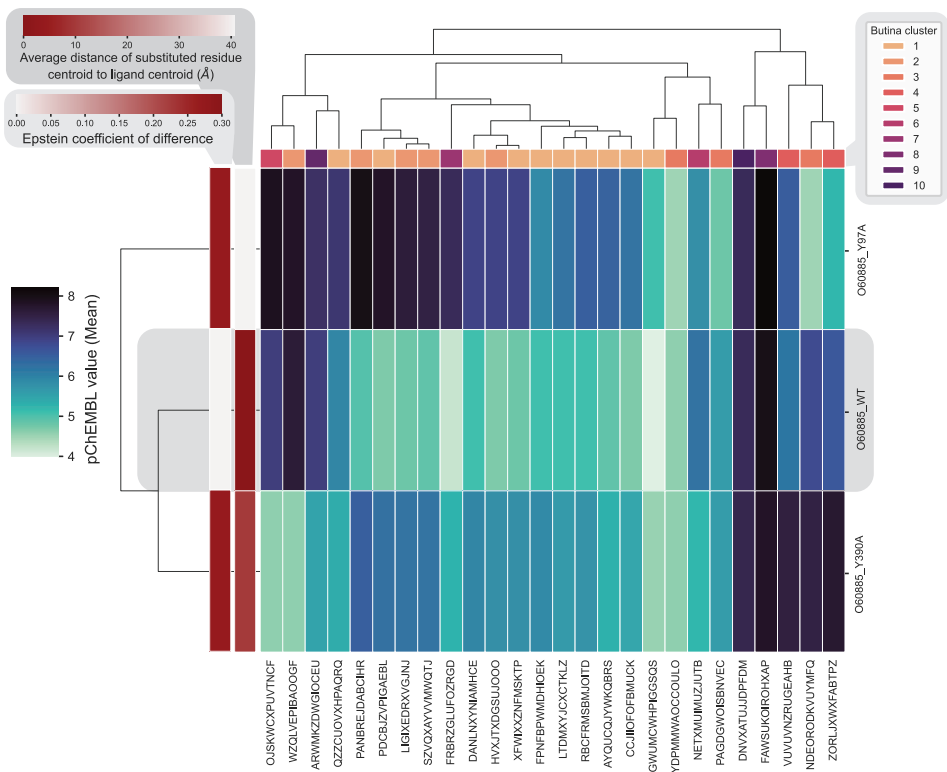
Supplementary Figure 4.6. EGFR (P00533) bioactivity analysis subset used to compute the bioactivity cluster map. Compounds identified by connectivity with their biggest ring systems are highlighted in red. Color coding represents the Butina cluster each compound was assigned to, using a cutoff of 0.7.



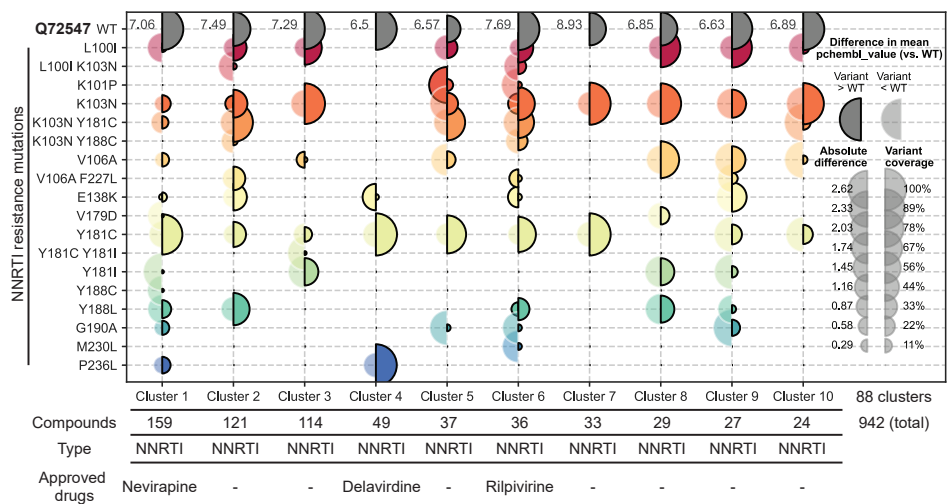
Supplementary Figure 4.7. Full-panel bioactivity analysis of the effect of HIV-1 RNaseH-RT (Q72574) variants. The bioactivity analysis subset was computed from a common subset for variants with a compound coverage greater than 3%. Bioactivity is represented in the heatmap as the pchembl value of different compounds, on the x-axis, tested on several variants, on the y-axis. Compounds are annotated by their connectivity and preferred name for approved drugs. Compounds are also divided between nucleoside (NRTI) and non-nucleoside reverse transcriptase inhibitors (NNRTI), which are orthosteric and allosteric inhibitors, respectively. Compounds and variants were clustered by their overall bioactivity profile. Compounds are further represented by their corresponding Butina clusters upon clustering of the subset with a cutoff of 0.5. Variants are further represented by the distance from the substituted residue to the centroid of the ligand in the structure of the protein and by the Epstein coefficient of difference calculated for the amino acid substitution. The distance to known NRTI and NNRTI resistance variants was calculated for co-crystallized ligands in the corresponding binding site. The structures of the four compounds in cluster 1 are displayed to exemplify the utility of this analysis to follow resistance variant selectivity in compounds with the same scaffold. In variants with multiple substitutions reported, the average distance and Epstein coefficient of difference are reported.



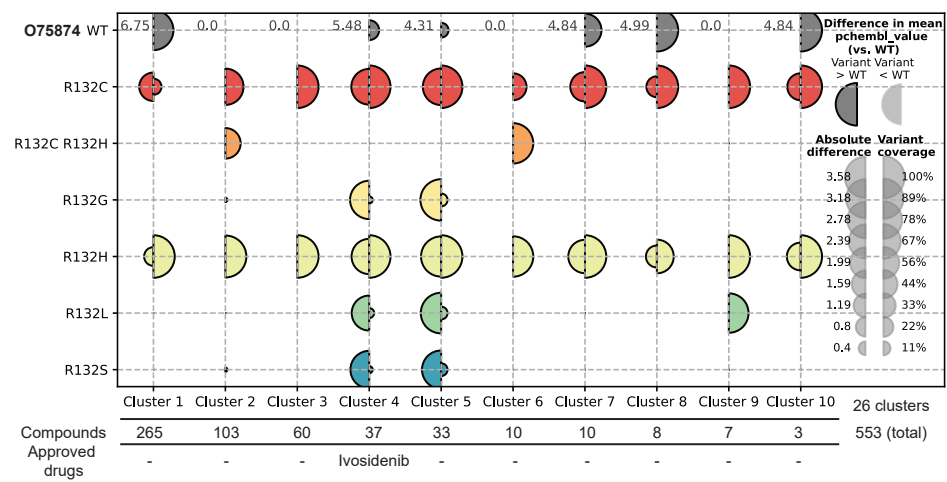
Supplementary Figure 4.8. Full-panel bioactivity analysis of the effect of oxidoreductase IDHC (O75874) variants. The bioactivity analysis subset was computed from a common subset for compounds tested on at least two variants and variants with a compound coverage greater than 20%. Bioactivity is represented in the heatmap as the pchembl value of different compounds, on the x-axis, tested on several variants, on the y-axis. Compounds are annotated by their connectivity. Compounds and variants were clustered by their overall bioactivity profile. Compounds are further represented by their corresponding Butina clusters upon clustering of the subset with a cutoff of 0.7. Variants are further represented by the distance from the substituted residue to the centroid of the ligand in the structure of the protein and by the Epstein coefficient of difference calculated for the amino acid substitution. The structures of three compounds from different clusters with similar bioactivity profiles across variants are highlighted to exemplify the applicability of this analysis to explore different scaffolds with similar selectivity profiles. Olutasidenib (cluster 1) is a clinical candidate IDHC inhibitor for patients with IDHC susceptible variants (R132X).



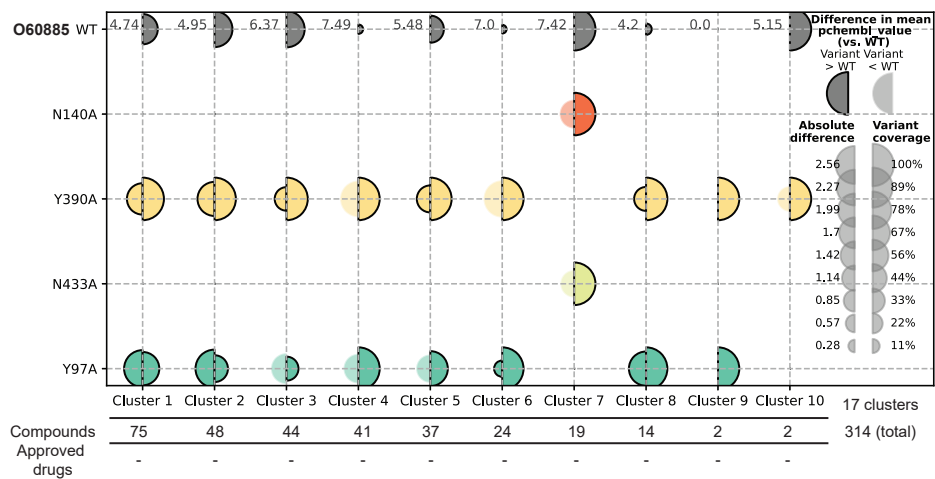
Supplementary Figure 4.9. Full-panel bioactivity analysis of the effect of epigenetic regulator BRD4 (O60885) variants. The bioactivity analysis subset was computed from a common subset for variants with a compound coverage greater than 2%. Bioactivity is represented in the heatmap as the pchembl value of different compounds, on the x-axis, tested on several variants, on the y-axis. Compounds are annotated by their connectivity. Compounds and variants were clustered by their overall bioactivity profile. Compounds are further represented by their corresponding Butina clusters upon clustering of the subset with a cutoff of 0.7. Variants are further represented by the distance from the substituted residue to the center of geometry (centroid) of the ligand in the structure of the protein and by the Epstein coefficient of difference calculated for the amino acid substitution. The variants displayed have no clinical significance in ClinVar and were likely tested in the context of alanine scanning strategies to elucidate the binding site of BRD4. Y97 is part of the bromodomain (BD) 1 domain, while Y390 is part of the BD2 domain. This analysis enables the identification of compounds with differential binding modes.



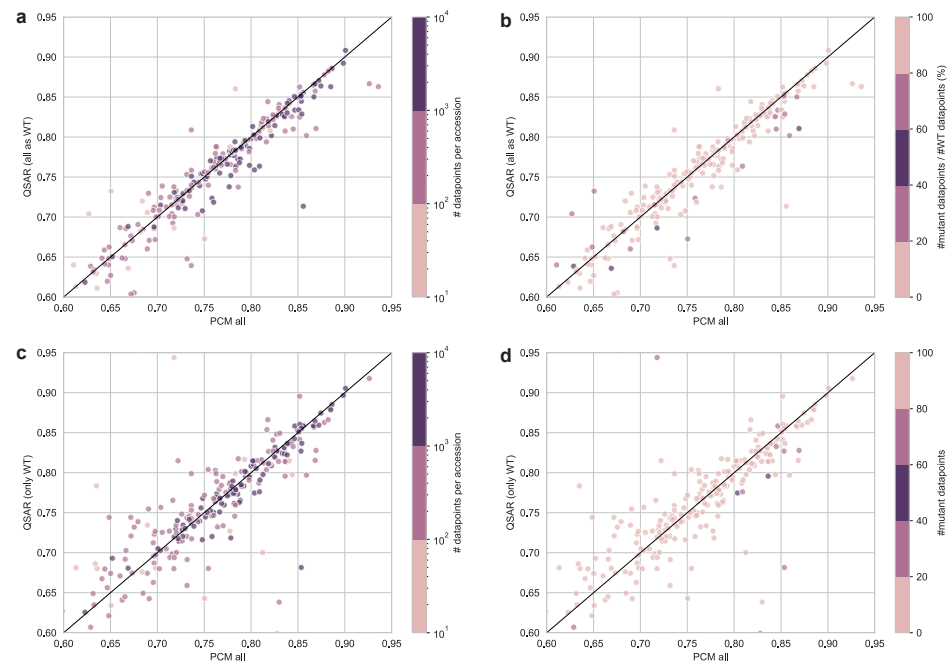
Supplementary Figure 4.10. HIV-1 RNaseH-RT (Q72574) bioactivity variability across variants compared to WT for compounds in the 10 most populated Butina Clusters upon clustering compounds tested on at least two variants with a clustering threshold of 0.5. Differences between the mean *pchembl_value* in WT, displayed in the first row as calculated for the compounds in each cluster, and the mean *pchembl_value* in each of the variants for the compounds in the same clusters. The left bubbles represent the result of subtracting the variant mean from the WT mean. The bubble size represents the absolute value of this difference (error). Opaque left bubbles represent a positive error (i.e. the mean calculated for the variant is higher than for WT), and translucent left bubbles represent a negative error (i.e. the mean calculated for the variant is lower than for WT). Right bubble sizes represent the variant coverage, in other words, the percentage of compounds in each cluster that was tested on a specific variant. All variants in this analysis are NNRTI resistance variants and the top 10 clusters also contain NNRTIs, some including approved drugs. This analysis facilitates the monitoring of NNRTI resistance variants across NNRTI scaffolds, represented by the different Butina clusters.



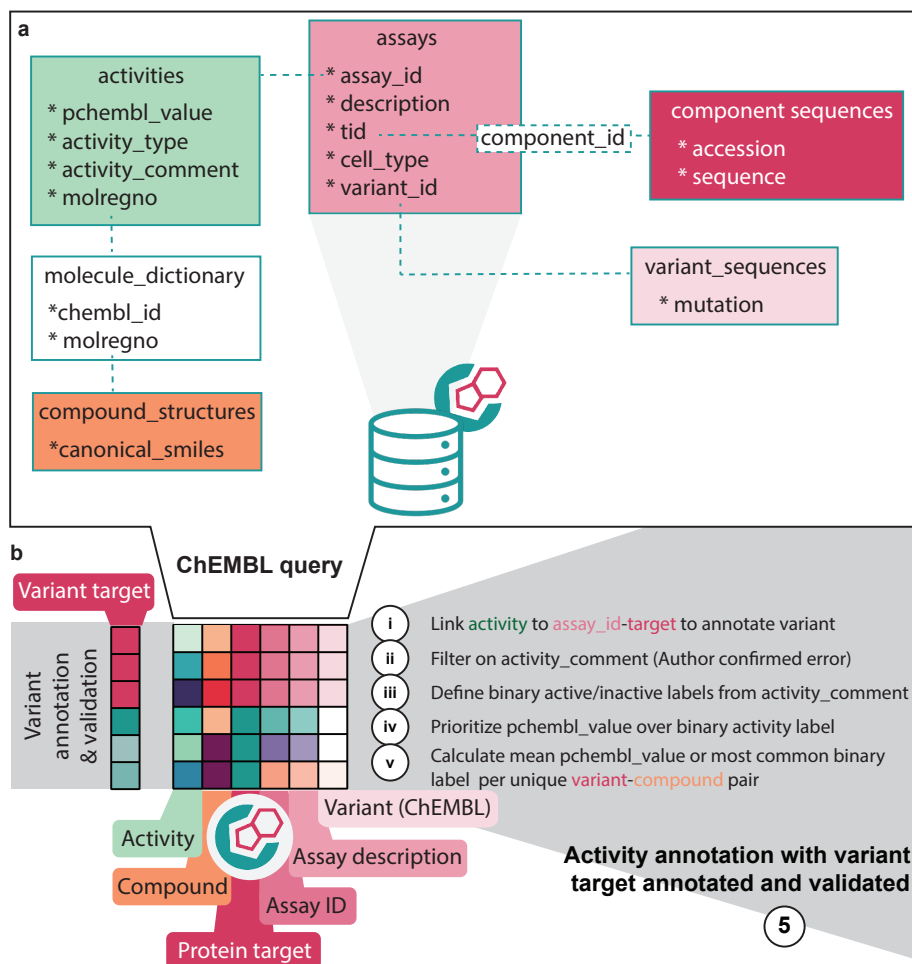
Supplementary Figure 4.11. Oxidoreductase IDHC (O75874) bioactivity variability across variants compared to WT for compounds in the 10 most populated Butina Clusters upon clustering compounds tested on at least two variants with a clustering threshold of 0.5. Differences between the mean *pchembl_value* in WT, displayed in the first row as calculated for the compounds in each cluster, and the mean *pchembl_value* in each of the variants for the compounds in the same clusters. The left bubbles represent the result of subtracting the variant mean from the WT mean. The bubble size represents the absolute value of this difference (error). Opaque left bubbles represent a positive error (i.e. the mean calculated for the variant is higher than for WT), and - if available - translucent left bubbles represent a negative error (i.e. the mean calculated for the variant is lower than for WT). Right bubble sizes represent the variant coverage, in other words, the percentage of compounds in each cluster that was tested on a specific variant. Ivosidenib (cluster 4) is an approved IDHC inhibitor for patients with IDHC susceptible variants (R132X).



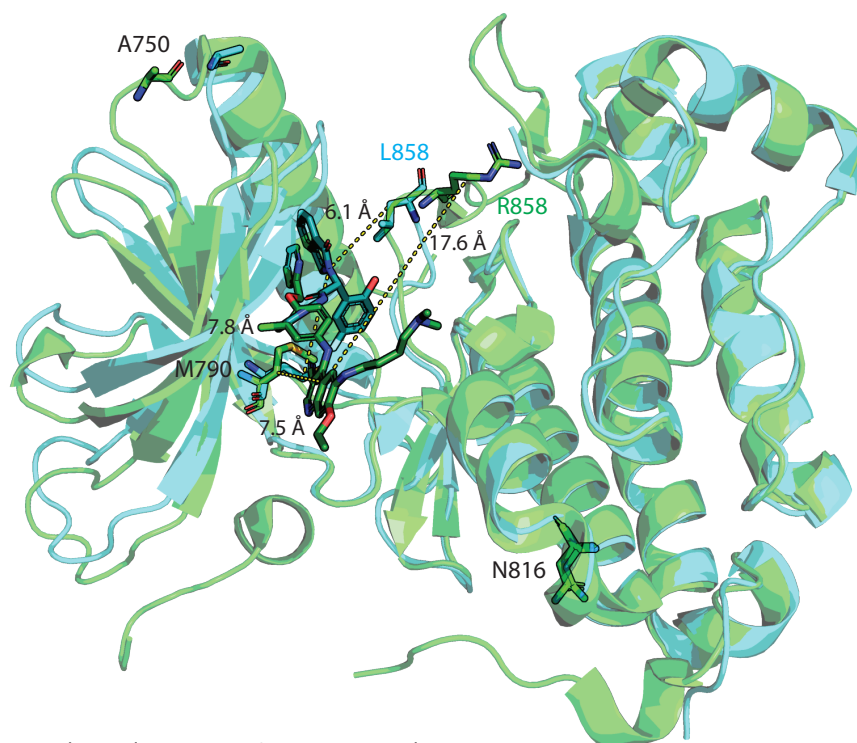
Supplementary Figure 4.12. Epigenetic regulator BRD4 (O60885) bioactivity variability across variants compared to WT for compounds in the 10 most populated Butina Clusters upon clustering compounds tested on at least two variants with a clustering threshold of 0.5. Differences between the mean *pchembl* value in WT, displayed in the first row as calculated for the compounds in each cluster, and the mean *pchembl* value in each of the variants for the compounds in the same clusters. The left bubbles represent the result of subtracting the variant mean from the WT mean. The bubble size represents the absolute value of this difference (error). Opaque left bubbles represent a positive error (i.e. the mean calculated for the variant is higher than for WT), and translucent left bubbles represent a negative error (i.e. the mean calculated for the variant is lower than for WT). Right bubble sizes represent the variant coverage, in other words, the percentage of compounds in each cluster that was tested on a specific variant. The variants displayed have no clinical significance in ClinVar and were likely tested in the context of alanine scanning strategies to elucidate the binding site of BRD4. Y97 and N40 are part of the BD1 domain, while Y390 and N433 are part of the BD2 domain. This analysis enables the identification of clusters of compounds with differential binding modes.



Supplementary Figure 4.13. Comparison of the Pearson correlation coefficients between PCM models built on the complete VEBD and QSAR models built either considering all bioactivity data points as having been obtained on WT proteins (**a, b**) or QSAR models built from bioactivity data points experimentally obtained on WT proteins only (**c, d**). Highlighted is the importance of the number of data points (a and c) and data balance of data points measured on variants and WTs (b and d) on the measured performance.



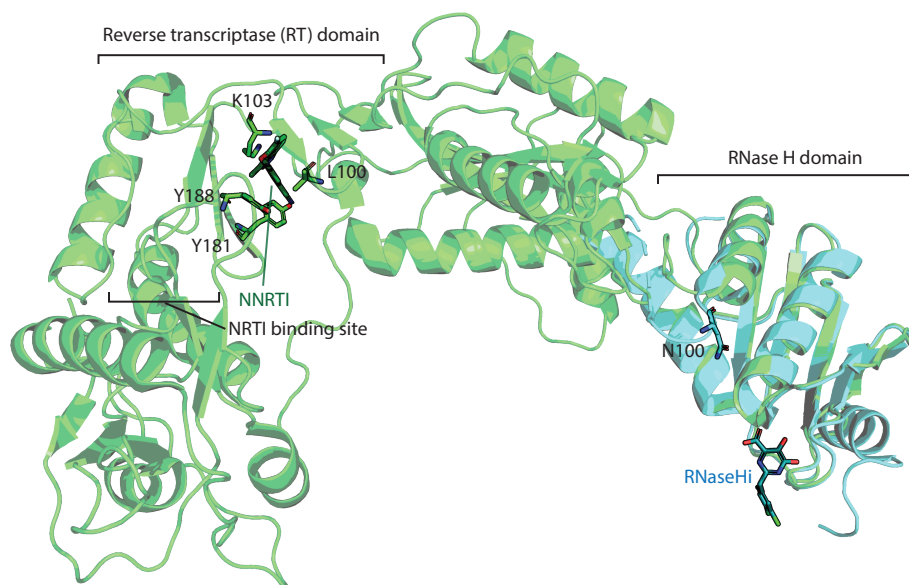
Supplementary Figure 4.14. ChEMBL query and activity variant annotation strategy. **a)** Bioactivity data is queried from ChEMBL via a SQL query that links six tables via primary and foreign keys. **b)** Upon variant annotation and validation, assay-target pairs are linked to bioactivity data for all available compounds as noted in Figure 4.1 step (5). Bioactivity data with negative activity comments is filtered out. Binary activity comments are then defined at threshold *pchembl_value* 6.5. Continuous data is however prioritized over binary labels. Unique bioactivity data is defined for each unique annotated variant-compound pair by computing the average *pchembl_value* or the most common binary label in the absence of continuous data.



3W2Q : EGFR kinase domain T790M/L858R mutant with HKI-272

5ZWJ : Crystal structure of EGFR 675-1022 T790M/C797S/V948R in complex with EAI045

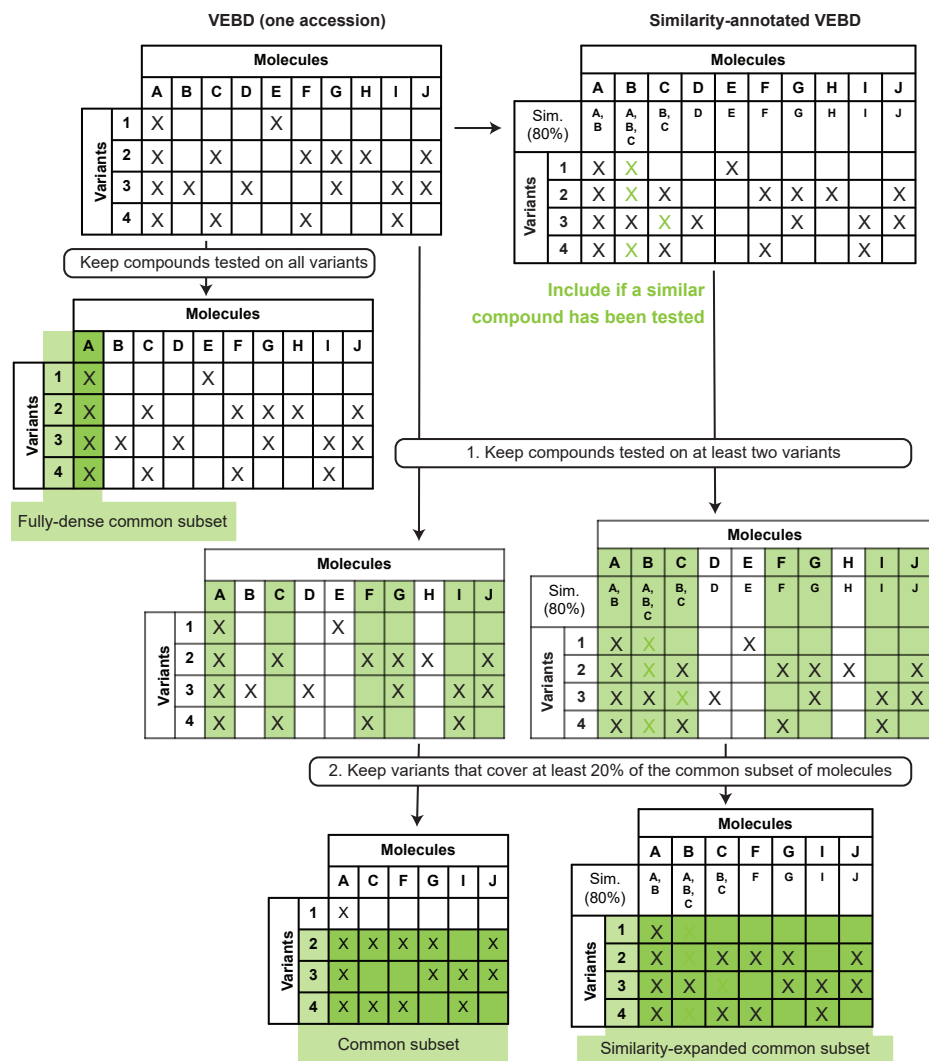
Supplementary Figure 4.15. Structural differences between two PDB structures of EGFR crystallized with ligands with distinct sizes and binding modes lead to different calculated distances to residues of interest. In green, PDB 3W2Q. The distance from the ligand's centroid to the centroid of M790 is 7.5 Å and to the centroid of R858 17.6 Å. In blue, PDB 5ZWJ. The distance from the ligand's centroid to the centroid of M790 is 7.8 Å and to the centroid of L858 6.1 Å.



2JLE : Novel indazole NNRTIs created using molecular template hybridization based on crystallographic overlays

3HYF : Crystal structure of HIV-1 RNase H p15 with engineered E. coli loop and active site inhibitor

Supplementary Figure 4.16. Structural differences of the two PDB structures with co-crystallized ligands linked to UniProt code Q72547 (HIV-1 RNaseH-RT). PDB 2JLE (green) contains both the reverse transcriptase (RT) domain – with a non-nucleoside RT inhibitor (NNRTI) bound – and the RNase H domain. PDB 3HYF contains only the RNase H domain – with an RNase H inhibitor (RNaseHi) bound.



Supplementary Figure 4.17. Common subset design strategy. When possible, fully dense common subsets were computed from the VEBD by keeping the compounds tested on all variants for the accession of interest (X in the data matrix represents that there is bioactivity data for a particular molecule-variant pair). Otherwise, non-fully dense common subsets (simply referred to as common subsets) were computed in two steps. Firstly, by keeping compounds tested on at least a threshold number of variants (by default two). Secondly, by keeping variants that cover at least a certain percentage (by default 20%) of the pre-selected compounds for the common subset. Similarity-expanded common subsets were computed similarly to common subsets but starting from a similarity-annotated VEBD, where each molecule was linked to other molecules in the dataset with Tanimoto similarity above a certain threshold (by default 80%). Steps 1 and 2 to generate the similarity-expanded common subset were the same as for the normal common subset but considering to calculate the statistics similar molecules tested on a different variant. For example, molecule B has only been tested on variant 3. However, B is similar to molecule A, which has been tested in all variants. Therefore, for steps 1 and 2 B is considered to have been tested in all variants, as represented by the green X in the data matrix.



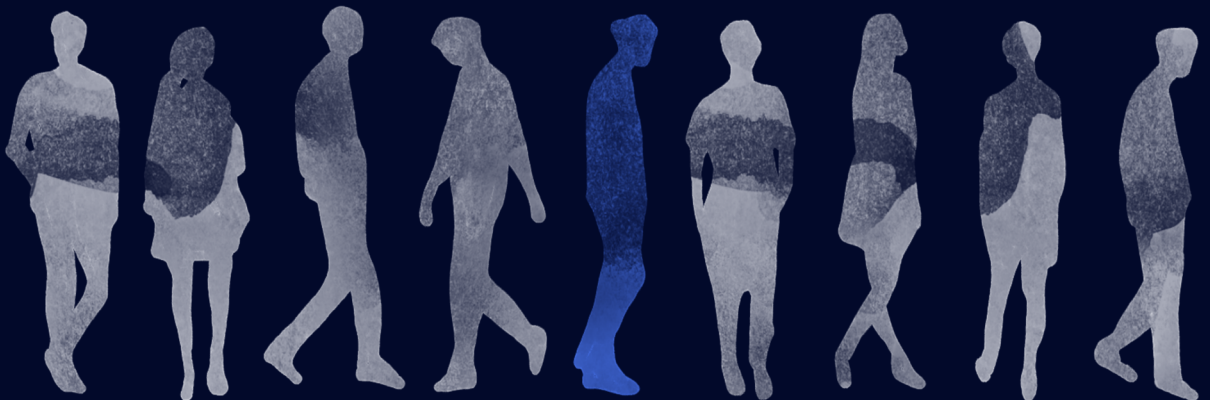
Chapter 5


Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors

Brandon J. Bongers[†], Marina Gorostiola González[†], Xuesong Wang, Herman W.T. van Vlijmen, Willem Jaspers, Hugo Gutiérrez-de-Terán, Kai Ye, Adriaan P. IJzerman, Laura H. Heitman, Gerard J.P. van Westen

Adapted from: *Scientific Reports* **12**, 21534 (2022)

[†]These authors contributed equally





G Protein-coupled receptors (GPCRs) are the most frequently exploited drug target family, moreover, they are often found mutated in cancer. Here we used a dataset of mutations found in patient samples derived from the Genomic Data Commons and compared it to the natural human variance as exemplified by data from the 1000 genomes project. We explored cancer-related mutation patterns in all GPCR classes combined and individually. While the location of the mutations across the protein domains did not differ significantly in the two datasets, a mutation enrichment in cancer patients was observed among class-specific conserved motifs in GPCRs such as the Class A “DRY” motif. A Two-Entropy Analysis confirmed the correlation between residue conservation and cancer-related mutation frequency. We subsequently created a ranking of high-scoring GPCRs, using a multi-objective approach (Pareto Front Ranking). Our approach was confirmed by the re-discovery of established cancer targets such as the LPA and mGlu receptor families, but also discovered novel GPCRs which had not been linked to cancer before such as the P2Y Receptor 10 (P2RY10). Overall, this study presents a list of GPCRs that are amenable to experimental follow-up to elucidate their role in cancer.



Introduction

Cancer is the second leading cause of death globally¹. Research on this multifactorial disease has expanded our knowledge significantly over the last two decades², leading to public databases containing patient-derived data³. Cancer is typically the result of compounding mutations that transform healthy cells into malignant ones⁴. Previous work involving large-scale mutational analysis picked up G Protein-coupled receptors (GPCRs) as the second most mutated class of proteins in the context of cancer after kinases⁵. Cancer cells are driven to proliferate and avoid the immune system. GPCRs have multiple functions in this process from increased growth (early stage) all the way to metastasis (late stage)⁶. Thus, any anomalies in GPCR functioning might be related to cancer growth. Another interesting property of GPCRs is that they are the most common drug target family with around 35% of drugs acting through a GPCR⁷, providing a diverse set of molecular tools to potentially combat cancer.

GPCRs consist of seven highly conserved transmembrane (TM) domains, typically harboring the ligand binding pocket for natural ligands, e.g. endogenous hormones or neurotransmitters. Human GPCRs are divided into several classes based on sequence similarity: A, B, C, D, F, and T (as used on GPCRdb)^{8,9}. The TM domains are connected via extra- and intracellular loops (ECL; ICL) displaying a lower degree of conservation. Most GPCRs also have an eighth TM domain that is connected by intracellular loop 4. The extracellular loops are known to also be involved in ligand recognition and activation, whereas the intracellular part of the receptor is linked to G protein recognition and activation. Finally, GPCRs contain an N- and C-terminus which are also relatively little conserved between and within classes^{9,10}.

In previous work, knock-down studies have been performed on several proteins to identify their role in the context of cancer, typically embarked upon after prior identification of the protein's role in cancer^{11,12}. One of the main reasons these *in vivo* studies are done is to identify whether a mutation is either a driver, providing a selective growth advantage and promoting cancer development, or a passenger mutation occurring co-incidentally. Moreover, these studies provide insight into whether a driver mutation is located on either an oncogene or a tumor suppressor gene¹³. The prioritization of point mutations for experimental characterization, when the role of the protein in cancer is still unknown, could accelerate the discovery of relevant oncogenic alterations.

Here, we focused on GPCRs in the context of cancer by using patient-derived data sets and specifically looked at trends and mutational patterns in this protein family. We performed a deeper investigation into several “motifs”, parts of the GPCR sequence that are conserved that contribute most to the stability and function of the GPCR^{14–19}. Class-specific motifs and several broad differences between classes were also considered. Moreover, we provided a list of GPCRs with known small molecule ligands (including approved drugs), ranked by interest for follow-up using multi-objective ranking. They were ranked on mutational count, mutations in regions of interest, availability of in-house expertise, and ability to perform virtual screening (by QSAR). Finally, we exemplified our findings in a more in-depth analysis of C-C chemokine receptor type 5

(CCR5) to show the feasibility of our approach.

Results

Overview of datasets

Missense mutations in all GPCR human classes were collected from the GDC and 1000 Genomes datasets (Table 5.1). The GDC dataset contained more subjects than the 1000 Genomes set, but both were in the same order of magnitude based on missense mutation count. However, as fewer unique missense mutations were found in natural variance, most cancer-related mutations had a small frequency. To account for differences in the datasets’ number of data points, the mutation ratio per dataset was used instead of absolute mutation frequency in the subsequent comparative analyses (see Methods).

Table 5.1. Overview of the composition of the GDC and 1000 Genomes datasets.

		GDC dataset (v 22.0)			1000 Genomes dataset (2020)		
Total subjects		10,179			3,202		
Total cancer types		53			n/a		
Missense mutations		2,129,235			2,943,276		
Class	Missense mutations in GPCRs	Total	Unique	Unique receptors	Total	Unique	Unique receptors
	All class	45,902	40,431	394	43,884	24,237	396
	Class A	26,342	23,122	284	20,528	11,454	286
	Class B	10,745	9,588	47	15,439	8,814	47
	Class B1	1,499	1,342	15	2,174	1,283	15
	Class B2	9,246	8,246	32	13,265	7,531	32
	Class C	5,592	4,842	22	5,273	2,644	22
	Class F	1,155	1,039	11	487	368	11
	Class T	1,675	1,494	24	1,639	719	24
	Other GPCRs	393	346	6	518	238	6

Two-Entropy Analysis

A two-entropy analysis (TEA) was performed on our dataset as was done previously¹⁹. This method was chosen primarily to evaluate residue conservation across GPCRs and within GPCR subfamilies. Secondly, we tried to leverage its ability to define residue functional characterization. Of note, we performed this analysis not only for Class A GPCRs but for all classes defined in GPCRdb; together and independently. Key to the TEA approach is that for each alignment position the Shannon entropy, which measures the level of conservation of amino acid residues at a certain position in a multiple

sequence alignment, is calculated both within a GPCR subfamily and within all GPCRs. Therefore, the combination of these can provide a measure for the position function. Multiple interesting groups were identified, such as residues relevant for receptor function/activation (type Q3). Type Q3 are positions with a low Shannon entropy both within GPCR subfamilies and for the entire GPCR superfamily, this high conservation is linked to involvement in GPCR-conserved working mechanisms. Separating the graph into quadrants (Q1-4), type Q3 residues are represented in the bottom left quadrant in **Figure 5.1**. A second group is residues relevant for ligand recognition (type Q2), made up of residues that are conserved within subfamilies, but not within the GPCR superfamily. Hence, these are associated with ligand recognition that is specific and conserved within a given subfamily. Type Q2 residues, represented in the top left quadrant were less noticeable in the all-class TEA (**Figure 5.1a**) since the inclusion of a larger number of subfamilies led to an increase in the overall entropy. However, it was more obvious in Classes A-C (**Figure 5.1b-d**). Finally, in the top right quadrant of the TEA plot a third group of residues, Q1, is represented that are conserved neither among all GPCRs nor GPCR subfamilies. These are more likely to have only a small implication in receptor functions.

Residue conservation was linked to absolute mutation count frequency per position with Ballesteros-Weinstein number in cancer patients (color coding in **Figure 5.1** and **Supplementary Figure 5.1**). Residues with a high mutation frequency were defined as those above the 90th percentile in the distribution of mutation counts by position. Conversely, residues with a low mutation frequency were defined as those under the 10th percentile. Absolute mutation count was (anti)correlated with entropy (**Figure 5.1**). We observed a trend where more conserved type Q3 residues (bottom left quadrant, low entropy) had a higher mutation rate in cancer compared to the less conserved Q1 residues (top right quadrant, high entropy). We illustrated this with the mean \pm SD entropy overall and across families (**Figure 5.1** and **Supplementary Table 5.1**). In the all-class TEA (**Figure 5.1a**), the low mutation range had mean entropy values of 0.45 ± 0.38 and 0.41 ± 0.27 (Shannon and Average entropy across families, respectively). The high mutation range had lower mean entropy values of 0.30 ± 0.10 and 0.28 ± 0.13 , respectively. On the contrary, this trend was not observed in natural variance data from the 1000 Genomes dataset (**Supplementary Figure 5.2**). There, mean entropy values for the low mutation range were 0.40 ± 0.30 and 0.33 ± 0.23 , respectively; and 0.34 ± 0.08 and 0.39 ± 0.12 , respectively, for the high mutation range. We observed an average downward shift in entropy values for highly mutated positions per subfamily (not in the overall Shannon entropy) and an upward shift for less frequently mutated positions. Combined this showed a pressure in the GDC data for mutations in subfamily-conserved positions at the expense of mutations in non-conserved positions. This trend was maintained across classes, although less marked for Classes B and C, and supported by the fact that from the type Q3 residues highlighted in **Figure 5.1a**, higher mutation frequencies were associated with the most conserved positions in TM domains 3, 4, and 7 (i.e. 3x50, 4x50, and 7x50). These are part of the “DRY” (TM3), “GWGxP” (TM4), and “NPxxY” (TM7) conserved GPCR functional motifs. The high amount of mutations in residues of these and other motifs was further investigated in the section *Mutation patterns within functionally conserved motifs*. Overall, cancer mutation frequency was correlated

with individual residue conservation, hence we investigated groups of residues as defined by GPCR domains to further explore cancer mutation patterns.

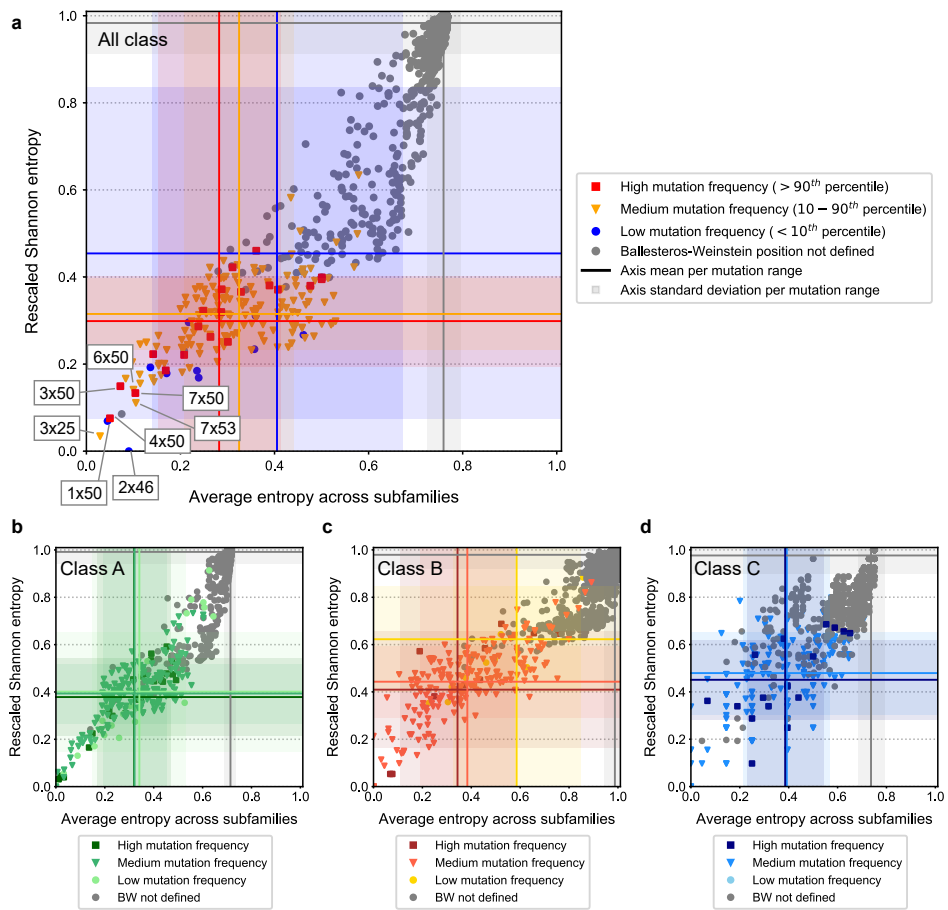


Figure 5.1. Shannon entropy across GPCR subfamilies versus Shannon global Entropy correlated to cancer-related mutations. A two-entropy analysis plot for all GPCRs with aligned positions. The average entropy across subfamilies (as defined by GPCRdb), i.e. conserved within a subfamily is on the x-axis, and the Shannon entropy is on the y-axis. **a)** Analysis for all GPCR classes combined. Residues are colored by the frequency of mutations found in the GDC dataset, with blue being low ($< 10^{\text{th}}$ percentile), orange medium ($10-90^{\text{th}}$ percentiles), and red high ($> 90^{\text{th}}$ percentile). Residues with no defined Ballesteros-Weinstein (BW) generic numbers are colored grey. Blue, orange, red, and grey lines represent the mean entropy values for each axis per mutation range (high, medium, low, and non-defined Ballesteros-Weinstein, respectively). Blue, orange, red, and grey shadows represent the standard deviation to the mean entropy values for each axis per mutation range (high, medium, low, and non-defined Ballesteros-Weinstein, respectively). **b)** Analysis for Class A GPCRs. **c)** Analysis for Class B GPCRs. **d)** Analysis for Class C GPCRs. The coloring scheme for panels (b)-(d) is equivalent to that of panel (a).

Mutation rates over GPCR structural domains

We hypothesized that mutations associated with altered function in the context of cancer would occur more frequently in domains with higher conservation (i.e. TM domains) where positive selective pressure would favor them. Conversely, we expected mutations to be distributed more randomly over the sequence among the 1000 Genomes set and to be underrepresented in the conserved TM domains. However, the distribution in both sets was quite similar (**Figure 5.2a,b**). Most mutations were in the N-terminus (~ 25% of the total across all classes), followed by the C-terminus (~ 15% of the total across all classes), which are on average the longest domains. The TM domains were next in mutation count, followed by ICL3 and ECL2. Finally, the remaining loops had the lowest

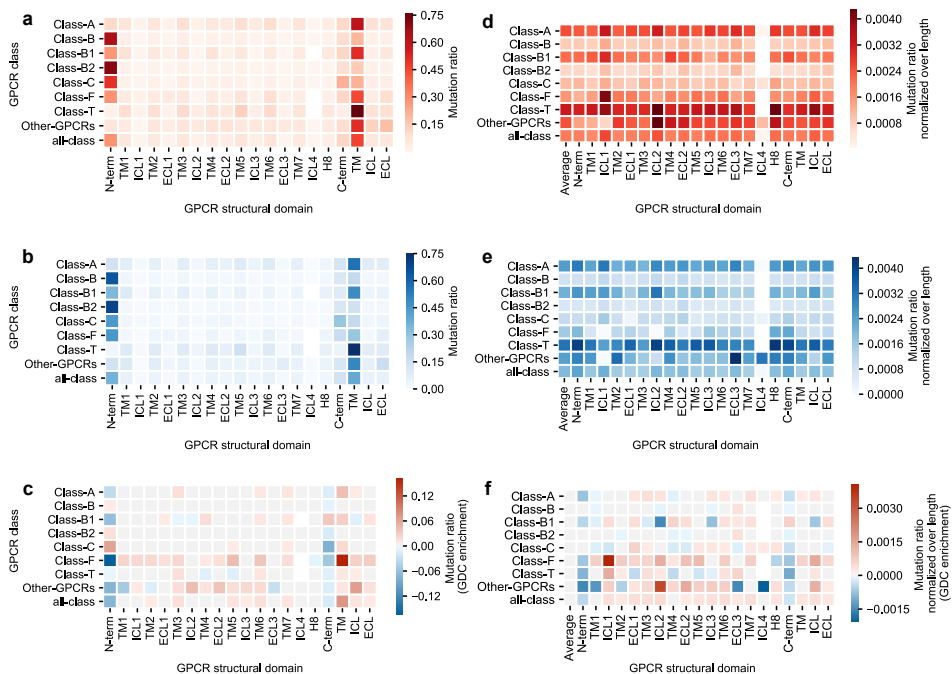


Figure 5.2. Distribution of mutation frequencies per GPCR structural domain. **a)** Mutation ratio found in each structural domain in the GDC dataset for GPCRs in all classes combined and independently. **b)** Mutation ratio found in each structural domain in the 1000 Genomes dataset for GPCRs in all classes combined and independently. **c)** Mutation ratio enrichment in the GDC dataset over the 1000 Genomes dataset. **d)** Mutation ratio normalized over average domain length found in each structural domain in the GDC dataset for GPCRs in all classes combined and independently. **e)** Mutation ratio normalized over average domain length found in each structural domain in the 1000 Genomes dataset for GPCRs in all classes combined and independently. **f)** Length-normalized mutation ratio enrichment in the GDC dataset over the 1000 Genomes dataset. “TM”, “ICL” and “ECL” represent the (normalized) mutation ratios in aggregated domains. In panels (d-f), “Average” represents the average ratio considering a domain as the whole protein. In panels (a) and (d), a darker shade of red represents a higher (normalized) mutation ratio in the GDC dataset. In panels (b) and (e), a darker shade of blue represents a higher (normalized) mutation ratio in the 1000 Genomes dataset. In panels (c) and (f), a darker shade of red represents a higher (normalized) mutation ratio enrichment towards the GDC dataset, while a darker shade of blue represents a higher (normalized) mutation ratio enrichment towards the 1000 Genomes dataset.

amount of mutations. Around 40% of the mutations were found in the aggregated 7TM domains across all classes. No major differences between GDC and 1000 Genomes were observed when we compared mutation ratios (**Figure 5.2c**), although there was enrichment observed in cancer-related mutations in the TM regions, as opposed to the N-terminus and C-terminus. To remove the bias caused by differences in the average length of the different domains, we calculated the mutation ratio normalized over average domain length.

After normalization mutation ratios were more consistent over domains for every class in both the GDC and 1000 Genomes datasets (**Figure 5.2d,e**). This correction was crucial to compare classes as observed in the N-terminus: Class B2 had a higher mutation ratio than Class T (**Figure 5.2a**) but after normalization (**Figure 5.2d**) a hotspot appeared in Class T. In general, all domains were slightly enriched in the GDC data except N-terminus and C-terminus (**Figure 5.2f**). Of note were the differences observed between classes. For example, ICL2 was enriched across all classes (except B1) and highly enriched in Class Other GPCRs. Conversely, Class B1 showed a cancer enrichment in the C-terminus that was not observed in any other class. Zooming into specific domains showed mutational hotspots in different classes that can result in a therapeutic advantage. We concluded that some domains may be more amenable to mutation in the context of cancer. To further investigate these incipient mutation patterns in protein domains, we proceeded to the analysis of previously identified motifs that have a conserved function in GPCRs and that were also highlighted in our two-entropy analysis.

Mutation patterns within functionally conserved motifs

Several highly conserved motifs relevant to GPCR function are known in different classes. They are “DRY”, “CWxP”, and “NPxxY” in Class A; “GWGxP”, “RE”, and “PxxG” in Class B; “HETx” in Class B2; and the “R/K” mutational hotspot in Class F (**Table 5.2**). Point mutations in these motifs usually cause a disruption or change in function^{14–18}. We therefore hypothesized that mutational pressure in these motifs would occur in cancer to disturb normal GPCR function. For direct comparison between motifs, we calculated a mutation ratio normalized over motif length. As a reference, the average normalized mutation rates obtained over the whole GDC and 1000 Genomes datasets are shown.

In each motif investigated the mutation rate in cancer patients was higher than the natural variation in that motif (**Figure 5.3a**). Moreover, in the GDC dataset (red bars) “DRY”, “RE”, and “R/K” motifs were enriched in cancer compared to the average mutation ratio, whereas for the 1000 Genomes (blue bars) there was a clear reduction for all motifs. The GDC enrichment is shown for the most populated classes (**Figure 5.3b**) and for all classes (**Supplementary Figure 5.3**). Class A-specific domains (i.e. “DRY”, “CWxP”, and “NPxxY”) were enriched in Class A. Class B-specific domains (i.e. “HETx”, “RE”, “GWGxP”, and “PxxG”) were enriched mostly in Class B but also in Class A. Interestingly, the enrichment pattern was very different in Class B1 and B2. Of note, the B2-specific motif “HETx” was more highly enriched for cancer mutations in Class B1. Finally, the “R/K” motif was slightly enriched in all classes except Class

B1, but highly enriched in Class F. Class C showed minimal cancer enrichment across all motifs. An absolute count of the mutations found in the motifs in both sets is shown in **Supplementary Figure 5.4**. We concluded that conserved motifs are increasingly mutated in cancer samples over natural variance, confirming their essential role and conservation.

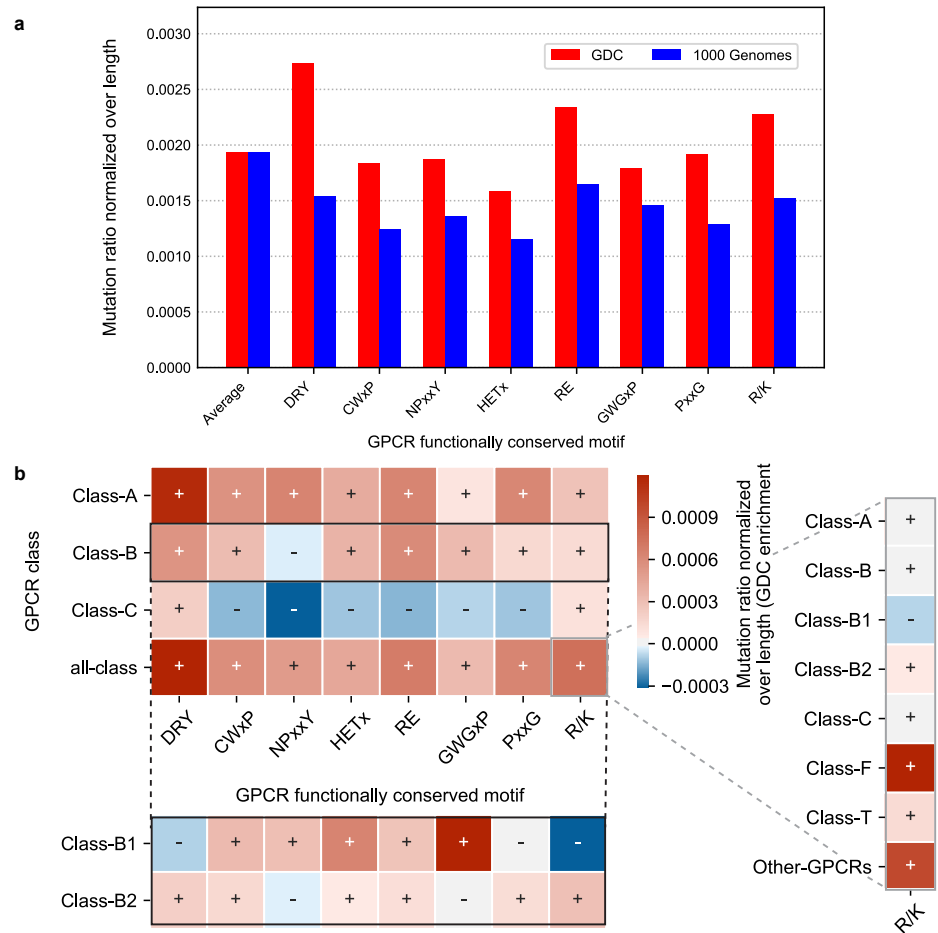


Figure 5.3. Distribution of mutation frequencies per functionally conserved motif. Mutation ratios normalized over motif length in GDC and 1000 Genomes datasets of conserved motifs found in different GPCR classes. Motifs analyzed are “DRY”, “CWxP”, and “NPxxY” (Class A); “HETx”, “RE”, “GWGxP”, and “PxxG” (Class B); and “R/K” (Class F). “Average” represents the average ratio considering the whole protein length. **a)** Analysis of all GPCR classes combined. Red bars show the normalized mutation ratio in the GDC dataset, while blue bars show the ratio of the 1000 Genomes dataset. **b)** Length-normalized mutation ratio enrichment in the GDC dataset over the 1000 Genomes dataset in all classes combined and independently. The most populated classes are included in the main heatmap for visualization purposes. An extension of Class B is provided by breaking the heatmap row into Class B1 and Class B2. An extension of the all-class enrichment of the “R/K” motif is also provided for all classes independently. A darker shade of red represents a higher enrichment over the GDC dataset, and a darker shade of blue represents a higher enrichment over the 1000 Genomes dataset. The intensity of shades can be compared within the main heatmap (Classes A–C and all-class), and across each extension separately.

To gain further insights we selected the most mutated individual positions in the GDC dataset corrected for mutation frequency in natural variance. We represented this for all classes together and for Class A-C in **Figure 5.4**. A count overview of unique GPCR cancer mutations is provided in **Supplementary Figure 5.5**, and an overview of the substitutions found in all of the mutations is in **Supplementary Figure 5.6**. Most of the mutations analyzed derived from Class A (**Figure 5.4**), hence proving the relevance of a per-class analysis. Overall and in Class A the most frequently mutated residue was 3x50 (BW numbering), part of the “DRY” motif. This was followed by 7x50 (“NPxxY” motif) in Class A. In Class B, 4x51 and 4x53 (“GWGxP” motif) and 6x45 (“PxxG” motif) were among the top 10. Interestingly, in Class A and Class C, several residues in H8 were highly mutated (i.e. 8x49, 8x51, and 8x53), and in Class C we found an ICL1 residue (12x48) in the top 10. Given the enrichment in cancer found in functionally conserved motifs (**Figure 5.3**), we suggest that the residues found among the most frequently mutated should be further functionally characterized since we hypothesize that they are relevant to receptor function.

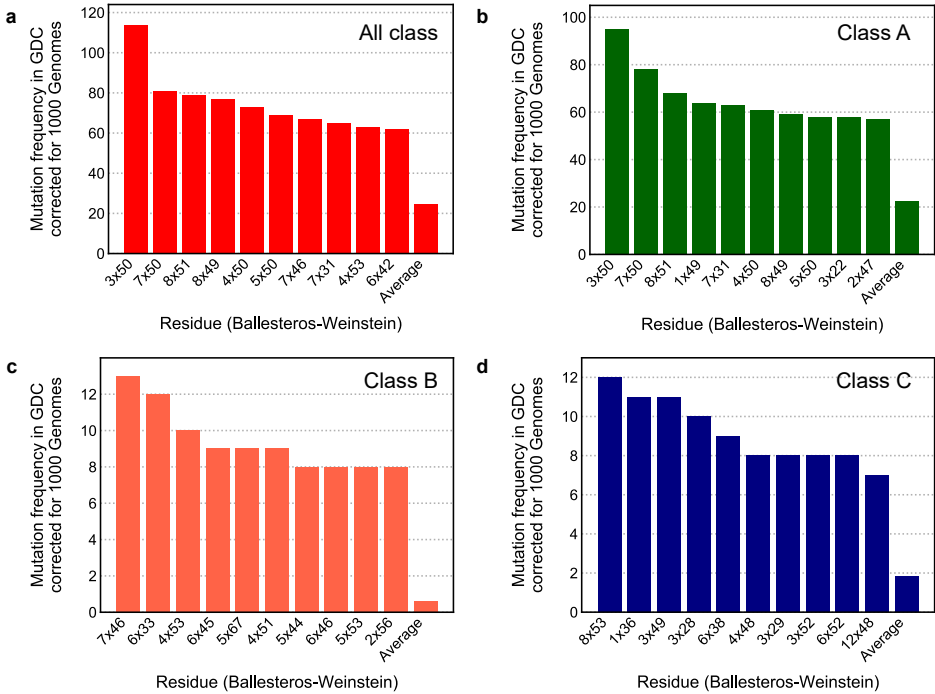


Figure 5.4. Most frequently mutated residues in GDC corrected for natural variance. The 10 positions with the highest mutation frequency in GPCRs in the GDC dataset corrected for the mutation frequency in the 1000 Genomes dataset. **a)** Analysis of all GPCR classes combined. **b)** Analysis of Class A GPCRs. **c)** Analysis of Class B GPCRs. **d)** Analysis of Class C GPCRs. The residue location in Ballesteros-Weinstein notation is shown on the x-axis, while on the y-axis the corrected mutation frequency of that residue is given. “Average” is the average mutation frequency per residue over all the data.

Ranking GPCRs for follow-up

Having confirmed that patterns can be identified in GPCR mutations in cancer, we ranked GPCRs for experimental follow-up. Pareto sorting was performed as a recommendation system to identify GPCRs with a suggested high impact in cancer biology amenable to small molecule intervention and follow-up. Pareto sorting is based on multiple (not always correlating) properties. The Pareto analysis was done in two ways. Firstly, we implemented Pareto ranking solely based on somatic mutation data. The four selected properties for Pareto ranking were: Mutations in highly conserved TEA Q3 residues in GDC (maximized) and 1000 Genomes (minimized), and mutation rate in TM domains in GDC (maximized) and in 1000 Genomes (minimized). Additionally, we introduced two practical objectives to bias the mutation-based recommendation towards a set of in-house objectives representing the feasibility of *in vitro* or *in silico* follow-up. The feasibility of small molecule intervention was assessed by training a machine-learning model (random forest) for each GPCR in our data set using bioactivity data from ChEMBL 27, with circular fingerprints as molecular descriptors. The two practical objectives introduced were the average R^2 of ChEMBL QSAR prediction models (maximized), and the in-house availability of proteins for experiments (maximized). The order of the properties determined the priority during the Pareto sorting.

The first front in the Pareto optimization is considered “dominating”, which means that this set of GPCRs scored better in the combined properties than any other set. For the remaining data points a second front can be calculated, with GPCRs that scored worse than those in the first front but better than the rest of the solutions. Therefore, we used the first and second fronts for a subsequent ranking based on crowding distances between the receptors (**Figures 5.5a** and **5.5b**, respectively). Crowding distances are a measure of how dense the environment is; denser environments mean more balance in the objectives and thus more interesting GPCRs. As the crowding distance can go up to near infinite, we used a cut-off at a value of 10.

Twenty-four GPCRs from the best scoring (first) front translated to the GPCRs with the most desirable scores in the combined objectives of the Pareto optimization including “practical objectives” (**Figure 5.5a**). The 13 receptors identified in the first front using exclusively mutation-derived objectives were contained in their totality in the first Pareto front with all objectives and, similarly, the 12 receptors in the mutation-only second front were entirely distributed between the first and second fronts (**Figure 5.5**). GPCRs previously linked to cancer showed up in the first front alongside others that have not been thoroughly investigated yet. This was confirmed in a similar ranking for GPCR subfamilies (**Supplementary Figure 5.7**). The second Pareto front (**Figure 5.5b**), contained 28 GPCRs. Hence, our recommendation system produced Pareto fronts that represented a list of potential candidates for follow-up experimental research. From the receptors of our first Pareto front, we selected one for which there was in-house expertise, CCR5, as a case study for further investigation using a crystal structure-based analysis to characterize the potential effects of the retrieved mutations in receptor function and/or ligand binding.

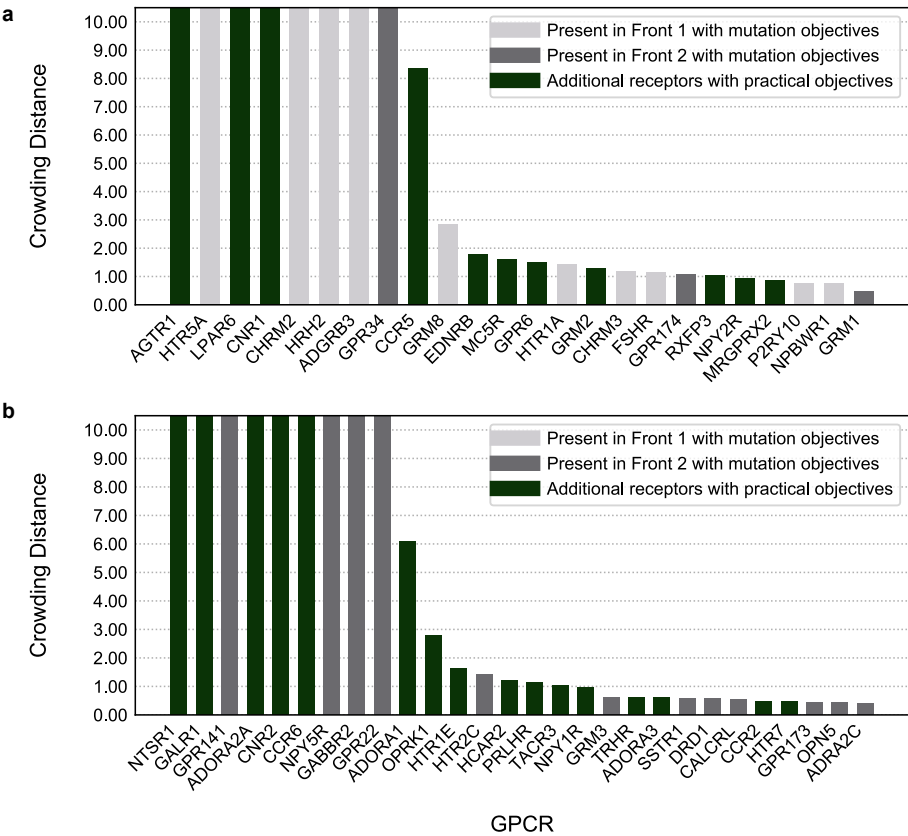


Figure 5.5. Crowding distances of the first and second Pareto fronts. **a)** First Pareto front, consisting of 24 GPCRs. **b)** Second Pareto front, consisting of 28 GPCRs. On the x-axis, the gene names of GPCRs are shown, while on the y-axis their crowding distance is shown. Crowding distance was cut off at 10, as the differences between these high-scoring receptors become negligible above that threshold. In grey, GPCRs detected by Pareto ranking using exclusively four mutation-derived objectives (light gray for the 1st front and darker grey for the 2nd front). In green, additional GPCRs that show up in the first two Pareto fronts by adding practical objectives to the recommendation system.

CCR5 structural analysis

Mutations found in the GDC dataset for CCR5 were cross-linked to GPCRdb data to find prior mutagenesis data. We then mapped the mutations onto the protein structure (PDB code 4MBS²⁰). We focused on regions relevant to protein function and ligand binding. These mutations are widely spread across the receptor’s structure (**Figure 5.6a**), including mutations in ECL2 – a region that largely contributes to chemokine ligand recognition (**Figure 5.6b**), G protein binding region (**Figure 5.6c**), and orthosteric binding site (**Figure 5.6d**). The crystal structure of CCR5 used as a reference in **Figure 5.6** (PDB code 4MBS) contains the thermostabilizing mutation A233^{6,33}E, which has been characterized for the inactive CCR5 conformation. In this structure, a

small molecule inhibitor – maraviroc – is co-crystallized in the orthosteric binding site (i.e. spanning the so-called major and minor binding pocket). Of note, some of the mutations found in the GDC dataset were in positions in close proximity to the inhibitor. Out of the 73 mutations found in our dataset, only 12 mutations had been previously annotated, while 37 mutations had no data available and 24 consisted of not-annotated data. Further analysis of previously annotated data shed some light on the functional implications of these mutations.

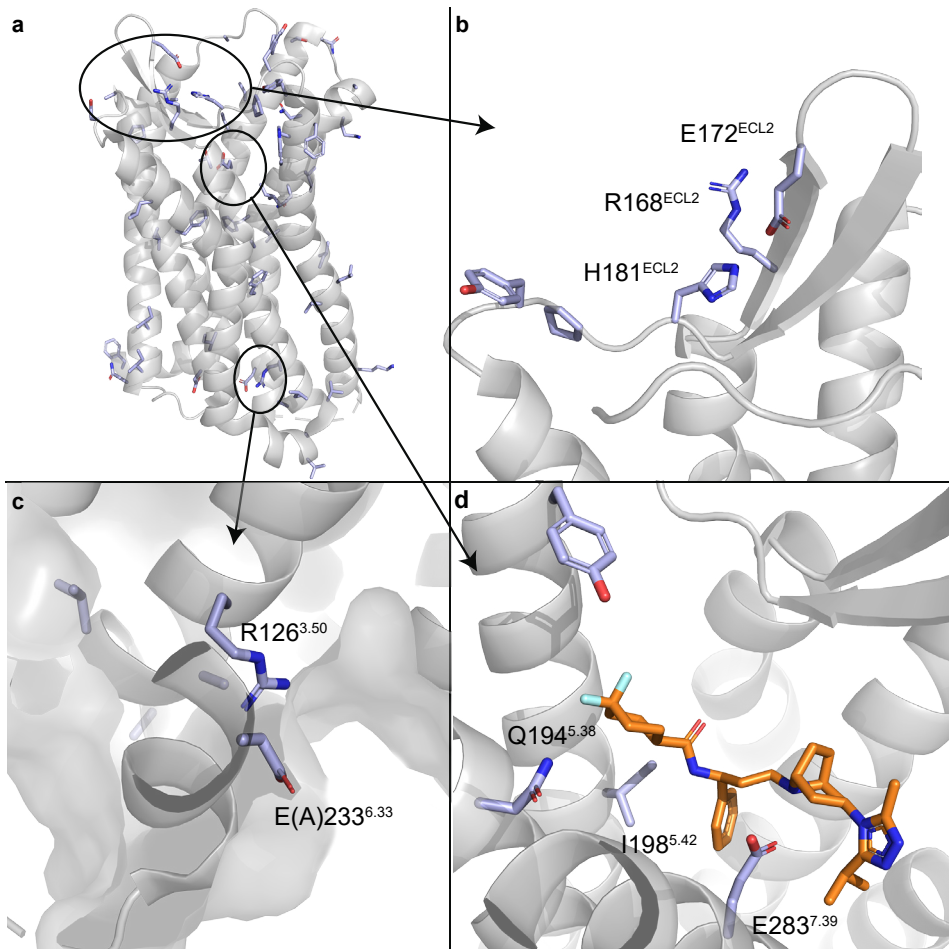


Figure 5.6. Cancer-derived mutation mapping in CCR5 structure. **a)** The mutations found in the GDC dataset for CCR5 mapped on the 3D structure of the receptor. **b)** Mutated residues found in the ECL2 region. **c)** G protein binding site, containing the mutation A233^{6.33}E, which has been characterized as a thermostabilizing mutation for the inactive CCR5 structure (PDB code 4MBS). **d)** The orthosteric binding site, with the small molecule inhibitor maraviroc (orange).

Discussion

Here we performed a comprehensive comparison of mutations found in cancer patients (GDC dataset) versus mutations found in natural variance (1000 Genomes dataset) in all classes of GPCRs together and independently. We followed this up by investigating several highly conserved motifs for an increase in mutation rate compared to the other residues. Finally, we performed a Pareto Front analysis to create a ranking of GPCRs that warrant follow-up for their context in cancer, and we analyzed some of the cancer-related mutations found for one of the top-ranking receptors from a functional-structural point of view.

Our original hypothesis was that more conserved residues (i.e. lower entropy in a two-entropy analysis of all residue positions in the GPCRdb alignment) would experience a higher mutational pressure in cancer patients. We confirmed a trend for the all-class analysis showing that positions with a low amount of mutations per position were assigned higher entropy values than positions with a high amount of mutations per position (**Figure 5.1a**). Conversely, the trend was not observed in a similar analysis in the 1000 Genomes dataset (**Supplementary Figure 5.2**). Overall, we identified an incipient pattern between functional conservation and mutation rates in the GDC set, which was maintained in class-specific analyses thus confining the applicability domain of the TEA originally established by Ye *et al.*¹⁹. However, subfamily-specific residues were not identified in the all-class analysis, possibly due to discrepancies in subfamily classification in GPCRdb. Other methods could be used to better distinguish functional residues across GPCR classes that, for example, are not dependent on a fixed subfamily classification (e.g. TEA-O also defined by Ye *et al.*¹⁹) or define the classification levels on the fly (e.g. TreeDet²¹).

We then studied mutation distribution after aggregating residues by protein (**Figure 5.2**) and subsequently compared these across all available classes. The total count of mutations found in the larger and less conserved domains (i.e. C- and N-terminus) is higher as the chance of mutations occurring is therefore higher. However, when corrected for average length most of them showed similar mutation rates. Of note, mutations in TM, ICL, and ECL domains showed an enrichment in cancer patients, while the contrary was observed for the C- and N-terminus (**Figure 5.2f**). The ICL and ECL domains are known to be important in receptor stabilization, signal transmission, and ligand and G protein recognition^{22,23}. However, they also represent the most variable domains in terms of length and motif composition explaining the lack of consistent enrichment across GPCR classes in cancer in these domains. This also aligns with the observation that GPCR mutation rates were not homogeneously distributed among cancer types. For example, some primary sites (e.g. Corpus uteri) showed a clear enrichment compared to others (see **Supplementary Figure 5.8**). Literature confirms this distribution with an emphasis on specific residue changes that affect the entire function of the protein^{24,25}.

A clearer pattern emerged in conserved motifs of GPCRs. We speculate that changes in these positions have a very high chance of disabling receptor function, supported by the observed higher mutation pressure in cancer compared to natural variance across

classes (**Figure 5.3a**). Thus, mutations might not be tolerated in healthy tissue but can be advantageous to cancer development. “DRY” mutations can decrease G protein coupling and recognition leading to reduced binding affinity of drugs²⁶. For both mutations in “DRY” and “Np_{xx}Y,” it has been shown that a decrease in ligand-receptor complex stability may occur, decreasing the response from the GPCR^{27,28}. These motifs have been shown to be collectively involved in a conserved Class A GPCR activation pathway¹⁴. As expected, “HET_x”, “RE”, “GWG_xP” and “P_{xx}G” all showed mutation enrichment in cancer in Class B GPCRs, but also in Class A GPCRs. These motifs are important for TM signaling, with those with a mutated motif showing loss of function¹⁵. The same principle is found for the mutational “R/K” hotspot, which is highly mutated in Class F GPCRs, serving as a switch for receptor activation¹⁸. Additionally, we found highly mutated H8 residues, in line with their recent identification as a functionally conserved motif in Class A GPCRs related to downstream signaling²⁹.

Subsequently, we ranked individual GPCRs for follow-up work via Pareto front analysis (**Figure 5.5**). Several of the top-ranked receptors had a known link to cancer. Notable entries include the C-C Chemokine receptor (CCR) type 5, which has been linked to regulatory T cells mediating tumor growth³⁰, and CCR type 2, a key player in microenvironment-derived tumor progression³¹, LPA (Lysophosphatidic acid) receptor LPAR6, upregulated in bladder cancer³², GRM (Metabotropic glutamate) receptors 2 (GRM2) and 8 (GRM8), known for dysregulating signaling pathways that are crucial in cancer prevention³³; serotonin receptors 5HT_{1A} (HTR1A), known to be involved in at least breast, ovarian, and pancreatic cancer, 5HT_{5A} (HTR5A), recently linked to breast cancer^{34,35}, and the adenosine A₁ (ADORA1) and A_{2A} (ADORA2A) receptors, linked to the progression and metastasis of a variety of cancer types as well as immune escape and immunotherapy^{36,37}. An example of a GPCR not previously linked directly to cancer was the P2Y receptor family member 10 (P2RY10), found in the first Pareto front. P2RY10 has been linked to chemotaxis via eosinophil degranulation, which could make it a potential target in cancer, although this is still highly speculative³⁸. Of note, cancer-related receptors were identified in our Pareto fronts both using exclusively somatic mutation-derived objectives and including practical objectives. The recommendation system proposed here is meant to allow user-specific objectives and therefore the practical objectives proposed here could be substituted by e.g. availability of crystal structures or cell lines overexpressing the receptor of interest.

Finally, the structural analysis of site-mutagenesis data in one of the top receptors from the first Pareto front (CCR5) shed light on the functional implication of some of the cancer-related mutations. This included a cluster of six residues in ECL2 found within the GDC dataset, from which four positions were previously shown to influence chemokine binding when mutated to Ala^{39,40}. In the G protein binding site, the Class A highly conserved R126^{3,50} was found to be mutated. This position is in the “DRY” motif and it is the most frequently mutated position in the GDC set, resulting in altered G protein coupling to the receptor in for instance the adenosine receptor family⁴¹. Some experimental evidence is available for CCR5 as well, where mutation to Asn abolished G protein signaling⁴². In the orthosteric site, four amino acids were previously investigated by a site-directed mutagenesis study by Garcia-Perez *et al.*, Y187^{5,31}, I198^{5,42}, N258^{6,58}, and

E283^{7,39,40} with variable effects. Mutating residue E283^{7,39}, to Ala or to the more conservative Gln, had the biggest effect on maraviroc affinity decrease. The structural effect of I198^{5,42} and E283^{7,39} mutations in maraviroc binding can be derived from the crystal structure of CCR5 with this negative allosteric modulator (**Figure 5.6c**). Mutations on these two positions had an important effect on the ligand binding of two other HIV-1 drugs – vicriviroc and aplaviroc – and clinical candidates – TAK-779 and TAK-220 – in two studies^{43,44}. Whilst E283^{7,39}A abolishes maraviroc binding, chemokine CCL5 binding is mildly (20-fold) affected⁴³. On the contrary, Y187^{5,31}A showed almost no effect on the binding affinity of maraviroc, while affecting chemokine recognition⁴⁰. These observations exemplify the relevance of our method to prioritize cancer-related mutations in site-mutagenesis studies and link them to receptor activation, endogenous ligand recognition, and the recognition of small (drug-like) molecules.

While completing this manuscript the TCGA dataset was used to identify significantly mutated GPCRs in cancer in a complementary extensive study by Wu *et al.*⁴⁵. In comparison, we elaborated on our findings through a motif analysis of highly conserved residues in GPCRs, a link to positional entropy, and a link to structural information (i.e. analyzing the CCR5 chemokine receptor). Moreover, we included the availability of chemical tools to study the selected GPCRs, as exemplified by our QSAR models. Another recent study by Huh *et al.*⁴⁶ focused on Class A GPCRs expressed in tumors reaching similar conclusions regarding Class A-specific functional motifs. There, a similar method was used to calculate mutation enrichment from natural variance which predicted the impact of mutations in specific sequence positions. Their results were validated *in vitro*, confirming the parallel effect of Class A GPCR mutations in receptor signaling. Our results extend to all GPCR class-specific functional motifs, opening novel paths to GPCR cancer research. Recently, we have published analyses of two other GPCRs, the Adenosine A₁ and A_{2B} receptors, for which cancer-related somatic mutations were identified similar to the analysis as presented here^{47,48}. There we used a yeast system to explore the effect said cancer-related mutations have on receptor function directly and found that there is a complex pattern of activation modulation. Similar approaches could be used to experimentally validate the relevance in cancer of somatic mutations in across all GPCR classes prioritized in this work.

While here the focus was on GPCRs, other receptor families can be investigated in a similar manner. Notable examples include solute carriers or receptor-tyrosine kinases, as highlighted in **Chapter 3** and through this thesis. The objectives in the Pareto optimization can also be adapted, providing a modified way of scoring the receptors depending on the scope of the study. While our analysis focused on differences in missense mutations occurring in cancer patients and natural variance, many other alterations (e.g. insertion/deletions, gene and protein expression levels) have been reported for GPCRs in the context of cancer^{6,49}, and complementary analyses could be executed focusing on these. Finally, this computational approach can become part of a targeted therapy pipeline, suggesting key locations for *in vitro* and *in vivo* cancer-associated studies.

Conclusions

We conclude that mutations found in GPCRs related to cancer are in general weakly correlated to specific domains in the protein or functional conservation. However, there is a higher mutational pressure in class-specific functionally conserved motifs in cancer patients (as shown in the GDC set) compared to healthy individuals. Moreover, we show that the role and mechanism of specific mutations can be elucidated using structural analysis as an intermediate step toward experimental validation. Finally, we provide a list of GPCRs that are amenable to experimental follow-up. The data may help in exploring new avenues in the design of cancer therapies, either by linking existing data to ligand binding and recognition, or the identification of potential new roles for residues not previously studied.

Materials and Methods

Cancer-related mutations

Cancer-associated mutations were obtained from the Genomic Data Commons (GDC), part of the US National Cancer Institute effort (version 22.0, January 16th, 2020)³. GDC contains multi-dimensional mapping of genomic changes in several cancer types, including the complete dataset from The Cancer Genomic Atlas project (TCGA)⁵⁰. We re-compiled part of the GDC database version 22.0 in a MySQL format to facilitate reproducible, version-consistent, big data cancer data analysis. Data was obtained from the GDC API engine and data transfer tool, depending on availability (unrestricted-access data only). The SQL database contains 19 tables distributed in eight different fields. Some data fields (i.e. gene expression data) contain analyzed data derived from GDC raw data files. A more extensive description of the database architecture, analyses performed, and the end-to-end mapping strategy is available in **Appendix A**. We used data on somatic missense mutations found in a diverse set of cancer types, which we will refer to as the “GDC” data set.

Natural variation

As a reference, we used the 1000 Genomes data⁵¹, including an additional data set released in 2020 by the New York Genome Center (NYGC). This is a dataset containing the natural variation of mutations in the genome. The dataset used in this study was obtained from the UniProt variance database in October 2020⁵². From this data, all somatic missense mutations were gathered. Subsequently, only mutations found in the 1000 Genomes subset were kept, removing cancer-derived mutations from COSMIC and known pathological mutations. We refer to this dataset as “1000 Genomes”.

Mutation dataset curation

We filtered both sets for GPCR-unique mutation pairs, along with the frequency. At the same time, we annotated the resulting GDC and 1000 Genomes datasets with identifiers from GPCRdb⁸. This set was used for two entropy analysis, domain-based analysis, and motif-based analysis. Subsequently, prior to QSAR modeling and Pareto sorting, both datasets were enriched with bioactivity data from ChEMBL (release 27)⁵³.

Bioactivity data

From ChEMBL (release 27)⁵³ ligand-protein interaction data was gathered for all GPCRs in GPCRdb⁸. Data points were filtered as follows: confidence score of 9, available pchembl value, and the protein belonging to the GPCR family (L2 protein class). A pchembl value is a standardized value that equals the negative logarithm of the measured activity for records with dose-response activity types.

Structural information

The data set was enriched with structural information from GPCRdb⁸ for GPCRs present in the GDC and 1000 Genomes dataset. Included were the family trees to find related proteins, the amino acid sequence of a protein, and sequence alignment data to add generic numbering to the residues. Finally, we used the HUGO Gene Nomenclature Committee (HGNC) identifiers for source-to-source mapping.

Multiple sequence alignment and generic numbering

The structurally supported multiple sequence alignment (MSA) provided by GPCRdb was used to study sequence conservation and link sequence positions to sequence- and structure-based generic GPCR numbering schemes. Generic numbering schemes (such as Ballesteros-Weinstein for Class A⁵⁴) can be used to compare positions between GPCRs but are often limited to the TM domains. There are two parts to the number separated by a decimal sign. The first identifies the domain (e.g. TM), and the second is relative to the most conserved residue in that TM. The most conserved residue is defined to be position 50, with downstream positions receiving a lower number (towards the N-terminus) and upstream positions receiving a higher number (towards the C-terminus). Other schemes are available for Class B, C, and F. Structure-based curations of these schemes have been developed by GPCRdb⁸. The GPCRdb generic values contain the same two parts but are separated by an “x” for differentiation purposes. We annotated the MSA with class-specific structure-based GPCRdb numbering schemes. Finally, we cross-linked the class-specific generic numbers with the more abundant class-A GPCRdb (GPCRdb(A)) equivalent to facilitate all-class analyses. For consistency, we refer to generic residue numbers in our work as Ballesteros-Weinstein, or BW, but give the GPCRdb(A) notation (i.e. 3x50 instead of 3.50) to denote the structural correction.

Investigated motifs

Several conserved motifs commonly found in GPCRs were investigated (**Table 5.2**). All are found in the literature to be functionally relevant in specific classes and often are referred to with the class-specific generic residue numbering schemes. To select these motifs across all classes, the Ballesteros-Weinstein residue numbering scheme was used.

Table 5.2. Investigated motifs, and their residues as noted by their generic residue numbering, both class-specific and Ballesteros-Weinstein.

Motif	Class	Generic residues (Class-specific)	Ballesteros-Weinstein generic residues
DRY	Class A	3.49, 3.50, 3.51*	3x49, 3x50, 3x51
CWxP	Class A	6.47, 6.48, 6.49, 6.50*	6x47, 6x48, 6x49, 6x50
nPxY	Class A	7.49, 7.50, 7.51, 7.52, 7.53*	7x49, 7x50, 7x51, 7x52, 7x53
HETx	Class B	2.50, 3.50, 6.42, 7.57 **	2x43, 3x46, 6x37, 7x53
RE	Class B	2.46, 8.49 **	2x39, 8x49
GWGxP	Class B	4.49, 4.50, 4.51, 4.52, 4.53 **	4x49, 4x50, 4x51, 4x52, 4x53
PxxG	Class B	6.47, 6.48, 6.49, 6.50 **	6x42, 6x43, 6x44, 6x45
R/K	Class F	6.32 ***	6x36

* Class-specific generic residue numbering scheme: Ballesteros-Weinstein^{8,54}

** Class-specific generic residue numbering scheme: Wootten⁸

*** Class-specific generic residue numbering scheme: Wang⁸

Two-Entropy Analysis

Two-entropy analysis (TEA) was performed as described previously in the literature¹⁹. We reimplemented the revised TEA algorithm, adjusted by Ye *et al.* to account for gaps in the multiple sequence alignment and for the differences in number of subfamily members. The reimplementation was validated by application to the synthetic dataset provided by Ye *et al.* (**Supplementary Figure 5.9**)¹⁹. We renamed “Total entropy” as “Rescaled Shannon entropy” and “Average entropy” as “Average entropy across subfamilies” for clarification. While the algorithm was not modified, two adaptations were made in the application, firstly using the GPCRdb hierarchy levels to define GPCR subfamilies, resulting in 83 subfamilies across all GPCR classes. From these, “Class A orphans” and “Class C orphans” were removed from the analysis. Secondly, we did not limit the entropy calculation to Class A GPCRS but applied it to all GPCR classes with more than one subfamily per class (**Supplementary Table 5.2**). However, contrary to previous work we included only human GPCR sequences.

Statistical analysis per position

The frequencies of mutations in both sets were analyzed per class and in combination (**Supplementary Table 5.2**). Mutation frequency was calculated as the sum of patients bearing any unique mutation in any receptor in a position of the multiple sequence

alignment included in:

- GPCR structural domains (i.e. N-terminus, TM domains, ECL and ICL loops, and C-terminus; also aggregated domains “TM”, “ECL”, and “ICL”)
- Functionally conserved motifs (**Table 5.2**)
- Individual alignment positions

To allow pairwise comparisons between sets, mutation ratios were calculated for cases (a) and (b), as defined in equations (1)-(3):

$$\tilde{M}_{s,d} = \frac{M_{s,d}}{M_s} \quad (1) \quad \langle l \rangle_{s,d} = \frac{\sum_{i=0}^{P_{s,d}} l_{s,d,i}}{P_{s,d}} \quad (2) \quad \tilde{M}'_{s,d} = \frac{\tilde{M}_{s,d}}{\langle l \rangle_{s,d}} \quad (3)$$

where M_s is the mutation frequency in a set s , $M_{s,d}$ is the mutation frequency in a set s per domain d , $\langle l \rangle_{s,d}$ is the average length per set s and domain d , $P_{s,d}$ is the number of proteins per set s and domain d , and $l_{s,d,i}$ is the length (number of residues) per set s and domain d in a protein i .

The mutation ratio, $\tilde{M}_{s,d}$, was visualized in **Figure 5.2a-c**. The mutation ratio normalized over average domain length, $\tilde{M}'_{s,d}$, was visualized in **Figure 5.2d-f** and in **Figure 5.3**. In **Figure 5.2d-f**, domains refer to GPCR structural domains and in **Figure 5.3** domains refer to functionally conserved GPCR motifs. In **Figures 5.2d-f** and **5.3**, a total mutation ratio, $\tilde{M}_{s,d=total}$, was calculated for reference. This represents the average mutation ratio in one residue if the totality of the protein sequence is taken into account and in **Figures 5.2d-f** and **5.3** is visualized as domain/motif “Average”. $\tilde{M}_{s,d=total}$ and $\tilde{M}'_{s,d=total}$ are derived from equations (1)-(3) as follows:

$$\begin{aligned} \tilde{M}_{s,d=total} &= \frac{M_{s,d=total}}{M_s} = \frac{M_s}{M_s} = 1 \\ \langle l \rangle_{s,d=total} &= \frac{\sum_{i=0}^{P_{s,d=total}} l_{s,d=total,i}}{P_{s,d=total}} \\ \tilde{M}'_{s,d=total} &= \frac{\tilde{M}_{s,d=total}}{\langle l \rangle_{s,d=total}} = \frac{1}{\langle l \rangle_{s,d=total}} \end{aligned}$$

In **Figures 5.2c,d** and **5.3b** we calculated GDC enrichments by subtracting $\tilde{M}_{s=GDC,d} - \tilde{M}_{s=1000\ G,d}$ and $\tilde{M}'_{s=GDC,d} - \tilde{M}'_{s=1000\ G,d}$, respectively.

For case (c) we calculated mutation frequency for each alignment position for the GDC and 1000 Genomes sets separately. Subsequently, we corrected the GDC frequency for natural variance by subtracting the 1000 Genomes frequency from the GDC frequency.

Pareto front

The multi-objective ranking was done within the Pareto method as implemented in Pipeline Pilot (version 18.1)⁵⁵. Two implementations were designed. The first one was based exclusively on mutation data and the following properties were used: Mutation rate in TM domains in GDC (maximized), mutation rate in TM domains in the 1000 Genomes set (minimized), GDC mutations in highly conserved TEA Q3 residues (maximized), and 1000 Genomes mutations in TEA Q3 residues (minimized). For this purpose, TEA Q3 residues were defined as those in the all-class TEA with “Rescaled Shannon entropy” < 0.5 and “Average entropy across subfamilies” < 0.5. The second implementation included two practical objectives to bias the ranking towards recommendations for subsequent *in vitro* or *in silico* studies. These practical objectives were the average R² of ChEMBL QSAR prediction models (maximized) and the in-house availability for experimental assays (maximized). The first and second fronts from each implementation were used in further analysis, but all data is provided as supporting information. The suitability of including practical objectives as part of a tunable recommendation system was evaluated by comparing the results of the two implementations. The performed QSAR models were Random Forest R models trained in Pipeline Pilot using 500 trees and a default seed of 12345. A 50/50 percent training/ hold-out test set was used in duplicate to create and validate these models, with ECFP6 used as molecular descriptors⁵⁶.

3D structural analysis

CCR5 crystal structure (PDB code 4MBS) was obtained from the Protein Data Bank²⁰. Mutagenesis data was retrieved from the GPCRdb and mapped onto the 3D crystal structure using PyMol⁵⁷.

Software

Accelrys Pipeline Pilot 2018 (version 18) was used for all the calculations and analysis⁵⁵. Any calculations performed were done in SI units, using the infrastructure provided in Pipeline Pilot. Data was written in plain text files and Excel. Graphs were created using Python’s module Matplotlib⁵⁸.

References

- Wild, C. P., Weiderpass, E. & Stewart, B. W. *World Cancer Report: Cancer Research for Cancer Prevention*. (2020).
- Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).
- Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
- Knudson, A. G. Two genetic hits (more or less) to cancer. *Nat Rev Cancer* **1**, 157–162 (2001).
- O'Hayre, M. *et al.* The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat Rev Cancer* **13**, 412–424 (2013).
- Arakaki, A. K. S., Pan, W. A. & Trejo, J. A. GPCRs in cancer: Protease-activated receptors, endocytic adaptors and signaling. *Int J Mol Sci* **19**, 2–24 (2018).
- Hausser, A. S. *et al.* Pharmacogenomics of GPCR Drug Targets. *Cell* **172**, 41–54.e19 (2018).
- Munk, C. *et al.* GPCRdb: the G protein-coupled receptor database – an introduction. *Br J Pharmacol* **173**, 2195–2207 (2016).
- Cvick, V., Goddard, W. A. & Abrol, R. Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications. *PLoS Comput Biol* **12**, e1004805 (2016).
- Congreve, M., de Graaf, C., Swain, N. A. & Tate, C. G. Impact of GPCR Structures on Drug Discovery. *Cell* **181**, 81–91 (2020).
- Thorpe, L. M., Yuzugullu, H. & Zhao, J. J. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat Rev Cancer* **15**, 7–24 (2015).
- Nairismägi, M.-L. *et al.* JAK-STAT and G-protein-coupled receptor signaling pathways are frequently altered in epitheliotropic intestinal T-cell lymphoma. *Leukemia* **30**, 1311–1319 (2016).
- Pon, J. R. & Marra, M. A. Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease* **10**, 25–50 (2015).
- Zhou, Q. *et al.* Common activation mechanism of class A GPCRs. *Elife* **8**, 1–31 (2019).
- Arimont, M. *et al.* Identification of Key Structural Motifs Involved in 7 Transmembrane Signaling of Adhesion GPCRs. *ACS Pharmacol Transl Sci* **2**, 101–113 (2019).
- Liang, Y. L. *et al.* Phase-plate cryo-EM structure of a class B GPCR-G-protein complex. *Nature* **546**, 118–123 (2017).
- Bortolato, A. *et al.* Structure of Class B GPCRs: New horizons for drug discovery. *Br J Pharmacol* **171**, 3132–3145 (2014).
- Wright, S. C. *et al.* A conserved molecular switch in Class F receptors regulates receptor activation and pathway selection. *Nat Commun* **10**, 1–12 (2019).
- Ye, K., Vriend, G. & IJzerman, A. P. Tracing evolutionary pressure. *Bioinformatics* **24**, 908–915 (2008).
- Tan, Q. *et al.* Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* **341**, 1387–1390 (2013).
- Carro, A. *et al.* TreeDet: A web server to explore sequence space. *Nucleic Acids Res* **34**, W110–W115 (2006).
- Semack, A., Sandhu, M., Malik, R. U., Vaidehi, N. & Sivaramakrishnan, S. Structural elements in the Gαs and Gβγ C termini that mediate selective G Protein-coupled Receptor (GPCR) signaling. *Journal of Biological Chemistry* **291**, 17929–17940 (2016).
- Lindner, D., Walther, C., Tennemann, A. & Beck-Sickinger, A. G. Functional role of the extracellular N-terminal domain of neuropeptide Y subfamily receptors in membrane integration and agonist-stimulated internalization. *Cell Signal* **21**, 61–68 (2009).
- Tao, Y. X. & Segaloff, D. L. Functional analyses of melanocortin-4 receptor mutations identified from patients with binge eating disorder and nonobese or obese subjects. *Journal of Clinical Endocrinology and Metabolism* **90**, 5632–5638 (2005).
- Stoy, H. & Gurevich, V. v. How genetic errors in GPCRs affect their function: Possible therapeutic strategies. *Genes Dis* **2**, 108–132 (2015).
- Kim, K.-M. & Caron, M. G. Complementary roles of the DRY motif and C-terminus tail of GPCRs for G protein coupling and β-arrestin interaction. *Biochem Biophys Res Commun* **366**, 42–47 (2008).
- Olivella, M., Caltabiano, G. & Cordero, A. The role of Cysteine 6.47 in class A GPCRs. *BMC Struct Biol* **13**, 3 (2013).
- Nomiyama, H. & Yoshie, O. Functional roles of evolutionary conserved motifs and residues in vertebrate chemokine receptors. *J. Leukoc. Biol* **97**, 39–47 (2015).
- Dijkman, P. M. *et al.* Conformational dynamics of a G protein-coupled receptor helix 8 in lipid membranes. *Sci Adv* **6**, 8207–8221 (2020).
- Schlecker, E. *et al.* Tumor-infiltrating monocytic myeloid-derived suppressor cells mediate CCR5-dependent recruitment of regulatory T cells favoring tumor growth. *J Immunol* **189**, 5602–11 (2012).
- Hao, Q., Vadgama, J. v. & Wang, P. CCL2/CCR2 signaling in cancer pathogenesis. *Cell Communication and Signaling* **18**, 1–13 (2020).

32. Houben, A. J. S. & Moolenaar, W. H. Autotaxin and LPA receptor signaling in cancer. *Cancer and Metastasis Reviews* **30**, 557–565 (2011).
33. Prickett, T. D. & Samuels, Y. Molecular Pathways: Dysregulated Glutamatergic Signaling Pathways in Cancer. *Clinical Cancer Research* **18**, 4240–4246 (2012).
34. Gwynne, W. D. *et al.* Antagonists of the serotonin receptor 5A target human breast tumor initiating cells. *BMC Cancer* **20**, 1–17 (2020).
35. Sarrouilhe, D. & Mesnil, M. Serotonin and human cancer: A critical view. *Biochimie* **161**, 46–50 (2019).
36. Masjedi, A. *et al.* Silencing adenosine A2a receptor enhances dendritic cell-based cancer immunotherapy. *Nanomedicine* **29**, 102240 (2020).
37. Ni, S., Wei, Q. & Yang, L. Adora1 promotes hepatocellular carcinoma progression via pi3k/akt pathway. *Oncotargets Ther* **13**, 12409–12419 (2020).
38. Hwang, S. M. *et al.* Lysophosphatidylserine receptor P2Y10: A G protein-coupled receptor that mediates eosinophil degranulation. *Clinical and Experimental Allergy* **48**, 990–999 (2018).
39. Blanpain, C. *et al.* The Core Domain of Chemokines Binds CCR5 Extracellular Domains while Their Amino Terminus Interacts with the Transmembrane Helix Bundle. *Journal of Biological Chemistry* **278**, 5179–5187 (2003).
40. Garcia-Perez, J. *et al.* Allosteric model of maraviroc binding to CC Chemokine Receptor 5 (CCR5). *Journal of Biological Chemistry* **286**, 33409–33421 (2011).
41. Jespers, W. *et al.* Structural Mapping of Adenosine Receptor Mutations: Ligand Binding and Signaling Mechanisms. *Trends Pharmacol Sci* **39**, 75–89 (2018).
42. Lagane, B. *et al.* Mutation of the DRY motif reveals different structural requirements for the CC chemokine receptor 5-mediated signaling and receptor endocytosis. *Mol Pharmacol* **67**, 1966–76 (2005).
43. Kondru, R. *et al.* Molecular interactions of CCR5 with major classes of small-molecule anti-HIV CCR5 antagonists. *Mol Pharmacol* **73**, 789–800 (2008).
44. Swinney, D. C. *et al.* A study of the molecular mechanism of binding kinetics and long residence times of human CCR5 receptor small molecule allosteric ligands. *Br J Pharmacol* **171**, 3364–3375 (2014).
45. Wu, V. *et al.* Illuminating the Onco-GPCRome: Novel G protein-coupled receptor-driven oncocrine networks and targets for cancer immunotherapy. *Journal of Biological Chemistry* **294**, 11062–11086 (2019).
46. Huh, E. *et al.* Recurrent high-impact mutations at cognate structural positions in class A G protein-coupled receptors expressed in tumors. *Proc Natl Acad Sci U S A* **118**, 1–12 (2021).
47. Wang, X. *et al.* Characterization of cancer-related somatic mutations in the adenosine A2B receptor. *Eur J Pharmacol* **880**, 173126 (2020).
48. Wang, X. *et al.* Cancer-related somatic mutations alter adenosine A1 receptor pharmacology—A focus on mutations in the loops and C-terminus. *The FASEB Journal* **36**, 1–16 (2022).
49. Sriram, K., Moyung, K., Corriden, R., Carter, H. & Insel, P. A. GPCRs show widespread differential mRNA expression and frequent mutation and copy number variation in solid tumors. *PLoS Biol* **17**, 1–43 (2019).
50. Broad Institute of MIT and Harvard. Firehose 2015_11_01 run. Available at <https://doi.org/10.7908/C1571BB1>.
51. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
52. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
53. ChEMBL27 Database Release. Available at <https://doi.org/10.6019/CHEMBL.database.27>.
54. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in Neuroscience* **25**, 366–428 (1995).
55. BIOVIA Pipeline Pilot | Scientific Workflow Authoring Application for Data Analysis.
56. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform* **9**, 45 (2017).
57. The PyMOL Molecular Graphics System, Version 1.4 Schrödinger, LLC.
58. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95 (2007).

Supplementary Information

Supplementary Table 5.1. Two-Entropy Analysis parameters for GDC and 1000 Genomes sets in all GPCR classes analyzed combined and independently. Shannon (Sh.) and Average group (Gr.) entropy mean and standard deviation (SD) values for all three levels of mutation rates: low (< 10th percentile), medium (10th - 90th percentile), and high (> 90th percentile).

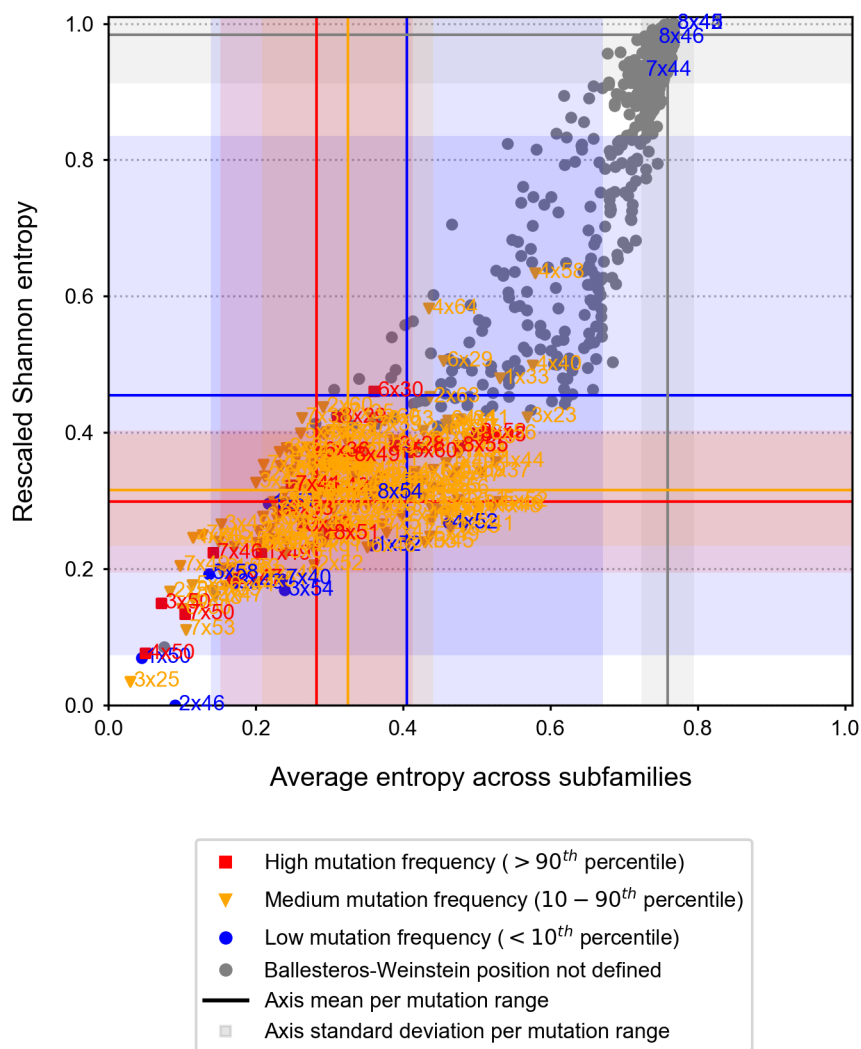
	GDC							1000 Genomes						
Class	10 th /90 th percentiles	Low Mean ± SD		Medium Mean ± SD		High Mean ± SD		10 th /90 th percentiles	Low Mean ± SD		Medium Mean ± SD		High Mean ± SD	
		Sh.	Gr.	Sh.	Gr.	Sh.	Gr.		Sh.	Gr.	Sh.	Gr.	Sh.	Gr.
All class	41/74	0.45 ± 0.38	0.41 ± 0.27	0.32 ± 0.08	0.32 ± 0.12	0.30 ± 0.10	0.28 ± 0.13	18/40	0.40 ± 0.30	0.33 ± 0.23	0.31 ± 0.09	0.31 ± 0.12	0.34 ± 0.08	0.39 ± 0.12
Class A	28/55	0.40 ± 0.25	0.34 ± 0.19	0.39 ± 0.13	0.32 ± 0.13	0.38 ± 0.16	0.32 ± 0.15	10/25	0.38 ± 0.22	0.28 ± 0.17	0.39 ± 0.14	0.32 ± 0.13	0.41 ± 0.10	0.38 ± 0.12
Class B1	1/5	-	-	0.41 ± 0.26	0.35 ± 0.30	0.39 ± 0.23	0.34 ± 0.28	1/5	-	-	0.42 ± 0.25	0.35 ± 0.29	0.53 ± 0.26	0.49 ± 0.29
Class B2	3/9	0.53 ± 0.17	0.45 ± 0.21	0.46 ± 0.18	0.43 ± 0.21	0.43 ± 0.23	0.37 ± 0.22	2/9	0.43 ± 0.18	0.40 ± 0.20	0.47 ± 0.18	0.43 ± 0.21	0.41 ± 0.14	0.39 ± 0.12
Class B	4/13	0.62 ± 0.22	0.59 ± 0.26	0.44 ± 0.15	0.38 ± 0.19	0.41 ± 0.25	0.34 ± 0.24	3/13	0.52 ± 0.25	0.47 ± 0.26	0.45 ± 0.16	0.39 ± 0.2	0.46 ± 0.14	0.40 ± 0.14
Class C	1/6	-	-	0.48 ± 0.17	0.39 ± 0.18	0.45 ± 0.17	0.39 ± 0.16	1/4	-	-	0.50 ± 0.18	0.40 ± 0.19	0.50 ± 0.14	0.46 ± 0.11
Class F *	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Class T *	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Other GPCRs *	-	-	-	-	-	-	-	-	-	-	-	-	-	-

* Two Entropy Analysis was not performed in classes with only one GPCRdb subfamily defined.

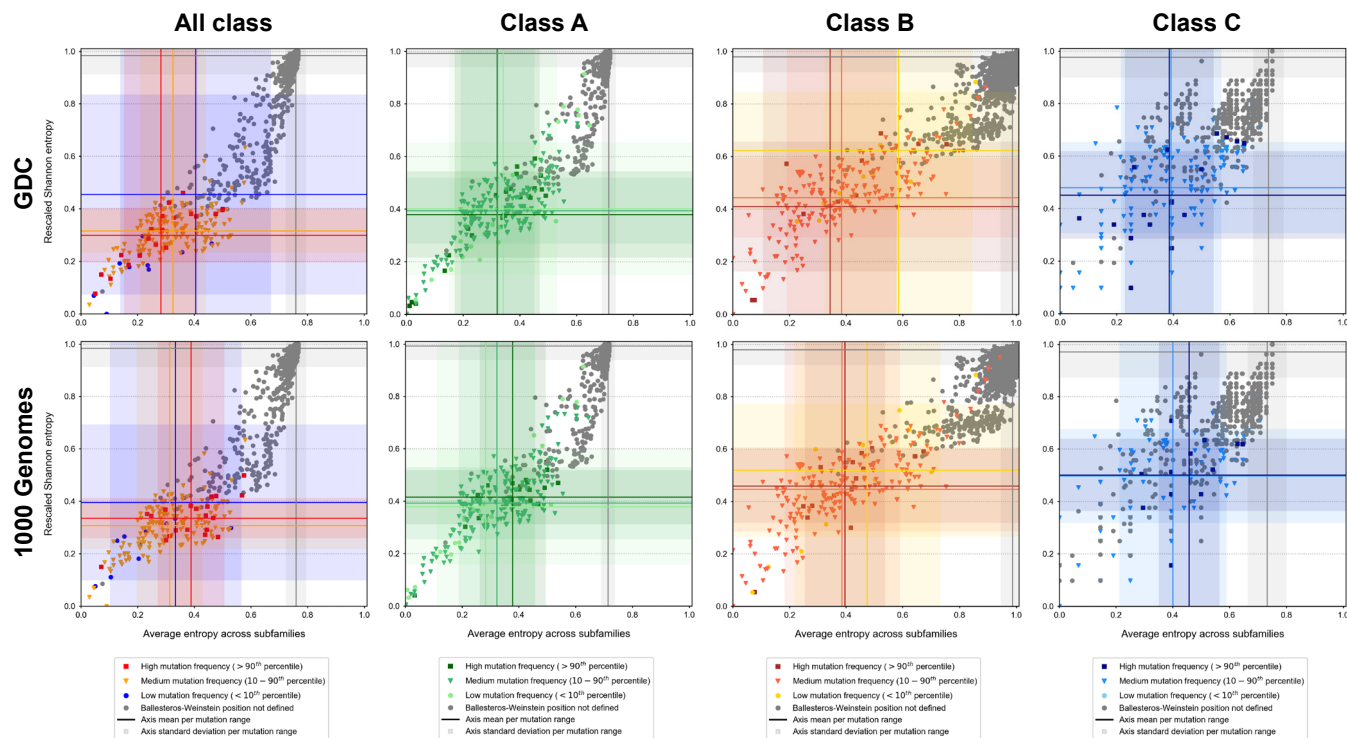
Supplementary Table 5.2. GPCR classes analyzed, number of members per class and GPCRdb sub-families defined in the Two-Entropy Analysis.

Class		Number of receptors in alignment	GPCRdb hierarchy levels (subfamilies)
All class		401	83
Class A (Rhodopsin)		289	61
Class B*		48	14
	Class B1 (Secretin)	15	5
	Class B2 (Adhesion)	33	9
Class C (Glutamate)		22	5
Class F (Frizzled)		11	1
Class T (Taste 2)		25	1
Other GPCRs		6	1

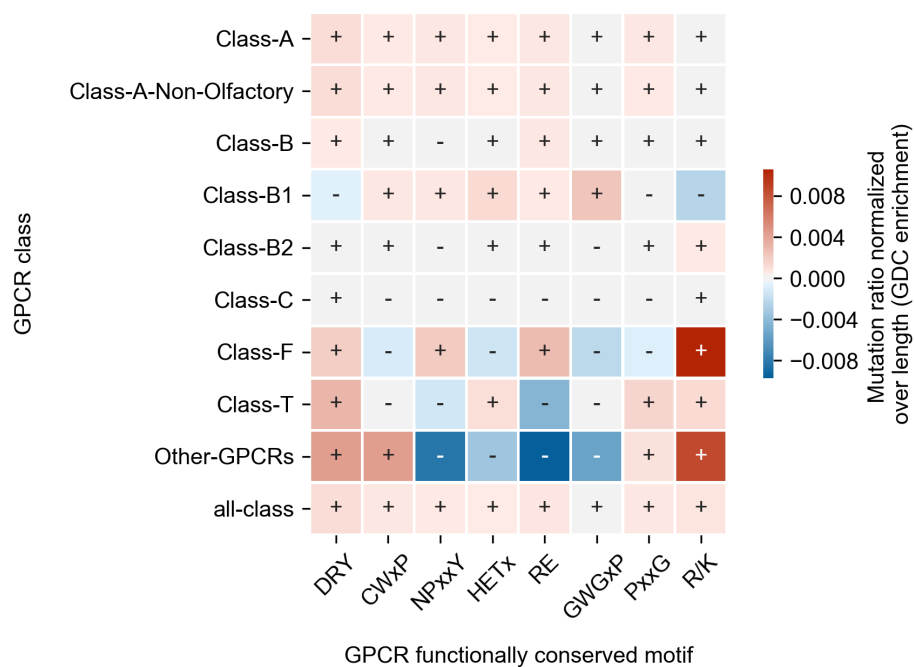
* Synthetic class formed by aggregation of Class B1 and Class B2 to facilitate the analysis of class-specific functional motifs described in the literature.



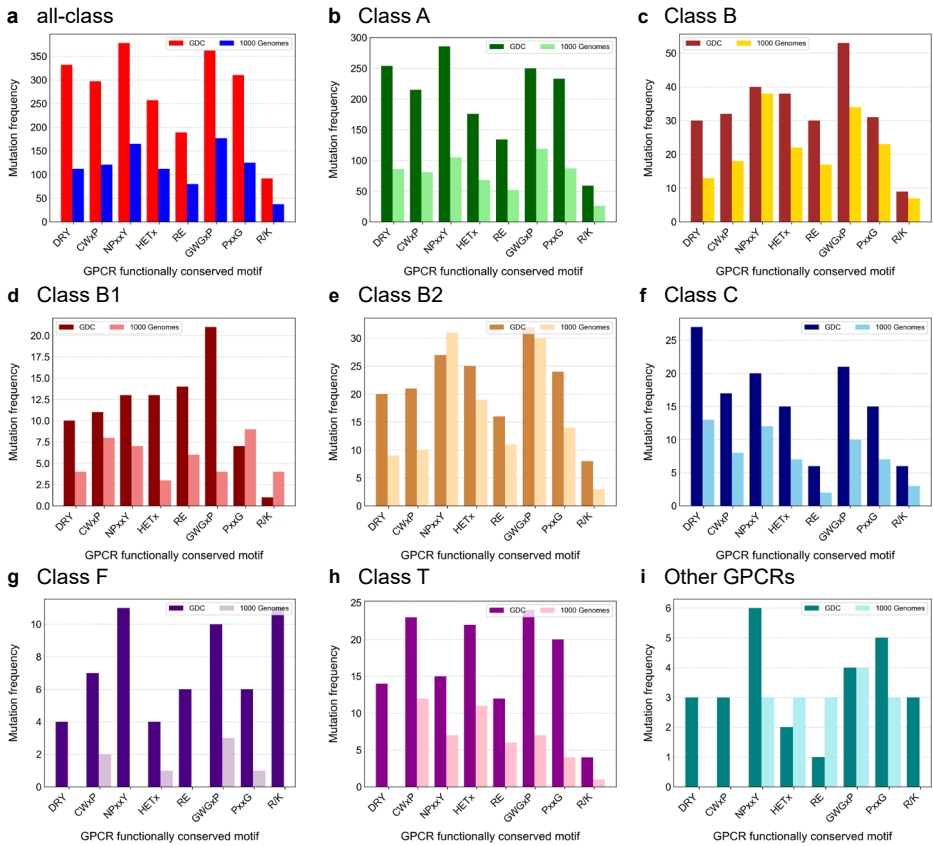
Supplementary Figure 5.1. Shannon entropy across GPCR subfamilies versus Shannon global Entropy correlated to cancer-related mutations, with residue and GDC labels. A two-entropy analysis plot for all GPCRs with aligned positions and labeled residues. The average entropy across families, i.e. conserved within a family is on the x-axis, and the Shannon entropy overall is on the y-axis. Residues are colored by the frequency of mutations found in the GDC dataset, with blue being low ($< 10^{th}$ percentile), orange medium ($10^{th} - 90^{th}$ percentiles), and red high ($> 90^{th}$ percentile). Residues with no defined Ballesteros-Weinstein labels are colored grey. Blue, orange, red, and grey lines represent the mean entropy values for each axis per mutation range (high, medium, low, and non-defined Ballesteros-Weinstein, respectively). Blue, orange, red, and grey shadows represent the standard deviation to the mean entropy values for each axis per mutation range (high, medium, low, and non-defined Ballesteros-Weinstein, respectively).



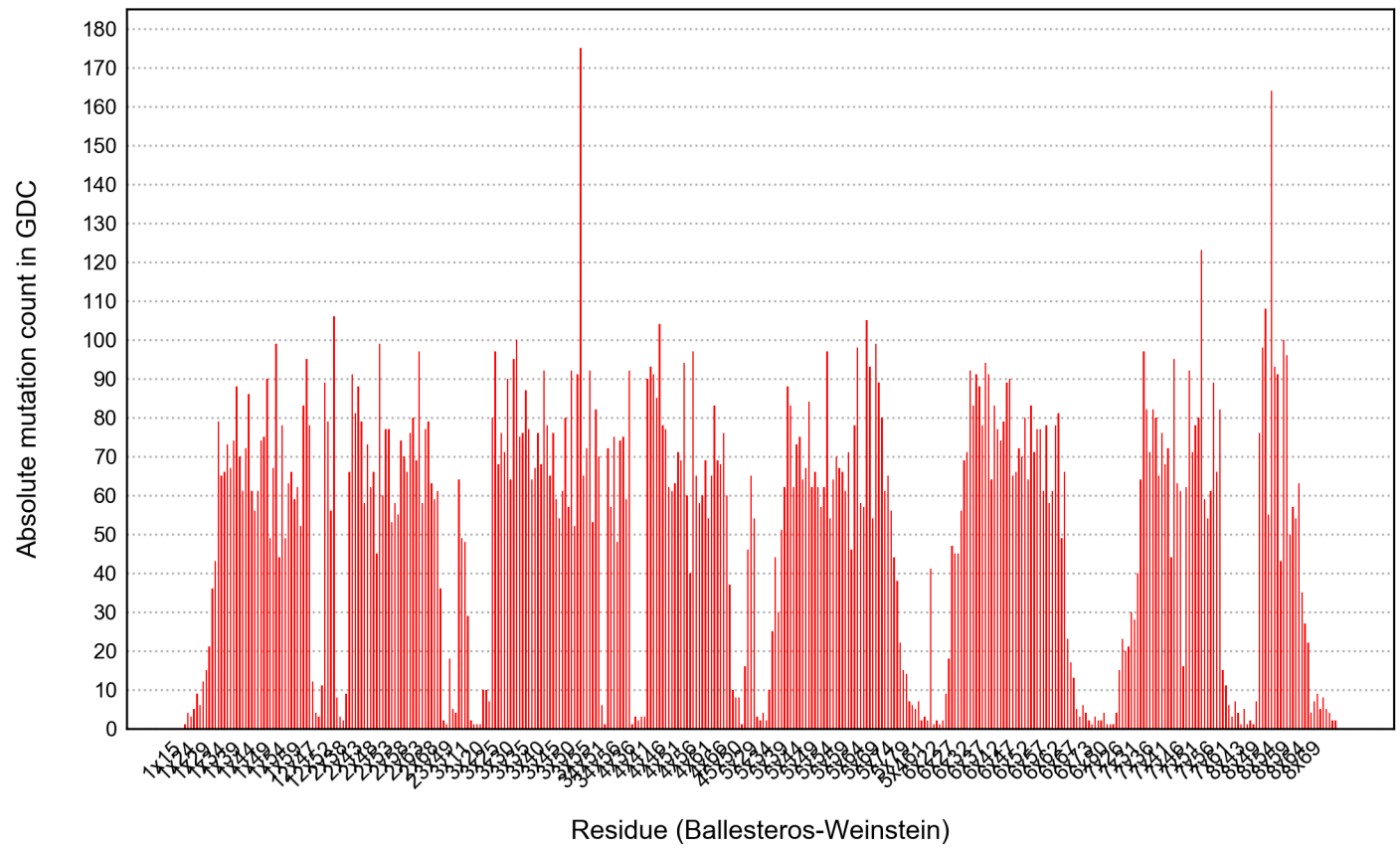
Supplementary Figure 5.2. Two-entropy analysis correlated to cancer-related mutations and natural variance across GPCR classes. The analysis is performed on all GPCR classes combined, as well as Class A-C independently. Residues are colored by the frequency of mutations found in the GDC dataset (top row), and the 1000 genomes dataset (bottom row). In the all-class analysis, blue is low (< 10th percentile), orange medium (10-90th percentiles), and red high (> 90th percentile) mutation frequency. Residues with no defined Ballesteros-Weinstein generic numbers are colored grey. Blue, orange, red, and grey lines represent the mean entropy values for each axis per mutation range (high, medium, low, and non-defined Ballesteros-Weinstein, respectively). Blue, orange, red, and grey shadows represent the standard deviation to the mean entropy values for each axis per mutation range (high, medium, low, and non-defined Ballesteros-Weinstein, respectively). The coloring scheme for classes A-C is equivalent to that of all classes combined.



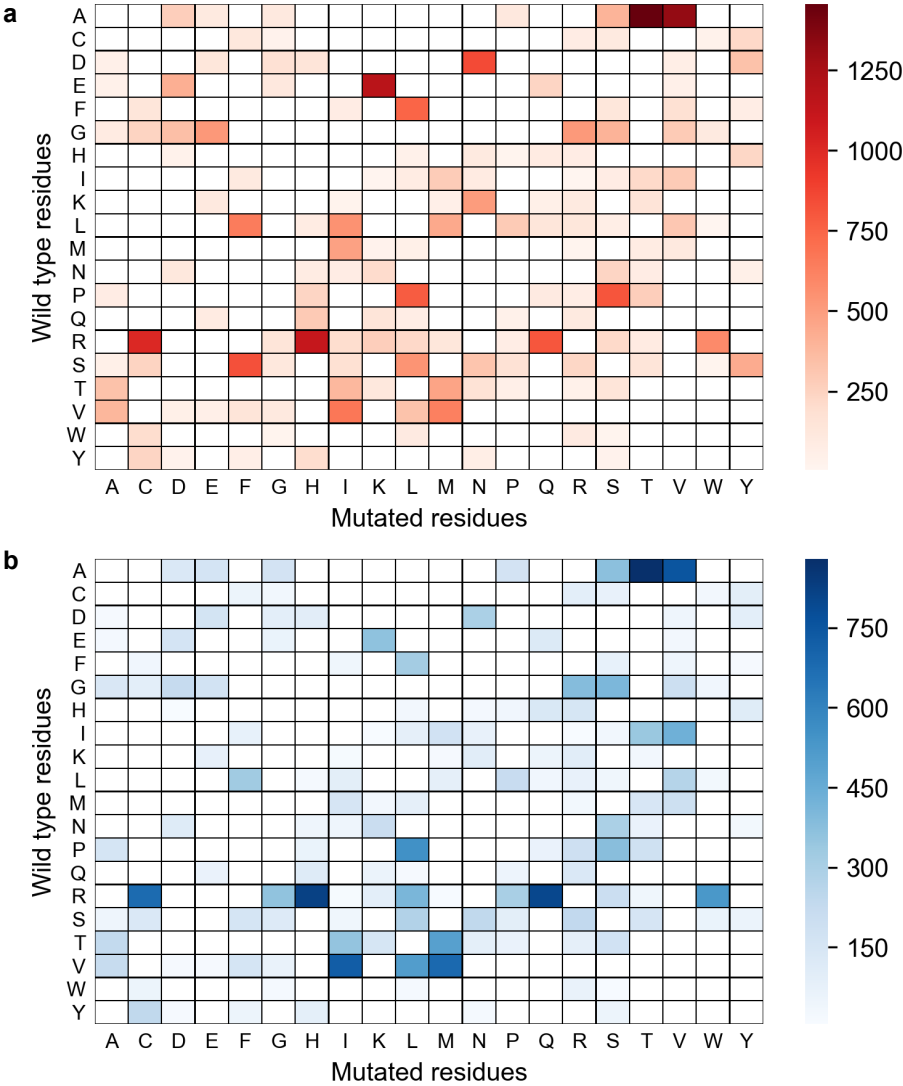
Supplementary Figure 5.3. Enrichment of mutation frequencies per GPCR functionally conserved motifs across all GPCR classes. Length-normalized mutation ratio enrichment in the GDC dataset over the 1000 Genomes dataset in all classes combined and independently. Motifs analyzed are “DRY”, “CWxP”, and “NPxxY” (Class A); “HETx”, “RE”, “GWGxP”, and “PxxG” (Class B); and “R/K” (Class F). “Average” represents the average ratio considering the totality of the protein length. A darker shade of red represents a higher enrichment over the GDC dataset, and a darker shade of blue represents a higher enrichment over the 1000 Genomes dataset.



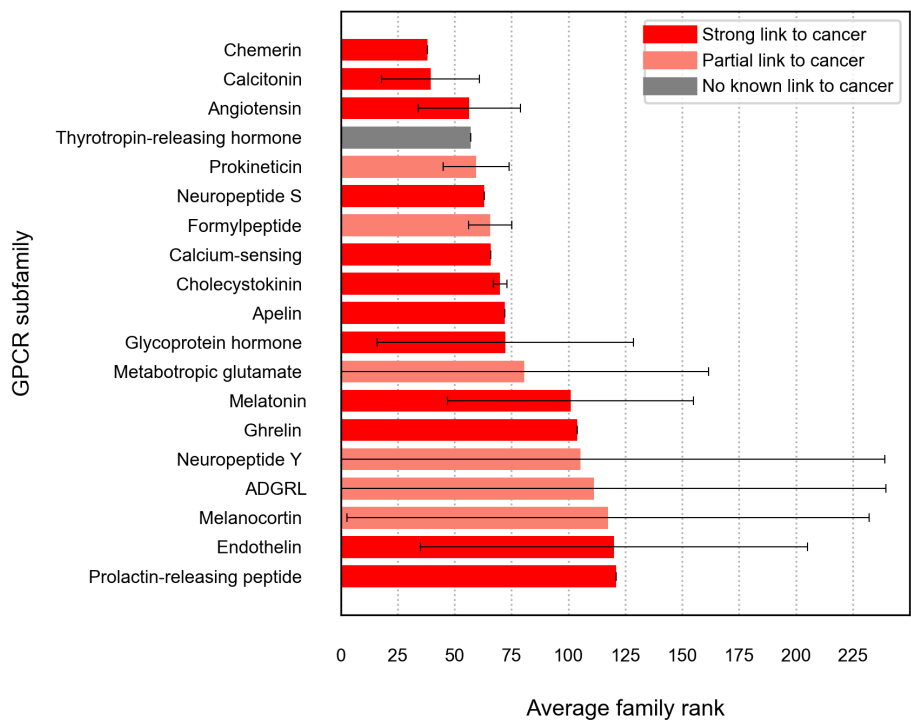
Supplementary Figure 5.4. Mutation frequency cancer and natural variance in GPCR functionally conserved motifs across GPCR classes. Motifs analyzed are “DRY”, “CWxP”, and “NPxxY” (Class A); “HETx”, “RE”, “GWGxP”, and “PxxG” (Class B); and “R/K” (Class F). **a)** Analysis of all GPCR classes combined. **b)** Analysis of Class A. **c)** Analysis of Class B. **d)** Analysis of Class B1. **e)** Analysis of Class B2. **f)** Analysis of Class C. **g)** Analysis of Class F. **h)** Analysis of Class T. **i)** Analysis of Class Other GPCRs.



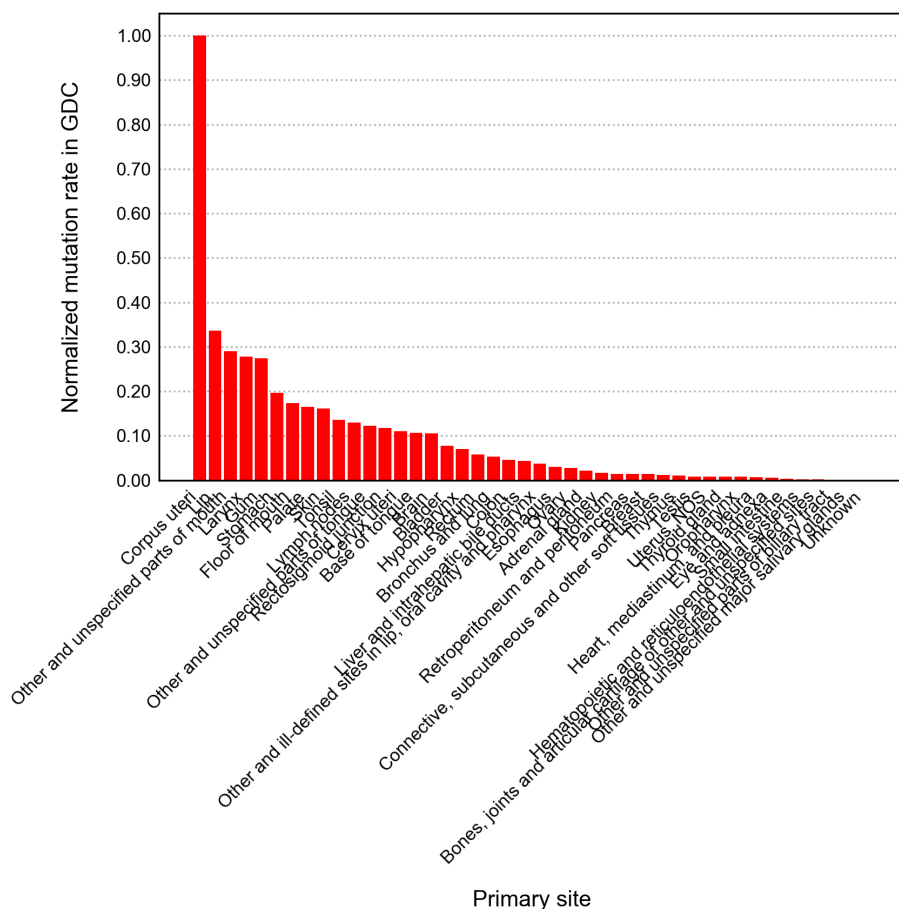
Supplementary Figure 5.5. GPCR cancer mutations on Ballesteros-Weinstein positions. GPCR cancer mutations plotted for the Ballesteros-Weinstein positions found in the GDC data. Positions are ordered from lowest to highest and X-axis labels are displayed every five residues for visualization purposes.



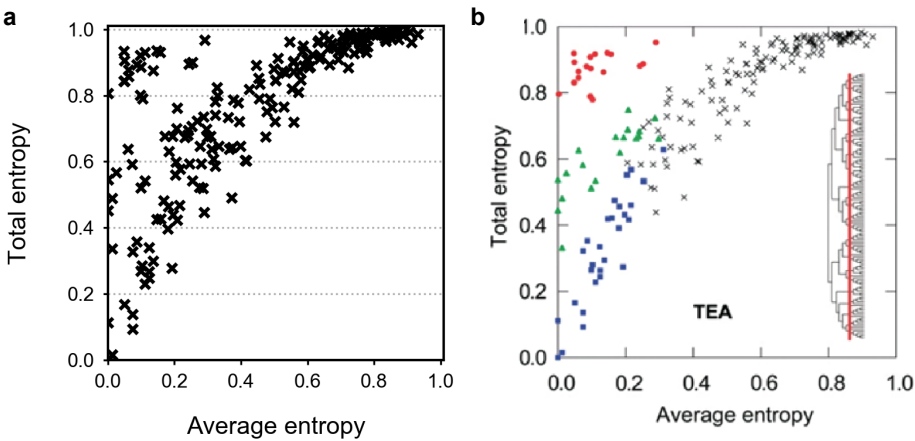
Supplementary Figure 5.6. Heat-map cancer substitutions. **a)** Heat-map showing the frequency of substitutions found in the GDC dataset. A darker shade of red means a higher frequency. **b)** Heat-map showing the frequency of substitutions found in the 1000 Genomes dataset. A darker shade of blue means a higher frequency.



Supplementary Figure 5.7. Average Rank of GPCR families and their link to cancer in the literature. Average rank of GPCR families related to the mutation ratio in individual family members. For each GPCR, the absolute mutation count was divided by receptor length, to provide a mutation rate for each. To identify patterns within GPCR families, a family-wide rank was calculated by averaging the ranking of each of the members in a family and subsequently compared to the other families. Shown on the y-axis are the different GPCR families as categorized by GPCRdb, while on the x-axis their average rank as a receptor family is given. The lower the average rank value, the better. The error bars represent the standard deviation of individual GPCR rankings within the family. Color coding represents the link to cancer in the literature for the family. Red represents a strong link (i.e. all members of the family have been linked to cancer), salmon represents a partial link (i.e. some members of the family have been linked to cancer), and grey represents no link to cancer reported.



Supplementary Figure 5.8. GPCR mutation rates by cancer type. Normalized GPCR mutation rate per primary site (i.e. cancer type). The mutation rate per primary site is normalized by the number of patients in GDC with that cancer type.



Supplementary Figure 5.9. Two-entropy analysis re-implementation. **a)** Re-implementation of two-entropy analysis in a synthetic dataset as defined by Ye *et al.* in ¹⁹. **b)** Original analysis, figure adapted from Ye *et al.* in ¹⁹.



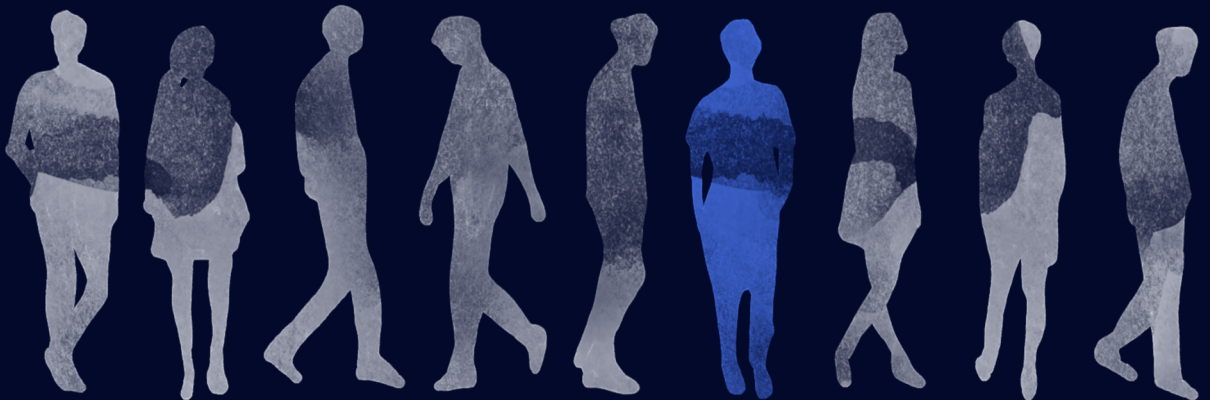
Chapter 6

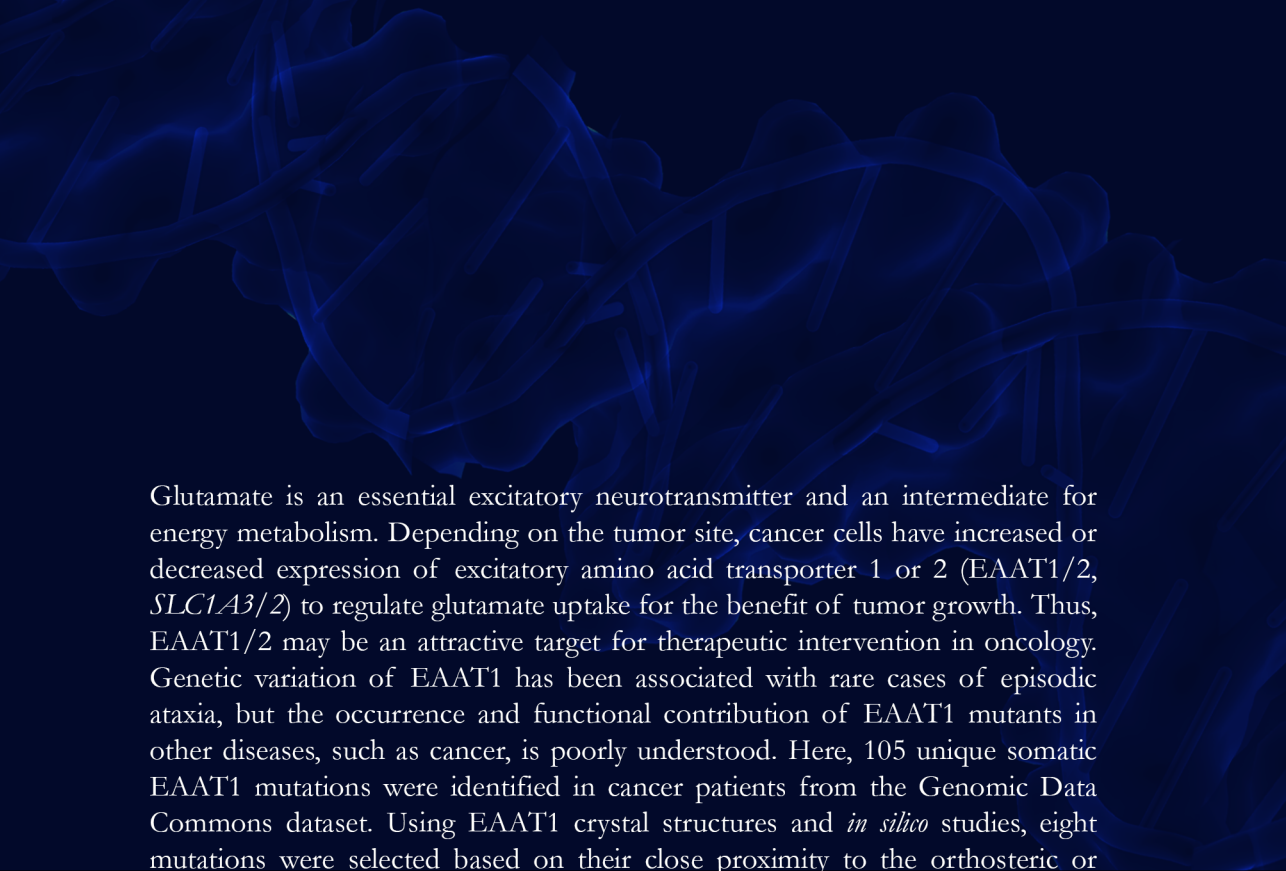
Molecular insights into disease-associated
glutamate transporter (EAAT1 / *SLC1A3*) variants
using *in silico* and *in vitro* approaches

Marina Gorostiola González[†], Hubert J. Sijben[†], Laura Dall'Acqua, Rongfang Liu,
Adriaan P. IJzerman, Laura H. Heitman, Gerard J.P. van Westen


Adapted from: *Frontiers in Molecular Biosciences* **10**, 3389 (2023)

[†]These authors contributed equally





Glutamate is an essential excitatory neurotransmitter and an intermediate for energy metabolism. Depending on the tumor site, cancer cells have increased or decreased expression of excitatory amino acid transporter 1 or 2 (EAAT1/2, *SLC1A3/2*) to regulate glutamate uptake for the benefit of tumor growth. Thus, EAAT1/2 may be an attractive target for therapeutic intervention in oncology. Genetic variation of EAAT1 has been associated with rare cases of episodic ataxia, but the occurrence and functional contribution of EAAT1 mutants in other diseases, such as cancer, is poorly understood. Here, 105 unique somatic EAAT1 mutations were identified in cancer patients from the Genomic Data Commons dataset. Using EAAT1 crystal structures and *in silico* studies, eight mutations were selected based on their close proximity to the orthosteric or allosteric ligand binding sites and the predicted change in ligand binding affinity. *In vitro* functional assessment in a live-cell, impedance-based phenotypic assay demonstrated that these mutants differentially affect L-glutamate and L-aspartate transport, as well as the inhibitory potency of an orthosteric (TFB-TBOA) and allosteric (UCPH-101) inhibitor. Moreover, two episodic ataxia-related mutants displayed functional responses that were in line with the literature, which confirmed the validity of our assay. Of note, ataxia-related mutant M128R displayed inhibitor-induced functional responses never described before. Finally, molecular dynamics (MD) simulations were performed to gain mechanistic insights into the observed functional effects. Taken together, the results in this work demonstrate 1) the suitability of the label-free phenotypic method to assess functional variation of EAAT1 mutants and 2) the opportunity and challenges of using *in silico* techniques to rationalize the *in vitro* phenotype of disease-relevant mutants.



Introduction

Glutamate is an abundant endogenous amino acid that acts as the major excitatory neurotransmitter in the central nervous system and serves as a key metabolite in energy homeostasis¹. In the synaptic cleft, glutamate is transported across the cell membrane via excitatory amino acid transporters (EAATs), which belong to subfamily 1 of the solute carrier (SLC) transporters². Glutamate transport is thermodynamically coupled to the transport of three Na⁺ ions and one proton, and the counter-transport of one K⁺ ion, where the binding of Na⁺ and/or substrate activates an uncoupled Cl⁻ conductive state³. Deregulated glutamate levels have been associated with a plethora of neurological diseases^{4,5} and more recently with cancer^{6,7}. As a result, pharmacological modulation of EAATs may be a promising therapeutic strategy for conditions that are associated with altered glutamate levels^{8,9}.

Depending on the location of the tumor, cancerous cells have been shown to exploit the uptake, metabolism, and signaling properties of glutamate as well as aspartate as fuel for tumor proliferation and expansion. Healthy glia cells abundantly express EAAT1 and EAAT2 to mediate the majority of glutamate clearance². However, expression levels of EAAT2 are vastly reduced in gliomas, which combined with increased efflux via the glutamate/cystine antiporter (xCT, *SLC7A11*) leads to elevated glutamate levels surrounding the glioma that induce cell death and allow further growth of the tumor^{10,11}. Moreover, EAAT1 was found to be overexpressed and cause glutamate efflux in aggressive glioblastomas, which indicates selective EAAT1 inhibitors as a potential treatment option for glioma¹². In several instances of cancer in peripheral tissues, EAAT1 expression has been linked to a poor disease prognosis. Under hypoxia or conditions that starve the tumor of glutamine, some cancer cells promote EAAT1 or EAAT2 expression to drive uptake of aspartate or glutamate which rescues cancer cell growth^{13–15}. As such, EAAT expression in such tumors could be a predictive biomarker and pharmacological modulation of glutamate transporter expression or activity could be of therapeutic interest.

Despite the clear advantages for tumor cells to regulate EAAT expression, little is known about human genetic variations of these transporters in cancer, although several mutations have been associated with other diseases. Thus far, reports have linked seven missense mutations in the coding region of EAAT1 to the etiology of extremely rare cases of episodic ataxia type 6 (EA6)¹⁶. These mutants vary in their degree of loss- or gain-of-function of substrate transport and/or anion conductivity¹⁶. Moreover, several other EAAT1 mutations and duplications have been associated with other neurological disorders including migraine, ADHD, autism, and Tourette's syndrome^{17–19}. To the best of our knowledge, there have been no reports so far that associate mutations of EAAT1 with the development and progression of cancer.

Over the last fifteen years, a growing number of 3D structures have been published for the archaeal glutamate transporter orthologues Glt_{ph}²⁰ and Glt_{tk}²¹, as well as human EAAT1^{22,23}, EAAT2²⁴, and EAAT3²⁵, in complex with the endogenous substrate L-aspartate, Na⁺ ions, and/or inhibitors. Glutamate transporters assemble in obligate

homo-trimers of which the protomers operate independently of each other. Each protomer consists of a rigid trimerization or scaffold domain (scaD) and a dynamic transport domain (tranD) that engages with the substrate and co-transported Na^+ ions²². Structures covering inward-facing, intermediate, and outward-facing conformations provide information on the movement of individual transmembrane helices (TMs). Specifically, the flexible helical hairpin 2 (HP2) in tranD controls the access of ligands to the substrate binding site and is an essential “gate” that upon opening and closing regulates the “elevator-like” translocation of tranD. Of note, these transport mechanisms have been elucidated in part thanks to molecular dynamic (MD) simulations^{26,27}. Thus, these structures may be used to gain mechanistic insight into the effects of genetic variability on transport function, as was previously demonstrated by mapping genetic variants of glucose (GLUT1) and nucleoside (ENT1) transporters to their respective crystal structures²⁸.

In this study, a series of EAAT1 somatic mutations that were identified from biopsy material of cancer patients represented in the Genomic Data Commons (GDC) dataset²⁹ were characterized. Using the reported ligand-bound crystal structures of EAAT1^{22,23}, predictions were made on which variants would most likely impact the binding of substrates (L-glutamate and L-aspartate). To determine whether these mutants would affect the binding of potential pharmacological modulators, the orthosteric inhibitor TFB-TBOA³⁰ and the allosteric inhibitor UCPH-101⁹ were included, which have been co-crystallized with EAAT1²². The selected eight mutations, together with two EA6-associated mutants (M128R, T318A), were tested *in vitro* for substrate uptake and inhibition using a label-free impedance-based phenotypic assay that was previously developed in our lab³¹. Mutants displayed divergent effects on EAAT1 function, which was apparent from an altered substrate and/or inhibitor potency. Finally, MD simulations and molecular docking were used to explore the mechanisms of the observed *in vitro* results. These *in silico* approaches mainly explored the effect of conformational changes on ligand and ion coordination stability. We demonstrate the application of a combined *in silico* and *in vitro* approach to characterize EAAT1 variants, which could aid drug discovery efforts.

Results

Cancer-related mutations are widespread across the EAAT1 structure

Somatic mutations in EAAT1 are found in cancer patients suffering from different cancer types. Across all cancer types in the Genomic Data Commons (GDC)²⁹, 105 unique EAAT1 mutations were identified primarily located in uterine cancer (29 mutations) followed by lung cancer and melanoma (21 mutations each) and colon cancer (11 mutations). The frequency of these unique mutations is comparable to natural variance occurrence (1.18% vs. 1.75%, respectively), and they are widespread across the EAAT1 structure without any specific mutational pattern observed per cancer type (**Supplementary Figure 6.1**). However, most EAAT1 mutations found in cancer patients are not present in natural variance, and some of them are found in structural domains in which conformational rearrangements could lead to transport function impairment. For example,

there are mutations located in the vicinity of the binding sites occupied by the substrate and coordinating Na^+ ions, as well as in the HP2 domain (**Supplementary Figure 6.1**). Moreover, certain mutations found in cancer patients are located in the binding pockets occupied by orthosteric and allosteric EAAT1 inhibitors, which could lead to changes in their binding affinity and potency. Twelve mutations not present in natural variance that were found in the functional and binding domains mentioned above (Y127C, V247F, C252F, R388K, F389L, V390M, P392L, I397V, A446E, A446V, L448Q, and R479W) were shortlisted to characterize their effect with a combination of *in silico* and *in vitro* methods (**Figure 6.1**).

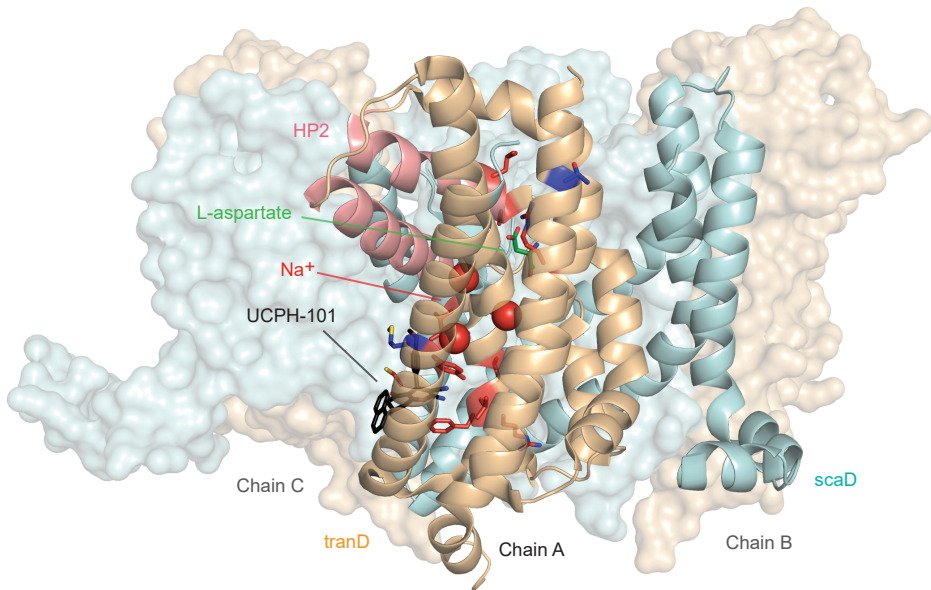


Figure 6.1. Structural distribution of cancer- and ataxia-related mutants in EAAT1 functionally relevant domains presented in this study. Cancer-related mutations (Y127C, V247F, C252F, R388K, F389L, V390M, P392L, I397V, A446E, A446V, L448Q, and R479W) are mapped in red onto chain A of the EAAT1 trimer (PDB 7AWM). Ataxia-related mutations (M128R and T318A) are mapped in dark blue onto chain A. Chains B and C are represented as surfaces. Protein domains are color-coded as follows: tranD domain (orange), scaD domain (cyan), and helical hairpin 2 (HP2) domain (red). The co-crystallized substrate, L-aspartate, is represented in green sticks in chain A. The three coordinated Na^+ ions are represented as red spheres in chain A. The allosteric inhibitor UCPH-101 is represented in black sticks.

EAAT1 mutants are predicted to have a local effect on substrate and inhibitor binding affinity

The effect on ligand binding affinity of cancer-related mutants found in the orthosteric and allosteric binding sites of EAAT1 was tested *in silico* to prioritize mutations for *in vitro* testing. Changes in binding energy $\Delta\Delta G_{\text{bind}}$ were calculated for two endogenous substrates (L-aspartate and L-glutamate), one competitive “orthosteric” inhibitor (TFB-TBOA), and one non-competitive “allosteric” inhibitor UCPH-101 (**Table 6.1**). Since the method employed short-range Monte Carlo sampling, the analysis was restricted to

mutants in the vicinity of the ligand of interest and classified the mutants as “orthosteric” (V247F, P392L, A446E, A446V, L448Q, and R479W, **Figure 6.2a,b**) and “allosteric” (Y127C, V247F, C252F, R388K, F389L, V390M, and I397V, **Figure 6.2c,d**). A positive $\Delta\Delta G_{\text{bind}}$ over 1 kcal/mol can be interpreted as a significant decrease in binding affinity, while a negative $\Delta\Delta G_{\text{bind}}$ below −1 kcal/mol can be interpreted as a significant increase in binding affinity (**Table 6.1**)³².

Table 6.1. Binding energy changes ($\Delta\Delta G_{\text{bind}}$) predicted in ICM-Pro for EAAT1 orthosteric and allosteric mutants. ^a $\Delta\Delta G_{\text{bind}}$ was calculated for the endogenous substrates L-aspartate and L-glutamate and for the competitive inhibitor TFB-TBOA for orthosteric EAAT1 mutants. The systems used were chain A of PDB 5LLU (with L-aspartate co-crystallized and L-glutamate docked), and chain A of PDB 5MJU (with TFB-TBOA co-crystallized). ^b For the allosteric mutants, $\Delta\Delta G_{\text{bind}}$ was calculated for the allosteric inhibitor UCPH-101 in Chain A of PDB 5MJU. ^c V247F is situated between the orthosteric and allosteric sites.

Orthosteric mutants				Allosteric mutants	
	$\Delta\Delta G_{\text{bind}}$ (kcal/mol) ^a				$\Delta\Delta G_{\text{bind}}$ (kcal/mol) ^b
	L-aspartate	L-glutamate	TFB-TBOA		UCPH-101
V247F ^c	0.52	0.08	-0.70	Y127C	5.82
P392L	0.04	-0.01	-0.70	V247F ^c	0.68
A446E	6.39	-0.90	1.86	C252F	-0.49
A446V	0.58	-1.73	2.23	R388K	-0.05
L448Q	-0.35	-1.88	1.79	F389L	3.83
R479W	7.13	6.42	42.19	V390M	-0.76
-	-	-	-	I397V	-0.62

Within the orthosteric mutants, a substantial increase in $\Delta\Delta G_{\text{bind}}$ values was observed in mutant R479W for both endogenous substrates and especially for the inhibitor TFB-TBOA, which indicates highly unfavorable binding of these ligands. V247F and P392L did not show significant changes as these residues are further away from the substrate’s binding site, but an incipient increased binding affinity towards TFB-TBOA was observed. A446V and L448Q, and to a lesser extent A446E, showed an increased binding affinity towards L-glutamate. Interestingly, while both A446 mutants displayed a reduced TFB-TBOA affinity, A446E and A446V showed a different profile for the two endogenous substrates. A substantial loss of binding affinity towards L-aspartate was observed in A446E, but not A446V. Within the allosteric mutants, Y127C and F389L showed a significant decrease in binding affinity towards UCPH-101. V390M showed the biggest increase in binding affinity, although this change in $\Delta\Delta G_{\text{bind}}$ was not significant.

Based on these results, five orthosteric (P392L, A446E, A446V, L448Q, and R479W) and two allosteric mutants (Y127C and V390M) were selected for *in vitro* testing based on their differential $\Delta\Delta G_{\text{bind}}$ profiles. Moreover, V247F was included in the selection since it was considered to be at the interface of both binding pockets. Of the selected residues, Y127, V390, P392, A446, L448, and R479 are fully conserved in mammalian EAATs, as well as the archaeal glutamate transporter homolog Glt_{ph} (except V390 and L448), which suggests the relative importance of these residues in protein function (**Supplementary Figure 6.2**). To validate the *in vitro* assay, two additional EA6-associated EAAT1 mutations were selected that have been reported to either completely abolish glutamate

transport (M128R) or have unaltered transport (T318A). Neither of these two residues are conserved in other glutamate transporters (**Supplementary Figure 6.2**). M128 is adjacent to Y127 and in close proximity to the binding site of UCPH-101, whereas T318 is not in the vicinity of ligand binding sites (**Figure 6.2**).

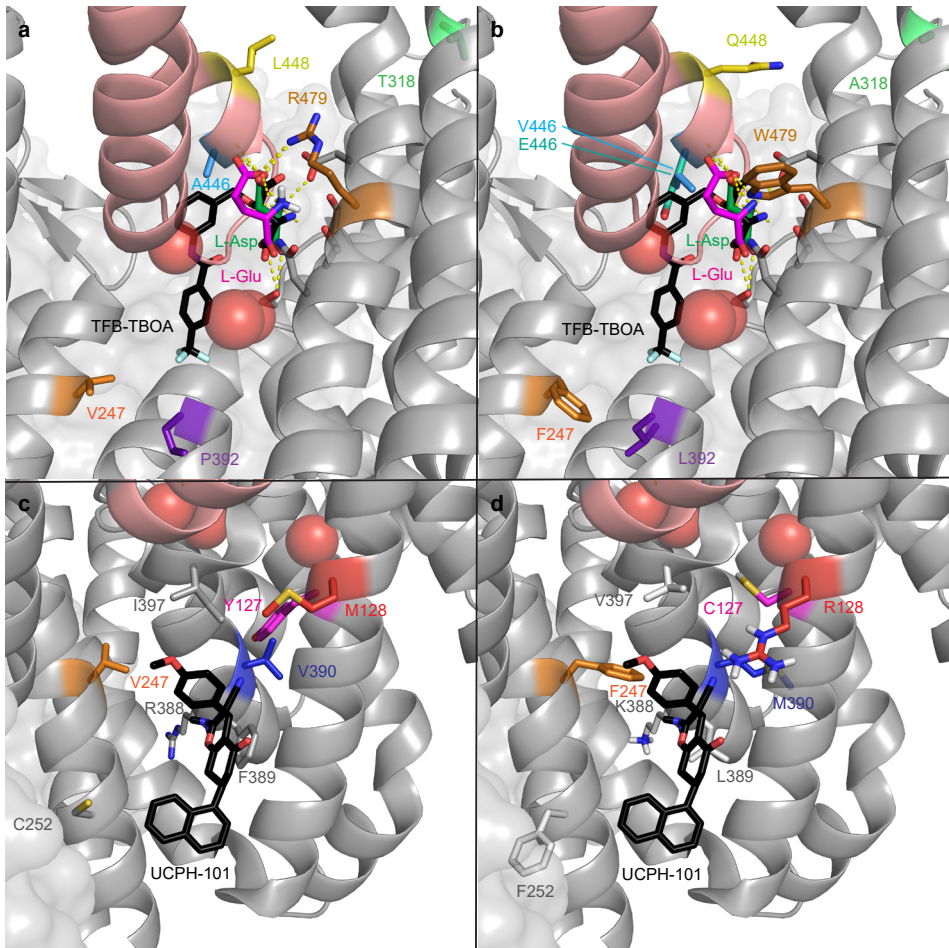


Figure 6.2. EAAT1 disease-related mutations in the orthosteric and allosteric binding sites. Mutations are mapped onto chain A of PDB 7AWM. Thermostabilizing mutations C252V and T318M were reverted in 7AWM for $\Delta\Delta G_{\text{bind}}$ calculation and visualization purposes. For spatial reference, the helical hairpin 2 (HP2) domain helices are colored salmon. The three coordinated Na^+ ions are represented as red spheres. **a)** WT residues where mutations have been found in cancer in the orthosteric binding site of EAAT1. Ataxia-related reference mutation T318A is visualized in light green. The co-crystallized substrate, L-aspartate, is represented as green sticks. The docked substrate, L-glutamate, is represented in magenta. The competitive inhibitor TFB-TBOA is represented as black sticks and superimposed to the 7AWM structure from its position in PDB 5MJU. Polar contacts between the substrate and EAAT1 are represented as dashed yellow lines. **b)** Mutated residues in the orthosteric binding site of EAAT1. **c)** WT residues where mutations have been found in cancer in the allosteric binding site of EAAT1. Ataxia-related reference mutation M128R is visualized in red. The co-crystallized allosteric inhibitor UCPH-101 is represented as black sticks. **d)** Mutated residues in the allosteric binding site of EAAT1.

EAAT1 mutants respond differentially to substrates in a phenotypic assay

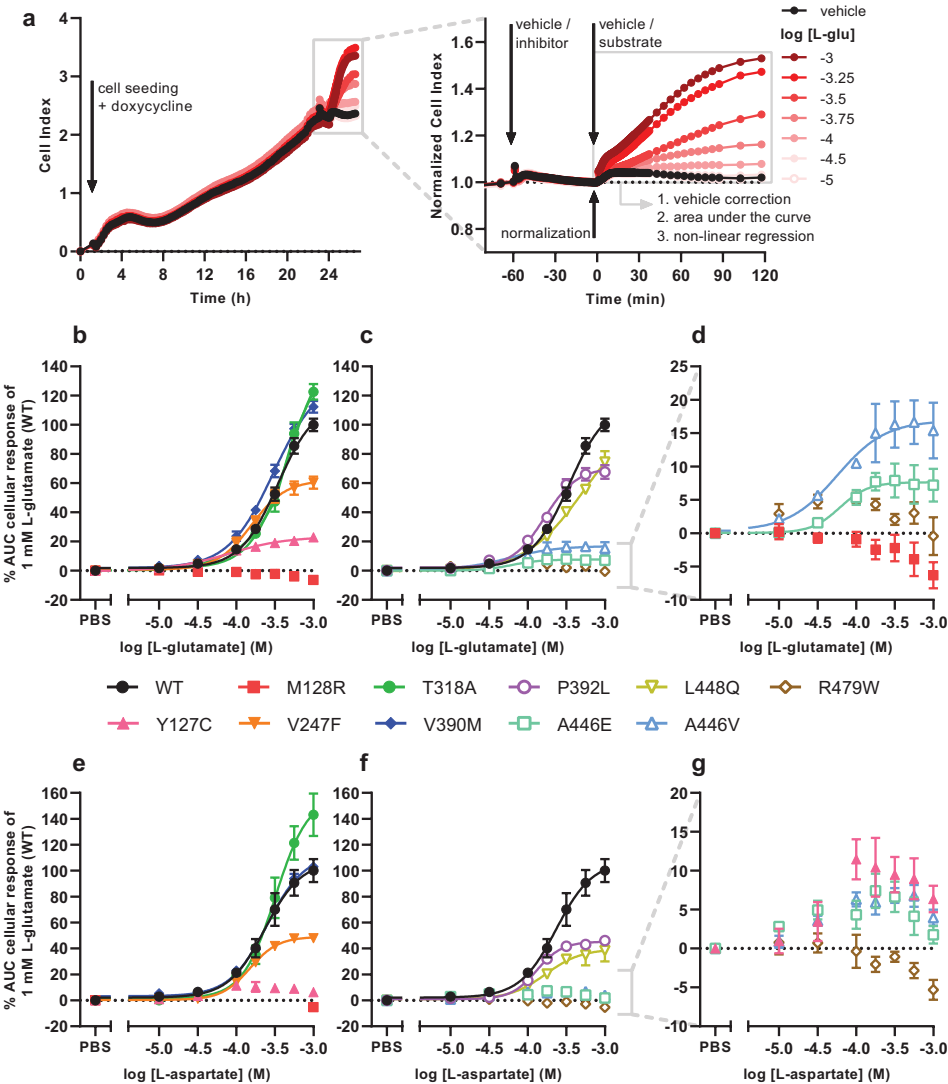


Figure 6.3. Cellular responses of L-glutamate and L-aspartate in an impedance-based phenotypic assay on EAAT1_{WT} and mutant cells. **a)** Illustrative graph of the assay and analysis procedure. EAAT1_{WT} cells are seeded and grown for 24 h in the presence of 1 µg/ml doxycycline to induce EAAT1 expression. Cells are pretreated with vehicle (PBS/DMSO) or inhibitor (TFB-TBOA or UCPH-101, only in Figure 6.4) for 60 min and subsequently stimulated with vehicle (PBS) or substrate (L-glutamate or L-aspartate) for 120 min. The Cell Index (CI) is normalized prior to substrate stimulation and the cellular response is quantified by analyzing the net area under the curve (AUC). **b-g)** Concentration-response curves of **(b-d)** L-glutamate and **(e-g)** L-aspartate on EAAT1_{WT} cells and **(b,e)** ataxia and allosteric site mutants and **(c,f)** orthosteric site mutants. **d,g)** Zoom-in on mutants with low maximal cellular responses. Cellular response is expressed as the net AUC of the first 120 min after L-glutamate or L-aspartate stimulation. Graphs are normalized to the response of 1 mM L-glutamate or L-aspartate on EAAT1_{WT} cells. Data are shown as the mean ± SEM of three to seven individual experiments each performed in duplicate.

To assess the selected mutants for their function *in vitro*, a series of HEK293 JumpIn cell lines were generated and modified to stably express either WT (EAAT1_{WT}) or mutant EAAT1 upon induction with 1 µg/ml doxycycline for 24 h. None of the ten mutants showed either a decreased or increased expression of the HA-tagged EAAT1 compared to EAAT1_{WT} after doxycycline treatment, indicating that the mutations did not affect the translation of the transgene (**Supplementary Figure 6.3**).

To assess whether the EAAT1 mutants affect transporter functionality, an impedance-based phenotypic assay was used. In this set-up, adherent cells (over)expressing EAAT1 are cultured on gold-plated electrodes in a 96-well E-plate. Upon stimulation with high concentrations (10 µM – 1 mM) of substrate (i.e., L-glutamate or L-aspartate) the cells started spreading as a result of Na⁺-dependent substrate uptake via EAAT1 and subsequent cell spreading. The expanded electrode coverage by the cells generated an increase in impedance over time, which was expressed as Cell Index (CI) and interpreted as a readout of EAAT1 function (**Figure 6.3a**). Growth curves were recorded prior to inhibitor pretreatment and substrate stimulation and all mutants displayed similar CI traces compared to EAAT1_{WT}, which suggested that the presence of mutant EAAT1 did not substantially affect cell adhesion or proliferation during the experiments (**Supplementary Figure 6.4**). L-glutamate induced a concentration-dependent cellular response in EAAT1_{WT} (pEC₅₀ = 3.5 ± 0.0), which was reflected by a gradual increase of the normalized Cell Index (nCI) in the first 120 min after substrate stimulation (**Figure 6.3a-d**, **Table 6.2**). A comparable L-glutamate potency was observed for the EA6 mutant T318A (pEC₅₀ = 3.3 ± 0.0) with a slightly increased maximal response (E_{max}), whereas the L-glutamate response was completely abolished for M128R (**Figure 6.3b,d**). The allosteric site mutants V247F (pEC₅₀ = 3.8 ± 0.0) and V390M (pEC₅₀ = 3.5 ± 0.0) produced similar L-glutamate potencies compared to EAAT1_{WT}, where V247F has a 62% reduced E_{max} (**Figure 6.3b**). The potency of L-glutamate on Y127C was enhanced (pEC₅₀ = 4.1 ± 0.1) but displayed a substantial drop (94%) in E_{max} (**Figure 6.3b**). The orthosteric site mutants P392L (pEC₅₀ = 3.8 ± 0.0) and L448Q (pEC₅₀ = 3.3 ± 0.1) showed no significant change in L-glutamate potency, although the concentration-effect curve for L448Q appeared more linear and shifted rightward and did not appear to reach a maximum within the tested concentration range (**Figure 6.3c**). Both A446E and A446V produced glutamate responses with a strongly reduced E_{max}, but with significantly enhanced L-glutamate potency (pEC₅₀ = 4.4 ± 0.3 and 4.3 ± 0.2, respectively), whereas no concentration-dependent L-glutamate response was observed for R479W (**Figure 6.3c,d**).

Next, the responsiveness of the EAAT1 mutants to the endogenous substrate L-aspartate was assessed. L-aspartate induced a concentration-dependent cellular response in EAAT1_{WT} (pEC₅₀ = 3.6 ± 0.1) similar to L-glutamate (**Figure 6.3e**). The potency of L-aspartate was comparable in the EA6 mutant T318A (pEC₅₀ = 3.5 ± 0.0) with an elevated E_{max}, whereas in M128R no L-aspartate response was observed at 1 mM (**Figure 6.3e**). The response of L-aspartate in V390M (pEC₅₀ = 3.6 ± 0.0) was identical to EAAT1_{WT} (**Figure 6.3e**). The mutants V247F (pEC₅₀ = 3.8 ± 0.0), P392L (pEC₅₀ = 3.9 ± 0.0) and L448Q (pEC₅₀ = 3.7 ± 0.1) produced similar L-aspartate potencies, but a substantially lowered E_{max} (~60%) compared to EAAT1_{WT} (**Figure 6.3e,f**). For

Y127C, A446E, and A446V the maximal L-aspartate response was reduced. Although the L-aspartate response increases at low substrate concentrations, it dropped at high concentrations, resulting in a bell-shaped concentration-effect curve from which no pEC_{50} and E_{max} were calculated (**Figure 6.3e-g**). Similar to L-glutamate, no L-aspartate response was observed for R479W (**Figure 6.3f,g**). Collectively, these data demonstrate that the selected EAAT1 mutants impact L-glutamate and L-aspartate transport.

Table 6.2. Potencies (pEC_{50}) of L-glutamate and L-aspartate and inhibitory potencies (pIC_{50}) of TFB-TBOA and UCPH-101 on Jumpln-EAAT1_{WT} and mutant cells in an impedance-based phenotypic assay. ^a Maximal responses (E_{max}) are normalized to the cellular response of 1 mM L-glutamate or L-aspartate (100%) on Jumpln-EAAT1_{WT} cells.

	L-glutamate		L-aspartate		TFB-TBOA	UCPH-101
	pEC_{50} (log M)	E_{max}^a (%)	pEC_{50} (log M)	E_{max}^a (%)	pIC_{50} (log M)	pIC_{50} (log M)
WT	3.5 ± 0.0	117 ± 5	3.6 ± 0.1	108 ± 9	6.7 ± 0.1	5.4 ± 0.0
Y127C	4.1 ± 0.1 ***	23 ± 3	N.D.	N.D.	6.2 ± 0.0 *	N.D.
M128R	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
V247F	3.8 ± 0.0	55 ± 9	3.8 ± 0.0	49 ± 1	5.7 ± 0.1 ****	5.3 ± 0.0
T318A	3.3 ± 0.0	156 ± 4	3.5 ± 0.0	158 ± 18	6.9 ± 0.1	5.4 ± 0.0
V390M	3.5 ± 0.0	132 ± 6	3.6 ± 0.0	112 ± 3	6.7 ± 0.0	5.4 ± 0.0
P392L	3.8 ± 0.0	71 ± 4	3.9 ± 0.0	46 ± 3	6.5 ± 0.1	N.D.
A446E	4.4 ± 0.3 ****	8 ± 2	N.D.	N.D.	7.4 ± 0.2 **	5.9 ± 0.2
A446V	4.3 ± 0.2 ****	16 ± 4	N.D.	N.D.	N.D.	N.D.
L448Q	3.3 ± 0.1	116 ± 25	3.7 ± 0.1	47 ± 13	7.9 ± 0.0 ****	5.9 ± 0.1 **
R479W	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.

EAAT1 inhibitors induce cellular response in M128R mutant

To assess whether the selected mutants modulated the effects of the competitive (“orthosteric”) inhibitor TFB-TBOA and the non-competitive (“allosteric”) inhibitor UCPH-101, the cells were pretreated for 1 h with increasing concentrations of inhibitor prior to stimulation with 1 mM L-glutamate. In EAAT1_{WT}, inhibitor pretreatment itself did not result in substantial changes in the nCI (**Supplementary Figure 6.5c-f**). Strikingly, the M128R pretreatment with TFB-TBOA resulted in a concentration-dependent sharp nCI increase which peaked after 10-30 min, whereas pretreatment with UCPH-101 induced a more gradual nCI increase that plateaued after 60 min (**Supplementary Figure 6.5a,b**). These inhibitor responses were not observed in any of the other mutants, although V247F, A446E, and A446V showed concentration-dependent decreases of the nCI upon TFB-TBOA pretreatment, which were substantially lower in magnitude compared to M128R (**Supplementary Figure 6.5d,f**). This suggests that M128R displays a distinct physiological phenotype compared to EAAT1_{WT} and other mutants.

To elucidate a potential mechanism behind the M128R response to both inhibitors, it was assessed whether the inhibitors displayed any interaction with each other or the

substrate L-glutamate. Indeed, cells pretreated with TFB-TBOA were responsive to subsequent stimulation with UCPH-101 and vice-versa, indicating that the cellular responses elicited by either inhibitor are additive and are constituted by independent mechanisms (**Supplementary Figure 6.6a,b**). Interestingly, the response caused by TFB-TBOA pretreatment was completely blocked after stimulation with 1 mM L-glutamate, and a TFB-TBOA response was prevented when cells were pretreated with L-glutamate, indicating that the TFB-TBOA response is transient and originates from interactions at the substrate binding site (**Supplementary Figure 6.6a,c**). In contrast, L-glutamate stimulation after UCPH-101 pretreatment does not reduce the nCI. The UCPH-101 response after L-glutamate pretreatment has a comparable magnitude to the UCPH-101 pretreatment on its own, suggesting that L-glutamate and UCPH-101 do not compete for the same binding site (**Supplementary Figure 6.6b,c**). In addition, the Na⁺/K⁺-ATPase (NKA) inhibitor ouabain prevented any inhibitor- or substrate-induced cellular responses in M128R cells, which indicates that TFB-TBOA and UCPH-101 responses are likely dependent on ion influx (**Supplementary Figure 6.6d**).

EAAT1 mutants alter TFB-TBOA and UCPH-101 inhibition

For EAAT1_{WT} and all other mutants, except M128R, the inhibitory potencies of TFB-TBOA and UCPH-101 were assessed by analyzing the response of 1 mM L-glutamate after 60 min pretreatment with increasing inhibitor concentrations. In EAAT1_{WT}, TFB-TBOA inhibited the L-glutamate response in a concentration-dependent manner ($pIC_{50} = 6.7 \pm 0.1$) (**Figure 6.4a,b**, **Table 6.2**). The EA6 mutant T318A ($pIC_{50} = 6.9 \pm 0.1$), allosteric site mutant V390M ($pIC_{50} = 6.7 \pm 0.0$) and orthosteric site mutant P392L ($pIC_{50} = 6.5 \pm 0.1$) did not affect the inhibitory potency of TFB-TBOA (**Figure 6.4a,b**). Both Y127C ($pIC_{50} = 6.2 \pm 0.0$) and V247F ($pIC_{50} = 5.7 \pm 0.1$) significantly decreased the potency, whereas L448Q ($pIC_{50} = 7.9 \pm 0.0$) significantly enhanced the inhibitory potency of TFB-TBOA (**Figure 6.4a,b**). Interestingly, A446E was susceptible to TFB-TBOA inhibition, showing an increased inhibitory potency ($pIC_{50} = 7.4 \pm 0.2$), whereas A446V as well as R479W did not display any sigmoidal concentration-dependent inhibition by TFB-TBOA (**Figure 6.4b,c**).

The effects of EAAT1 mutants on UCPH-101 inhibition were different from TFB-TBOA. In EAAT1_{WT}, UCPH-101 could inhibit the response of L-glutamate in a concentration-dependent manner ($pIC_{50} = 5.4 \pm 0.0$) (**Figure 6.4d,e**, **Table 6.2**). V247F ($pIC_{50} = 5.3 \pm 0.0$), T318A ($pIC_{50} = 5.4 \pm 0.0$) and V390M ($pIC_{50} = 5.4 \pm 0.0$) did not affect L-glutamate response inhibition by UCPH-101 (**Figure 6.4d**). In Y127C, P392L, A446V, and R479W UCPH-101 was unable to inhibit the L-glutamate response at any of the tested concentrations, indicating a loss of the UCPH-101 interaction (**Figure 6.4d-f**). Similar to TFB-TBOA, both L448Q ($pIC_{50} = 5.9 \pm 0.1$) and A446E ($pIC_{50} = 5.9 \pm 0.2$) enhanced the inhibitory potency of UCPH-101, although this was not significant for A446E ($p = 0.0919$) (**Figure 6.4e,f**). Taken together, these data imply that the selected EAAT1 mutants differentially modulate both substrate and EAAT1 inhibitor interactions.

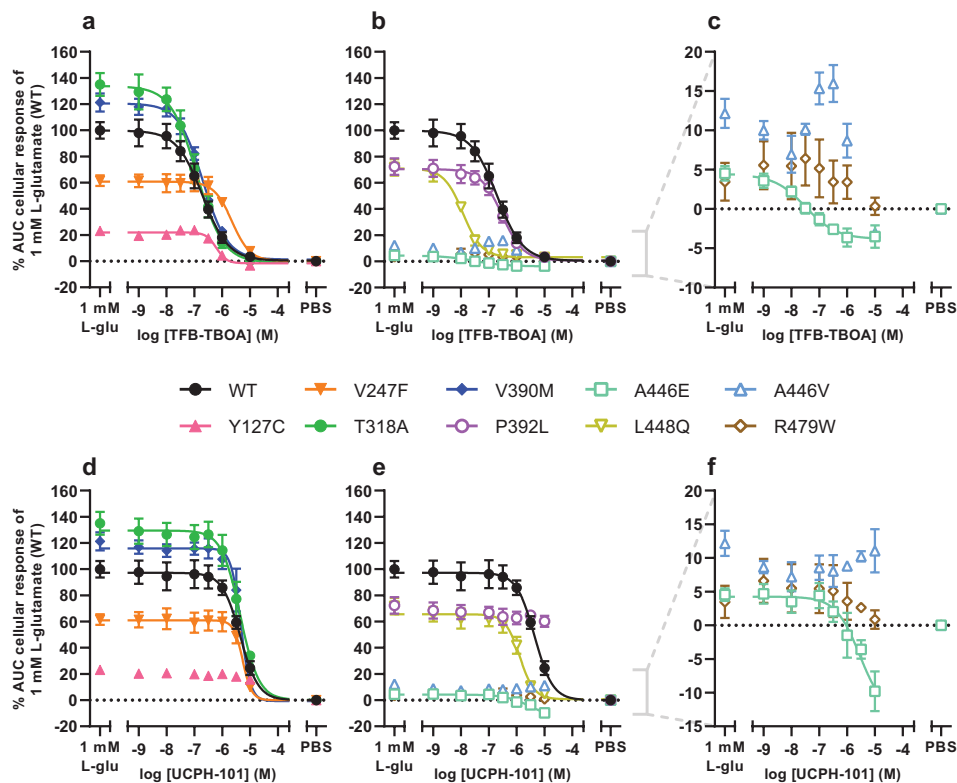


Figure 6.4. Inhibition of L-glutamate responses by TFB-TBOA and UCPH-101 in an impedance-based phenotypic assay on EAAT1_{WT} and mutant cells. **a-f)** Concentration-inhibition curves of **(a-c)** TFB-TBOA and **(d-f)** UCPH-101 on EAAT1_{WT} cells and **(a,d)** ataxia and allosteric site mutants, and **(b,e)** orthosteric site mutants. **c,f)** Zoom-in on mutants with low maximal cellular responses. Cells were pretreated with TFB-TBOA, UCPH-101, or vehicle (PBS/DMSO) for 60 min and stimulated with a submaximal concentration (EC₈₀) of 1 mM L-glutamate or vehicle (PBS) for 120 min. Cellular response is expressed as the net AUC of the first 120 min after L-glutamate stimulation and graphs are normalized to the response of 1 mM L-glutamate on EAAT1_{WT} cells. Data are shown as the mean ± SEM of three individual experiments each performed in duplicate.

EAAT1 mutants alter transporter conformation and substrate stability over time

To assess the effect of EAAT1 mutants in transporter and substrate stability, ten replicates of 500 ns MD trajectories were simulated for the WT and seven mutants that showed differential behavior *in vitro* (Y127C, M128R, P392L, A446E, A446V, L448Q, and R479W). The simulations started from the endogenous substrate L-aspartate-bound conformation, with coordinated Na⁺ ions in sites Na1-3 and closed HP2 domain. This represents the transporter conformation prior to its transition to the inward-facing conformation. The stability of this conformation was followed over time in regards to the system overall (i.e. protein RMSD), the substrate in the binding site (i.e. ligand RMSD in respect to protein), the opening of the HP2 domain (i.e. distance between the HP1 and HP2 domain tips), and coordination of the Na⁺ ions (i.e. distance between Na⁺ ion and

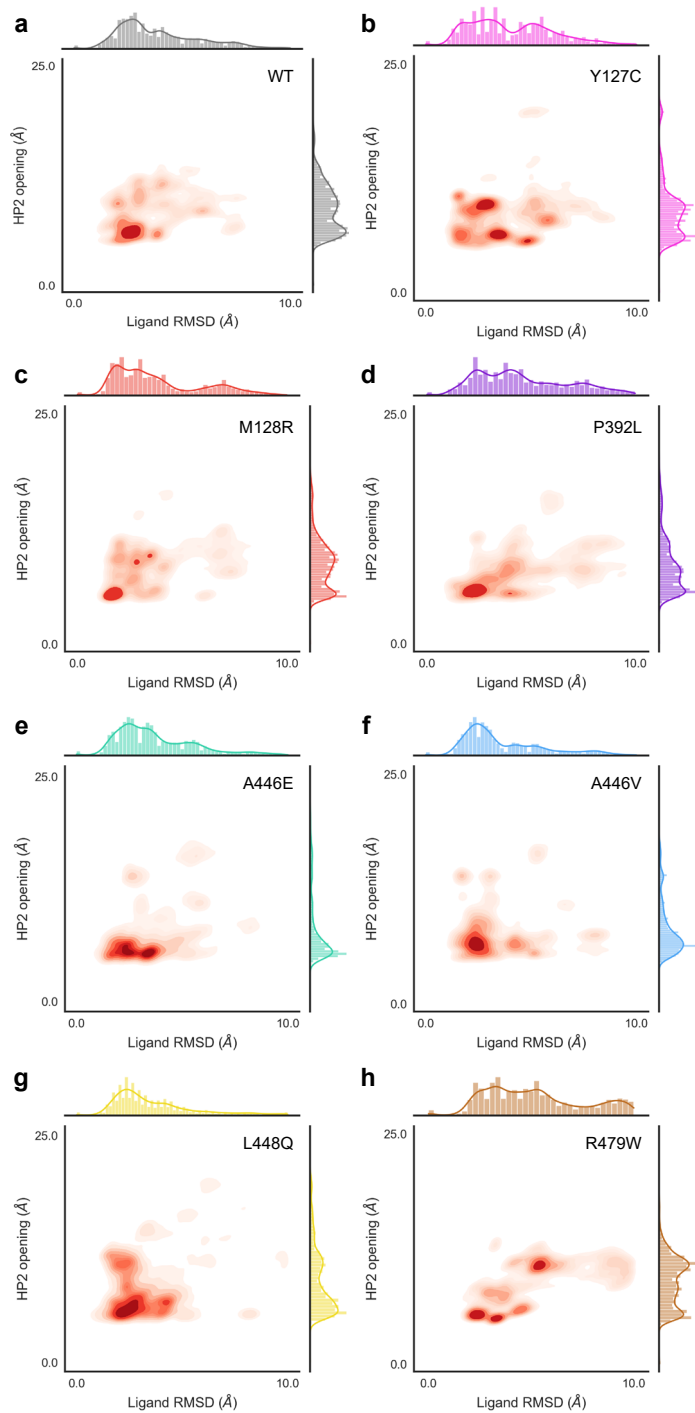


Figure 6.5 (caption on the following page)

► **Figure 6.5.** HP2 domain opening and L-Asp substrate stability sampling density derived from MD simulations on EAAT1_{WT} and mutants. HP2 opening was calculated as the distance between S366 C α (HP1 tip) and G442 C α (HP2 tip). Substrate (L-Asp) stability is represented by ligand RMSD respective to the protein. Sampling density was calculated across all frames in all replicates simulated for HP2 opening and substrate stability in combination (inside the axes box) and independently (outside the axes) for EAAT1_{WT} **(a)** and mutants **(b-h)**.

one coordinating atom). Compared to WT (**Figure 6.5a**), mutants A446E, A446V, and L448Q (**Figure 6.5e-g**) showed a similar high ligand stability (i.e. low ligand RMSD), which correlated with a stabilized “closed” HP2 conformation. HP2 domain closure was especially pronounced in A446E and A446V mutants compared to WT. On the contrary, ligand instability was higher in mutants Y127C, P392L, and R479W (**Figure 6.5b,d,h**), which correlated with increased opening of the HP2 domain, particularly in R479W. In R479W, substrate instability was also directly linked to the loss of key interactions of L-aspartate in the binding pocket, mainly with R479 and T402 (**Supplementary Figure 6.7, 6.8**). Mutant M128R (**Figure 6.5c**) showed a very similar distribution to WT both in terms of HP2 opening and ligand stability, which suggests that the mutation in M128 does not directly affect the conformation of the orthosteric binding site.

While the mutant effects on transporter conformation (i.e. HP2 opening) affected ligand stability, they barely had an impact on Na⁺ ion coordination. Firstly, from the MD simulations, it was observed that the Na⁺ ions coordinated in sites Na1 and Na3 were extremely stable in the WT system and all mutants simulated (**Supplementary Figure 6.9a-h**). In particular, mutant M128R seemed to heavily restrict movement for the Na⁺ ion coordinated in position Na3 compared to the rest of the mutants (**Supplementary Figure 6.9c**). On the contrary, the ion occupying site Na2, which is coordinated in the last place before HP2 closure, was highly unstable across the board (**Supplementary Figure 6.9i-p**). Compared to WT, Na2 was more unstable in mutants A446V and L448Q (**Supplementary Figure 6.9n,o**). However, Na⁺ coordination instability in the Na2 site was not correlated to HP2 opening, since ion instability was observed both at lower and higher HP2 opening distances.

EAAT1 mutant-driven conformational changes impact inhibitor docking binding poses

To evaluate whether the conformational changes in the HP2 domain observed upon mutation affect inhibitor binding as they do substrate coordination, molecular docking was performed per mutant in a representative selection of five frames from the MD trajectories (**Figure 6.6**). The selected frames represented the most common HP2 opening distances per mutant: 6.0 ± 0.2 Å (WT), 6.6 ± 0.7 Å (Y127C), 5.2 ± 0.1 Å (M128R), 7.0 ± 0.2 Å (P392L), 5.4 ± 0.2 Å (A446E), 5.4 ± 0.1 Å (A446V), 5.6 ± 0.2 Å (L448Q), and 10.5 ± 0.1 Å (R479W), but had different orthosteric and allosteric pocket conformations (**Supplementary Table 6.1,6.2**). The highest scoring poses in TFB-TBOA docking roughly maintained the position and polar interactions of the aspartic acid moiety observed in the co-crystallized conformation (**Supplementary Figure 6.10**). The rest of the molecule, however, could be flipped around the two contiguous chiral centers

to different positions depending on the exact conformation of the HP2 domain. This behavior was observed for the WT (**Figure 6.6a**) and mutants Y127C (**Figure 6.6b**), A446V (**Supplementary Figure 6.10g**), and L448Q (**Supplementary Figure 6.10h**). The lower scoring pose on mutant A446E (**Supplementary Figure 6.10f**) also maintained the aspartic acid moiety position, but the rest of the molecule was forced into a less stable conformation due to the HP2 configuration induced by E446 interactions. None of the lowest-scoring poses in mutants M128R, P392L, and R479W maintained the aspartic acid moiety position. In mutant R479W (**Figure 6.6c**) this effect was due to the less flexible and bulkier side chain of W479, which pushed TFB-TBOA deeper in the pocket causing the loss of key interactions (**Supplementary Figure 6.7,6.8**).

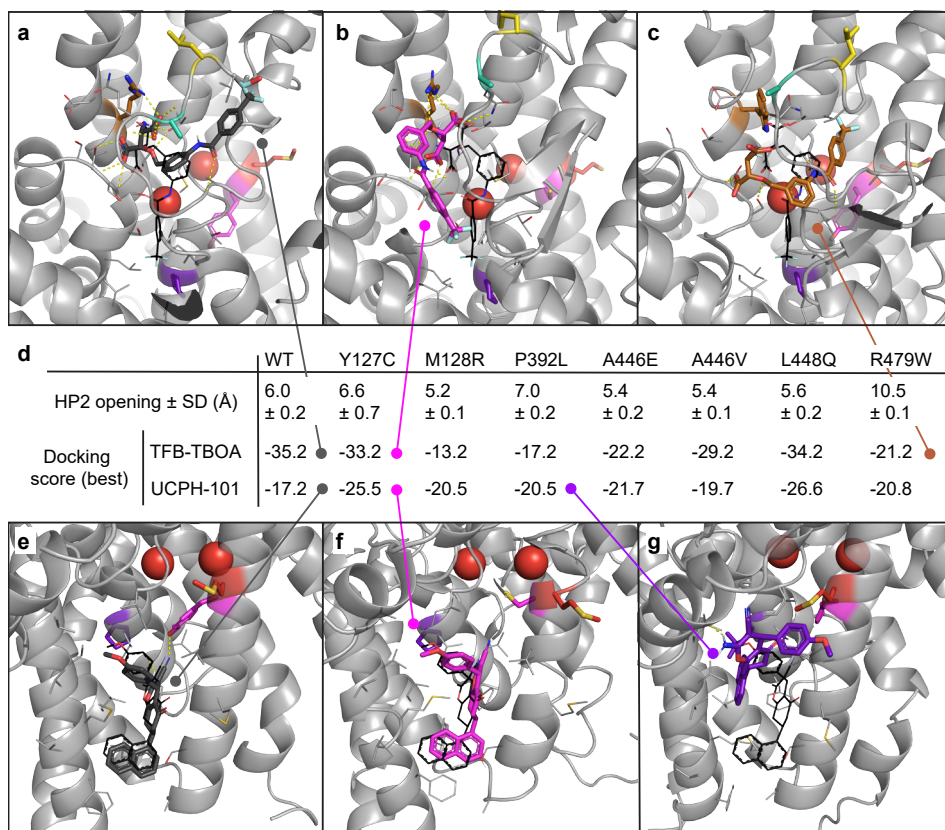


Figure 6.6. Molecular docking of inhibitors TFB-TBOA and UCPH-101 in EAAT1 MD frames with most representative HP2 opening distances. Docking was performed in chain A of a random selection of frames with the top five most common HP2 opening distances across all replicates and frames. **a-c**) Top docking poses of orthosteric inhibitor TFB-TBOA in EAAT_{WT} (**a**) and mutants Y127C (**b**) and R479W (**c**). TFB-TBOA binding pocket was derived from its co-crystallized pose in PDB 5MJU, represented in black for reference. **d**) Mean HP2 opening distance in the five frames selected from MD for docking. Docking scores of the top poses in EAAT_{WT} and mutants. **e-g**) Top docking poses of allosteric inhibitor UCPH-101 in EAAT_{WT} (**e**) and mutants Y127C (**f**) and P392L (**g**). UCPH-101 binding pocket was derived from its co-crystallized pose in PDB 7AWM, represented in black for reference. Na⁺ ions are represented as red spheres. Hydrogen bonds are represented with dashed yellow lines.

Compared to TFB-TBOA, the binding of allosteric inhibitor UCPH-101 was less affected by mutations as represented by the range in docking scores (**Figure 6.6d**) and poses (**Supplementary Figure 6.11**). The pose observed in co-crystallized structures was maintained in the top docking poses in WT (**Figure 6.6e**) and mutants Y127C (**Figure 6.6f**) and L448Q (**Supplementary Figure 6.11h**). The top poses in mutants Y127C and L448Q also showed a higher docking score (-25.5 and -26.6, respectively) compared to WT (-17.2), although only the pose on L448Q maintained one of the two hydrogen bonds in the co-crystallized pose to P389. UCPH-101 docked in mutant A446E (**Supplementary Figure 6.11f**) occupied the same region but the pose was flipped compared to WT. Docking poses in mutants M128R, P392L, A446V, and R479W (**Supplementary Figure 6.11d,e,g,i**) did not reach the allosteric pocket deeply enough to make relevant interactions. In the case of mutants M128R and R479W, there seemed to be a closure of the binding pocket entrance flanked by TM4c (ScaD) and TM3 (TranD). For P392L, the lower part of the pocket seemed not accessible based on the best docking pose (**Figure 6.6g**). The mutation to Leu in P392 reverted the helix kink that was produced by Pro in that position in the TM7a domain and that stabilized the allosteric binding pocket (**Supplementary Figure 6.12**). Taken together, these results suggest that EAAT1 conformational changes triggered by disease-related mutations affect the way inhibitors TFB-TBOA and UCPH-101 bind to the orthosteric and allosteric pockets, respectively.

Discussion

The role of glutamate and aspartate in cancer is increasingly appreciated³³. Indeed, the regulation of intra- and extracellular levels of these amino acids by EAATs and other transporters, with respect to the tumor microenvironment, is the subject of ongoing investigations. So far, the altered function of EAAT1 as a result of single missense mutations has been linked to several extremely rare cases of episodic ataxia type 6 (EA6)¹⁶. However, there have been no reports on the contribution of genetic variants of EAATs to the development of cancer, and it remains a question to what degree loss- or gain-of-function mutations in these transporters are relevant for disease progression. In this study, 105 unique somatic mutations were identified in cancer patients, none of which occurred as natural variants. Eight cancer-associated and two reference EA6-related EAAT1 missense mutants were analyzed in a label-free phenotypic assay, which together with structural insights provides an initial understanding of altered transporter function and cell behavior.

All EAAT1 mutants were expressed at similar relative levels compared to EAAT1_{WT}, therefore not affecting protein translation (**Supplementary Figure 6.3**). Interestingly, in previous studies, several EAAT1 mutants displayed attenuated or increased glutamate uptake activity as a result of reduced (P290R, M128R^{16,34}) or enhanced (E219D, T318A^{16,19}) surface membrane density, respectively. Indeed, in our functional assay, T318A showed a considerable increase in substrate E_{\max} (**Figure 6.3b,e**, **Table 6.2**), which may be attributed to enhanced membrane insertion of EAAT1¹⁶. Most other mutants displayed a substantial decrease in substrate E_{\max} , with the maximal response being

generally lower for L-aspartate than L-glutamate.

Tyr at position 127 is located in TM3 and is conserved in all human EAATs and the archaeal Glt_{ph} (**Supplementary Figure 6.2**), where the backbone carboxylate of Tyr is part of the third Na⁺ binding site (Na3)^{23,25,27}. Substitution of Y127 to Cys does not affect the ability of EAAT1 to translocate substrate, albeit with a substantially reduced E_{\max} (**Figure 6.3b**). In addition to forming Na3, Y127 forms a hydrogen bond with the carbonitrile group of UCPH-101²². The docking studies suggest that this bond cannot form in Y127C (**Figure 6.6f**). However, the Y127C mutation seems to lead to an opening of the TM3 helix and widening of the tranD-scaD interface pocket that makes it less suitable for blocking the elevator mechanism, which might be related to the loss of UCPH-101 inhibition (**Figure 6.4d**). In line with this, mutation of Y127 to Phe, Leu, Ile, or Arg showed a significant drop in pIC₅₀ of UCPH-101 in a [³H]-D-aspartate uptake assay³⁵.

M128 is adjacent to Y127 and is exposed to membrane lipids. The M128R mutation was found in an EA6 patient and patch clamp experiments demonstrated that M128R shows a complete loss of glutamate uptake as well as abolished anion currents that could not be explained by slightly reduced surface expression levels¹⁶. Indeed, no L-glutamate or L-aspartate responses in M128R (**Figure 6.3b,e**) were detected, which suggests that this mutant is likely transport-incompetent. Surprisingly, substantial concentration-dependent positive cellular responses were observed when M128R cells were treated with TFB-TBOA or UCPH-101, which were not observed in EAAT1_{WT} or other mutants (**Supplementary Figure 6.5**). Although our computational studies did not shed any light on the potential mechanism of the observed behavior (**Figure 6.5c, Supplementary Figure 6.9c,k**), a recent study demonstrated that mutation of M128 to Arg may inflict two potential disruptions to EAAT1³⁶. The positively charged Arg could flip towards the “inside” of the protein and disrupt the binding of Na⁺ to Na3. Occupation of this site by Na⁺ is crucial to initiate substrate binding and translocation³⁷, which may explain the absence of glutamate transport in M128R. In our simulations, however, a tighter coordination in Na3 was observed. Secondly, the Arg in M128R could flip “outward” towards the lipid bilayer. Other MD studies revealed a local membrane deformation, which recruited a density of water molecules halfway into the bilayer³⁶. This may provide a pathway for Na⁺ ions that enter the Na3 site to leak into the cytosol, which could result in cell volume increase and subsequent morphological changes³¹. Thus, we hypothesize that binding of TFB-TBOA or UCPH-101 to EAAT1 M128R stabilizes an Arg “outward” conformation that allows uncoupled Na⁺ influx, which results in a phenotypic response in the absence of substrate (**Supplementary Figure 6.5, 6.6**). To our knowledge, this is the first report of inhibitor-induced functional responses in glutamate transporters, which warrants further investigation and could hold promise for future therapeutic strategies.

The second episodic ataxia-derived mutant, T318A, showed no signs of affecting EAAT1 transporter function other than an increased substrate E_{\max} , in line with the lack of evidence of its pathogenicity^{36,38}. In other studies, mutation to Ala increased glutamate uptake and anion currents as a result of increased surface expression of the

transporter^{16,36}. A similar conservative effect was found for both Val 247 and 390, which are located adjacently to hydrophobic residues conferring the selectivity of UCPH-101 towards EAAT1²². However, mutations V247F and V390M did not affect substrate translocation (**Figure 6.3**) or UCPH-101 binding (**Figure 6.4d**), indicating that these residues are not crucial for inhibitor binding. Interestingly, TFB-TBOA's inhibitory potency was reduced in V247F (**Figure 6.3**), possibly due to the increased residue bulkiness affecting the hydrophobic cavity size.

The Pro at position 392 is located in TM7a near V390 and is completely conserved throughout the SLC1 family and Glt_{ph}²². P392 is part of the scaD–tranD interface that lines the hydrophobic cavity of the chloride conductive pathway^{39,40}. Mutation of P392 to small hydrophobic residues (Ala, Val) resulted in slightly increased substrate affinities and anion conductances⁴¹, which may be reflected by a small increase in pEC₅₀ for L-glutamate and L-aspartate in P392L (**Table 6.2**). Strikingly, while TFB-TBOA binding is unaffected, P392L causes a complete loss of UCPH-101's inhibition of the L-glutamate response (**Figure 6.4b,e**, **Table 6.2**). As observed in MD simulations, mutation to a slightly bulkier Leu corrects the disruption in the helical turn caused by Pro in TM7a (**Supplementary Figure 6.12**) and promotes an increase in helix rigidity that displaces the location of the nonpolar residues in this region. This substantially reduces the affinity of UCPH-101 for this site, as observed by the loss of the original binding pose in the docking results (**Figure 6.6g**). Interestingly, other EAAT1 Pro mutations have been shown not to revert the kink, as opposed to the original hypotheses⁴².

Three mutations (A446E, A446V, and L448Q) are located in HP2, which is an important structural element that regulates the access of Na⁺ and substrate to their binding sites^{20,22}. In our phenotypic assay, both A446E and A446V displayed vastly reduced maximal substrate responses but significantly increased affinities (**Table 6.2**), which could be the result of low surface expression or a reduced turnover rate⁴³. Tracking the HP2 opening over time suggests that mutations in the HP2 domain increase the stability of a “closed” conformation in the presence of bound L-Asp compared to WT (**Figure 6.5e-g**). Such “closed” conformation could be the result of tighter interactions with the endogenous substrate and lead to reduced transport rate⁴⁴. Notably, mutation to Val at this position abrogates L-glutamate response inhibition, whereas a Glu substitution results in a significantly enhanced potency of TFB-TBOA (**Figure 6.4c,f**, **Table 6.2**). The stabilization of a “closed” HP2 conformation might reduce access to the orthosteric pocket for competitive inhibitors such as TFB-TBOA or, alternatively, induce a higher inhibitory potency by locking in place the aspartic acid moiety²². The differential effects observed for mutants A446E and A446V, however, cannot be explained by the current *in silico* studies, where a more favorable TFB-TBOA binding pose is predicted for A446V compared to A446E (**Supplementary Figure 6.10f,g**). A clear hindrance here is docking the orthosteric inhibitor in a marked HP2 “closed” conformation, when TFB-TBOA is known to stabilize an “open” HP2 conformation in the transporter²².

The adjacent HP2 residue L448 is involved in HP2 backbone flexibility, which is essential for K⁺-dependent re-translocation of the tranD during the transport cycle²⁶. Strikingly, the pIC₅₀ for both TFB-TBOA and UCPH-101 are markedly increased in L448Q. These

results are also supported by the favorable poses generated from the docking studies for both inhibitors (**Supplementary Figure 6.10h, 6.11h**). In a previous study, mutation of L448 to Cys reduced L-glutamate affinity and maximal transport rate, but it significantly enhanced the inhibitory potency of the competitive inhibitor DL-TBOA⁴⁵. The enhanced pIC_{50} for both UCPH-101 and TFB-TBOA may be the result of a reduced affinity of L-glutamate in the orthosteric site, which could augment the apparent inhibitory potency.

The Arg at position 479 confers substrate selectivity and is conserved among glutamate/aspartate transporters. The guanidinium group of R479 forms a hydrogen bond with the sidechain carboxylate of the substrate during translocation²². Moreover, R479 forms a salt bridge with E406 in TM7 during K^+ re-translocation, which sterically hinders closure of HP2 and substrate binding^{23,26}. Neutralization of R479 (i.e., mutation to Ala) renders EAAT1 K^+ -independent and results in drastically reduced glutamate/aspartate affinity²⁶, which was also observed in Glt_{ph} upon mutation of Arg to Cys⁴⁶. As observed in MD simulations, the bulkiness of the indole moiety pushes the HP2 domain to an “open” conformation (**Figure 6.5h**) and disrupts the electrostatic interactions in the binding site (**Supplementary Figure 6.8**), which leads to a loss of substrate activity (**Figure 6.2b**). This local effect was already evident from the relatively high $\Delta\Delta G_{bind}$ values for R479W compared to other mutated residues (**Table 6.2**), which indicates a substantially reduced ligand binding affinity.

Discrepancies observed between the *in vitro* and *in silico* experiments likely arise from the fact that the simulations focused only on a small part of the complex elevator transport cycle and cannot therefore provide a complete mechanism for all the analyzed mutants. Adding to the complexity of the system, heterogeneity was observed among the dynamic behavior of the three protomers, which has been described for glutamate transporter analogs to trigger heterogeneous substrate binding⁴⁷. These results warrant follow-up *in vitro* or *in silico* experiments that investigate alterations in protein solvation, anion conductivity, and substrate transport kinetics^{36,48}, which could help to further explain our functional observations. Moreover, while mutations in a ligand binding site may disrupt or stabilize ligand interactions, they could potentially lead to allosteric effects via disruption of conserved interaction networks⁴⁹.

Conclusions

Taken together, divergent effects of EAAT1 disease-related variants were observed on substrate-induced cellular responses, as well as orthosteric and allosteric inhibition, in an impedance-based phenotypic assay. Subsequent MD simulations and docking studies aided in the formulation of hypotheses that could substantiate the observed *in vitro* effects. Importantly, to allocate these missense variants to a substantial involvement in cancer development and progression translational studies that link genotype to phenotype would be required. Thus, the methods presented in this study may aid in the identification and characterization of pathogenic transporter variants, which may have implications for the development of selective and efficacious therapeutics.

Materials and Methods

Materials

Modified Jump In T-REx HEK 293 (JumpIn) cells overexpressing human wild-type (WT, EAAT1_{WT}) or mutant EAAT1 were kindly provided by the RESOLUTE consortium (Research Center for Molecular Medicine, Medical University of Vienna, Austria). L-glutamic acid monosodium salt monohydrate, L-aspartic acid monosodium salt monohydrate, doxycycline hyclate, Dulbecco's modified Eagle's medium (DMEM) and Dulbecco's phosphate-buffered saline (PBS) were purchased from Sigma Aldrich (St. Louis, MO, USA). 2-amino-4-(4-methoxyphenyl)-7-(naphthalen-1-yl)-5-oxo-5,6,7,8-tetrahydro-4H-chromene-3-carbonitrile (UCPH-101) was purchased from Santa Cruz Biotechnology (Dallas, TX, USA). (2S,3S)-3-[3-[4-(trifluoromethyl)benzoylamino]benzyloxy] aspartate (TFB-TBOA) was purchased from Axon Medchem (Groningen, The Netherlands). Lipofectamine 3000, P3000 buffer, Gateway LR Clonase II enzyme mix, and Proteinase K solution were purchased from ThermoFischer (Waltham, MA, USA). QuikChange II kit was purchased from Agilent Technologies (Santa Clara, CA, USA). QIAprep Spin Miniprep Kit was purchased from QIAGEN (Hilden, Germany). xCEL-Ligence PET E-plates 96 (Agilent Technologies, Santa Clara, CA, USA) were purchased from Bioké (Leiden, The Netherlands). All other chemicals were of analytical grade and obtained from standard commercial sources.

Selection of cancer-related mutations

Cancer-related mutations were obtained from the Genomic Data Commons²⁹ version 22.0 released on January 16th, 2020, as re-compiled in **Chapter 5**⁵⁰. Somatic missense mutations were retrieved for gene *SLC1A3* (EAAT1) in all cancer types. The 105 unique mutations found were mapped onto the 3D structure of EAAT1 (PDB 5LLU, 5MJU²² and 7AWM²³), with particular attention to the functional motifs and binding sites defined by Canul-Tec *et al.*^{22,23}. Two sets of mutations of interest were defined by visual inspection in the proximity (i.e., 5 Å from co-crystallized ligands) of the orthosteric binding site – occupied by the substrate L-aspartate – and allosteric binding site – occupied by allosteric inhibitor UCPH-101. The “orthosteric” set of mutations included P392L, A446E, A446V, L448Q, and R479W. The “allosteric” set of mutations included Y127C, C252F, R388K, F389L, V390M, and I397V. Additionally, mutation V247F is located at the interface of the two sites and was therefore included in both sets.

As a reference, *SLC1A3* (EAAT1) mutations found in natural variance in the 1000 Genomes dataset⁵¹ were retrieved. This dataset was obtained from the UniProt variance database in October 2020⁵². For the purpose of comparison, the percentage of mutations in EAAT1 found in cancer patients and natural variance was calculated by dividing the number of mutations in EAAT1 by the number of patients in each dataset (10,179 and 3,202, respectively) and multiplying it by 100.

System preparation and molecular docking

The monomeric EAAT1 systems for binding affinity change predictions were prepared from chain A in PDB codes 5LLU and 5MJU²² in ICM-Pro version 3.9-2c (Molsoft LLC, San Diego)^{53,54}. The systems were prepared by optimizing the protonation states and orientation of histidine and cysteine residues, and the orientation of glutamine and asparagine residues. Moreover, the position of hydrogen atoms was sampled and optimized. Stabilizing mutations in residues selected for further analysis were reverted (i.e., C252V, T318M). Subsequently, L-glutamate was prepared by adding hydrogen atoms and assigning atomic charges, and docked it into the orthosteric binding site of PDB 5LLU, originally occupied by L-aspartate. Upon removal of L-aspartate from the binding site, docking was performed with default settings and 10 poses stored by defining the residues surrounding L-aspartate as the binding site. The poses were analyzed in light of the experimental data available, docking scores, and interaction patterns. The pose with the highest docking score was selected for further analysis.

EAAT1 trimeric systems with L-aspartate bound were prepared for MD simulations from the biological assembly of PDB 7AWM, containing chains A-C. This preparation step was performed directly in the academic version of the Desmond program, release 2021.1⁵⁵, and is described in detail in the corresponding MD section.

Binding affinity change predictions

To prioritize mutations for *in vitro* testing, changes in EAAT1 binding affinity were predicted to endogenous substrates L-aspartate and L-glutamate, and the inhibitors TFB-TBOA (competitive) and UCPH-101 (allosteric) caused by point mutations. This analysis was performed in ICM-Pro as follows. The difference in binding energy ($\Delta\Delta G_{\text{bind}}$, in kcal/mol) is calculated as the difference between the Gibbs binding energy (ΔG_{bind} , in kcal/mol) in the mutant and the WT. ΔG_{bind} is calculated for fixed backbone and Monte Carlo-sampled flexible side chains in the vicinity of the mutated residue as the energy of the protein-ligand complex minus the energy of the protein and ligand separately.

For the cancer-related mutations found in the orthosteric binding site (P392L, A446E, A446V, L448Q, and R479W), $\Delta\Delta G_{\text{bind}}$ was calculated for endogenous ligands L-aspartate and L-glutamate (previously docked) in system 5LLU. Moreover, $\Delta\Delta G_{\text{bind}}$ was calculated for the competitive inhibitor TFB-TBOA in system 5MJU. For the cancer-related mutations found in the allosteric binding site (Y127C, C252F, R388K, F389L, V390M, and I397V), $\Delta\Delta G_{\text{bind}}$ was calculated for the allosteric inhibitor UCPH-101 in system 5MJU. For V247F, which is at the interface of both ligand binding sites, $\Delta\Delta G_{\text{bind}}$ was calculated for L-glutamate, L-aspartate, TFB-TBOA, and UCPH-101 as described above.

Structural visualization

All visualizations of EAAT1 structures were generated in PyMOL using PDB 7AWM. Where TFB-TBOA was visualized, PDB 5MJU was superimposed on 7AWM.

Mutagenesis

DNA primers for EAAT1 mutants were designed with a single or double base pair substitution for the resultant amino acid using the QuikChange Primer Design Program and synthesized by Integrated DNA Technologies (IDT, Leuven, Belgium) (Table 6.3). Site-directed mutagenesis was performed using a QuikChange II kit. In brief, per mutant 50 ng template DNA (codon-optimized ORF for EAAT1 (*SLC1A3*) in a pDONR221 vector (pDONR221-*SLC1A3*, Addgene #131889)) together with 10 μM forward and reverse primer, 1 μl dNTP mix, 2.5 μl 10x reaction buffer and 2.5 U DNA polymerase were run in a PCR thermal cycler for 22 cycles (each cycle consisted of 30 s 95°C, 1 min 55°C, 10 min 68°C). Non-mutated DNA was removed by addition of 5 U DpnI restriction enzyme for 2 h at 37°C. Mutant DNA was transformed into XL1-Blue competent cells in the presence of 50 μg/ml kanamycin for selection. Plasmid was isolated using a QIAprep Spin Miniprep Kit verified by Sanger sequencing (Leiden Genome Technology Center, Leiden, The Netherlands).

Table 6.3. DNA primers (forward and reverse) that were used to generate eight cancer-related and two ataxia-related EAAT1 mutants. Mutated bases are bold and underlined.

Mutant	Forward primer (5')	Reverse primer (5')
Y127C	GAGAGCCGTGGTGTACT G TATGACCACAACCATCA	TGATGGTTGTGGTCATA C AGTACACCACGGCTCTC
M128R	TGAGAGCCGTGGTGTACTATA GG ACCACAACCAT	ATGGTTGTGGTC C TATAGTACACCACGGCTCTCA
V247F	AATGCCCTGGGCCTG TTC GTGTTCAGCATGTGC	GCACATGCTGAACAC GA ACAGGCCAGGGCAIT
T318A	CAGCTGGCCATGTAC GCC GTGACAGTGTATCG	CGATCACTGTACGG C CTATCATGGCCAGCTG
V390M	GACAAGCGGGTGACCAGATT T ATGCTGCCAGTG	CACTGGCAGCA T AAATCTGGTCACCCGCTTGTC
P392L	CAGATTGTGTGCTG C TAGTGGGCGCCACCA	TGGTGGCGCCCACT AG CAGCACAAATCTG
A446E	CAGGCATCCCACAGG AA GGCCTGGTGACCATG	CATGGTCACCAGGCC TT CCTGTGGGATGCCTG
A446V	GCATCCCACAGG TCC GCCTGGTGAC	GTCACCAGGCCG A CCTGTGGGATGC
L448Q	CACAGGCGCGCC AG GTGACCATGGT	ACCATGGTCACC TGG CGGCCTGTG
R479W	GGTTTCTGGATAGGCTG TG GACAACCACAACGTGCT	AGCACGTTTGTGGTTGT CA CAGCCTATCCAGAAACC

Gateway cloning

To allow stable transfection into JumpIn cells, the WT and mutant pDONR221-*SLC1A3* plasmids were cloned into a pJTI R4 DEST CMV TO pA expression vector with a C-terminal Twin-Strep-tag and a hemagglutinin (HA)-tag using Gateway cloning. The expression vector contains a tet-operon (TO) that allows doxycycline (dox)-inducible expression of the transgene. In brief, 150 ng pDONR221-*SLC1A3* plasmid and 150 ng pJTI R4 DEST CMV TO pA in TE buffer (10 mM Tris, 1 mM EDTA) were incubated with Gateway LR Clonase II enzyme mix at 25°C for 1 h. To remove endogenous nucleases, the mixture was incubated with a Proteinase K solution for 10 min at 37°C. The resulting vectors (WT or mutant pJTI-*SLC1A3*) were transformed into XL1-Blue competent cells in the presence of 100 μg/ml ampicillin for selection. Plasmid was isolated and sequenced as described in the previous section.

Cell culture

JumpIn-EAAT1 cells were split twice per week into 10 cm dishes in culture medium (high glucose DMEM containing 10% fetal calf serum, 2 mM Glutamax, 100 IU/ml penicillin and 100 µg/ml streptomycin) at 37°C and 5% CO₂. After thawing and recovery, cells were grown for 3-5 days in culture medium with 5 µg/ml blasticidin and 2 mg/ml G418 before switching to culture medium.

Generation of stably transfected WT and mutant JumpIn-EAAT1 cells

JumpIn cells were seeded at 90,000 cells/well in culture medium onto a 24-well culture plate and grown within 24 h to 60-70% confluence. Per mutant or WT, a mix of 1.8 µl P3000 buffer, 450 µg pJTI R4 Integrase plasmid and 450 µg pJTI-*SLC1A3* plasmid in OptiMEM was added to a mix of 2.1 µl Lipofectamine 3000 in OptiMEM (90 µl total per condition) and incubated for 5 min at RT. As a control for antibiotic selection, one dish of cells was incubated with sterile water instead of pJTI-*SLC1A3*. Cells were transfected with 60 µl of the total mix. On the next day, the transfection medium was replaced by fresh culture medium. After 24 h cells were trypsinized and seeded onto 6 cm culture dishes at 200,000 cells/well to grow for 3-4 days. When 70% confluence was reached medium was replaced with selection medium (culture medium with 1 mg/ml G418) to select for successfully transfected cells. Selection medium was refreshed every 2-3 days for 2 weeks until non-transfected cells were all dead and colonies had grown in the transfected dishes. Colonies were resuspended in selection medium and grown to confluence before cryofreezing pools of transfected cells. Prior to use in experiments, cells were cultured in regular culture medium for at least 24 h.

Whole cell HA-tag ELISA

To determine the relative amount of C-terminal HA-tagged protein expressed in doxycycline (dox)-induced JumpIn-EAAT1 WT and mutant cells, an enzyme-linked immunosorbent assay (ELISA) was performed on whole, permeabilized cells. Each condition was tested in quintuplicate per experiment. Cells were seeded in culture medium onto a 96-well culture plate coated with 0.1 mg/ml poly-D-lysine at 60,000 cells/well in the presence or absence of 1 µg/ml dox (100 µl total volume) and were grown for 22-24 h at 37°C and 5% CO₂. Cells were washed with PBS and fixed with 4% formaldehyde for 10 min, then washed with Tris-buffered saline (TBS). To allow access of the antibodies to the intracellular HA-tag, cells were incubated with permeabilization buffer (TBS + 0.5% Tween-20 (TBST), 2% bovine serum albumin (BSA) and 0.2% saponin) for 60 min at RT. After blocking and permeabilization, cells were incubated with 1:2500 rabbit anti-HA polyclonal antibody (Invitrogen, Carlsbad, CA, USA) for 60 min at RT and washed with TBST. Subsequently, cells were incubated for with 1:3000 goat anti-rabbit horse radish peroxidase (HRP)-conjugated IgG antibody (Brunschwig Chemie, Amsterdam, The Netherlands) for 30 min at RT and washed with TBS. Immunoreactivity was visualized by addition of 3,3',5,5'-tetramethylbenzidine (TMB) for 2.5 min at RT and subsequent quenching with 1 M H₃PO₄. Absorbance was measured at 450 nm using a Wallac

EnVision multimode plate reader (PerkinElmer, Groningen, The Netherlands).

Impedance-based phenotypic assay

To measure functional substrate responses and substrate inhibition on WT and mutant JumpIn-EAAT1 cells, a label-free impedance-based cell swelling assay was employed as described previously by our lab³¹. An xCELLigence real-time cell analyzer (RTCA) system (Agilent Technologies, Santa Clara, CA, USA) was used to record real-time changes in cell morphology. The assay principle is that EAAT1-mediated, Na⁺-dependent substrate influx induces cell swelling, which leads to cell spreading. This results in an increased cellular impedance over time and as such is a readout of transporter function. For the assay, JumpIn-EAAT1 cells are cultured in medium onto gold-plated electrodes of a 96-well E-plate and for each well the impedance is measured on predefined time intervals at 10 kHz. The impedance is converted to the unitless parameter Cell Index (CI), which can be plotted over time:

$$CI = \frac{(Z_i - Z_0)\Omega}{15\Omega}$$

where Z_i is the impedance at any given time point and Z_0 is the baseline impedance measured at the start of each experiment⁵⁶.

Assays were performed at 37°C and 5% CO₂ in a final volume of 100 µl/well. Baseline impedance was measured in 40 µl culture medium prior to cell seeding. Cells grown to 70-80% confluence were seeded in 50 µl at 60,000 cells/well in the presence of 1 µg/ml dox to induce EAAT1 expression and left at RT for 30 min prior to placement of the E-plate in the RTCA recording station. After 22 h, cells were pretreated with 5 µl vehicle (PBS/DMSO) or, in inhibitor experiments, 1 nM – 10 µM of TFB-TBOA or UCPH-101 or 1 µM ouabain, and impedance was recorded for 60 min. Subsequently, cells were stimulated with 5 µl vehicle (PBS), 10 µM – 1 mM L-glutamate (submaximal concentration [EC₈₀, 1 mM] in inhibitor experiments) or L-aspartate, 200 nM TFB-TBOA (EC₅₀) or 6.3 µM UCPH-101 (EC₅₀), and impedance was recorded for 120 min. Each condition was tested in duplicate per experiment and levels of DMSO were kept constant at 0.1% for all assays and wells.

Data analysis and statistics

Whole cell HA-tag ELISA

In each experiment, the mean absorbance for each condition was divided over the mean absorbance of non-induced (–dox) JumpIn-EAAT1_{WT} cells to obtain fold expression over –dox cells. To assess whether the total protein expression of dox-induced (+dox) JumpIn-EAAT1 mutant cells was significantly different from +dox JumpIn-EAAT1_{WT} cells, a one-way ANOVA with Dunnett's post-hoc test was done for cells that were tested on the same ELISA plate.

Impedance-based phenotypic assay

Data was recorded using RTCA Software v2.0 or v2.1.1 (ACEA Biosciences). Depending on the part that was used for analysis, the CI values were normalized to the time of inhibitor pretreatment or substrate stimulation yielding normalized CI (nCI) values for all subsequent data points. The nCI values were exported and analyzed in GraphPad Prism v9 (GraphPad Software, San Diego, CA, USA). Vehicle-only conditions were subtracted from all other conditions to correct for vehicle-induced, ligand-independent effects. The remaining nCI curves were quantified by analyzing the net area under the curve (AUC) of the first 120 min after substrate stimulation. The AUC values, which are expressed as the cellular response, were fitted to a sigmoidal concentration-effect curve with a variable slope to determine the potencies of the EAAT1 substrates and inhibitors. Data are shown as the mean \pm standard error of the mean (SEM) of at least three separate experiments each performed in duplicate, unless stated otherwise. Comparison of multiple mean values to a control (i.e., EAAT1_{WT}) was done using a one-way ANOVA with Dunnett's post-hoc test. Differences were considered statistically significant when p-values were below 0.05.

Molecular dynamics

Conformational changes were sampled over time in WT and mutant EAAT1 trimeric systems with MD simulations. The simulations were performed using the academic version of the Desmond program, release 2021.1⁵⁵. The OPLS-2005 force field and SPC water model were used. EAAT1 was simulated with L-aspartate bound as substrate, directly derived from PDB 7AWM biological assembly, where the co-crystallized substrate L-aspartate and Na⁺ ions were kept during preparation and UCPH-101 and Ba²⁺ ions were removed. WT and seven mutants with differential *in vitro* results (Y127C, M128R, P392L, A446E, A446V, L448Q, and R479W) were sampled. All systems were prepared in four steps: (a) the mutation of interest was introduced; (b) default protein preparation wizard was run; (c) the system was stripped to contain the protein trimer and the ligands and ions of interest; (d) the system was embedded in a POPC lipid bilayer respect to the α -helices, solvated with SPC water molecules, the charge was neutralized with Cl⁻ ions, and NaCl was added in physiological concentration (0.15 M). Subsequently, the systems were relaxed with the default protocol, which includes a restrained minimization followed by an unrestrained minimization and four stages of MD runs with decreasing constraints. The production runs were simulated for 500 ns with a recording interval of 500 ps (1000 frames) in an NPT ensemble with a temperature of 300 K and a pressure of 1 bar. Each system was run for ten replicates with velocities randomly initialized with random seeds (**Supplementary Table 6.3**).

MD trajectory analysis

The analysis of MD trajectories was performed in Desmond and PyMOL version 2.5.2 (Schrödinger LTD). Using Desmond analysis scripts, the trajectories' Root Mean Square Deviation (RMSD) was calculated for the protein α carbon (C α) atoms and for the

ligand (L-aspartate) with respect to the protein. These RMSD values represent the stability of the protein system and the ligand, respectively, over the time of the simulation. Moreover, Root Mean Square Fluctuation (RMSF) was calculated for the protein C α atoms. The RMSF values represent the stability/flexibility through the simulation of each of the protein residues. RMSD and RMSF values were calculated independently for each chain in the trimeric system. Protein RMSD was used as an overall measure of the system's stability. Therefore, (chain) systems with protein RMSD reaching 10 Å were excluded from further analysis.

In PyMOL, the trajectories were loaded and fitted to the first frame in the simulation to correct for rotations and translations. Subsequently, the distance in each frame was calculated between a pair of atoms to obtain four measures (1-4). (1) HP2 opening: distance between HP1 and HP2 domain tips, as defined by Alleva *et al.* for Glt_{ph}²⁷. In EAAT1, the distance was measured between S366 C α (HP1 tip) and G442 C α (HP2 tip). The atoms corresponding to the HP1 and HP2 domain tips in EAAT1 were defined via sequence alignment with Glt_{ph}. (2) Na⁺ coordination in Na1 site: distance between Na⁺ ion originally coordinated in Na1 site and one of the Na1 coordinating atoms. The distance was measured between Na⁺ with residue number 601 and D487 C α . (3) Na⁺ coordination in Na2 site: distance between Na⁺ ion originally coordinated in Na2 site and one of the Na2 coordinating atoms. The distance was measured between Na⁺ with residue number 603 and T396 C α . (4) Na⁺ coordination in Na3 site: distance between Na⁺ ion originally coordinated in Na3 site and one of the Na3 coordinating atoms. The distance was measured between Na⁺ with residue number 602 and D400 C α . These distances were measured independently for each chain in the trimeric system.

The sampling density of MD metrics computed per frame (i.e. RMSD and distances) was plotted in Python 3.8 using Matplotlib and Seaborn libraries^{57–59}. The sampling density maps were calculated with data from the 1002 frames in each chain (A, B, C) sampled in the ten replicates simulated per mutant. Unstable systems (i.e. protein RMSD reaching 10 Å) were not included in the density maps. These included Y127C replicate 9 (all chains) and replicate 10 (chain A); M128R replicate 1 (all chains); P392L replicate 1 (chain C) and replicate 10 (chain C); A446E replicate 3 (chain B) and replicate 4 (all chains); A446V replicate 3 (all chains); L448Q replicate 4 (chain C); and R479W replicate 3 (chain A), replicate 5 (chain A), and replicate 7 (chains A, B).

Inhibitor docking in MD trajectory frames

The orthosteric (TFB-TBOA) and allosteric (UCPH-101) inhibitors were docked in a representative selection of MD frames with the most frequent HP2 opening distances but different binding pocket conformations. Chain A was selected for docking because it showed the highest substrate stability in EAAT1_{WT}. For EAAT1_{WT} and each simulated mutant, five random frames with the most frequent HP2 opening distances in the distribution across all replicates and frames were selected (**Supplementary Table 6.4**). Frames were extracted from the trajectories using PyMOL, including the chain A protein atoms and originally coordinated Na⁺ ions. Binding pocket residues were defined as those in the 5 Å neighborhoods of the co-crystallized inhibitors in PDB 5MJU

(TFB-TBOA) and 7AWM (UCPH-101). The characteristics of the different pocket conformations used for ensemble docking were further analyzed by predicting all possible pockets using the pocket finder tool in ICM-Pro and visually selecting the orthosteric and allosteric sites. Pocket volume, hydrophobicity, buriedness, and DLID score were used to confirm pocket conformational variability. ICM-Pro implementation of flexible docking (4D docking) was performed using the five extracted frames to build a receptor map complex per mutant. The rest of the docking setup and parameters followed the general framework described in the section *System preparation and molecular docking*. In 4D docking, the stack of receptor conformations provided is considered as a single receptor object, and ten docking poses are generated on the most favorable conformations. The best docking pose in terms of docking score for each mutant was selected for analysis.

References

1. Tzingounis, A. V. & Wadiche, J. I. Glutamate transporters: confining runaway excitation by shaping synaptic transmission. *Nat Rev Neurosci* **8**, 935–947 (2007).
2. Vandenberg, R. J. & Ryan, R. M. Mechanisms of Glutamate Transport. *Physiol Rev* **93**, 1621–1657 (2013).
3. Alleva, C., Machtens, J. P., Kortzak, D., Weyand, I. & Fahlke, C. Molecular Basis of Coupled Transport and Anion Conduction in Excitatory Amino Acid Transporters. *Neurochem Res* **47**, 9–22 (2022).
4. Peterson, A. R. & Binder, D. K. Astrocyte glutamate uptake and signaling as novel targets for antiepileptogenic therapy. *Front Neurol* **11**, 1006 (2020).
5. Lewerenz, J. & Maher, P. Chronic glutamate toxicity in neurodegenerative diseases-What is the evidence? *Front Neurosci* **9**, 1–20 (2015).
6. Yi, H., Talmon, G. & Wang, J. Glutamate in cancers: from metabolism to signaling. *The Journal of Biomedical Research* **34**, 1 (2020).
7. Freidman, N. *et al.* Amino Acid Transporters and Exchangers from the SLC1A Family : Structure , Mechanism and Roles in Physiology and Cancer. *Neurochem Res* **45**, 1268–1286 (2020).
8. Kortagere, S. *et al.* Identification of Novel Allosteric Modulators of Glutamate Transporter EAAT2. *ACS Chem Neurosci* **9**, 522–534 (2018).
9. Jensen, A. A. *et al.* Discovery of the first selective inhibitor of excitatory amino acid transporter subtype 1. *J Med Chem* **52**, 912–915 (2009).
10. Takano, T. *et al.* Glutamate release promotes growth of malignant gliomas. *Nat Med* **7**, 1010–1015 (2001).
11. Robert, S. M. & Sontheimer, H. Glutamate transporters in the biology of malignant gliomas. *Cellular and Molecular Life Sciences* **71**, 1839–1854 (2014).
12. Corbetta, C. *et al.* Altered function of the glutamate–aspartate transporter GLAST, a potential therapeutic target in glioblastoma. *Int J Cancer* **144**, 2539–2554 (2019).
13. Garcia-Bermudez, J. *et al.* Aspartate is a limiting metabolite for cancer cell proliferation under hypoxia and in tumours. *Nat Cell Biol* **20**, 775–781 (2018).
14. Tajan, M. *et al.* A Role for p53 in the Adaptation to Glutamine Starvation through the Expression of SLC1A3. *Cell Metab* **28**, 721–736.e6 (2018).
15. Bacci, M. *et al.* Reprogramming of Amino Acid Transporters to Support Aspartate and Glutamate Dependency Sustains Endocrine Resistance in Breast Cancer. *Cell Rep* **28**, 104–118.e8 (2019).
16. Chivukula, A. S., Suslova, M., Kortzak, D., Kovermann, P. & Fahlke, C. Functional consequences of SLC1A3 mutations associated with episodic ataxia 6. *Hum Mutat* **41**, 1892–1905 (2020).
17. Kovermann, P. *et al.* Impaired K⁺ binding to glial glutamate transporter EAAT1 in migraine. *Sci Rep* **7**, 13913 (2017).
18. van Amen-Hellebrekers, C. J. M. *et al.* Duplications of SLC1A3: Associated with ADHD and autism. *Eur J Med Genet* **59**, 373–376 (2016).
19. Adamczyk, A. *et al.* Genetic and functional studies of a missense variant in a glutamate transporter, SLC1A3, in Tourette syndrome. *Psychiatr Genet* **21**, 90–97 (2011).
20. Boudker, O., Ryan, R. M., Yernool, D., Shimamoto, K. & Gouaux, E. Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* **445**, 387–393 (2007).
21. Guskov, A., Jensen, S., Faustino, I., Marrink, S. J. & Slotboom, D. J. Coupled binding mechanism of three sodium ions and aspartate in the glutamate transporter homologue Glt Tk. *Nat Commun* **7**, 1–6 (2016).
22. Canul-Tec, J. C. *et al.* Structure and allosteric inhibition of excitatory amino acid transporter 1. *Nature* **544**, 446–451 (2017).
23. Canul-Tec, J. C. *et al.* The ion-coupling mechanism of human excitatory amino acid transporters. *EMBO J* **41**, e108341 (2022).
24. Kato, T. *et al.* Structural insights into inhibitory mechanism of human excitatory amino acid transporter EAAT2. *Nat Commun* **13**, 4714 (2022).
25. Qiu, B., Matthies, D., Fortea, E., Yu, Z. & Boudker, O. Cryo-EM structures of excitatory amino acid transporter 3 visualize coupled substrate, sodium, and proton binding and transport. *Sci Adv* **7**, 1–10 (2021).
26. Kortzak, D. *et al.* Allosteric gate modulation confers K⁺ coupling in glutamate transporters . *EMBO J* **38**, 1–17 (2019).
27. Alleva, C. *et al.* Na⁺-dependent gate dynamics and electrostatic attraction ensure substrate coupling in glutamate transporters. *Sci Adv* **6**, eaba9854 (2020).
28. Schaller, L. & Lauschke, V. M. The genetic landscape of the human solute carrier (SLC) transporter superfamily. *Hum Genet* **138**, 1359–1377 (2019).
29. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
30. Shimamoto, K. *et al.* Characterization of novel L-threo- β -Benzyloxyaspartate derivatives, potent

- blockers of the glutamate transporters. *Mol Pharmacol* **65**, 1008–1015 (2004).
31. Sijben, H. J. *et al.* Impedance-Based Phenotypic Readout of Transporter Function: A Case for Glutamate Transporters. *Front Pharmacol* **13**, 1–18 (2022).
 32. Cournia, Z., Allen, B. & Sherman, W. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. *J Chem Inf Model* **57**, 2911–2937 (2017).
 33. Vettore, L., Westbrook, R. L. & Tennant, D. A. New aspects of amino acid metabolism in cancer. *Br J Cancer* **122**, 150–156 (2020).
 34. Winter, N., Kovermann, P. & Fahlke, C. A point mutation associated with episodic ataxia 6 increases glutamate transporter anion currents. *Brain* **135**, 3416–3425 (2012).
 35. Abrahamsen, B. *et al.* Allosteric modulation of an excitatory amino acid transporter: the subtype-selective inhibitor UCPH-101 exerts sustained inhibition of EAAT1 through an intramonomeric site in the trimerization domain. *Journal of Neuroscience* **33**, 1068–1087 (2013).
 36. Wu, Q. *et al.* Ataxia-linked SLC1A3 mutations alter EAAT1 chloride channel activity and glial regulation of CNS function. *Journal of Clinical Investigation* **132**, e154891 (2022).
 37. Bastug, T. *et al.* Position of the third Na⁺ site in the aspartate transporter Glt Ph and the human glutamate transporter, EAAT1. *PLoS One* **7**, 13–17 (2012).
 38. Choi, K.-D. *et al.* Genetic Variants Associated with Episodic Ataxia in Korea. *Sci Rep* **7**, 13855 (2017).
 39. Seal, R. P. & Amara, S. G. A reentrant loop domain in the glutamate carrier EAAT1 participates in substrate binding and translocation. *Neuron* **21**, 1487–1498 (1998).
 40. Chen, I. *et al.* Glutamate transporters have a chloride channel with two hydrophobic gates. *Nature* **591**, 327–331 (2021).
 41. Cater, R. J., Vandenberg, R. J. & Ryan, R. M. The Domain Interface of the Human Glutamate Transporter EAAT1 Mediates Chloride Permeation. *Biophys J* **107**, 621–629 (2014).
 42. Colucci, E. *et al.* Mutation in glutamate transporter homologue GltTk provides insights into pathologic mechanism of episodic ataxia 6. *Nat Commun* **14**, 1799 (2023).
 43. Trinco, G. *et al.* Kinetic mechanism of Na⁺-coupled aspartate transport catalyzed by GltTk. *Commun Biol* **4**, 751 (2021).
 44. Ciftci, D. *et al.* Linking function to global and local dynamics in an elevator-type transporter. *PNAS* **118**, e20255220118 (2021).
 45. Leighton, B. H., Seal, R. P., Shimamoto, K. & Amara, S. G. A hydrophobic domain in glutamate transporters forms an extracellular helix associated with the permeation pathway for substrates. *J Biol Chem* **277**, 29847–29855 (2002).
 46. Scopelliti, A. J., Font, J., Vandenberg, R. J., Boudker, O. & Ryan, R. M. Structural characterisation reveals insights into substrate recognition by the glutamine transporter ASCT2/SLC1A5. *Nat Commun* **9**, 1–12 (2018).
 47. Reddy, K. D., Ciftci, D., Scopelliti, A. J. & Boudker, O. The archaeal glutamate transporter homologue GltPh shows heterogeneous substrate binding. *Journal of General Physiology* **154**, e202213131 (2022).
 48. Wang, J., Li, P., Yu, X. & Grever, C. Observing spontaneous, accelerated substrate binding in molecular dynamics simulations of glutamate transporters. *PLoS One* **16**, 1–19 (2021).
 49. Levine, M. V., Cuendet, M. A., Khelashvili, G. & Weinstein, H. Allosteric mechanisms of molecular machines at the membrane: Transport by sodium-coupled symporters. *Chem Rev* **116**, 6552–6587 (2016).
 50. Bongers, B. J. *et al.* Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors. *Sci Rep* **12**, 21534 (2022).
 51. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
 52. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
 53. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* **15**, 488–506 (1994).
 54. Neves, M. A. C., Totrov, M. & Abagyan, R. Docking and scoring with ICM: The benchmarking results and strategies for improvement. *J Comput Aided Mol Des* **26**, 675–686 (2012).
 55. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. in *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06* (2006).
 56. Kho, D. *et al.* Application of xCELLigence RTCA biosensor technology for revealing the profile and window of drug responsiveness in real time. *Biosensors* **5**, 199–222 (2015).
 57. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95 (2007).
 58. Waskom, M. Seaborn: Statistical Data Visualization. *J Open Source Softw* **6**, 3021 (2021).
 59. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, Scotts Valley, CA, 2009)

Supplementary Information

Supplementary Table 6.1. Orthosteric pocket properties across the five random frames selected per mutant for 4D docking. Pockets were predicted and characterized using the ICM pocket finder tool and visually inspected to match the orthosteric pocket (*: some orthosteric pockets seem to be divided into two pockets in ICM). Pocket properties are volume (Å³), Hydrophobicity (representing the percentage of the pocket surface in contact with hydrophobic residues, ranges from 0 to 1), Buriedness (calculated based on solvent accessibility, ranges from 0.5 – completely open and surface flat – to 1.0 – completely buried), and DLID (Merck’s Drug-like density score (see Sheridan *et al.* *JCIM* 2010). Values above zero and those with slightly negative values are considered “druggable”).

Frame	Pocket	Volume	Hydrophobicity	Buriedness	DLID
7AWM	Orthosteric	147.01	0.59	1.00	0.26
5MJU	Orthosteric	607.78	0.42	0.69	-0.31
wt_1A_203	Orthosteric	467.11	0.62	0.90	0.78
wt_9A_125	Orthosteric	462.99	0.63	0.90	0.79
wt_9A_231	Orthosteric	400.09	0.58	0.86	0.44
wt_9A_617	Orthosteric	493.75	0.59	0.84	0.52
wt_9A_731	Orthosteric	722.43	0.65	0.88	1.11
Y127C_1A_522	Orthosteric	784.45	0.51	0.76	0.37
Y127C_3A_154	Orthosteric	475.74	0.48	0.81	0.15
Y127C_3A_618	Orthosteric	596.56	0.62	0.94	1.13
Y127C_3A_786	Orthosteric	389.43	0.55	0.90	0.48
Y127C_3A_892	Orthosteric	437.20	0.44	0.80	-0.06
M128R_4A_15	Orthosteric	187.94	0.81	0.99	0.92
M128R_4A_103	Orthosteric	388.81	0.64	0.87	0.58
M128R_4A_546	Orthosteric	504.84	0.60	0.89	0.77
M128R_4A_769	Orthosteric	194.63	0.73	0.97	0.66
M128R_4A_769	Orthosteric*	101.61	0.32	0.63	-2.06
M128R_4A_941	Orthosteric	186.41	0.81	0.99	0.91
P392L_4A_347	Orthosteric	341.24	0.48	0.76	-0.29
P392L_4A_609	Orthosteric	395.64	0.61	0.89	0.59
P392L_4A_817	Orthosteric	568.70	0.61	0.90	0.91
P392L_4A_880	Orthosteric	505.91	0.57	0.84	0.50
P392L_4A_971	Orthosteric	341.05	0.47	0.76	-0.32
A446E_2A_63	Orthosteric	670.60	0.55	0.78	0.45
A446E_2A_63	Orthosteric*	135.49	0.72	0.96	0.33
A446E_2A_440	Orthosteric	740.40	0.61	0.90	1.12
A446E_8A_69	Orthosteric	360.90	0.53	0.89	0.34
A446E_8A_386	Orthosteric	584.50	0.52	0.84	0.49
A446E_10A_254	Orthosteric	670.88	0.55	0.78	0.43

Supplementary Table 6.1 (continues)

A446E_10A_254	Orthosteric*	135.76	0.72	0.96	0.34
A446V_1A_163	Orthosteric	405.01	0.58	0.87	0.48
A446V_5A_35	Orthosteric	613.30	0.59	0.81	0.57
A446V_5A_508	Orthosteric	323.34	0.48	0.79	-0.24
A446V_5A_530	Orthosteric	304.85	0.59	0.84	0.17
A446V_5A_778	Orthosteric	631.34	0.56	0.81	0.52
L448Q_1A_110	Orthosteric	482.15	0.54	0.87	0.51
L448Q_2A_301	Orthosteric	327.78	0.61	0.94	0.66
L448Q_5A_82	Orthosteric	557.76	0.64	0.91	1.03
L448Q_5A_455	Orthosteric	373.75	0.61	0.85	0.40
L448Q_6A_93	Orthosteric	553.21	0.66	0.91	1.05
R479W_4A_175	Orthosteric*	571.68	0.62	0.82	0.65
R479W_4A_175	Orthosteric	234.19	0.56	0.86	0.02
R479W_5A_431	Orthosteric	529.61	0.62	0.86	0.71
R479W_5A_491	Orthosteric	714.29	0.57	0.85	0.80
R479W_5A_550	Orthosteric	565.14	0.53	0.78	0.24
R479W_5A_710	Orthosteric	578.60	0.61	0.82	0.63
R479W_5A_710	Orthosteric*	235.80	0.54	0.84	-0.12

Supplementary Table 6.2. Allosteric (UCPH-101) pocket properties across the five random frames selected per mutant for 4D docking. Pockets were predicted and characterized using the ICM pocket finder tool and visually inspected to match the allosteric pocket. The allosteric pocket was missing in some frames, which are not recorded in the table. Pocket properties are volume (Å³), Hydrophobicity (representing the percentage of the pocket surface in contact with hydrophobic residues, ranges from 0 to 1), Buriedness (calculated based on solvent accessibility, ranges from 0.5 – completely open and surface flat – to 1.0 – completely buried), and DLID (Merck’s Drug-like density score (see Sheridan et al. *JCIM* 2010). Values above zero and those with slightly negative values are considered “druggable”).

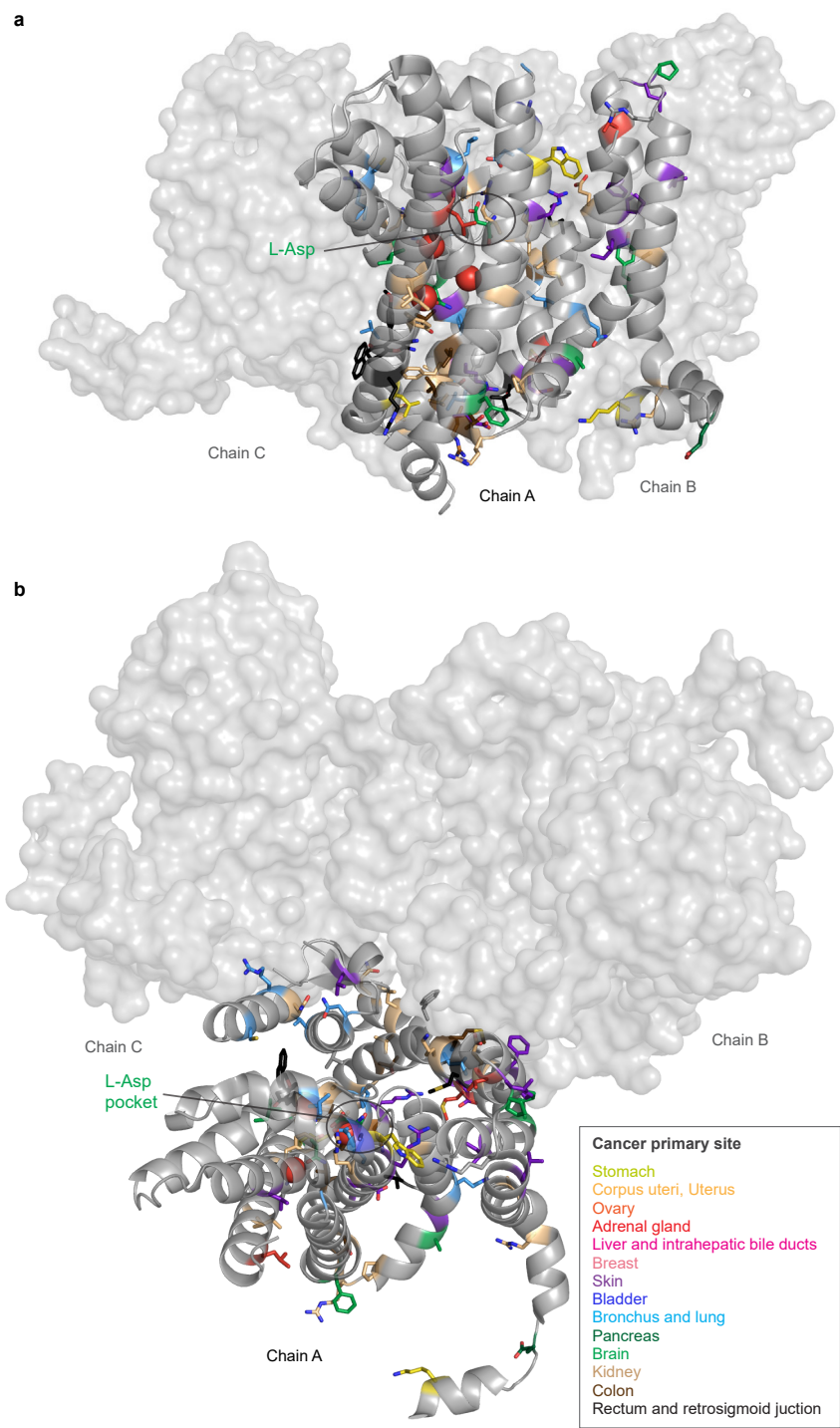
Frame	Pocket	Volume	Hydrophobicity	Buriedness	DLID
5MJU	Allosteric	129.07	0.67	0.77	-0.55
7AWM	Allosteric	114.40	0.41	0.55	-2.10
wt_9A_231	Allosteric	119.49	0.63	0.65	-1.17
wt_9A_617	Allosteric	166.22	0.66	0.72	-0.58
wt_9A_731	Allosteric	110.85	0.40	0.53	-2.25
Y127C_3A_154	Allosteric	165.31	0.62	0.68	-0.84
Y127C_3A_618	Allosteric	235.42	0.72	0.73	-0.13
Y127C_3A_786	Allosteric	169.85	0.73	0.80	-0.08
Y127C_3A_892	Allosteric	166.86	0.64	0.69	-0.77
M128R_4A_769	Allosteric	108.86	0.51	0.57	-1.85
P392L_4A_347	Allosteric	287.89	0.71	0.78	0.18
P392L_4A_817	Allosteric	160.23	0.54	0.61	-1.31
P392L_4A_880	Allosteric	177.89	0.67	0.72	-0.50
P392L_4A_971	Allosteric	290.36	0.72	0.78	0.20
A446E_2A_440	Allosteric	119.92	0.49	0.59	-1.74
A446E_8A_69	Allosteric	118.66	0.37	0.46	-2.54
R479W_5A_550	Allosteric	106.79	0.50	0.55	-1.92

Supplementary Table 6.3. Random seeds used to generate initial velocities in Molecular Dynamics simulations.

Mutant	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Replicate 6	Replicate 7	Replicate 8	Replicate 9	Replicate 10
WT	1613	1825	8414	8636	6037	8843	5519	5522	973	9694
Y127C	5409	5655	9816	9194	2819	5369	3230	5695	2910	6457
M128R	112	8172	8063	4417	9724	4479	4263	6945	668	2947
P392L	8167	3981	2959	52	5364	7026	1533	5596	3822	6974
A446E	782	9717	8228	1096	3085	2107	3786	8496	1711	5788
A446V	7735	3445	1885	3556	6824	9192	4487	2489	4094	9957
L448Q	7248	2175	8437	8704	3512	2870	4162	2289	78	3219
R479W	8385	4603	128	9635	5711	1994	1530	7953	3132	4046

Supplementary Table 6.4. Frames with the five most common HP2 opening distances across replicates of EAAT1 chain A MD simulations selected for docking.

Mutant	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
WT	Replicate 1 frame 203	Replicate 9 frame 125	Replicate 9 frame 231	Replicate 9 frame 617	Replicate 9 frame 731
Y127C	Replicate 1 frame 522	Replicate 3 frame 154	Replicate 3 frame 618	Replicate 3 frame 786	Replicate 3 frame 892
M128R	Replicate 4 frame 15	Replicate 4 frame 103	Replicate 4 frame 546	Replicate 4 frame 769	Replicate 4 frame 941
P392L	Replicate 4 frame 347	Replicate 4 frame 609	Replicate 4 frame 817	Replicate 4 frame 880	Replicate 4 frame 971
A446E	Replicate 2 frame 6	Replicate 2 frame 440	Replicate 8 frame 63	Replicate 8 frame 386	Replicate 10 frame 254
A446V	Replicate 1 frame 163	Replicate 5 frame 35	Replicate 5 frame 508	Replicate 5 frame 530	Replicate 5 frame 778
L448Q	Replicate 1 frame 110	Replicate 2 frame 301	Replicate 5 frame 82	Replicate 5 frame 455	Replicate 6 frame 93
R479W	Replicate 4 frame 175	Replicate 5 frame 431	Replicate 5 frame 491	Replicate 5 frame 550	Replicate 5 frame 710

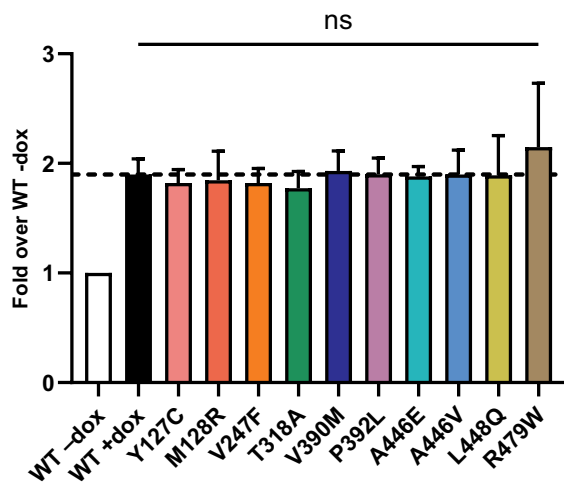


Supplementary Figure 6.1 (caption on the following page)

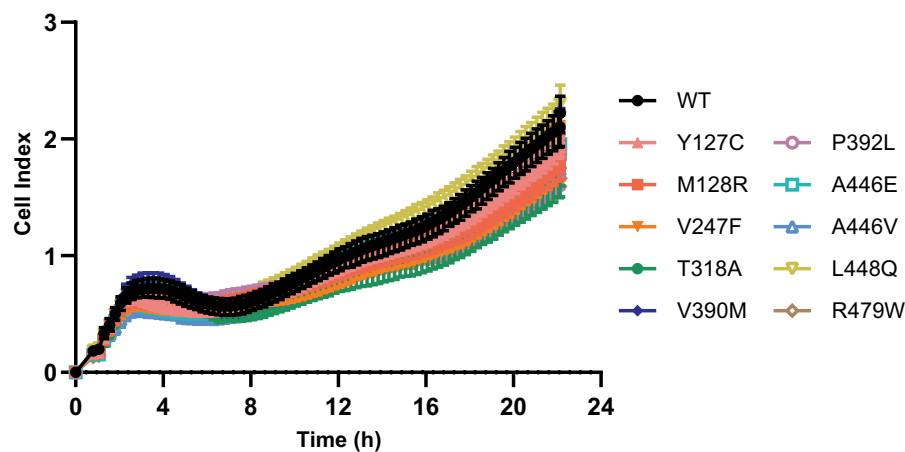
► **Supplementary Figure 6.1.** Structural distribution of cancer-related mutants per cancer type. Mutations from the Genomic Data Commons mapped onto the biological assembly of EAAT1 (PDB 7AWM). Chain A is represented as a grey cartoon, while chains B and C are represented as grey surfaces. The co-crystallized substrate, L-aspartate, is represented as green sticks in chain A. The three coordinated Na⁺ ions are represented as red spheres in chain A. Residues that have been observed mutated in cancer patients are colored by cancer primary site following the colors in the key. **a)** Frontal view, as aligned with cellular membrane. **b)** Top view, as seen from the extracellular side.

sp P43003 EAA1_HUMAN/I-542	1	-----MTKSNGEPMKMGGRMERFOOGVRKRTLLAKKVKQNIITKEDVSKYSLFRNAFVLLTAVTAVIGTILG	65
sp P43004 EAA2_HUMAN/I-574	1	MASTEGANNMPK-----QVEVRMHDSHLGSEEPKRRHLGLRLCDKLGKLLLTITVFGVILGAVGC	61
sp P43005 EAA3_HUMAN/I-524	1	-----MGKPKARKCEWKRFLKNNWVLLSTVAAYVLGITTG	35
sp P48664 EAA4_HUMAN/I-564	1	-----MSSHGNSLFLRESGQRLGRVQWLORLOESLQORALRTLRLQTLTLEHVLRLRRNAFILLTYSAVVIGVSLA	73
sp O00341 EAA5_HUMAN/I-560	1	-----MYPHAILARGRDVCRNRGLLILSVLSVIVGCLGL	34
sp O59010 GLT_PYRHO/I-425	1	-----MGLYRKYIEYPLVQKILIGLILGAIIVG	27
sp P43003 EAA1_HUMAN/I-542	66	FTLRP- YRMSYREVKYFSFGPELLMRMLQMLVPLIISSSLVTGMAALDSKASGKMGMAVVYVTTTTIIIAVVI	138
sp P43004 EAA2_HUMAN/I-574	62	GLRLASPIHPDVMLIAFPDILMRMLKMLILPLIISSLTIGLSGDAKASGRGLTRAMVYVNSTTIIIAVVLG	135
sp P43005 EAA3_HUMAN/I-524	36	VLVRHNSLSTLEKFFAFPGELMRMLKMLILPLIISSMTI GVAALDSNVSGKILGRAVYVYCTTLIIAVGL	109
sp P48664 EAA4_HUMAN/I-564	74	FALRP- YQLTYROIKYFSFGPELLMRMLQMLVPLIVSSSLVTGMAALDNKATGRMGMAVVYVTTTTIIIAVVI	146
sp O00341 EAA5_HUMAN/I-560	35	FLRLT- RRLSPQEISYFOFPGELMRMLKMLILPLVSSSLMSGLSDAKTSRLGLVLTVAIYVTTTMAVIG	107
sp O59010 GLT_PYRHO/I-425	28	LILGH- YGYADAVKTYKFGDGLVRLKMLVMPVIFASLVGGAASIPARLGRVGKIVVYVLTSAFAVLT	100
sp P43003 EAA1_HUMAN/I-542	139	IIIVIIIPHGKGTKE- NMHREGKIVRVTAADAFDLIRNMFPPNLEACFKQFKTNYKRSFKVPVIOANETL	209
sp P43004 EAA2_HUMAN/I-574	136	VILVLAIHPGNPKLKKQLGPGKKNDEVSSDAFDLIRNLFPPNLEACFKQIQVTQVTLVAPPDDEAN	207
sp P43005 EAA3_HUMAN/I-524	110	ILVLVSIKPGVTQKVGELIARTGSTPEVSTVDAMDLIRNMFPPNLEACFKQFKTNYKRSFKVPVIOANETL	178
sp P48664 EAA4_HUMAN/I-564	147	ILMVTIIPHGKGSKE- GLHREGRIETIPTADAFMDLIRNMFPPNLEACFKQFKTNYKRSFKVPVIOANETL	219
sp O00341 EAA5_HUMAN/I-560	108	IFMVSIIPHGSAQK- ETTEQSGKPISSADALDLIRNMFPPNLEACFKQFKTNYKRSFKVPVIOANETL	178
sp O59010 GLT_PYRHO/I-425	101	IIMARLFNPGAGIHLAVG-----GQGFQPKQAPPLVKILLDIPVTNPF-----	143
sp P43003 EAA1_HUMAN/I-542	210	V-----GAVI-NNVSEAMETLTR-----ITEELVPVPGSVNGVNALGLVFSMCFGFVIG-----N	259
sp P43004 EAA2_HUMAN/I-574	208	-----TSAVSLINETVTEVPEE- TKMVIKGLFEKQDMNVGLLIGFFIAFGIAMG-----K	258
sp P43005 EAA3_HUMAN/I-524	179	-----MTE-ESFTAVMTTIAI SKNKTKEYKIVGMSYSDGIVNLGLVFCFLVGLVIG-----K	228
sp P48664 EAA4_HUMAN/I-564	229	PGASMPPPFSEVNGTSL- ENVTALRGLQEMLSFEETVPVPGSANGIALGLVFSFATGVLIG-----G	284
sp O00341 EAA5_HUMAN/I-560	179	PRRILIIYGVEEN- GSHV- QNFALDTPPE-----VVKSEPGTSDNNVGLIIFVFSATGVLIG-----B	238
sp O59010 GLT_PYRHO/I-425	144	-----GA- LANGQVLTPIIFAIILGIAITYLMNSENEK	175
sp P43003 EAA1_HUMAN/I-542	260	MKEGQALREFFDLSNEAIMRLVAVIMWYAPVGLFLIAGKIVEMEDMGVIGGLAMYTIVTVIGLLIHAVI	333
sp P43004 EAA2_HUMAN/I-574	259	MGDAQKLMDVDFNILEIMVKLVIIMWYSPGLIACLICGKI IAKDLEVARGLQGMVTVIIGLIHGGI	332
sp P43005 EAA3_HUMAN/I-524	228	MGEKGQILVDFDNALSDATMKIVQIMCYMPLGILFLIAGKILEVEDWEI- F- RKLGLYMATVLTGLAIFHS	310
sp P48664 EAA4_HUMAN/I-564	285	MKHKGRVLRFDFDLSNEAIMRLVGI I WYAPVGLFLIAGKILEMEDMAVGLQGLMYTLTVIIGLHAGI	358
sp O00341 EAA5_HUMAN/I-560	239	MGDSGAPLVSFCQCLNESVMKIVAVAVWYFPFGVFLIAGKILEMDPRAVGKLGFGYVTVVCGVLVHGL	312
sp O59010 GLT_PYRHO/I-425	176	VRKSAETLDAIINLEAEAMYKIVNGVMQYAPIGVFALIAVYMAEQ- GVKKV- GELAKVIAAVYVGLTQL	244
sp P43003 EAA1_HUMAN/I-542	334	PLLYFL- - -VTRKNPWVFIGGLLOALITAGTSSSSATLPITFKCLLEENGVDRKRVTRFVLVPGVATINMDGTAL	404
sp P43004 EAA2_HUMAN/I-574	333	PLLYFV- - -VTRKNPFSFFAGIFQAWITALGTASSAGTLPTVTFRCLEENGLDKRVRFLVPGVATINMDGTAL	403
sp P43005 EAA3_HUMAN/I-524	302	PLLYFI- - -VVRKNPFRFAMGMAQALITALMISSSSATLPVTFRCLEENGNQDKRITRFLVPGVATINMDGTAL	372
sp P48664 EAA4_HUMAN/I-564	332	PLLYFL- - -VTHRNPPFFIGGMLQALITAMGTSSSSATLPITFRCLEEGVDORRITRFLVPGVATINMDGTAL	429
sp O00341 EAA5_HUMAN/I-560	313	PLLYFF- - -ITKKNPFI VFRIGILOALLALATSSSSATLPITFKCLLENNHIDRRIFARFLVPGVATINMDGTAL	383
sp O59010 GLT_PYRHO/I-425	245	-LVYFVLKLYGIDPISFIIKKAKDAMLTAFVTRSSSGTLPTVMRVAKE- MGISEGIYSFTLPLGATINMDGTAL	316
sp P43003 EAA1_HUMAN/I-542	405	YEALAAIFIAQVNNFELNFGQIITISITATAASIGAAGIPQAGVITVMVILVTSVGLPTDDI-----TLIIAVD	472
sp P43004 EAA2_HUMAN/I-574	404	YEAAVAIFIAQMNQVVLDDGGQIVTVSLTATLASVGAASIPSAAGVITMMLILTAVGLPTEDI-----SLLVAVD	471
sp P43005 EAA3_HUMAN/I-524	373	YEAAVAIFIAQLNDLDLIGQIITISITATSASIGAAGVPQAGVITMVIIVLSAVGLPAEDV-----TLIIAVD	440
sp P48664 EAA4_HUMAN/I-564	404	YEALAAIFIAQVNNYELNLGQIITISITATAASVGAAGIPQAGVITMVIIVLSVGLPTEDI-----TLIIAVD	497
sp O00341 EAA5_HUMAN/I-560	384	YEAAVAIFIAQVNNYELDFGQIITISITATAASIGAAGIPQAGVITMVIIVLSVGLPTDDI-----TLIIAVD	451
sp O59010 GLT_PYRHO/I-425	317	YQGVCTFFIANALGSHLTVGQQLTIVLTAVLASIGTAGVPAGAGIIMLAMVLSVGLPLTDPNVAAYAMAILGID	390
sp P43003 EAA1_HUMAN/I-542	473	WFLDRLRTITNVLGDSLGAGIVEHLSRHELKNRDVEMGNSVIEENEMKKPYO-----LIAQNETE- -KPI-DS-E	539
sp P43004 EAA2_HUMAN/I-574	472	WLDLDRMTSVNNVGDGSGAGIVYHLSKSELDITDSOHRVH- -EDIEMTKTQSIYDDMKNNHRESNOCYVAAHN	543
sp P43005 EAA3_HUMAN/I-524	441	WLDLDRFTMVNVLGDAFGTGI VEKLSKKELEQMDSEVN-----LVNPFALSTILDNEDSD- TKKSYVNGN	507
sp P48664 EAA4_HUMAN/I-564	498	WFLDRLRTMTNVLGDSIGAAVIEHLSQRELEQEAELT-----LPSLGKPKY- -SLMAQEKGA- SRGRGN-E	561
sp O00341 EAA5_HUMAN/I-560	452	WALDRLRTMTNVLGDAAGIMAHICRKDFARDTGTEKL-----LPCETKPVSLQEVIAAQNGC- VKSVAEASE	520
sp O59010 GLT_PYRHO/I-425	391	AILDMDRMTMVNVTGDLTGTAIVAKTEG- ELEKGVIA-----	425
sp P43003 EAA1_HUMAN/I-542	540	TKM-----	542
sp P43004 EAA2_HUMAN/I-574	543	SVIIVDECKVTLA-----ANGKSADCSVEEPPWKREK	574
sp P43005 EAA3_HUMAN/I-524	508	FAYD-----KSDTISFTQTSQF-----	524
sp P48664 EAA4_HUMAN/I-564	562	SAM-----	564
sp O00341 EAA5_HUMAN/I-560	521	LTLGPTCPHHVPVQVEQDEELPAASLNHCTIQISELETNV	560
sp O59010 GLT_PYRHO/I-425			

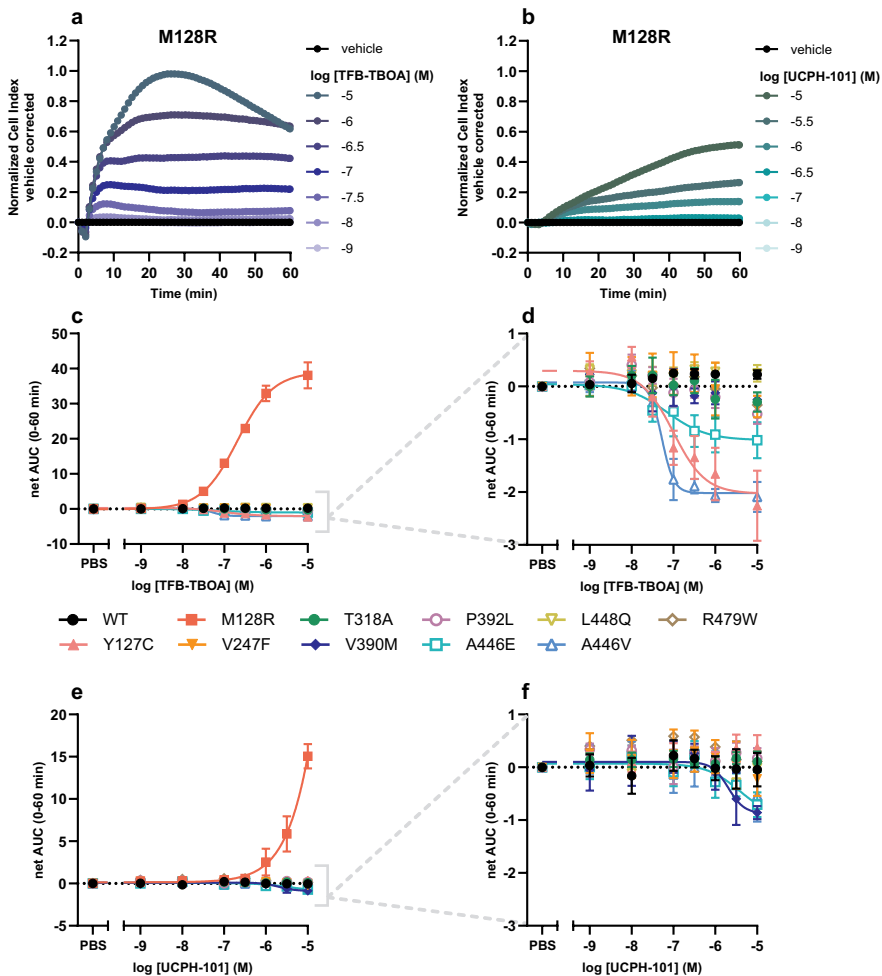
Supplementary Figure 6.2. Conservation of selected cancer-related mutants in EAAT family. Multiple sequence alignment of human EAATs (EAAT1-5) and *Pyrococcus horikoshii* homolog GlT_{ph} computed in Clustal-Omega. Colored, the positions of the cancer-related mutants analyzed *in vitro*: Y127C (pink), V247F (orange), V390M (dark blue), P392L (purple), A446V/E (blue), L448Q (yellow), R479W (brown). For reference, ataxia-related reference mutants are also colored: M128R (red) and T318A (green).



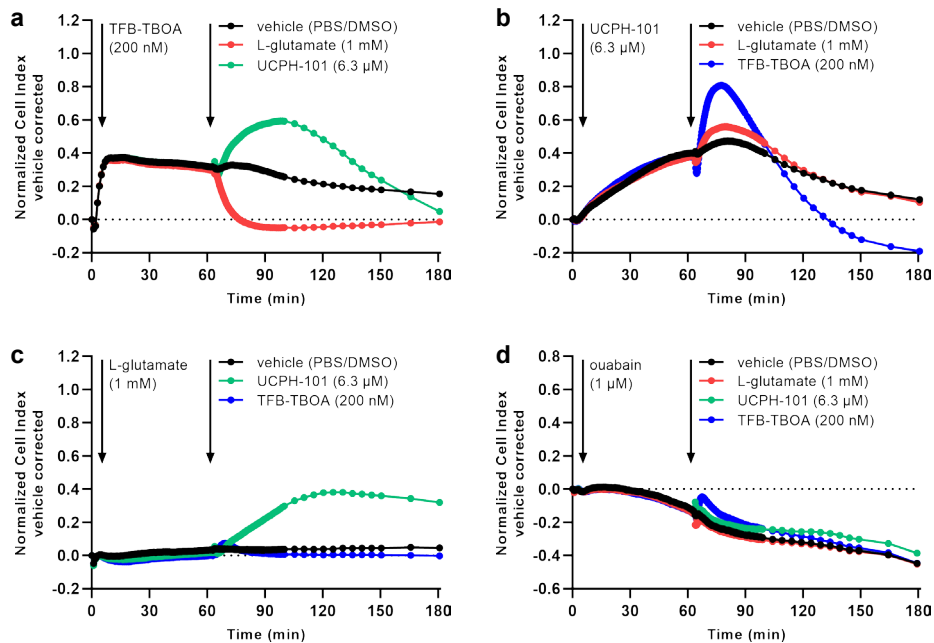
Supplementary Figure 6.3. Whole cell HA-tag ELISA on EAAT1_{WT} and mutant cells. Cells were grown for 24 h in the absence (-dox, WT only) or presence (+dox, WT, and mutants) of 1 μ g/ml doxycycline. Presence of total HA-tagged protein (plasma membrane and cytosolic) was determined in permeabilized cells. Absorbance for each condition is expressed as fold expression over WT (-dox). Data are shown as the mean \pm SEM of twelve (WT), six (M128R) or three (rest) individual experiments each performed in quintuplicate. Significant differences between EAAT1_{WT} and mutant cells were determined using one-way ANOVA with Dunnett's post-hoc test. ns = not significant for all mutants.



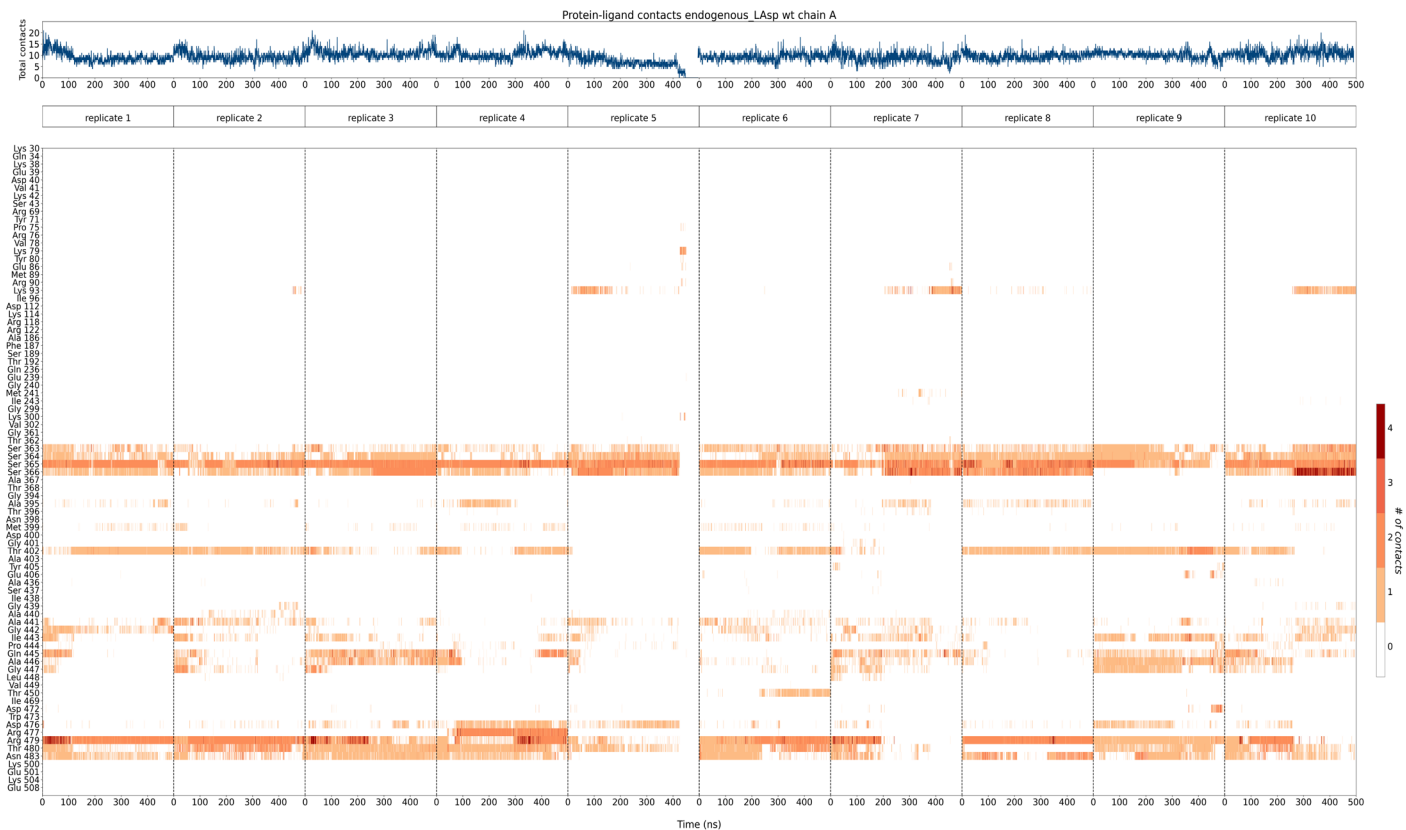
Supplementary Figure 6.4. Representative growth curves of EAAT1_{WT} and EAAT1 mutant cells in an impedance-based phenotypic assay. Data are shown as the mean \pm SD of eight replicates from a representative experiment.



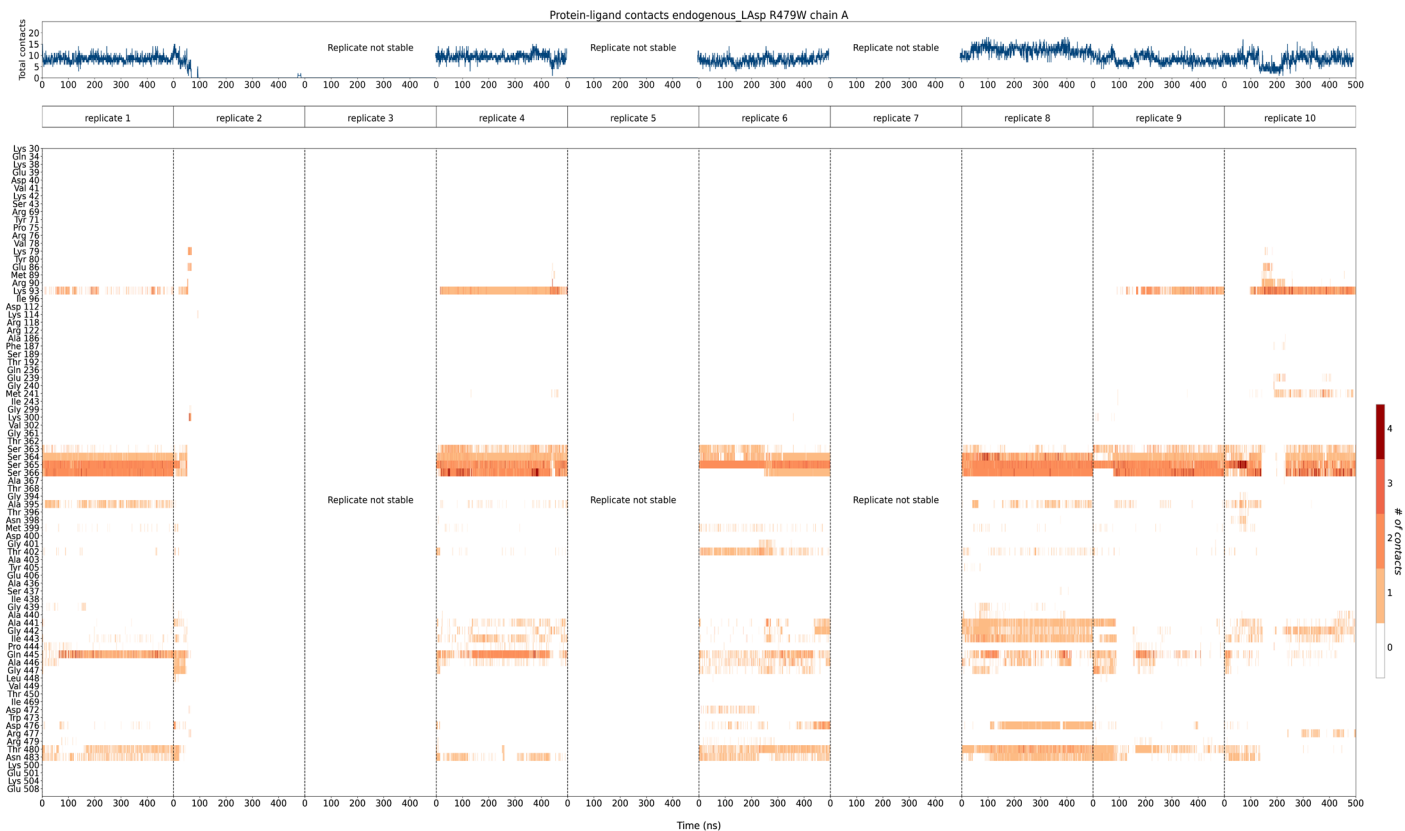
Supplementary Figure 6.5. Cellular responses of TFB-TBOA and UCPH-101 during pretreatment in an impedance-based phenotypic assay on EAAT_{WT} and mutant cells. **a,b**) Vehicle-corrected normalized Cell Index traces of M128R cells pretreated with **(a)** TFB-TBOA or **(b)** UCPH-101 from a representative experiment. **c**) Concentration-response curves of TFB-TBOA on M128R cells and **d**) zoom-in on EAAT_{WT} and other mutant cells. **e**) Concentration-response curves of UCPH-101 on M128R cells and **f**) zoom-in on EAAT_{WT} and other mutant cells. Cellular response is expressed as the net AUC of the first 60 min after inhibitor pretreatment. Data are shown as the mean \pm SEM of three individual experiments each performed in duplicate.



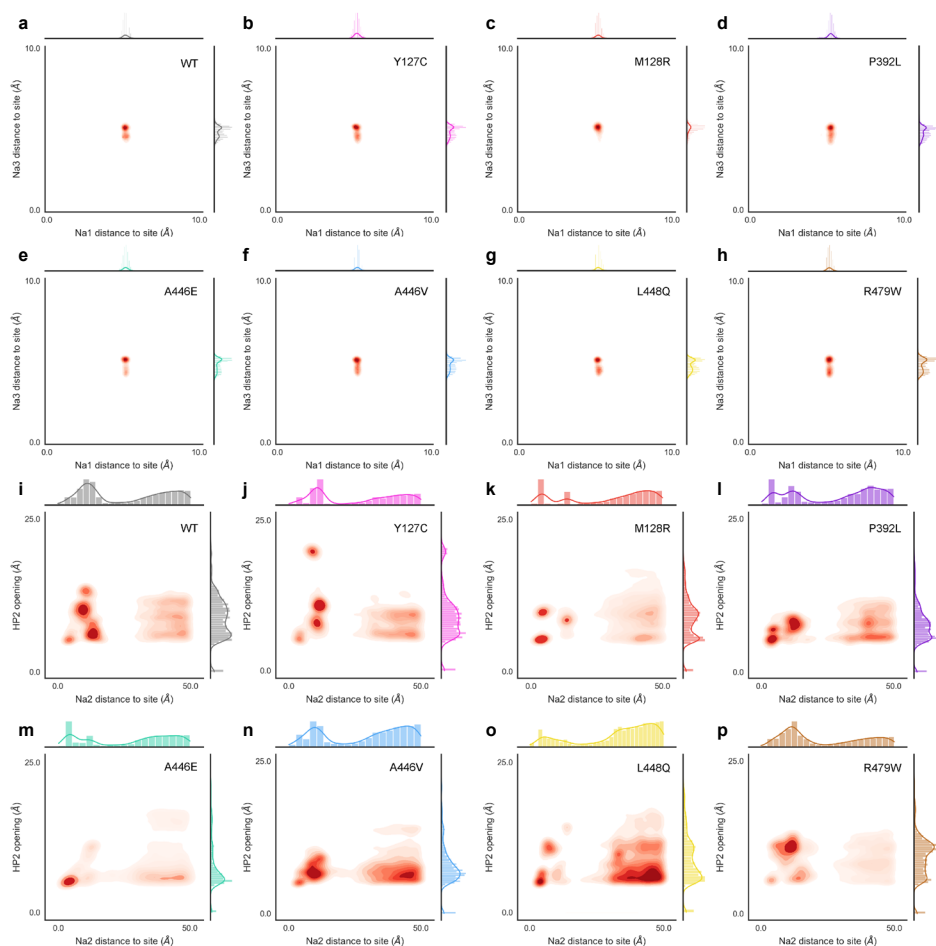
Supplementary Figure 6.6. Modulation of cellular responses by L-glutamate and inhibitors in an impedance-based phenotypic assay on M128R cells. **a)** Pretreatment with EC₅₀ (200 nM) TFB-TBOA and stimulation with vehicle, 1 mM L-glutamate or EC₅₀ (6.3 μM) UCPH-101. **b)** Pretreatment with 6.3 μM UCPH-101 and stimulation with vehicle, 1 mM L-glutamate or 200 nM TFB-TBOA. **c)** Pretreatment with 1 mM L-glutamate and stimulation with vehicle, 6.3 μM UCPH-101 or 200 nM TFB-TBOA. **d)** Pretreatment with 1 μM ouabain (Na⁺/K⁺-ATPase inhibitor) and stimulation with vehicle, 1 mM L-glutamate, 6.3 μM UCPH-101 or 200 nM TFB-TBOA. Data show vehicle-corrected normalized Cell Index traces of M128R cells pretreated for 60 min and subsequently stimulated for 120 min. Traces were normalized at the time point prior to pretreatment. Cells pretreated and stimulated with vehicle (PBS/DMSO) were used for vehicle correction.



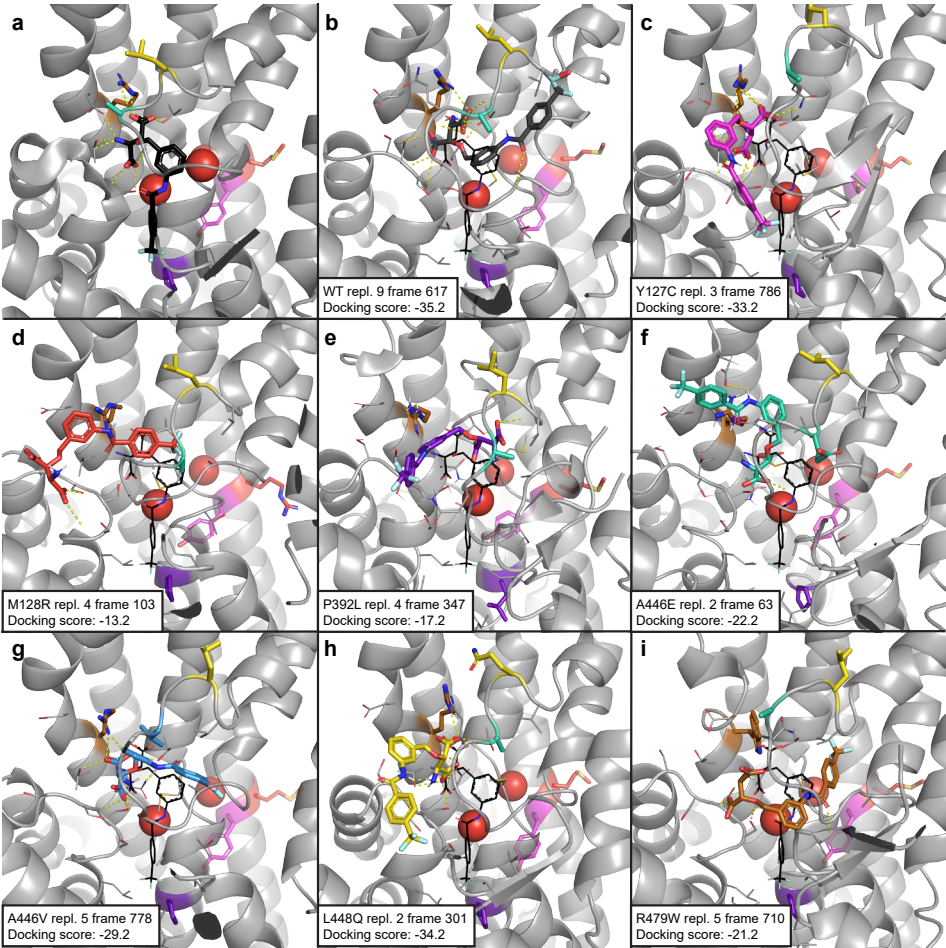
Supplementary Figure 6.7. Substrate–protein interactions across MD simulation replicates in EAAT1_{WT} chain A over time. In blue, the total number of L-Asp contacts with EAAT1 measured over the simulation time per replicate (500 ns). In different shades of orange, the number of contacts recorded with each residue at each simulation point.



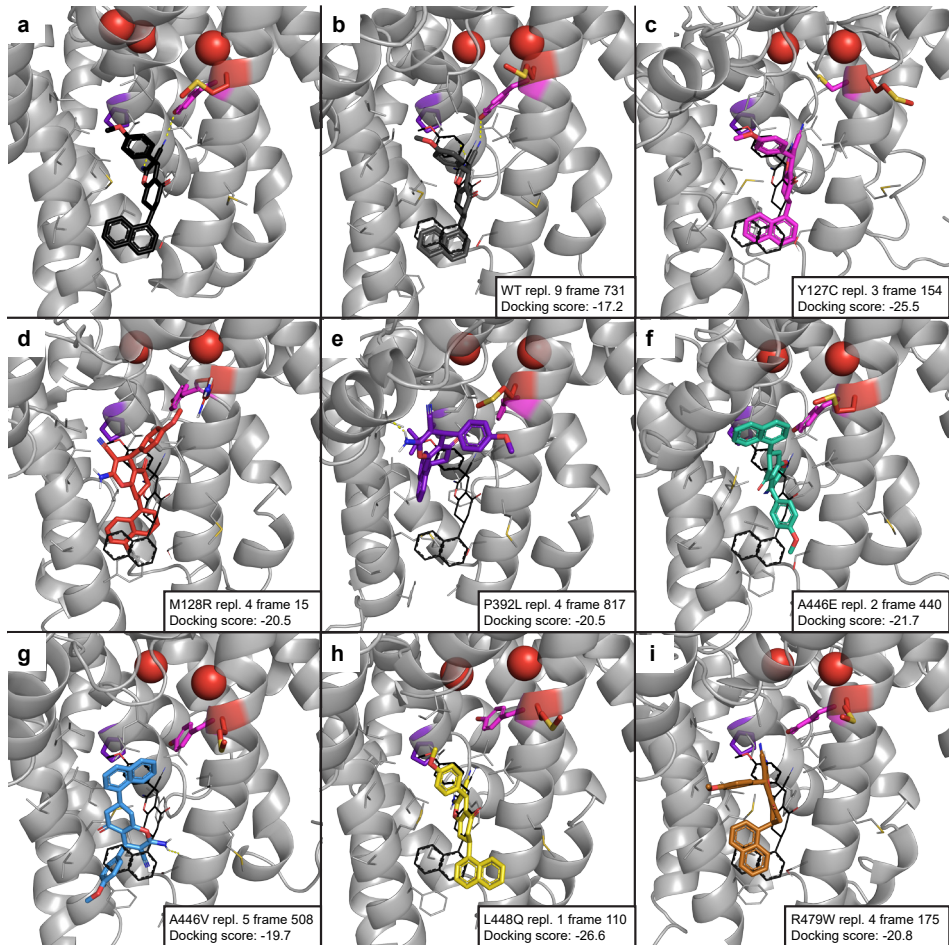
Supplementary Figure 6.8. Substrate–protein interactions across MD simulation replicates in EAAT mutant R479W chain A over time. In blue, the total number of L-Asp contacts with EAAT1 measured over the simulation time per replicate (500 ns). In different shades of orange, the number of contacts recorded with each residue at each simulation point. Replicates where protein RMSD reached 10 Å are labeled as “Replicate not stable” and the contacts are not reported.



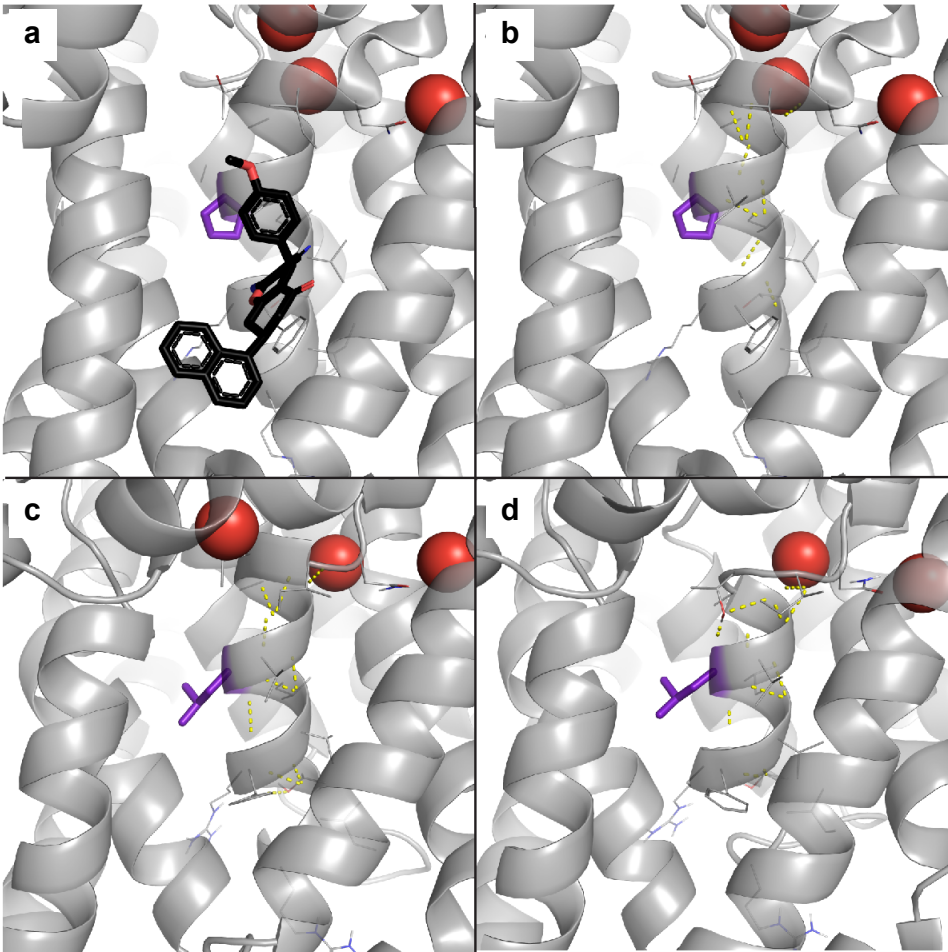
Supplementary Figure 6.9. Sodium ion coordination stability and HP2 domain opening sampling density derived from Molecular Dynamics simulations on EAAT1_{WT} and mutants. Sodium ion coordination stability is represented by the distance from the Na⁺ atom to the C α of one of its coordinating residues in sites Na1 (D487), Na2 (T396), and Na3 (D400). HP2 opening was calculated as the distance between S366 C α (HP1 tip) and G442 C α (HP2 tip). Sampling density was calculated across all frames in all replicates simulated for Na1-Na3 ion coordination stability (**a-h**) and Na2 coordination stability-HP2 opening (**i-p**). Density was analyzed for both pairs in combination (inside the axes box) and independently (outside the axes) for EAAT1_{WT} (**a,i**) and mutants (**b-h;j-p**).



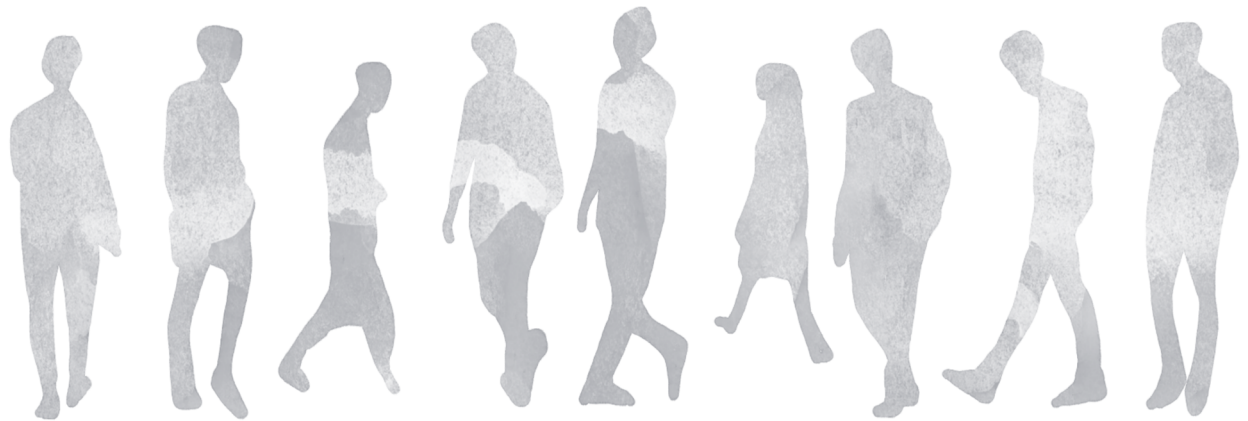
Supplementary Figure 6.10. Molecular docking top poses of orthosteric inhibitor TFB-TBOA in EAAT1 MD frames with most representative HP2 opening distances. Docking performed in chain A of a random selection of frames with the top five most common HP2 opening distances across all replicates and frames. TFB-TBOA binding pocket was derived from its co-crystallized pose in PDB 5MJU (**a**), represented in black for reference next to the docking poses generated in EAAT1_{WT} (**b**) and mutants (**c-i**). (Mutated) residues of interest are represented in the following colors: Y127 pink, M128 red, P392 purple, A446 green (or blue for A446V mutant), L448 yellow, and R479 brown. Coordinated Na⁺ ions are represented as red spheres. Hydrogen bonds are represented as dashed yellow lines.



Supplementary Figure 6.11. Molecular docking top poses of allosteric inhibitor UCPH-101 in EAAT1 MD frames with most representative HP2 opening distances. Docking performed in chain A of a random selection of frames with the top five most common HP2 opening distances across all replicates and frames. UCPH-101 binding pocket was derived from its co-crystallized pose in PDB 7AWM (a), represented in black for reference next to the docking poses generated in EAAT1_{WT} (b) and mutants (c-i). (Mutated) residues of interest are represented in the following colors: Y127 pink, M128 red, P392 purple, A446 green (or blue for A446V mutant), L448 yellow, and R479 brown. Coordinated Na⁺ ions are represented as red spheres. Hydrogen bonds are represented as dashed yellow lines.



Supplementary Figure 6.12. Effect of P392L mutant in Pro-induced TM7a helix kink. **a)** TM7a helix stabilizes the allosteric pocket where UCPH-101 inhibitor binds. Visualization in chain A of PDB 7AWM. **b)** P392 (purple) induces a kink in the TM7a helix that is represented by a lack of an additional hydrogen bond (dashed yellow lines) in that helix turn. **c-d)** P392L mutation reverts the Pro-induced kink, as represented by an additional hydrogen bond in that helix turn. Visualization in chain A of replicate 4 MD trajectory frames 347 (c) and 817 (d).

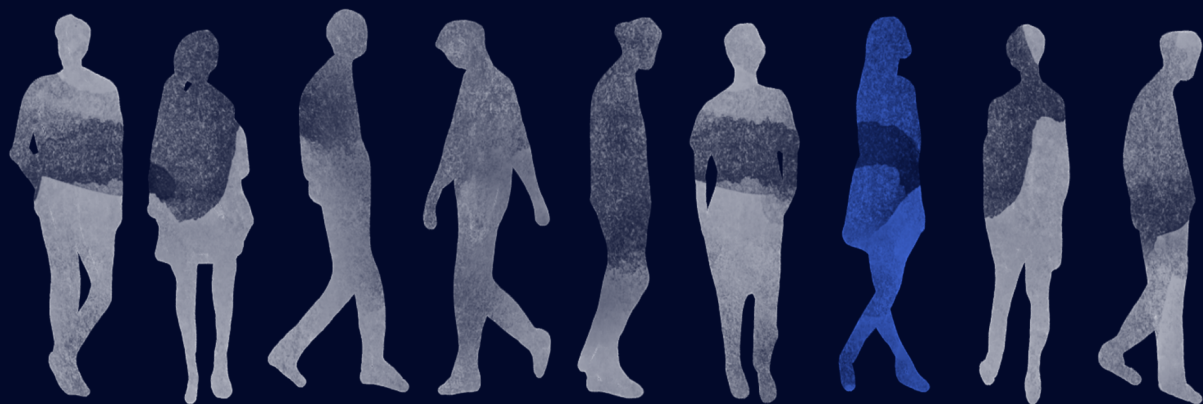


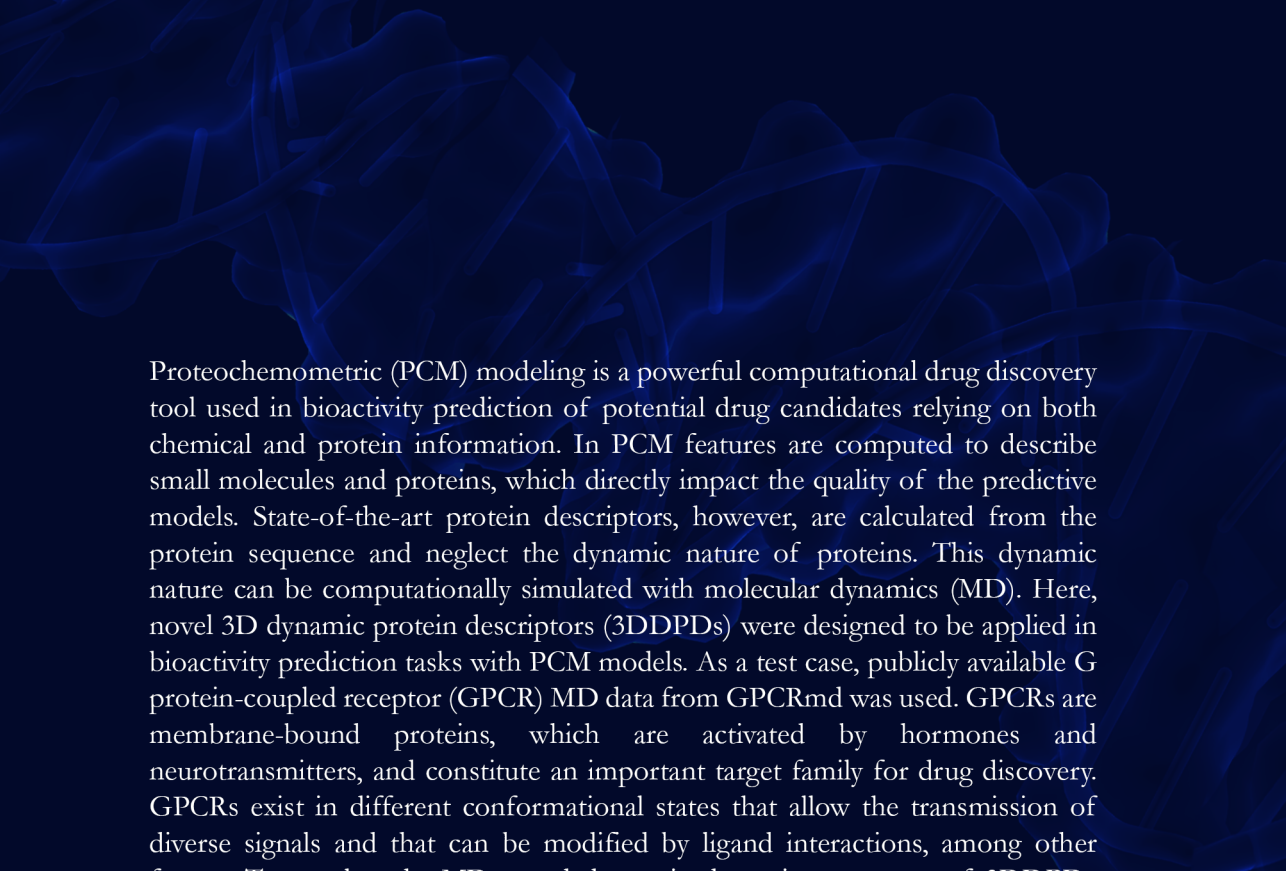
Chapter 7

3DDPDs: Describing protein dynamics for
proteochemometric bioactivity prediction. A case
for (mutant) G protein-coupled receptors


Marina Gorostiola González, Remco L. van den Broek, Thomas G.M. Braun,
Magdalini Chatzopoulou, Willem Jaspers, Adriaan P. IJzerman, Laura H. Heitman,
Gerard J.P. van Westen

Adapted from : *Journal of Cheminformatics* **15**, 74 (2023)



An abstract graphic consisting of several overlapping, flowing blue ribbons or strands that create a sense of movement and complexity, resembling a protein structure or a molecular pathway. It is positioned in the upper half of the page, behind the main text.

Proteochemometric (PCM) modeling is a powerful computational drug discovery tool used in bioactivity prediction of potential drug candidates relying on both chemical and protein information. In PCM features are computed to describe small molecules and proteins, which directly impact the quality of the predictive models. State-of-the-art protein descriptors, however, are calculated from the protein sequence and neglect the dynamic nature of proteins. This dynamic nature can be computationally simulated with molecular dynamics (MD). Here, novel 3D dynamic protein descriptors (3DDPDs) were designed to be applied in bioactivity prediction tasks with PCM models. As a test case, publicly available G protein-coupled receptor (GPCR) MD data from GPCRmd was used. GPCRs are membrane-bound proteins, which are activated by hormones and neurotransmitters, and constitute an important target family for drug discovery. GPCRs exist in different conformational states that allow the transmission of diverse signals and that can be modified by ligand interactions, among other factors. To translate the MD-encoded protein dynamics two types of 3DDPDs were considered: one-hot encoded residue-specific (rs) and embedding-like protein-specific (ps) 3DDPDs. The descriptors were developed by calculating distributions of trajectory coordinates and partial charges, applying dimensionality reduction, and subsequently condensing them into vectors per residue or protein, respectively. 3DDPDs were benchmarked on several PCM tasks against state-of-the-art non-dynamic protein descriptors. Our rs- and ps3DDPDs outperformed non-dynamic descriptors in regression tasks using a temporal split and showed comparable performance with a random split and in all classification tasks. Combinations of non-dynamic descriptors with 3DDPDs did not result in increased performance. Finally, the power of 3DDPDs to capture dynamic fluctuations in mutant GPCRs was explored. The results presented here show the potential of including protein dynamic information on machine learning tasks, specifically bioactivity prediction, and open opportunities for applications in drug discovery, including oncology.

A row of stylized, dark blue silhouettes of people of various ages and ethnicities walking from left to right. The silhouettes are simple and modern, representing a diverse group of individuals. They are located at the bottom of the page, below the main text.

Introduction

Proteins are complex biological units that constitute the basis for cellular function. As such, studying their structure and interaction with the environment is a key aspect of preclinical drug discovery¹. In computational drug discovery, the information encoded in proteins can be extracted and leveraged for several applications using machine learning². These include, among others, target identification³, computational mutagenesis⁴, protein-protein interaction studies^{5,6}, and small molecule-target binding affinity prediction^{7,8}. The latter, also referred to as bioactivity proteochemometric modeling (PCM), is an extension of the widely employed quantitative structure-activity relationship (QSAR) models enriched with protein descriptors⁷.

Several types of protein descriptors are available for PCM modeling and similar applications⁷⁻⁹. These can be broadly classified between sequence-based and structure-based descriptors. Descriptors derived from the protein sequence include discrete features calculated per residue (one-hot encoding)¹⁰ or protein¹¹ capturing physicochemical properties or amino acid composition. Additionally, deep learning applications of natural language processing have prompted the generation of protein embeddings from sequences¹². Structure-based descriptors can be derived from molecular graphs or the protein 3D structure by measuring connectivity, distances, and physicochemical properties among others^{8,9}. Moreover, ligand-protein interaction fingerprints can be derived from protein structures in complex with small molecules¹³ or from combinations of ligand and protein descriptors¹⁴.

While the goal of protein descriptors is to capture the full complexity of the protein, they largely fail to depict protein dynamism. At physiological temperatures, proteins exist in an equilibrium of structural conformations, which can be studied experimentally or simulated with Molecular Dynamics (MD)¹⁵. Changes in metabolite or ligand concentrations, as well as mutations and other structural alterations, can impact protein dynamics^{15,16}. These, in turn, directly influence protein function and interactions^{15,17}. The inclusion of dynamic information in protein descriptors could therefore increase performance in some of the machine learning applications listed above. Positive effects have already been reported in target and functional site identification¹⁸, but this potential is yet to be explored in PCM bioactivity modeling.

G protein-coupled receptors (GPCRs) have extensively been explored as targets in bioactivity prediction, including PCM, due to their biological and therapeutic relevance^{19,20}. GPCRs as a family share a highly conserved structure with seven transmembrane (TM) domains that exists in a dynamic equilibrium between active and inactive conformations^{21,22}. In the last decades, the scientific community has seen an increasing interest in the dynamic aspects of GPCRs, resulting in community efforts such as the GPCRmd database, where curated GPCR MD simulations are publicly available²³. Simultaneously, GPCR research in the context of oncological therapies is gaining momentum as explored in **Chapter 5**²⁴, with several *in vitro* studies showing how cancer-related somatic mutations affect receptor function and/or pharmacological intervention²⁵⁻²⁷. Some of the physiological effects observed in mutants have been associated with changes in

receptor dynamics thanks to MD simulations²⁸.

Here, 3D dynamic protein descriptors (3DDPDs) were developed leveraging atom coordinates and partial charges from publicly available single replicate MD simulations from GPCRmd. Two descriptor architectures were explored: embedding-like (protein specific – ps3DDPD), and one-hot encodings (residue specific – rs3DDPD). The performance in PCM GPCR bioactivity prediction of these novel protein descriptors was benchmarked against and in combination with a panel of state-of-the-art protein descriptors. Finally, the ability of our 3DDPDs to capture dynamic changes driven by (cancer-related) somatic point mutations in GPCRs was tested. These results highlight 3DDPDs as a stepping stone for further research on protein descriptors used for predicting drug-target interactions based on protein dynamics.

Results

3DDPDs generation and optimization

3D dynamic protein descriptors (3DDPDs) were designed to capture the dynamic behavior of proteins in MD simulations. For this purpose, atomic coordinates were first extracted from the MD trajectories, and their variability over a certain number of frames calculated. As proof of concept, 3DDPDs were conceived for single MD trajectory replicates in this work. In order to account not only for the position but also for the type of atoms in the protein, atomic partial charges were computed. Next, two strategies were developed to condense the dense atomic information into protein descriptors (**Figure 7.1**). These strategies correspond to the two types of 3DDPDs envisioned. The residue-specific (rs)3DDPD is closer to classical one-hot encoded protein descriptors and defines each residue in the protein with a fixed number of features. The rs3DDPD was designed to capture the differences across different sections of the target. The second type, protein-specific (ps)3DDPD, is closer to whole sequence protein embeddings and was designed to capture the differences between targets in a set. Consequently, atomic data were aggregated per target for rs3DDPDs and for all targets for ps3DDPDs and its dimensionality was reduced via principal component analysis (PCA). Several principal components (PCs) for each atom were selected and, in the case of rs3DDPDs, grouped per residue. A second dimensionality reduction step was applied to residue data and the selected PCs were placed in their matching sections corresponding to a multiple sequence alignment (MSA) of the targets of interest. For ps3DDPDs, the PCs selected per atom were grouped per target, resulting in the final descriptor.

The 3DDPD generation strategy described above was optimized by comparing the descriptors' performance on PCM modeling tasks. GPCRs were selected as the protein family for this case study given the availability of a large number of MD trajectories freely in the GPCRmd database²³. Particularly, the focus laid on Class A GPCR apo structures in the inactive or intermediate conformations, more broadly represented at the time of the analysis. The PCM dataset contained 26 GPCRs with available MD trajectories in GPCRmd and high-quality data in the Papyrus bioactivity dataset²⁹, in total

38,701 datapoints. Although two data split strategies (i.e. random and temporal) were applied in both regression and classification PCM tasks, the optimization strategy was driven mostly by the results in the most demanding task, regression with a temporal split.

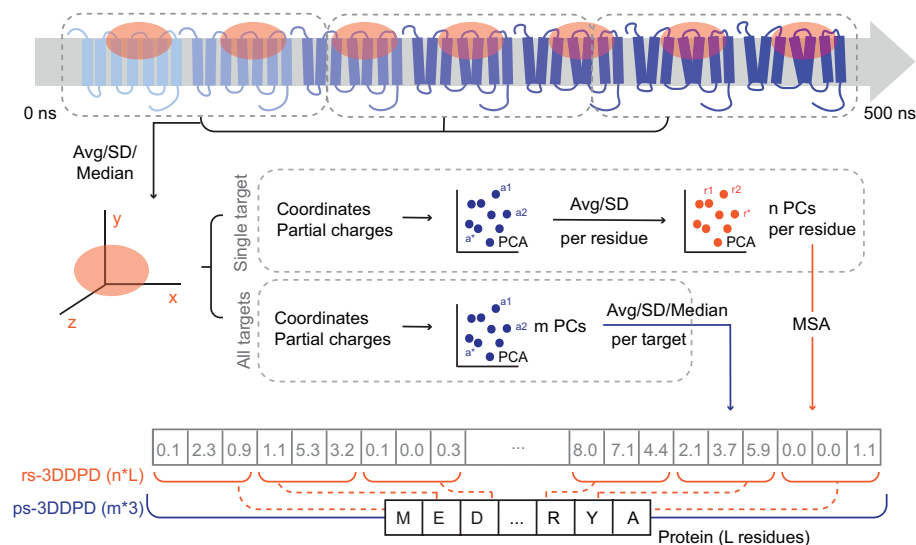


Figure 7.1. 3D dynamic descriptor (3DDPD) generation overview. First, a selection of residues and atoms is made. XYZ coordinates are collected for the selected atoms over all frames of the trajectory. The full simulation ranging from 0 to 500 ns is divided into sub-trajectories and atomic coordinate statistics (average, SD, and median) are computed for each of them. Two routes are possible from this point to generate either one-hot encoded residue-specific rs3DDPDs or embedding-like protein-specific ps3DDPDs. Respectively, atomic data is grouped and standardized either per target or for all targets and PCA is computed. A number of PCs for each atom are then selected and, in the case of rs3DDPDs, grouped per residue by calculating the average and SD. A second dimensionality reduction step is applied to residue data and the selected n number of PCs are mapped to their corresponding positions in an MSA of the targets of interest. This results in a vector rs3DDPD of length $n \times L$, where L is the length of the protein or the MSA. For ps3DDPDs, the m number of PCs selected per atom are grouped per target by calculating average, median, and SD, therefore resulting in the final vector descriptor of length $m \times 3$.

First, the “dynamic” properties derived from atomic coordinates were optimized. Here, the use of mean, median, and standard deviation from the mean (SD) or just the SD, representing the “rigidity” of each atomic coordinate was benchmarked. For rs3DDPDs, using SD resulted in better performance (**Figure 7.2a**), contrary to ps3DDPDs (**Figure 7.2b**). The number of frames included in each trajectory split was also optimized, where 100 or 500 frames yielded similarly better results (**Figure 7.2a**), so 100 frames were selected further. The variance explained by the selected number of PCs on atom data was optimized and set at 95% for both rs3DDPDs and ps3DDPDs (**Figure 7.2b**), and similarly, the number of PCs on residue data was optimized and set to 5 not to explode the number of features (**Figure 7.2a**).

Furthermore, the inclusion of atomic data from all heavy atoms or non-carbon atoms only was tested. The former option was significantly better for both rs3DDPDs (**Figure 7.2a**) and ps3DDPDs. Finally, residue selection strategies were tested to focus

the 3DDPDs on the protein binding site (**Figure 7.2c**). These selections were based on structural-driven MSAs at different protein family levels, starting from the full sequence, then the binding pocket of class A GPCRs, then specific GPCR families, such as nucleotide receptors, then GPCR subfamilies, such as adenosine receptors, and finally, target-specific binding pocket such as the adenosine A₁ receptor. To ensure a consistent number of features per descriptor, in rs3DDPDs only the first two options could be tested, where the class A binding pocket performed significantly worse than the full sequence (**Figure 7.2a**). In ps3DDPDs all selection methods performed similarly except for the family and target pockets, which performed significantly worse (**Figure 7.2b**).

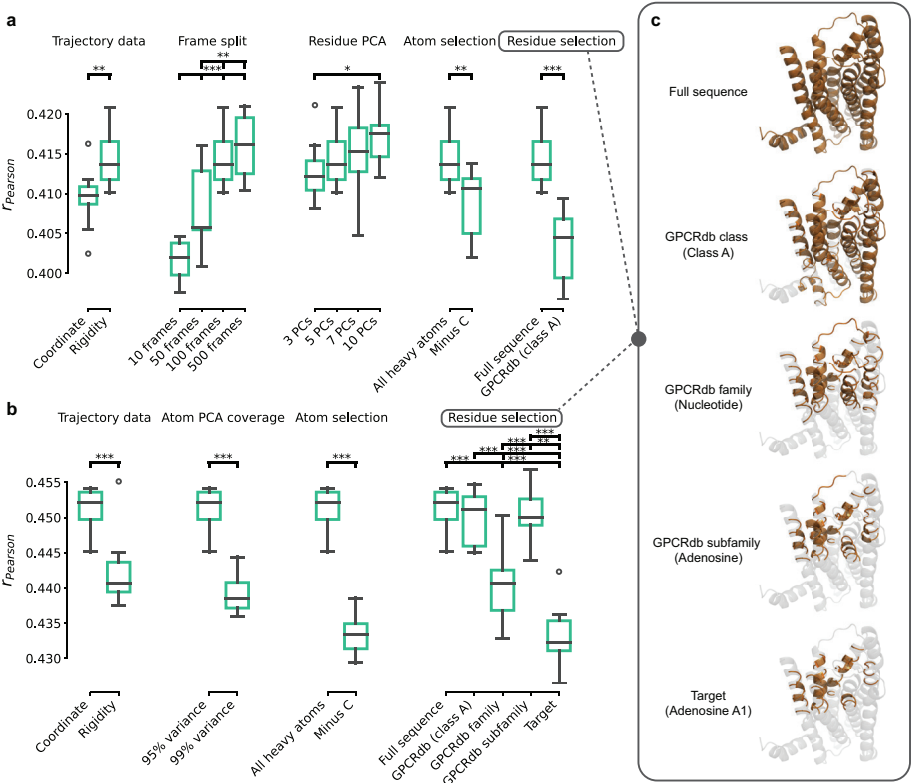


Figure 7.2. Optimization of the 3DDPD generation strategy. Ten PCM regression tasks with temporal split were trained with each variation of the 3DDPDs to select the optimal parameters. Pairwise differences were analyzed by their statistical significance in a Student's T test, represented by asterisks in (a,b): * = p-value < 0.05; ** = p-value < 0.01; *** = p-value < 0.001. **a)** rs3DDPDs were optimized by testing different options for trajectory data (i.e. choices of statistical metrics for sub-trajectory grouped coordinate atomic data: “coordinate” includes all, “rigidity” only SD), number of frames in the sub-trajectory frame splits, number of PCs from the residue PCA, atom selection (i.e. all heavy atoms or “minus C”: non-carbon), and residue selection (i.e. full sequence or class A GPCRdb-annotated binding pocket). **b)** ps3DDPDs were optimized based on trajectory data, variance covered by the selected number of atom PCA components, atom selection, and residue selection. **c)** Residue selection options exemplified on the structure of adenosine A1 receptor PDB 5UEN. In orange, the residues that would be selected by each of the five possible definitions of a structural-driven binding pocket selection approach: full sequence, class A, family, subfamily, and target.

The optimized rs3DDPD included “Rigidity” coordinate data calculated from 100-frame splits, where all atomic data was included for all residues in the protein sequence. In the atomic PCA, 95% of the variability was kept and 5 PCs in the residue PCA. This resulted in a vector of 3,785 features for the class A GPCRdb MSA used, of length 757. The optimized ps3DDPD included all coordinate data statistics calculated from 100-frame splits, where all atomic data was included for all residues in the protein sequence, and 95% of the variability was kept in the atomic PCA. This resulted in a vector of 30 features.

3DDPDs reflect the GPCR dynamic fluctuations

From the publicly available MD database for GPCRs, GPCRmd, a subset of 26 trajectories for class A GPCRs with sufficient bioactivity data for PCM modeling was selected, as described in the *Materials and Methods* section. Apo inactive conformations were selected to avoid bias towards a specific ligand-triggering activation mode. The targets selected covered 17 subfamilies within four class A families: aminergic, lipid, nucleotide, and peptide receptors. The analysis of the MD trajectories showed similarities between dynamic behaviors but also differences that can be potentially captured and exploited using 3DDPDs. Such differences can be better observed by aligning the Root Mean Square Fluctuation (RMSF) values to a GPCR class A MSA (**Figure 7.3a** and **Supplementary Figure 7.1**). Across GPCRs, there is a shared pattern of reduced mobility in the TM domains compared to extracellular (ECL) and intracellular (ICL) loops or N- and C-terminus. However, deviations from this pattern are common when comparing i) members of different families (e.g. adrenergic 5-hydroxytryptamine receptor 1_B (5HT1B) and nucleotide adenosine A₁ receptor (AA1R) in their overall dynamic behavior), ii) members of the same family but different subfamilies (e.g. nucleotide receptors adenosine A_{2A} (AA2AR) and P2Y purinoceptor 1 (P2RY1) in TM2, ICL2, ECL2, ICL3, and C-terminus), or iii) even members of the same subfamily (e.g. 5-hydroxytryptamine receptors 5HT1B and 2_B (5HT2B) in N-terminus, TM3, TM4, ECL2, ICL3, and ECL3). Importantly, the main dynamic patterns described above were highly conserved for the three different replicates of the same system available on GPCRmd (**Supplementary Figure 7.2**), suggesting that the omission of MD replicates in the current 3DDPD pipeline did not have a major impact on the results presented here.

The observed similarities and differences in dynamic behaviors between GPCRs were effectively captured by the optimized rs3DDPDs (**Figure 7.3b** and **Supplementary Figure 7.3**) and ps3DDPDs (**Figure 7.3c** and **Supplementary Figure 7.4**). In the translation from RMSF to rs3DDPD and ps3DDPD, positive and negative values appeared that represented inter- and intra-target variability, respectively. While rs3DDPDs reflected the dynamic fluctuations on a residue level that resembled more closely the RMSF pattern itself, ps3DDPDs showed a more generalized embedding of each protein dynamics compared to all the targets in the set thus enhancing the differences among targets. Of note, rs3DDPDs did not represent merely a transform of the RMSF values, as exemplified for the positions corresponding to the N-terminus and TM1 in P2RY1 and P2RY12 (**Figure 7.3a,b**). This suggests that information other than the atom

coordinate variability, such as the type of atoms and residues encoded by partial charges, was picked up by the 3DDPDs. In part, such an effect was likely possible thanks to the dimensionality reduction process that introduced several opportunities to exploit atomic and residue similarities and differences as opposed to the RMSF calculation.

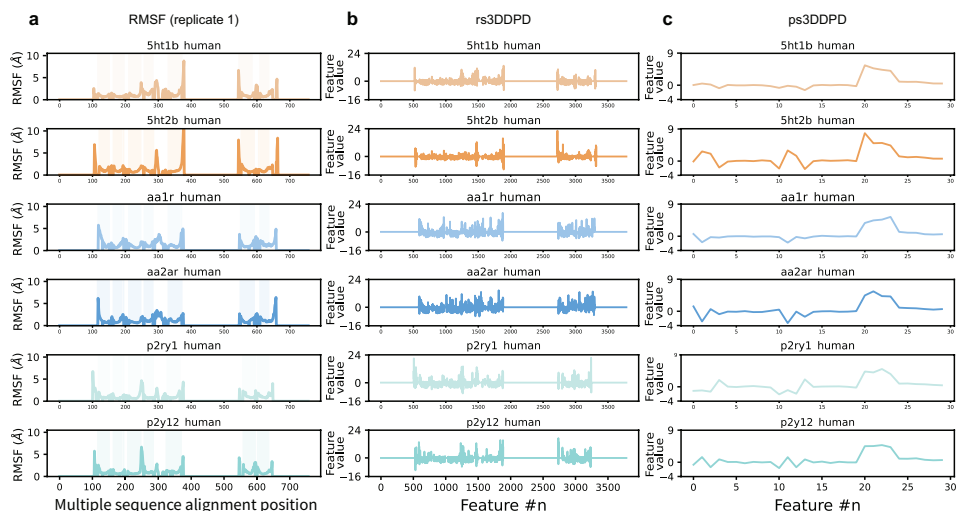


Figure 7.3. Representation of the GPCRs dynamic behavior by 3DDPDs. **a)** Dynamic fluctuations of the residues of six GPCRs from the set, represented by their RMSF (Å). The RMSF values are mapped to their corresponding positions in the MSA later used for rs3DDPD and non-dynamic descriptor calculation, for easier visualization. The regions in the MSA corresponding to domains TM 1-7 are shadowed for reference. Data for the complete set of 26 GPCRs is available in Supplementary Figure 7.1. **b)** Representation of the rs3DDPD feature values for the same subset of GPCRs. Data for the complete set of 26 GPCRs is available in Supplementary Figure 7.3. **c)** Representation of the ps3DDPD feature values for the same subset of GPCRs. Data for the complete set of 26 GPCRs is available in Supplementary Figure 7.4.

3DDPDs outperform non-dynamic protein descriptors in PCM regression tasks

The use of 3DDPDs as protein descriptors in PCM bioactivity modeling tasks was tested for our GPCR dataset. For this purpose, the performance of random forest (RF) models was benchmarked using 3DDPDs in combination with ECFP6 molecular fingerprints against models using as protein descriptors one of five other one-hot encoded descriptors (i.e. Zscale in two modalities, STscale, MS-WHIM, and PhysChem) or one protein embedding (i.e. UniRep). The benchmark was carried out for classification and regression tasks using two different types of training-test splits: 80:20 random split and temporal split with 2013 as a cutoff year for the test set. The temporal split was introduced as a more accurate representation of a drug discovery campaign where data from the past is used to predict novel chemical entities developed later in time and indeed showed a considerable decrease in chemical bias compared to the random split (0.051 vs. 0.279).

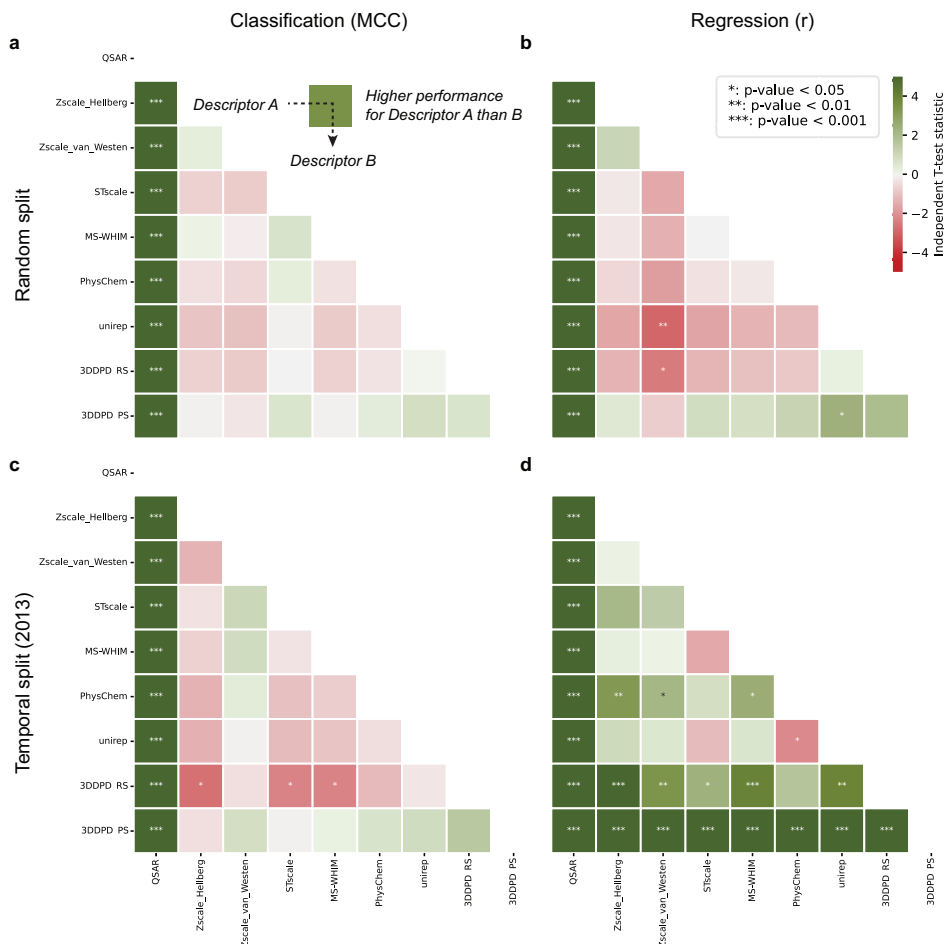


Figure 7.4. Benchmark of 3DDPD performance in PCM bioactivity modeling tasks against non-dynamic descriptors. Ten RF models with random seeds were trained and validated for each combination of protein descriptors with ECFP6 molecular fingerprints. A shade of green (the darker the better) represents better performance using a descriptor A instead of a descriptor B, as read in panel a. A shade of red (the darker the worse) represents worse performance using a descriptor A instead of a descriptor B. The statistical significance of the differences is derived from pairwise Student T-test and represented by asterisks: * = p-value < 0.05; ** = p-value < 0.01; *** = p-value < 0.001. Four PCM tasks were benchmarked: **a)** Classification with validation based on an 80:20 random split. In classification tasks, MCC was used as an evaluation metric on the test set. **b)** Regression with validation based on an 80:20 random split. In regression tasks, Pearson r was used as an evaluation metric on the test set. **c)** Classification with validation based on a temporal split, with 2013 as the cutoff year. **d)** Regression with validation based on a temporal split, with 2013 as the cutoff year.

The bioactivity dataset compiled for bioactivity modeling contained 38,701 bioactivity datapoints heterogeneously distributed across the 26 targets (**Supplementary Table 7.1**). Active data for classification was defined with a cutoff of 6.5 pchembl value. Firstly, the need for PCM modeling in such a set was assessed by comparing the performance of the PCM models to the average performance of individual QSAR models for each

of the GPCRs in the set. In all of the modeling scenarios, the worst performing PCM model outperformed significantly the QSAR models: Matthews correlation coefficient (MCC) 0.643 ± 0.005 (UniRep) vs. 0.578 ± 0.007 in random split classification, MCC 0.273 ± 0.003 (rs3DDPD) vs. 0.192 ± 0.009 in temporal split classification, Pearson r 0.832 ± 0.003 (UniRep) vs. 0.775 ± 0.005 in random split regression, and Pearson r 0.410 ± 0.003 (Zscale Hellberg) vs. 0.343 ± 0.004 in temporal split regression.

In PCM, models using 3DDPDs performed similarly to using other protein descriptors in classification tasks regardless of the split type (**Figure 7.4a,c**). One exception was the temporal split classification task, here rs3DDPDs produced slightly worse performance than models using Zscale Hellberg, Stscale, and MS-WHIM (MCC 0.273 ± 0.003 vs. 0.273 ± 0.005 , 0.278 ± 0.005 and 0.277 ± 0.004 , respectively, **Figure 7.4c**). In the regression task with random split, models using 3DDPDs performed again similarly to models using other protein descriptors (**Figure 7.4b**), with the exception of rs3DDPDs performing slightly but significantly worse than Zscale van Westen (Pearson r 0.832 ± 0.004 vs. 0.836 ± 0.004 , respectively) and ps3DDPDs performing slightly better than the UniRep protein embedding (Pearson r 0.835 ± 0.003 vs. 0.832 ± 0.003 , respectively). In the regression task with temporal split, however, both types of 3DDPDs outperformed the rest of the descriptors (**Figure 7.4d**). The performance of models trained with non-dynamic protein descriptors measured as Pearson r ranged from 0.410 ± 0.003 (Zscale Hellberg) to 0.415 ± 0.004 (PhysChem) passing by 0.410 ± 0.006 (Zscale van Westen), 0.410 ± 0.004 (MS-WHIM), 0.411 ± 0.004 (UniRep), and 0.413 ± 0.005 (Stscale). One-hot encoded rs3DDPDs performed significantly better than most of the other descriptors, except for PhysChem, with a Pearson r of 0.417 ± 0.004 . Embedding-like ps3DDPDs, however, significantly outperformed all the other descriptors, including rs3DDPDs, with a Pearson r of 0.451 ± 0.003 . These results were also confirmed in terms of Root Mean Square Error (RMSE), which was the lowest for ps3DDPDs (1.154 ± 0.003) and then QSAR models on average (1.168 ± 0.004), followed by rs3DDPDs (1.214 ± 0.005) and then the rest of non-dynamic protein descriptors (from 1.124 ± 0.005 to 1.221 ± 0.006). A summary of all validation metrics is given in **Supplementary Table 7.2** (random split) and **Supplementary Table 7.3** (temporal split).

In order to test the complementarity of the 3DDPDs with other protein descriptors, a set of regression models was trained with temporal splits with pairs of dynamic and non-dynamic protein descriptors (**Figure 7.5**). In all cases, the addition of a 3DDPD on top of a non-dynamic descriptor resulted in similar performance to the models trained exclusively using non-dynamic descriptors, or even slightly worse in the case of PhysChem + rs3DDPD. Moreover, the combination yielded statistically worse performance than using the dynamic descriptors alone, particularly in the case of ps3DDPD. This non-complementarity was further confirmed for ps3DDPDs by their exclusion from the most important features for the combination models (e.g. ps3DDPD + PhysChem, **Supplementary Figure 7.5d**), where only non-dynamic protein descriptor features and ECFP6 compound fingerprint bits were picked up as the top 25 most important for the model. For rs3DDPDs, however, there seemed to be a certain complementarity as both dynamic and non-dynamic protein descriptor features showed up among the top 25 most important for the model (e.g. rs3DDPD + Zscale van Westen,

Supplementary Figure 7.5c), even if this did not translate into an improvement in model performance.

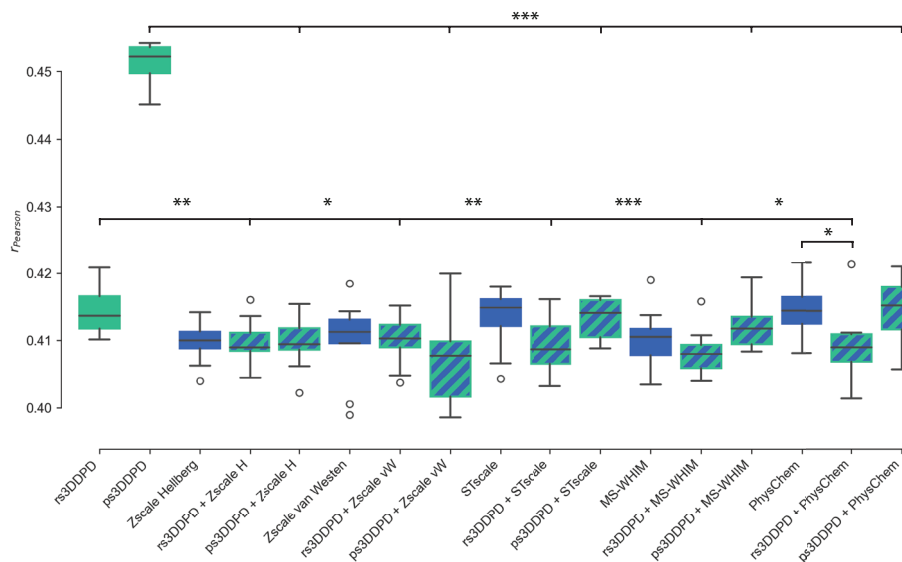


Figure 7.5. PCM model performance with dynamic and non-dynamic protein descriptor combination in regression tasks with a temporal split. In green, the performance of RF models trained on 3DDPDs. In blue, RF models trained on non-dynamic protein descriptors. In green and blue, RF models trained on a combination of both types. Zscale Hellberg and van Westen are abbreviated to Zscale H and vW, respectively. The statistical significance of the differences is derived from pairwise Student T-test and represented by asterisks: * = p-value < 0.05; ** = p-value < 0.01; *** = p-value < 0.001.

rs3DDPD features can be traced back to generic GPCR positions

A specific trait of one-hot encoded protein descriptors is that every feature can be traced back to specific protein sequence residues or MSA positions. For class A GPCRs, the aligned positions can additionally be linked to generic positions in the GPCR structure with known functional relevance. The most widely used generic position identifier for class A GPCRs is the Ballesteros-Weinstein (BW) schema³⁰, which consists of a first number identifying the TM domain followed by a second number that represents the level of conservation in that helix around the most conserved position that gets the value 50. Using the GPCRdb³¹ MSA mapping to BW positions, the most important rs3DDPD features in regression models were traced back to their generic GPCR positions.

In the models built with a temporal split, four rs3DDPD features were among the top 25 most important (**Figure 7.6a**). The most important feature overall, *A4223_PC3*, corresponded to the BW position 3.32 in TM3. For further interpretability, this generic position can also be directly mapped to a specific residue in a protein of interest. As an example, in AA1R 3.32 it translated to Val 87 (**Figure 7.6b**). The other three important rs3DDPD features did not correspond to any BW positions, as two of them were located in the ECL2 and one in the ECL3. From the three loop positions, only one exists in

adenosine receptor A1, Asn 147 (*AA292_PC3*). The two other ECL positions are only available in other receptors (**Supplementary Figure 7.1**). In the models built with a random split, the two most important rs3DDPD features, *AA128_PC2* and *AA576_PC5*, corresponded to TM1 1.38 and TM6 6.46 BW positions, respectively (**Figure 7.6c**). In AA1R, these translated to Ile 15 and Leu 245 (**Figure 7.6d**). The other two important rs3DDPD features correspond to positions in ICL3. Of note, the consensus between seeds on the importance of specific rs3DDPD features was less marked on the models with random split than on the models with temporal split (**Figure 7.6a,c**). This analysis was further applied to discuss the relevance of specific GPCR positions in ligand binding.



Figure 7.6. GPCR generic position mapping of most important rs3DDPD features in PCM regression tasks. **a)** Top 25 most important features in PCM regression models using a temporal split validation for the GPCR set. The importance was averaged across the ten random seeds trained and the SD represented as error bars. Rs3DDPD features are mapped to their corresponding GPCR Ballesteros-Weinstein number or, if not available, region of the protein. **b)** Representation of the most important rs3DDPD features in regression temporal split in the adenosine A1 receptor (PDB 5UEN). **c)** Top 25 most important features in PCM regression models using a random split validation. **d)** Representation of the most important rs3DDPD features in regression random split in the adenosine A1 receptor).

Dynamic fluctuations in mutants can be captured with 3DDPDs

To assess the viability of dynamic descriptors to capture differences between mutants in a potential mutant PCM model, a subset of 28 mutants from five of the GPCRs in our set was gathered: AA1R and AA2AR, muscarinic acetylcholine receptor 2 (ACM2), beta-2 adrenergic receptor (ADRB2), and CC chemokine receptor 5 (CCR5). The selection of mutations was done for the original set of 26 GPCRs when there was available mutagenesis data in GPCRdb (**Table 7.2**), from which the point mutation's effect in bioactivity was projected for the five resulting receptors (**Supplementary Figure 7.6**). Additionally, five mutations on these GPCRs present in cancer patients from the Genomic Data Commons (GDC) database were included that also had mutagenesis data in GPCRdb: AA1R R291C^{7.56} and R296C^{8.51}, AA2AR H278N^{7.42}, ACM2 V421L^{7.33}, and ADRB2 V317A^{7.43}. The cancer-related mutants, however, did not seem to have an effect on bioactivity given the limited amount of mutagenesis data available.

The selected mutations were introduced in equilibrated wild-type (WT) receptor systems from GPCRmd, which were subsequently re-equilibrated to run production 500 ns MD simulations following the GPCRmd pipeline. One of the selected mutations did not run successfully therefore it was discarded from the analysis (AA2AR H278N^{7.42}). Most mutant trajectories showed deviations from WT trajectories in terms of RMSF (**Supplementary Figure 7.7**), with the exception of AA1R and CCR5 mutants. The deviations were sometimes in the vicinity of the mutation (i.e. AA2AR M177A^{5.40}, N181A^{5.43}, Y271A^{7.35}; ADRB2 D130N^{3.49}, S203A^{5.43}, V317A^{7.43}; ACM2 D103E^{3.32}, V421L^{7.33}), but most commonly spawned across the whole sequence or altered stability in distant regions. For example, in AA2AR L85A^{3.33} increased flexibility in ICL2 and ECL2 and S91A^{3.39} in ICL3 and TM6. Moreover, adjacent mutations that triggered different effects were observed. For example, in ADRB2, S203A^{5.43} decreased stability in TM1, ICL2, and ECL3, while S204A^{5.44} decreased stability in TM2 and TM4 while increasing stability in ICL3. Of note, in ACM2 D103E^{3.32} and D103N^{3.32} triggered similar higher flexibility in ECL1 and ECL2, with an overall differential pattern of lower stability in D103E^{3.32}. In general, the mutations with smaller dynamic fluctuations from the WT also corresponded to those with a smaller effect in bioactivity, such as AA1R R291C^{7.56} and R296C^{8.51}, and ADRB2 V317A^{7.43} (**Supplementary Figures 7.6, 7.7**).

Next, the power of 3DDPDs to distinguish between mutants was tested. rs3DDPDs and ps3DDPDs were computed for the mutant trajectories and used to cluster the mutants based on the distance between descriptors. As rs3DDPDs are computed independently for each trajectory and reflect all atoms in the system, all mutants of the same target clustered together (**Figure 7.7a**). Within targets, clusters of mutants with similar overall dynamic behavior compared to WT were observed, for example, ADRB2 D79N^{2.50} and D130N^{3.49}, or with similar fluctuations from WT in specific regions, such as AA1R R291C^{7.56} and R296C^{8.51} in TM7 and H8/C-terminus (**Supplementary Figure 7.7**). For targets with unique differential dynamic patterns from WT for each mutant, like ACM2, the clusters discerned the most different patterns (e.g. D103N^{3.32} shows certain receptor stabilization compared to D103E^{3.32} and V421L^{7.33}, and is therefore excluded from the cluster). These results supported the ability of rs3DDPDs to capture

dynamic fluctuations in mutants. Nevertheless, the mutant discriminatory power of rs3DDPDs did not correlate directly to that of using directly RMSF (**Supplementary Figure 7.8a**) or RMSF differences to WT (**Supplementary Figure 7.8b**), which reinforced the notion that rs3DDPDs are not merely a transform of RMSF and include other non-dynamic atomic information.

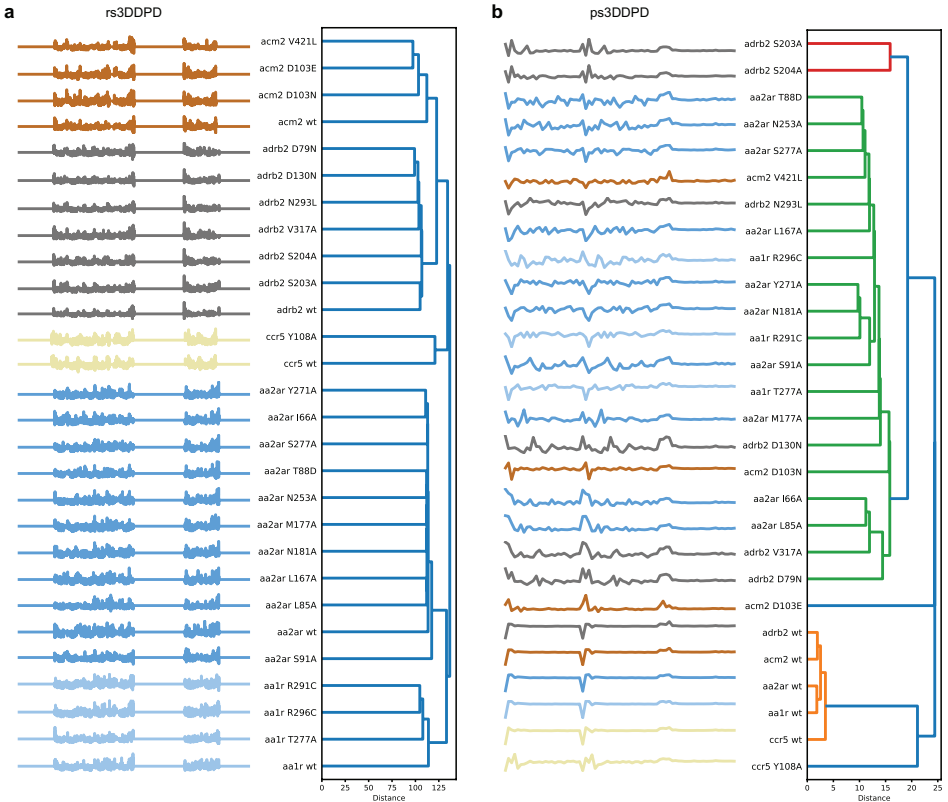


Figure 7.7. Discrimination of GPCR mutants using 3DDPDs as descriptors. Hierarchical clustering of GPCR variants based on their Euclidean distance between descriptor vectors. **a)** Mutants represented as rs3DDPDs. **b)** Mutants represented as ps3DDPDs. Individual clusters generated under a distance threshold of 70 % of the final merge are represented in different colors in the dendrograms.

Using ps3DDPDs, mutants were clustered based on overall similarities and differences in their dynamic behavior and residue composition across the set (**Figure 7.7b**). This way, the five WT targets clustered together because they had the most stable trajectories overall, and CCR5 Y108A^{3,32} was close by because overall it showed small differences to the WT trajectory (**Supplementary Figure 7.7**). However, some discrepancies with the expected results based on RMSF differences were found. For example, ADRB2 S203A^{5,43} and S204A^{5,44} formed their own cluster despite showing differential RMSF peaks. This and other examples suggest that ps3DDPD values for this set of mutants were heavily influenced by fluctuations in the N- and C-terminus, which were the most accentuated. Therefore ps3DDPDs did capture mutant fluctuations, but using them in their optimized form for WT GPCRs seemed suboptimal to discriminate mutants.

Discussion

PCM is a modality of bioactivity modeling that leverages similarities and differences between targets by combining them in the same model represented by protein descriptors⁷. The most commonly used protein descriptors in PCM characterize different properties of the sequence of residues¹⁰ but do not consider an important factor for protein-ligand binding: protein dynamics. Here, 3D dynamic protein descriptors (3DDPDs) were developed leveraging publicly available single-replicate MD simulations. This information was condensed into multiple steps that were optimized to produce a one-hot encoding residue-specific (rs3DDPD) and an embedding-like protein-specific (ps3DDPD) descriptor. The optimized 3DDPDs were subsequently benchmarked against non-dynamic protein descriptors in PCM tasks for a bioactivity set of 26 class A GPCRs. Finally, the use of 3DDPDs to describe point mutations was explored, which are otherwise under-represented by sequence-based non-dynamic descriptors.

The strategy to develop 3DDPDs borrows ingredients from other types of descriptors. Firstly the calculation of 3DDPDs starts from the collection of coordinate data for each atom, to which atomic partial charges were added to represent the electrostatic component over time (**Figure 7.1**). Other MD fingerprints for small molecules have used as starting properties the potential energy, solvent-accessible surface area, or radius of gyration³², ultimately similarly representing electrostatic and conformational changes of the molecule over time. More computationally expensive partial charges than Gasteiger could be explored, although the simpler implementation chosen here has been shown to be a cost-efficient option in other modeling tasks³³. Further down in our pipeline, PCA is used to reduce dimensionality, which is a common resource in protein descriptor calculation. However, for non-dynamic one-hot encoded descriptors, it is often used to calculate fixed features for each residue type (e.g. Zscale, MS-WHIM, Stscale^{10,34}) rather than specific features for each residue in the sequence, as was done for rs3DDPDs given the heavy influence of the environment in the dynamic behavior of single residues. On the other hand, protein embeddings are often the byproduct of a machine or deep learning model using a protein sequence as input^{12,35}, unlike the approach followed for ps3DDPDs. Here, instead, a common main framework was kept to increase the interpretability and interoperability of the resulting descriptors. This allowed us to follow a similar optimization route for both descriptor types (**Figure 7.2**). In terms of residue composition, for our particular dataset the full sequence was favored. In a less diverse GPCR set, however, the use of family- or subfamily-specific alignments and binding pocket selections would provide more relevant information to the model given the differential activation-induced conformational changes reported for GPCRs binding different ligand types²¹.

Next, the performance of our optimized 3DDPDs in PCM regression and classification tasks was tested using both random and temporal validation splits (**Figure 7.4**). The performance of our models was in line with other PCM models trained in similar conditions for subfamilies of GPCRs²⁹. In our set, 3DDPDs performed similarly to non-dynamic protein descriptors in classification tasks and regression tasks with a random split. These results suggest that the performance of these models had already reached its peak

and small differences in the way to represent the protein space did not make a difference. Nevertheless, the best-performing models in classification tasks did not reach a high MCC. Models reached 0.646 ± 0.009 in the random split (Zscale van Westen), and 0.278 ± 0.005 in the temporal split (Zscale Hellberg), hence questioning the relevance of this dataset for such task. Interestingly, protein embeddings (UniRep) showed lower performance across the board, which has also been shown in other datasets compared to sequence- and 3D-based protein descriptors³⁶. In the regression task with temporal split, however, 3DDPDs significantly outperformed non-dynamic descriptors. Given the more challenging form of validation introduced by the temporal split, the 3DDPDs represent an advantage. These results are likely also the result of performing 3DDPD optimization using this particular task. Nevertheless, similar behaviors have been observed in other benchmarks when using temporal splits compared to random splits^{29,37}. Moreover, in our PCM benchmark ps3DDPDs performed better than rs3DDPDs overall. One reason for this could be the difference in descriptor length: for the GPCR WT set, rs3DDPDs contained 3,785 features and ps3DDPDs 30 features. Moreover, the MSA used to compute rs3DDPD contained many gaps as it accounted for all class A GPCRs and not only the ones in the set. Therefore, lengthy rs3DDPDs with a large number of zeroes likely introduced noise in the model compared to the more compact ps3DDPDs. While this aspect would be corrected in practice by feature selection techniques prior to modeling, those were not applied here, similarly to hyperparameter optimization, to be able to explicitly benchmark the calculated descriptor with the least degrees of freedom. Finally, ps3DDPDs represent the overall differences between proteins in the set, which seems to be beneficial in agreement with the observation from Rackovsky and Scheraga that the description of the overall mobility of the protein correlates better with its structure than the description of individual residue mobility³⁸.

Subsequently, the biological relevance of the information contained in the 3DDPDs was investigated. One-hot encoding rs3DDPDs are calculated independently for each target and ps3DDPDs together for the targets in a particular set. Respectively, they exploit differences in atom coordinates and partial charges across positions in a target or a number of targets, representing the most relevant aspects of the protein dynamics, as defined by the RMSF fluctuations (**Figure 7.3, Supplementary Figures 7.1-7.4**). An advantage of rs3DDPDs is the possibility to be traced back to particular residues, alignment positions, or GPCR generic positions. This allowed us to investigate whether the 3DDPDs capture biologically relevant information from the MD simulation. To this end, the most important rs3DDPD features in regression PCM models were extracted and mapped to their corresponding GPCR generic positions (**Figure 7.6**). The most important feature in a temporal split corresponded to the BW position 3.32 in TM3. As an example, in AA1R this translated to Val 87, which lies within the orthosteric binding pocket and makes hydrophobic interactions with the endogenous ligand adenosine (PDB 7LD4³⁹). Other important rs3DDPD features were located in the ECL2 and ECL3, which as expected showed high flexibility in the MD simulations and are regions whose conformational changes are known to be relevant for ligand binding⁴⁰ and activation⁴¹. In the models built with a random split, the two most important rs3DDPD features corresponded to TM1 1.38 and TM6 6.46 BW positions, respectively. In AA1R, these translated to Ile 15 and Leu 245, which flank the binding site of non-endogenous co-crystallized antagonists

(PDB 5UEN⁴²). The other two important rs3DDPD features correspond to positions in ICL3, which are close to the G protein interface (PDB 7LD3³⁹). These results confirm that 3DDPDs capture relevant changes for GPCR ligand binding and activation and could help elucidate functional sites in orphan proteins. Similar approaches have previously leveraged MD information to identify relevant functional sites using deep learning models¹⁸ or graph-based approaches⁴³.

Finally, the use of 3DDPDs beyond WT proteins was showcased by applying them to GPCR mutant MD simulations computed for a selection of 28 variants from five targets in our set with varied *in vitro* effects on ligand binding (**Supplementary Figure 7.6**). The analysis of the MD trajectories showed major dynamic fluctuations compared to WT across the protein sequence, and not necessarily in the vicinity of the amino acid change, contrary to expectation (**Supplementary Figure 7.7**). Such allosteric effects on the protein dynamics dependent on the 3D organization of the protein have been previously shown to be able to explain the pathogenic mechanism of disease-driving variants^{44,45}, as well as cancer mutational drivers⁴⁶, and are therefore relevant to encode. Since 3DDPDs could not be applied to predict mutant bioactivity due to the lack of available data for our set, the power of the dynamic descriptors to discriminate between variants was investigated by clustering them based on the distance between descriptor vectors. To this end, rs3DDPDs were able to cluster all variants of the same target together, and smaller clusters were formed for mutants with similar dynamic behaviors compared to the WT (**Figure 7.7a**, **Supplementary Figure 7.7**). Nevertheless, the clusters created based on rs3DDPDs did not fully represent the clusters based on RMSF (**Supplementary Figure 7.8**), further supporting that 3DDPDs include non-dynamic information on top of dynamic information. These results make us confident to propose the use of rs3DDPDs as mutant descriptors in machine learning tasks. Other works have highlighted the use of dynamic information to predict differences between mutants, such as by extracting normal modes⁴⁷, or time series of changing geometrical features⁴⁸. However, as the changes in protein dynamics did not fully match the *in vitro* effects from the limited mutagenesis data available, the value in mutant bioactivity prediction needs to be further validated. Mutant clusters generated based on ps3DDPDs captured the most different dynamic changes between variants (**Figure 7.7b**), but this did not result in the expected clustering. The biggest differences in RMSF between mutants were observed in the N- and C-terminus, which are the most flexible regions of the GPCR together with the loops. While the termini have a function in the receptor, in the context of ps3DDPDs it seems to be blown out of proportion. An alternative would be to compute ps3DDPDs for particular regions of interest. For instance, we suggest analyzing functionally relevant residues derived from rs3DDPD feature importance, from observations in the RMSF analysis, or the literature (for example for cancer-related mutants as highlighted in **Chapter 5** for GPCRs²⁴).

One of the main limitations of our current approach is the reliability of MD simulations as input data for the computation of 3DDPDs. Firstly, the issue of MD stochastic stability is not addressed here⁴⁹, as different replicates are not used to compute our 3DDPDs. This was acceptable for the GPCR case study given the low inter-replicate variability found for MD simulations in GPCRmd. In the future, an analysis of the

impact of additional replicates in the data collection phase should be conducted. The introduction of replicas could be done twofold, either by directly using the average of the atomic coordinates as a starting point, or by using a bigger stack of individual atomic coordinates in the first PCA. Secondly, MD simulations are computationally expensive to generate, which can be a bottleneck. Similar publicly available repositories to those existing for GPCRs (i.e. GPCRmd) would help increase the applicability domain of dynamic descriptors to other protein families in the future. Finally, by extracting features from the MD trajectory, there is a constant need to make informed decisions to leave out data and reduce the amount of information available. Recently, graph neural networks (GNNs) have been used to represent MD trajectories⁵⁰. The network embeddings could be used as dynamic descriptors instead, letting the machine decide which features are more relevant, although such approaches do not necessarily produce better results⁵¹. As a last note on applicability, in our current work the description of the dynamic behavior of a protein is tackled, but the conformational changes introduced by ligand binding are not taken into account. Running MD simulations for every complex in the dataset would not be advisable, but the dynamic binding space could be represented for example by an additional term describing dynamic pharmacophores⁵² or computing cross-terms between dynamic protein and ligand descriptors¹⁴.

Conclusions

In this work, 3D dynamic protein descriptors (3DDPDs) were developed that capture the dynamic fluctuations of GPCRs as observed in MD simulations. Our one-hot encoding (rs3DDPDs) and embedding-like (ps3DDPDs) descriptors matched the performance in PCM tasks of non-dynamic state-of-the-art protein descriptors, outperforming them in regression tasks with a more challenging temporal split validation. Moreover, by mapping the most important rs3DDPD features in regression models to their GPCR generic positions it was shown that 3DDPDs represent biologically relevant information for ligand binding and activation. Finally, 3DDPDs were employed to discriminate mutant GPCRs based on their dynamic behavior with promising results that could be translated to the field of oncological drug discovery.

Materials and Methods

Wildtype GPCR MD trajectory selection and analysis

The MD simulations for the construction of 3D dynamic protein descriptors (3DDPDs) were obtained from GPCRmd²³ following the first official data deposit on November 14th, 2019. Given the positive bias towards inactive conformations, apo simulations in inactive conformation were selected for class A GPCRs with available bioactivity data (see PCM bioactivity modeling). When more than one system was available PDB codes with true apo structure with the highest resolution were selected (**Table 7.1**). Most selected MD trajectories had been simulated in triplicate for 500 ns over 2,500 frames following the GPCRmd standardized pipeline. The exceptions were GPCRmd ID 87 with 1,250

frames and ID 154 with 2,000 frames. For the generation of 3DDPDs, the first replicate was selected for each system.

Table 7.1. Wildtype GPCR MD trajectories selected from GPCRmd.

GPCR	PDB	GPCRmd ID	Resolution (Å)
5HT1B	4IAR	87	2.80
5HT2B	4IB4	92	2.70
AA1R	5UEN	165	3.20
AA2AR	5IU4	49	1.72
ACM1	5CXV	154	2.70
ACM2	3UON	111	3.00
ACM4	5DSG	157	2.60
ADRB2	2RH1	11	2.40
AGTR1	4ZUD	189	2.80
CCR5	4MBS	118	2.71
CNR1	5U09	163	2.60
CXCR4	3ODU	101	2.50
DRD3	3PBL	105	2.89
EDNRB	5GLH	158	2.80
FFAR1	4PHU	75	2.33
HRH1	3RZE	108	3.10
LPAR1	4Z35	184	2.90
OPRD	4N6H	73	1.80
OPRK	4DJH	59	2.90
OPRX	5DHH	155	3.00
OX1R	4ZJ8	186	2.75
OX2R	4S0V	91	2.50
P2RY1	4XNV	179	2.20
P2Y12	4PXZ	77	2.50
PAR1	3VW7	128	2.20

Python library MDtraj⁵³ was used to compute the Root Mean Square Deviation (RMSD) and RMSF of MD trajectories to assess the stability of the simulations and account for differences in the dynamic behavior of the selected GPCRs in different protein segments. RMSD was calculated for the protein atoms in reference to the first frame in the production run. RMSF was calculated for the protein C α backbone atoms over the total length of the simulation. To allow direct comparison between receptors, RMSF values were aligned based on their corresponding residue number to the class A GPCR MSA obtained from GPCRdb³¹. The location of TM domains in the RMSF plots was mapped based on the generic BW³⁰ residue numbers obtained from GPCRdb. BW numbers were also used throughout the manuscript to refer to equivalent locations in the GPCR structure.

3DDPD generation and optimization

Atomic coordinates were extracted from GPCRmd trajectories with MDtraj. Each trajectory was divided into sub-trajectories of a defined number of frames, f , and the mean, median, and SD of the x, y, and z coordinates were calculated for each sub-trajectory. Additionally, atomic partial charges were generated for each atom in the system with RDkit Gasteiger charges calculator⁵⁴. The next steps are tailored for the two flavors of 3DDPDs generated: one-hot encoding residue-specific (rs) 3DDPDs, and whole sequence embedding-like protein-specific (ps) 3DDPDs (**Figure 7.1**).

For rs3DDPDs, coordinate statistics and partial charges per atom were collected for each target and standardized between 0 and 1. Subsequently, dimensionality reduction was applied in the form of PCA. A number of PCs for each atom were selected and grouped per residue as average and SD. A second dimensionality reduction step was applied to residue data and the selected PCs were placed in their matching sections corresponding to an MSA of the targets of interest.

Protein-specific ps3DDPDs were generated similarly to rs3DDPDs with some differences. Firstly, coordinate statistics and partial charges per atom were collected for all targets together and standardized between 0 and 1. Secondly, atom PCA was not grouped per residue and no second PCA was applied. Instead, the PCs selected per atom were grouped per target as average, median, and SD, constituting the final descriptor.

The generation parameters for the descriptors were randomly initialized and sequentially optimized. The parameters optimized included (in the following order):

- i) Trajectory data: the use of all statistical values derived from the x, y, and z coordinates was compared to just the SD, representing the “rigidity” of each atomic coordinate.
- ii) Frame split: number of frames included in each trajectory split, for which 10, 50, 100, and 500 frames were tested. This parameter was optimized on rs3DDPDs and the results were applied to ps3DDPDs.
- iii) Residue PCA (only for rs3DDPDs): number of PCs selected after residue data PCA, either 3, 5, 7, or 10.
- iv) Atom PCA coverage: variance explained by the selected number of PCs on atom data, either 95% or 99%.
- v) Atom selection: inclusion of atomic data from all heavy atoms or just non-carbon atoms.
- vi) Residue selection: strategies to focus the 3DDPDs on the protein binding site. These selections were based on structural-driven MSAs at different protein family levels, starting from using the full sequence, then the binding pocket of class A GPCRs, then of specific GPCR families, then GPCR subfamilies,

and finally, target-specific binding pocket. To ensure a consistent number of features per descriptor, in rs3DDPDs only the first two options were tested.

- vii) Combination with classical protein descriptors: tested sequentially and, for the case of rs3DDPDs also embedded on the descriptor via the residue PCA.

The optimization of 3DDPDs was done by comparing their performance with different parameters on PCM Bioactivity regression modeling on a temporal split.

3DDPD and MD hierarchical clustering

Hierarchical clustering dendrograms were computed to visualize similarities and differences between 3DDPD descriptors and dynamic behavior (represented by MD's RMSF) across targets. Python package Scipy⁵⁵ was used to compute hierarchical clusters based on the Euclidean distance between non-null bits of 3DDPD or RMSF vectors. The accompanying representation of the descriptor or RMSF includes null bits that are derived from their mapping to the GPCR class A MSA. Plotting was done in Python using the package Matplotlib⁵⁶.

PCM Bioactivity modeling

The bioactivity dataset for PCM modeling was constructed starting from the highly curated Papyrus dataset version 5.50²⁹. For the regression task, high-quality datapoints with continuous data (pchembl values) were extracted for all available GPCRs. Receptors with MD inactive/intermediate apo trajectories available on GPCRmd and over 100 bioactivity datapoints were selected for the PCM set, resulting in 26 GPCRs and a total number of 38,701 bioactivity datapoints (**Supplementary Table 7.1**).

PCM modeling was implemented in Python 3.8⁵⁷ using the modeling capabilities of the Papyrus scripts Python package²⁹. Random Forest models from Scikit-learn⁵⁸ were used in regression and classification tasks as the state-of-the-art in bioactivity prediction. A pchembl value of 6.5 was considered as a cutoff between active and inactive compounds for classification tasks. Hyperparameters were set as default and not optimized during the training of the different models to reduce degrees of freedom in the comparison of the effect of different protein descriptors.

The compound descriptors used were Morgan fingerprints of radius 3 (ECFP6) and length 1024⁵⁴, pre-calculated in the Papyrus dataset. The protein descriptors used to benchmark the performance of 3DDPDs were one-hot encodings and protein embeddings. The former included MS-WHIM, STscale, PhysChem, and two flavors of Zscale (Hellberg and van Westen, with 5 and 3 PCs per residue each)^{10,34}. One-hot encodings were calculated using the Python package ProDEC⁵⁹ based on the class A GPCR MSA obtained from GPCRdb for our protein set. As protein embeddings UniRep⁶⁰ were used, pre-calculated in the Papyrus dataset. 3DDPDs were benchmarked as protein descriptors on their own and in combination with non-dynamic protein descriptors. The

best-performing rs3DDPDs and ps3DDPDs in the optimization phase were used for combination. Additionally, QSAR models were trained on each of the targets in the set with the same options and analysis as the PCM models to benchmark the use of protein descriptors.

Two methods were used to split the PCM dataset into training and test sets. Firstly a random split was used, where 80% of the data was allocated to the training set and 20% of the data to the test set. Data for all targets was present in both the training and the test set. Secondly, a temporal split was used to provide the model with a more challenging validation task than the random split, where compound-target pairs first recorded before 2013 were allocated to the training set, and newer datapoints to the test set. The cutoff year was selected to make sure that all targets were represented in the test set. This resulted in a test set with 39% of the data, which was not equally distributed per target but showed considerably reduced chemical bias between training and test set compared to the random split. Chemical bias was computed as the asymmetric validation embedding (AVE) bias defined by Wallach & Heifets⁶¹ using as active-inactive cutoff a pchembl value of 6.5.

All RF models were trained using 5-fold cross-validation, and the performance of the models was evaluated on the test set. The evaluation metrics reported were MCC for classification and Pearson r and RMSE for regression tasks. Other metrics are available in the Supplementary Information. For comparison purposes, a single average performance metric was calculated for QSAR RF models trained and tested on each target of the set independently.

Ten model replicates were trained for each protein descriptor benchmarked with random seeds 1234, 2345, 3456, 4567, 5678, 6879, 7890, 8901, 9012, and 9999. The seed was used for resampling, both in the form of K-Fold shuffling in cross-validation and train/test splitting, the latter only in the case of a random split. Moreover, each model was initialized with a random seed as per default in Scikit-learn RF. The statistical significance of the differences in performance when using different protein descriptors was calculated by performing an independent T-test of the average performance metrics in the pool of model replicates. Differences were considered significant when p-value < 0.05. Performance comparison plots were generated in Python using the packages Matplotlib⁵⁶ and Seaborn⁶².

Selection of GPCR (cancer-related) somatic mutants

In order to test the usage of 3DDPDs in mutants, several mutations for the GPCRs in the 3DDPD set were selected. To simulate a real application scenario, a mutant PCM dataset was created, gathering available mutagenesis data from GPCRdb for the GPCR 3DDPD set. Mutations with datapoints available for more than ten different ligands were selected.

To extend the applicability domain, somatic mutations in cancer patients were extracted from the GDC database v22.0⁶³ for the five GPCRs with selected mutagenesis data.

Cancer-related mutations with mutagenesis data available on GPCRdb, regardless of the magnitude, were added to the mutation selection list in order to include a subsample of mutations present in cancer patients (Table 7.2).

Table 7.2. GPCR mutations selected.

GPCR	PDB	GPCRmd ID	Mutation	GPCRdb (ligands / datapoints)	GDC patients	Motif
AA1R	5UEN	165	T277A ^{7.41}	13 / 36	0	-
			R291C ^{7.56}	4 / 4	1	NpxxY (ext)
			R296C ^{8.51}	4 / 4	1	-
AA2AR	5IU4	49	I66A ^{2.64}	20 / 22	0	-
			L85A ^{3.33}	21 / 21	0	-
			T88D ^{3.36}	14 / 16	0	-
			S91A ^{3.39}	12 / 16	0	-
			L167A ^{45.51}	20 / 20	0	-
			M177A ^{5.40}	22 / 24	0	-
			N181A ^{5.43}	20 / 20	0	-
			W246A ^{6.48}	37 / 52	0	CWxP
			N253A ^{6.55}	22 / 22	0	-
			Y271A ^{7.35}	20 / 22	0	-
			S277A ^{7.41}	29 / 33	0	-
			H278N ^{7.42}	3 / 3	1	-
ACM2	3UON	111	D103E ^{3.32}	32 / 42	0	-
			D103N ^{3.32}	12 / 15	0	-
			V421L ^{7.33}	1 / 1	1	-
ADRB2	2RH1	11	D79N ^{2.50}	12 / 12	0	-
			D130N ^{3.49}	11 / 11	0	DRY
			S203A ^{5.43}	12 / 12	0	-
			S204A ^{5.44}	13 / 13	0	-
			N293L ^{6.55}	12 / 12	0	-
			V317A ^{7.43}	5 / 5	1	-
CCR5	4MBS	118	Y108A ^{3.32}	12 / 20	0	-

Mutant MD simulations and 3DDPDs

Mutant MD simulations were performed according to the GPCRmd pipeline²³. Equilibrated GPCRmd WT systems were obtained from the first frame of the first simulation replicate available online for the GPCRmd IDs defined in Table 7.1. Using the HTMD package⁶⁴, the mutations of interest were introduced and the systems were re-equilibrated using AceMD MD engine⁶⁵ and default GPCRmd parameters.

Consecutively, the re-equilibrated trajectories were wrapped and 500ns production runs were simulated in triplicate with different random initialization seeds following the GPCRmd framework. Finally, the production trajectories were wrapped and rs3DDPDs and ps3DDPDs were generated from the first replicate.

3D visualization

Representations of proteins in 3D were generated using PyMOL 2.5.2⁶⁶.

References

- Burley, S. K. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. *J Biol Chem* **296**, 100559 (2021).
- Carracedo-Reboredo, P. *et al.* A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J*. **19**, 4538–4558 (2021).
- You, Y. *et al.* Artificial intelligence in cancer target identification and drug discovery. *Signal Transduct Target Ther* **7**, 156 (2022).
- Sankar, K. *et al.* A Descriptor Set for Quantitative Structure-property Relationship Prediction in Biologics. *Mol Inform* **41**, 2100240 (2022).
- Torkamannia, A., Omid, Y. & Ferdousi, R. A review of machine learning approaches for drug synergy prediction in cancer. *Brief Bioinform* **23**, 1–19 (2022).
- Satake, H., Osugi, T. & Shiraishi, A. Impact of Machine Learning-Associated Research Strategies on the Identification of Peptide-Receptor Interactions in the Post-Omics Era. *Neuroendocrinology* **113**, 251–261 (2021).
- Bongers, B. J., IJzerman, A. P. & Van Westen, G. J. P. Proteochemometrics – recent developments in bioactivity and selectivity modeling. *Drug Discov Today Technol* **32**, 89–98 (2019).
- Du, B. X. *et al.* Compound–protein interaction prediction by deep learning: Databases, descriptors and models. *Drug Discov Today* **27**, 1350–1366 (2022).
- Fernández-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M. & Aloy, P. Connecting chemistry and biology through molecular descriptors. *Curr Opin Chem Biol* **66**, 102090 (2022).
- Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets. *J Cheminform* **5**, 41 (2013).
- Ismail, H., White, C., AL-Barakati, H., Newman, R. H. & KC, D. B. FEPS: A tool for feature extraction from protein sequence. *Methods mol. biol.* **2499**, 65–104 (2022).
- Ibtehaz, N. & Kihara, D. Application of Sequence Embedding in Protein Sequence-Based Predictions. Preprint at *ArXiv* doi:10.48550/arXiv.2110.07609 (2021).
- Wang, D. D. *et al.* Structure-based protein-ligand interaction fingerprints for binding affinity prediction. *Comput Struct Biotechnol J* **19**, 6291–6300 (2021).
- Subramanian, V., Prusis, P., Pietilä, L. O., Xhaard, H. & Wohlfahrt, G. Visually interpretable models of kinase selectivity related features derived from field-based proteochemometrics. *J Chem Inf Model* **53**, 3021–3030 (2013).
- Miller, M. D. & Phillips, G. N. Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. *J Biol Chem* **296**, 100749 (2021).
- Abriata, L. A., Spiga, E. & Peraro, M. D. Molecular Effects of Concentrated Solutes on Protein Hydration, Dynamics, and Electrostatics. *Biophys J* **111**, 743–755 (2016).
- Stank, A., Kokh, D. B., Fuller, J. C. & Wade, R. C. Protein Binding Pocket Dynamics. *Acc. Chem. Res.* **49**, 809–815 (2016).
- Zhu, F. *et al.* Leveraging Protein Dynamics to Identify Functional Phosphorylation Sites using Deep Learning Models. *J. Chem. Inf. Model.* **62**, 3331–3345 (2022).
- Gao, J. *et al.* Study on human GPCR-inhibitor interactions by proteochemometric modeling. *Gene* **518**, 124–131 (2013).
- Odoemelam, C. S. *et al.* G-Protein coupled receptors: structure and function in drug discovery. *RSC Adv.* **10**, 36337 (2020).
- Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **117**, 139–155 (2017).
- Lee, Y., Lazim, R., Macalino, S. J. Y. & Choi, S. Importance of protein dynamics in the structure-based drug discovery of class A G protein-coupled receptors (GPCRs). *Curr Opin Struct Biol* **55**, 147–153 (2019).
- Rodríguez-Espigares, I. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat Methods* **17**, 777–787 (2020).
- Bongers, B. J. *et al.* Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors. *Sci Rep* **12**, 21534 (2022).
- Wang, X. *et al.* Cancer-related somatic mutations alter adenosine A1 receptor pharmacology—A focus on mutations in the loops and C-terminus. *FASEB J* **36**, 1–16 (2022).
- den Hollander, L. S. *et al.* Impact of cancer-associated mutations in CC chemokine receptor 2 on receptor function and antagonism. *Biochem Pharmacol* **208**, 115399 (2023).
- Feng, C. *et al.* Cancer-Associated Mutations of the Adenosine A2A Receptor Have Diverse Influences on Ligand Binding and Receptor Functions. *Molecules* **27**, 4676 (2022).
- Jespers, W. *et al.* Structural Mapping of Adenosine Receptor Mutations: Ligand Binding and Signaling Mechanisms. *Trends Pharmacol Sci* **39**, 75–89 (2018).
- Béguignon, O. J. M. *et al.* Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J Cheminform* **15**, 3 (2023).
- Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional

- models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in Neurosciences* **25**, 366–428 (1995).
31. Isberg, V. *et al.* GPCRdb: An information system for G protein-coupled receptors. *Nucleic Acids Res* **44**, D356–D364 (2016).
 32. Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data to Predict Free-Energy Differences. *J Chem Inf Model* **57**, 726–741 (2017).
 33. Bolcato, G., Heid, E. & Boström, J. On the Value of Using 3D Shape and Electrostatic Similarities in Deep Generative Methods. *J Chem Inf Model* **62**, 1388–1398 (2022).
 34. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets. *J Cheminform* **5**, 42 (2013).
 35. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34**, 2642–2648 (2018).
 36. Lim, H. *et al.* Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in SARS-CoV-2 spike glycoprotein. *Comput. Struct. Biotechnol. J.* **20**, 788–798 (2022).
 37. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform* **9**, 45 (2017).
 38. Rackovsky, S. & Scheraga, H. A. The structure of protein dynamic space. *Proc Natl Acad Sci U S A* **117**, 19938–19942 (2020).
 39. Draper-Joyce, C. J. *et al.* Positive allosteric mechanisms of adenosine A1 receptor-mediated analgesia. *Nature* **597**, 571–576 (2021).
 40. Lee, S. M., Booe, J. M. & Pioszak, A. A. Structural insights into ligand recognition and selectivity for classes A, B, and C GPCRs. *Eur J Pharmacol* **763**, 196–205 (2015).
 41. Hauser, A. S. & Kooistra, A. J. GPCR activation mechanisms across classes and macro/microscales. *Nat Struct Mol Biol* **28**, 879–888 (2021).
 42. Glukhova, A. *et al.* Structure of the Adenosine A1 Receptor Reveals the Basis for Subtype Selectivity. *Cell* **168**, 867–877 (2017).
 43. Bondar, A.-N. Graphs of Hydrogen-Bond Networks to Dissect Protein Conformational Dynamics. *J. Phys. Chem. B* **126**, 3973–3984 (2022).
 44. Ose, N. J. *et al.* Dynamic coupling of residues within proteins as a mechanistic foundation of many enigmatic pathogenic missense variants. *PLoS Comput Biol* **18**, e1010006 (2022).
 45. Li, B., Roden, D. M. & Capra, J. A. The 3D mutational constraint on amino acid sites in the human proteome. *Nat. Commun* **13**, 3273 (2022).
 46. Kumar, S., Clarke, D. & Gerstein, M. B. Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. *Proc Natl Acad Sci U S A* **116**, 18962–18970 (2019).
 47. Rodrigues, C. H. *et al.* DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* **46**, W350–W355 (2018).
 48. Wang, D. D., Ou-Yang, L., Xie, H., Zhu, M. & Yan, H. Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods. *Comput Struct Biotechnol J* **18**, 439–454 (2020).
 49. Knapp, B., Ospina, L. & Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J. Chem. Theory Comput.* **14**, 6127–6138 (2018).
 50. Li, Z., Meidani, K., Yadav, P. & Farimani, A. B. Graph Neural Networks Accelerated Molecular Dynamics. *J. Chem. Phys.* **156**, 144103 (2022).
 51. Volkov, M. *et al.* On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *J. Med. Chem* **2022**, 7946–7958 (2022).
 52. Janežič, M. *et al.* Dynophore-Based Approach in Virtual Screening: A Case of Human DNA Topoisomerase II α . *Int. J. Mol. Sci.* **22**, 13474 (2021).
 53. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**, 1528–1532 (2015).
 54. RDKit: Open-source cheminformatics; <http://www.rdkit.org>.
 55. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).
 56. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95 (2007).
 57. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, Scotts Valley, CA, 2009).
 58. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).
 59. Béguignon, O. J. M. ProDEC v1.0.2. Available at <https://doi.org/10.5281/zenodo.7007058>. Date accessed: 20/08/2022.
 60. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* **16**, 1315–1322 (2019).
 61. Wallach, I. & Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J Chem Inf Model* **58**, 916–932 (2018).
 62. Waskom, M. Seaborn: Statistical Data Visualization. *J Open Source Softw* **6**, 3021 (2021).

63. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
64. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J Chem Theory Comput* **12**, 1845–1852 (2016).
65. Harvey, M. J., Giupponi, G. & De Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* **5**, 1632–1639 (2009).
66. The PyMOL Molecular Graphics System, Version 1.4 Schrödinger, LLC.

Supplementary Information

Supplementary Table 7.1. Papyrus bioactivity data distribution across the set of 26 WT GPCRs.

Target ID	Activity datapoints	pchembl value (Mean)				
		Min	Max	Median	Mean	SD
P29274_WT	3991	4.00	11.00	6.82	6.88	1.17
P21554_WT	3741	4.00	10.52	6.82	6.91	1.20
P30542_WT	3519	4.00	12.20	6.48	6.58	1.01
P35462_WT	3152	3.10	10.54	7.62	7.49	1.17
P41145_WT	2910	4.09	11.52	6.85	7.02	1.41
P41143_WT	2219	4.00	10.74	7.00	6.89	1.37
P21453_WT	2038	4.03	10.80	7.82	7.66	1.46
O43614_WT	1901	4.30	10.05	6.91	6.85	1.17
O43613_WT	1820	4.19	9.80	6.09	6.34	1.11
O14842_WT	1304	4.16	9.52	6.60	6.53	0.92
P11229_WT	1273	4.03	10.85	6.50	6.67	1.20
P51681_WT	1252	4.04	11.52	7.28	7.13	1.41
P41146_WT	1155	4.32	10.43	7.54	7.51	1.08
P41595_WT	1125	4.19	9.96	6.69	6.75	0.87
P07550_WT	1002	3.85	10.92	7.68	7.53	1.54
Q9H244_WT	988	4.24	9.60	7.17	7.13	1.04
P30556_WT	876	4.01	10.00	5.23	5.90	1.72
P35367_WT	817	4.01	10.13	7.00	7.02	1.17
P08172_WT	791	4.02	10.36	6.92	6.98	1.30
P25116_WT	665	4.02	9.00	7.16	6.95	0.97
P08173_WT	584	4.00	10.75	6.41	6.49	1.03
P28222_WT	524	4.99	10.05	7.80	7.65	1.21
P61073_WT	402	4.15	9.21	7.04	6.91	0.94
P47900_WT	370	4.35	10.52	6.90	6.95	1.17
Q92633_WT	156	4.75	8.96	6.76	6.70	0.82
P24530_WT	126	4.00	9.39	6.01	6.08	1.00
Total	38,701					

Supplementary Table 7.2. Performance metrics of QSAR and PCM models with random validation split trained with different protein descriptors. QSAR model performance represents the average over the individual target models trained and validated without protein descriptors (NA: non-applicable).

Model	Split	Protein descriptor	Metric	mean	std
QSAR	random	NA	MCC	0.577714	0.007181
QSAR	random	NA	RMSE	0.705380	0.005661
QSAR	random	NA	r	0.774895	0.004677
QSAR	random	NA	R2	0.601279	0.007420
QSAR	random	NA	MAE	0.523834	0.004150
PCM	random	3DDPD_PS_all_f100_pc95_fs_aa	MCC	0.644716	0.00675
PCM	random	3DDPD_PS_all_f100_pc95_fs_aa	RMSE	0.704416	0.006187
PCM	random	3DDPD_PS_all_f100_pc95_fs_aa	r	0.835304	0.002997
PCM	random	3DDPD_PS_all_f100_pc95_fs_aa	R2	0.693622	0.004761
PCM	random	3DDPD_PS_all_f100_pc95_fs_aa	MAE	0.527470	0.004374
PCM	random	3DDPD_RS_std_f100_pc10_fs_aa	MCC	0.642643	0.008263
PCM	random	3DDPD_RS_std_f100_pc10_fs_aa	RMSE	0.710025	0.006935
PCM	random	3DDPD_RS_std_f100_pc10_fs_aa	r	0.832272	0.003893
PCM	random	3DDPD_RS_std_f100_pc10_fs_aa	R2	0.688707	0.006327
PCM	random	3DDPD_RS_std_f100_pc10_fs_aa	MAE	0.531428	0.005215
PCM	random	MS-WHIM	MCC	0.645082	0.007837
PCM	random	MS-WHIM	RMSE	0.706699	0.00666
PCM	random	MS-WHIM	r	0.834099	0.003879
PCM	random	MS-WHIM	R2	0.691614	0.006109
PCM	random	MS-WHIM	MAE	0.529213	0.004668
PCM	random	PhysChem	MCC	0.643758	0.00737
PCM	random	PhysChem	RMSE	0.707704	0.005797
PCM	random	PhysChem	r	0.833691	0.003228
PCM	random	PhysChem	R2	0.690748	0.005079
PCM	random	PhysChem	MAE	0.530157	0.004242
PCM	random	STscale	MCC	0.642903	0.007576
PCM	random	STscale	RMSE	0.706985	0.006782
PCM	random	STscale	r	0.834184	0.002843
PCM	random	STscale	R2	0.691390	0.004603
PCM	random	STscale	MAE	0.529546	0.005007
PCM	random	Zscale_Hellberg	MCC	0.644947	0.006025
PCM	random	Zscale_Hellberg	RMSE	0.706032	0.007275
PCM	random	Zscale_Hellberg	r	0.834601	0.004027
PCM	random	Zscale_Hellberg	R2	0.692197	0.006316
PCM	random	Zscale_Hellberg	MAE	0.529160	0.004882

Supplementary Table 7.2 (continues)

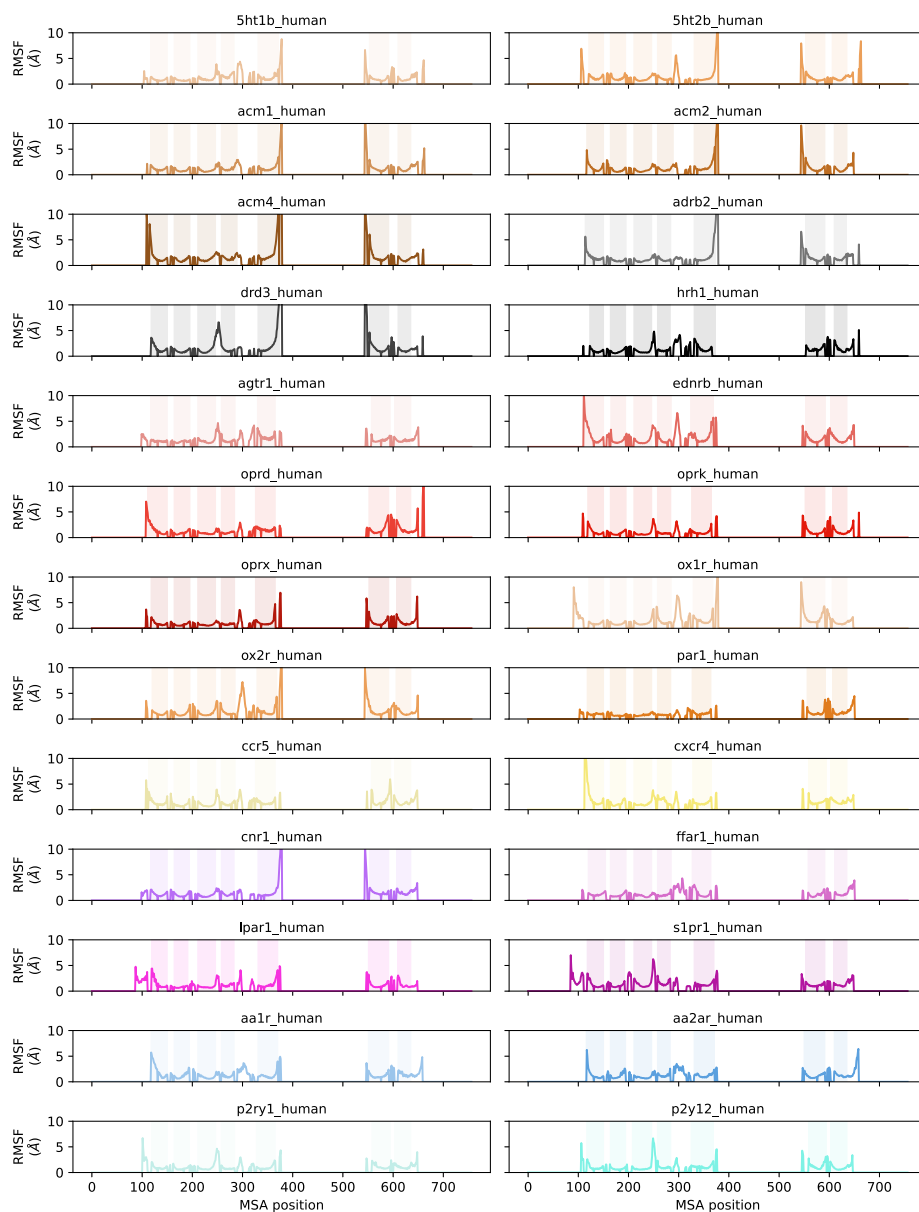
PCM	random	Zscale_van_Westen	MCC	0.645763	0.008597
PCM	random	Zscale_van_Westen	RMSE	0.703212	0.006583
PCM	random	Zscale_van_Westen	r	0.836407	0.003532
PCM	random	Zscale_van_Westen	R2	0.694661	0.005435
PCM	random	Zscale_van_Westen	MAE	0.527693	0.004787
PCM	random	unirep	MCC	0.642587	0.004844
PCM	random	unirep	RMSE	0.710875	0.006126
PCM	random	unirep	r	0.831989	0.003264
PCM	random	unirep	R2	0.687973	0.005269
PCM	random	unirep	MAE	0.532397	0.004853

Supplementary Table 7.3. Performance metrics of QSAR and PCM models with temporal validation split trained with different protein descriptors. QSAR model performance represents the average over the individual target models trained and validated without protein descriptors (NA: non-applicable).

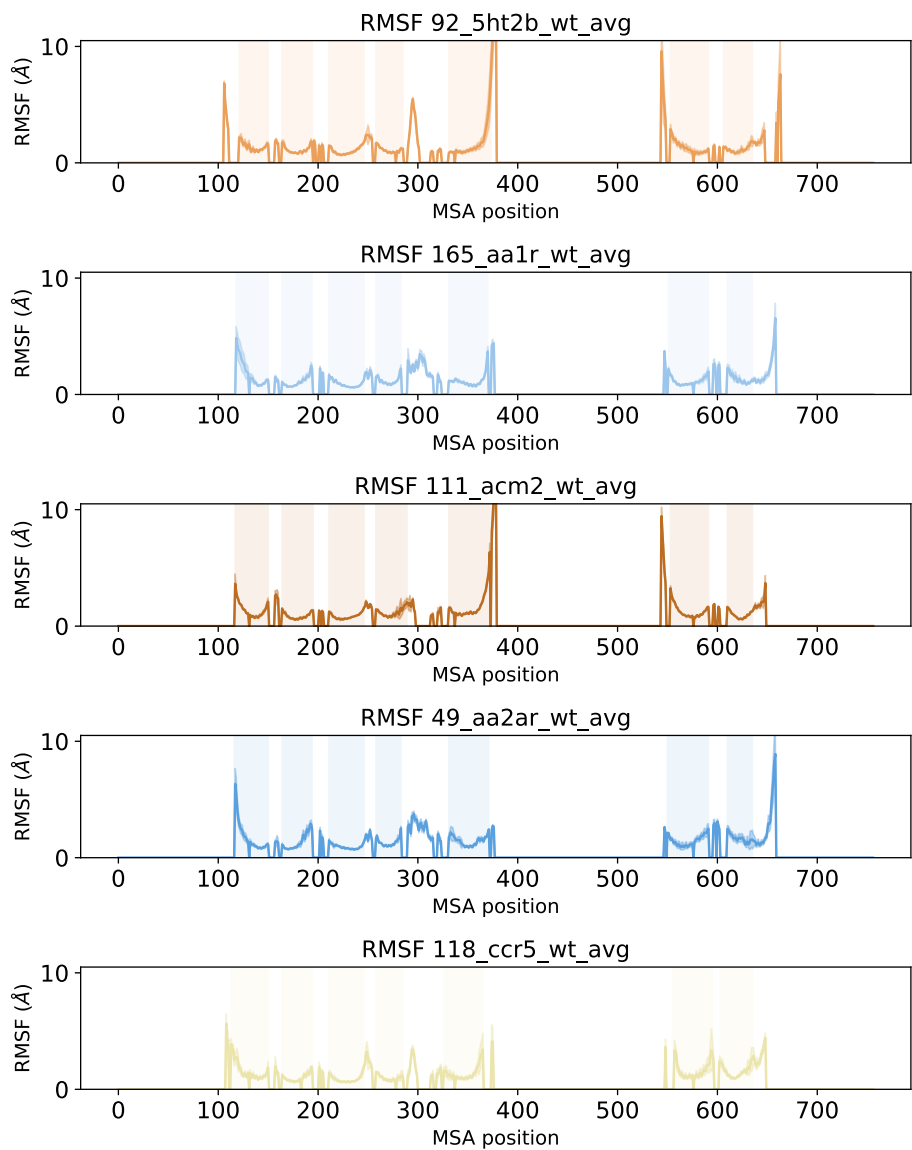
Model	Split	Protein descriptor	Metric	mean	std
QSAR	temporal	NA	MCC	0.191889	0.009093
QSAR	temporal	NA	RMSE	1.168438	0.003839
QSAR	temporal	NA	r	0.343006	0.004332
QSAR	temporal	NA	R2	-0.171235	0.008062
QSAR	temporal	NA	MAE	0.931834	0.002768
PCM	temporal	3DDPD_PS_all_f100_pc95_fs_aa	MCC	0.277186	0.00761
PCM	temporal	3DDPD_PS_all_f100_pc95_fs_aa	RMSE	1.153923	0.002905
PCM	temporal	3DDPD_PS_all_f100_pc95_fs_aa	r	0.451019	0.003336
PCM	temporal	3DDPD_PS_all_f100_pc95_fs_aa	R2	0.154453	0.004147
PCM	temporal	3DDPD_PS_all_f100_pc95_fs_aa	MAE	0.919372	0.002084
PCM	temporal	3DDPD_RS_std_f100_pc10_fs_aa	MCC	0.273142	0.003223
PCM	temporal	3DDPD_RS_std_f100_pc10_fs_aa	RMSE	1.213864	0.005166
PCM	temporal	3DDPD_RS_std_f100_pc10_fs_aa	r	0.41746	0.003671
PCM	temporal	3DDPD_RS_std_f100_pc10_fs_aa	R2	0.064317	0.007955
PCM	temporal	3DDPD_RS_std_f100_pc10_fs_aa	MAE	0.954875	0.003743
PCM	temporal	MS-WHIM	MCC	0.276817	0.003861
PCM	temporal	MS-WHIM	RMSE	1.218501	0.004445
PCM	temporal	MS-WHIM	r	0.410101	0.004479
PCM	temporal	MS-WHIM	R2	0.057159	0.006687
PCM	temporal	MS-WHIM	MAE	0.959843	0.003392
PCM	temporal	PhysChem	MCC	0.27533	0.004877
PCM	temporal	PhysChem	RMSE	1.21395	0.005249
PCM	temporal	PhysChem	r	0.414679	0.003797
PCM	temporal	PhysChem	R2	0.064184	0.007884
PCM	temporal	PhysChem	MAE	0.958551	0.003209
PCM	temporal	STscale	MCC	0.277505	0.004956
PCM	temporal	STscale	RMSE	1.217626	0.007125
PCM	temporal	STscale	r	0.413211	0.004509
PCM	temporal	STscale	R2	0.058495	0.010720
PCM	temporal	STscale	MAE	0.960353	0.004845
PCM	temporal	Zscale_Hellberg	MCC	0.278328	0.005163
PCM	temporal	Zscale_Hellberg	RMSE	1.22066	0.003245
PCM	temporal	Zscale_Hellberg	r	0.409729	0.003008
PCM	temporal	Zscale_Hellberg	R2	0.053820	0.004890
PCM	temporal	Zscale_Hellberg	MAE	0.962210	0.001808

Supplementary Table 7.3 (continues)

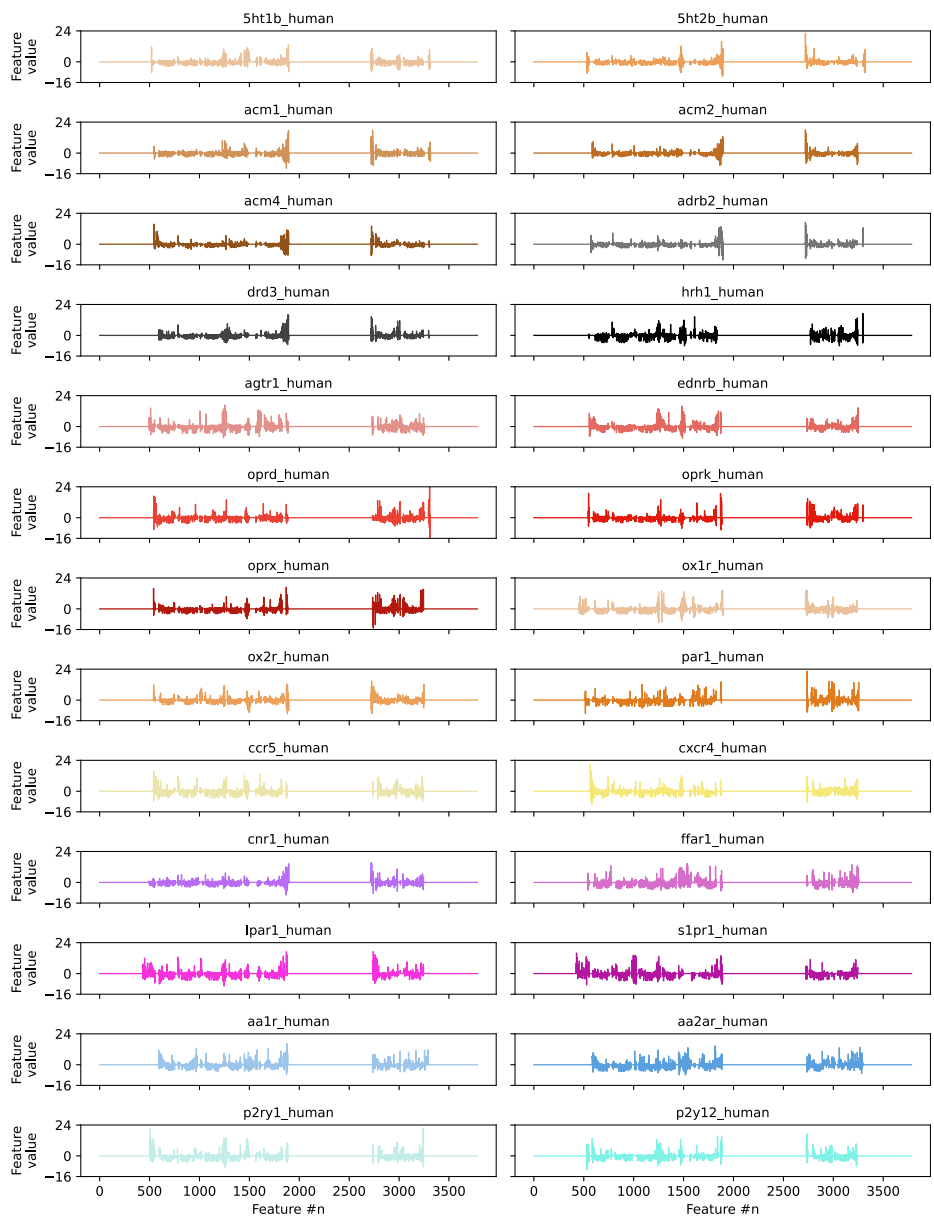
PCM	temporal	Zscale_van_Westen	MCC	0.274272	0.008416
PCM	temporal	Zscale_van_Westen	RMSE	1.221101	0.006088
PCM	temporal	Zscale_van_Westen	r	0.409944	0.005973
PCM	temporal	Zscale_van_Westen	R2	0.053121	0.009200
PCM	temporal	Zscale_van_Westen	MAE	0.962816	0.004110
PCM	temporal	unirep	MCC	0.273962	0.00843
PCM	temporal	unirep	RMSE	1.219178	0.004602
PCM	temporal	unirep	r	0.411132	0.003577
PCM	temporal	unirep	R2	0.056110	0.007120
PCM	temporal	unirep	MAE	0.959212	0.003379



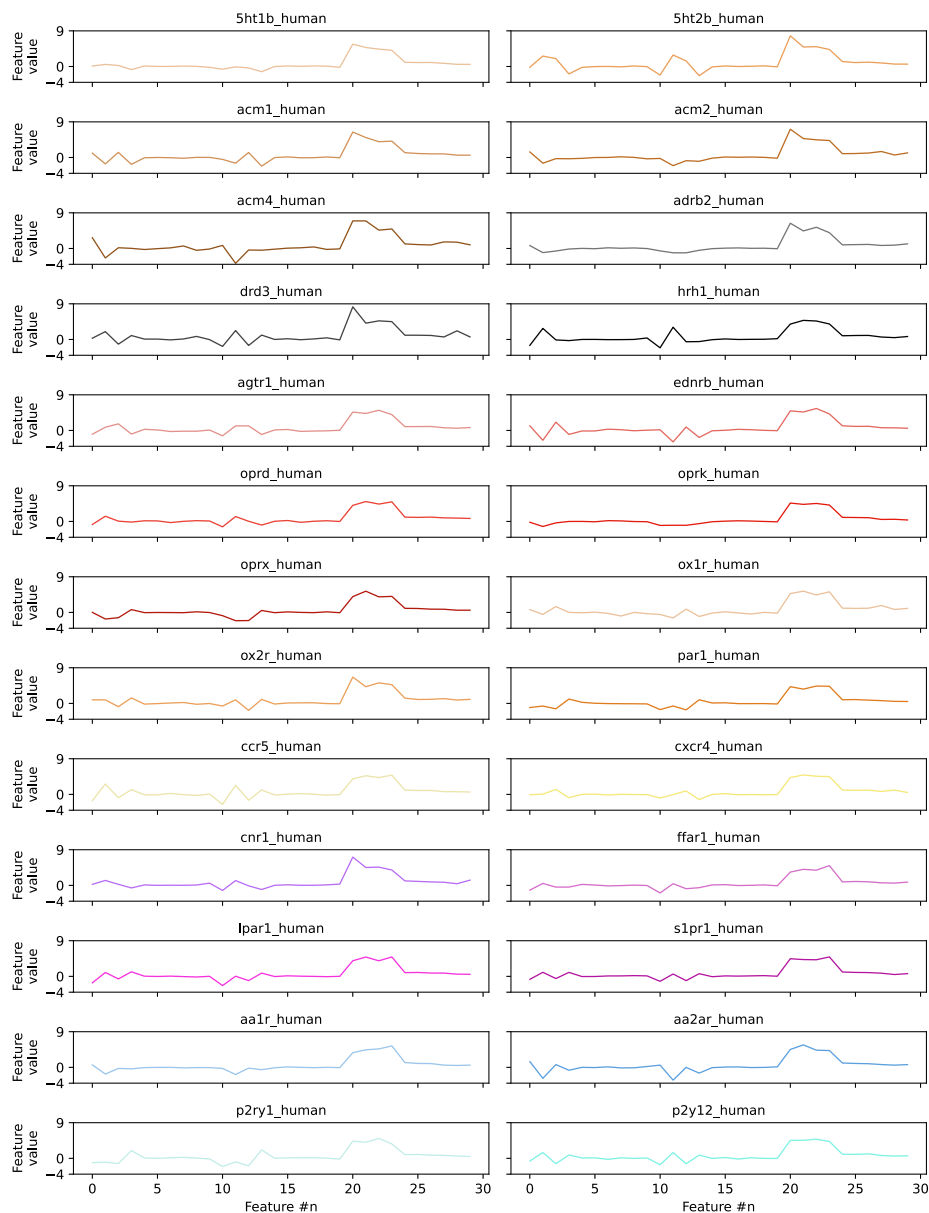
Supplementary Figure 7.1. RMSF values for the MD GPCRmd trajectories of the 26 GPCRs in the WT set. The RMSF values are mapped to their corresponding positions in the MSA later used for rs3DDPD and non-dynamic descriptor calculation, for easier visualization. The regions in the MSA corresponding to domains TM 1-7 are shadowed for reference. Each receptor is represented in a different color and receptors from the same subfamily/family are represented in the same color palette.



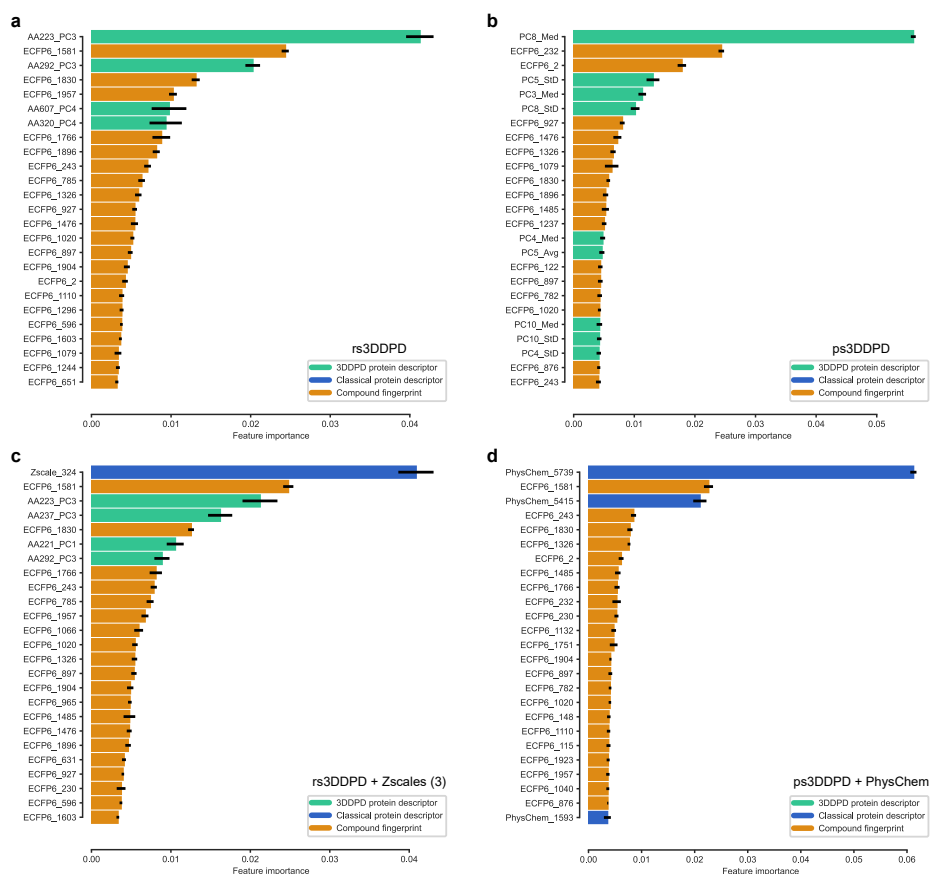
Supplementary Figure 7.2. RMSF average and variability over the three GPCRmd trajectory replicates for GPCRs 5HT2B, AA1R, ACM2, AA2AR, and CCR5. The average RMSF is represented as a line and the standard deviation of the mean is represented as a shade around the average. For easier comparison between targets, RMSF is aligned to the reference MSA, and the transmembrane domains TM1-7 are shaded.



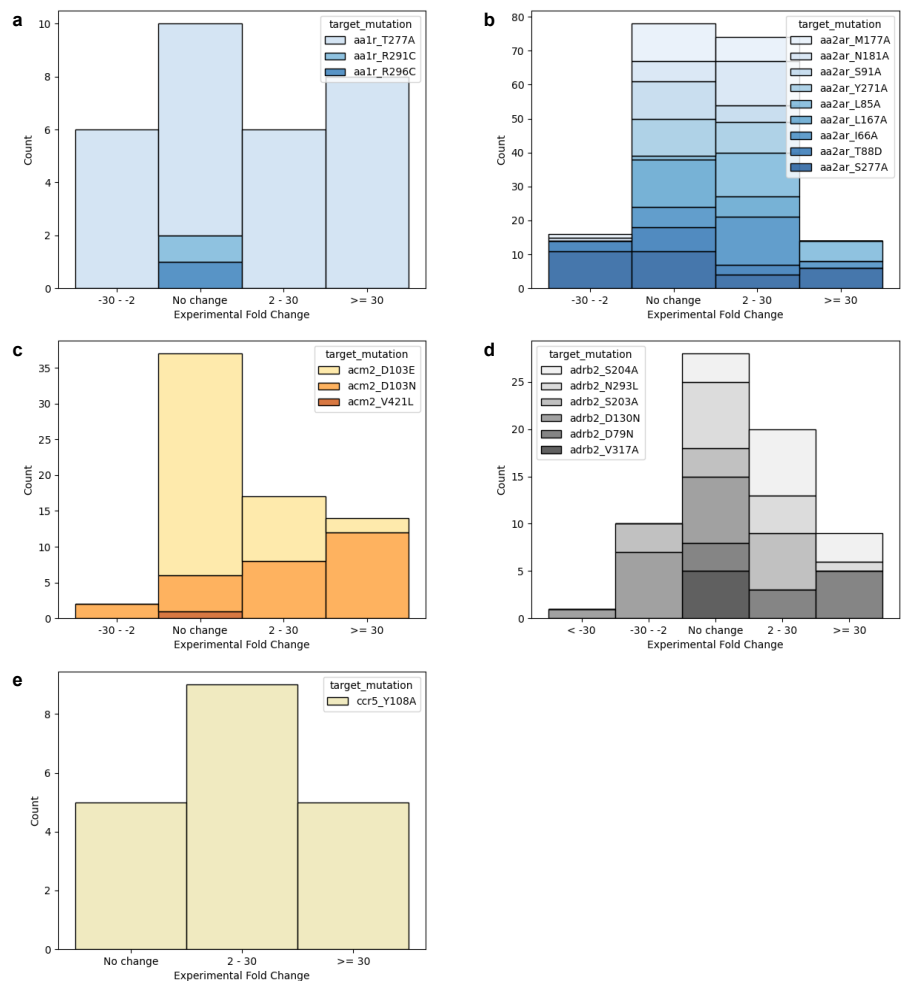
Supplementary Figure 7.3. Representation of rs3DDPD feature values for the 26 GPCRs in the WT set. Each receptor is represented in a different color and receptors from the same subfamily/family are represented in the same color palette.



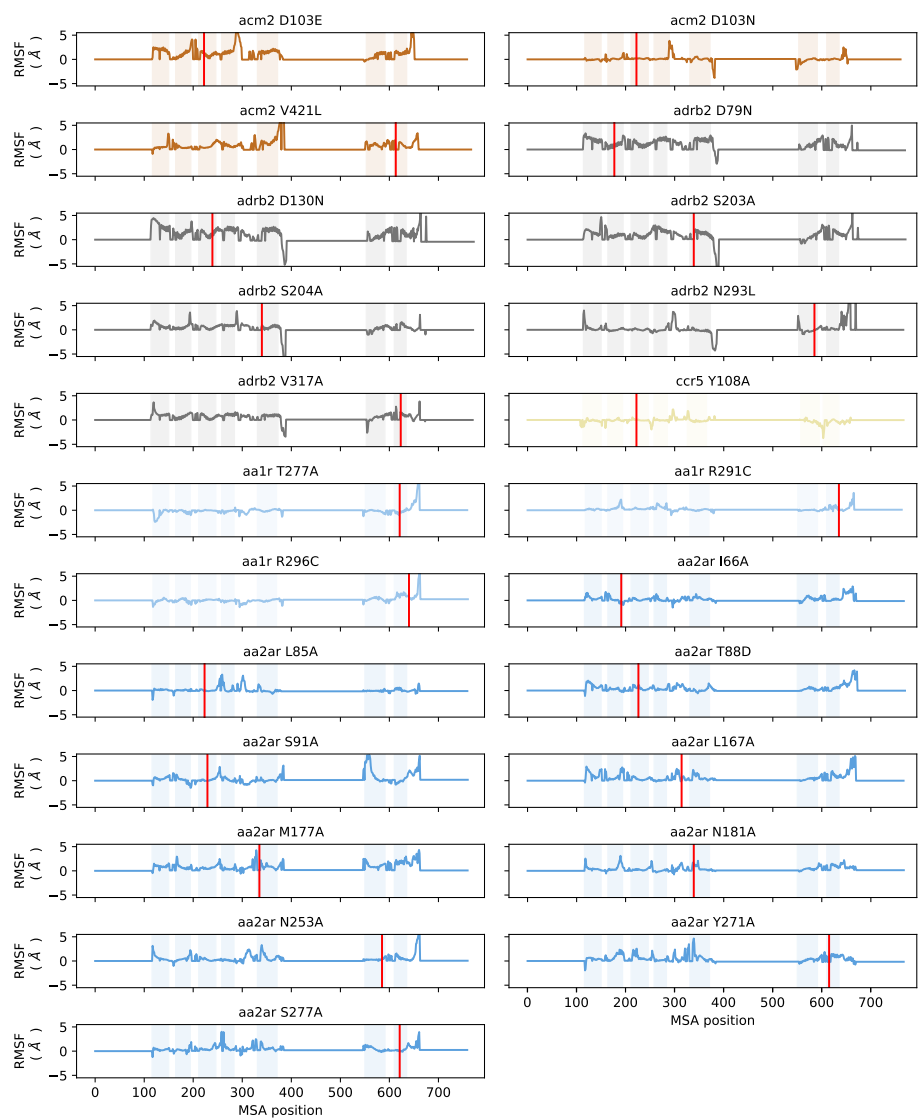
Supplementary Figure 7.4. Representation of ps3DDPD feature values for the 26 GPCRs in the WT set. Each receptor is represented in a different color and receptors from the same subfamily/family are represented in the same color palette.



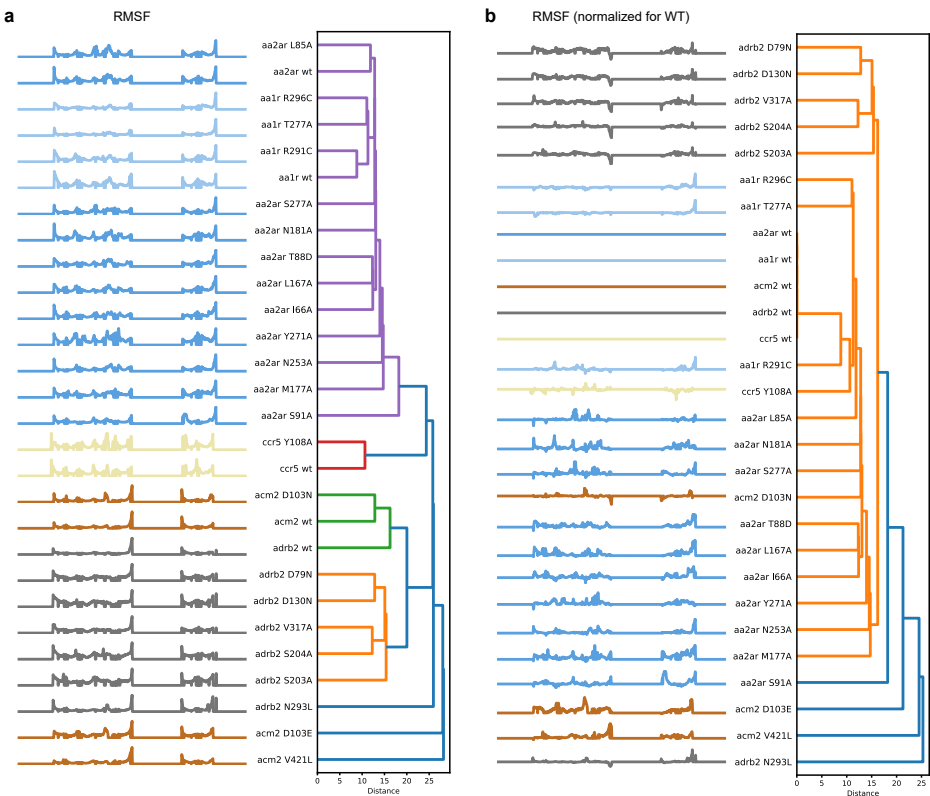
Supplementary Figure 7.5. Top 25 most important features in PCM regression models using a temporal split validation. The importance was averaged across the ten random seeds trained and the SD represented as error bars. The models were trained on the following protein descriptors: **a)** rs3DDPD, **b)** ps3DDPD, **c)** combination of rs3DDPD and Zscale van Westen, **d)** combination of ps3DDPD and PhysChem.



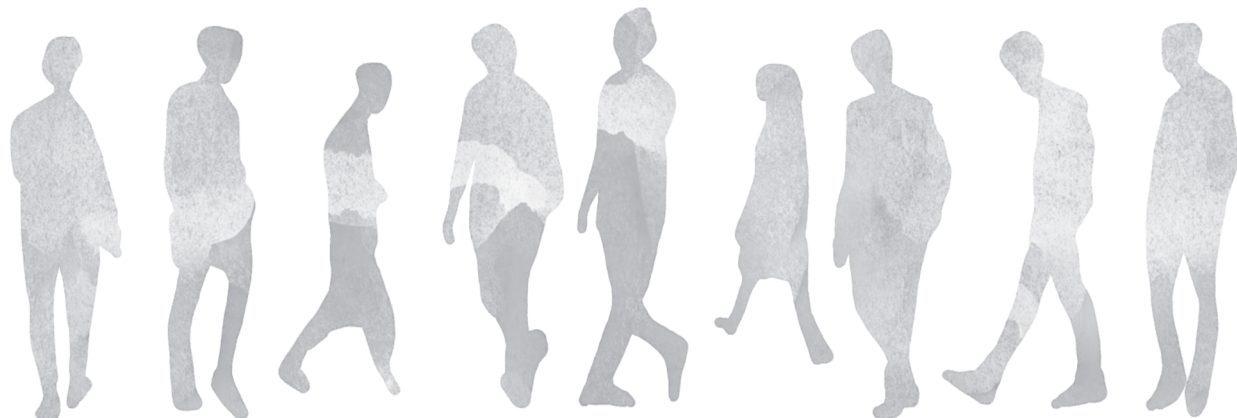
Supplementary Figure 7.6. Distribution of in vitro consequences from available mutagenesis data for the GPCR mutant set in GPCRdb. In the y axis, it is represented the number of ligands with available experimental fold change of virtually no change (between -2 and 2), positive or negative change (between absolute 2 and 30 fold change), or big positive or negative change (bigger than absolute 30 fold change). Bars are stacked for each mutant of the five targets in the set: **a)** adenosine A1 receptor (AA1R), **b)** adenosine A2A receptor (AA2AR), **c)** muscarinic acetylcholine receptor 2 (ACM2), **d)** beta-2 adrenergic receptor (ADRB2), **e)** CC chemokine receptor 5 (CCR5).



Supplementary Figure 7.7. Mutant GPCR RMSF normalized to WT. RMSF values are aligned to the MSA for easier comparison between targets. Domains representing TM 1-7 are shadowed. The location of the mutation in the MSA is highlighted in red.



Supplementary Figure 7.8. Discrimination of GPCR mutants using RMSF. Hierarchical clustering of GPCR variants based on their Euclidean distance between RMSF vectors. **a)** Mutants represented as MSA-aligned RMSF. **b)** Mutants represented as MSA-aligned normalized to WT. Individual clusters generated under a distance threshold of 70% of the final merge are represented in different colors in the dendrograms.

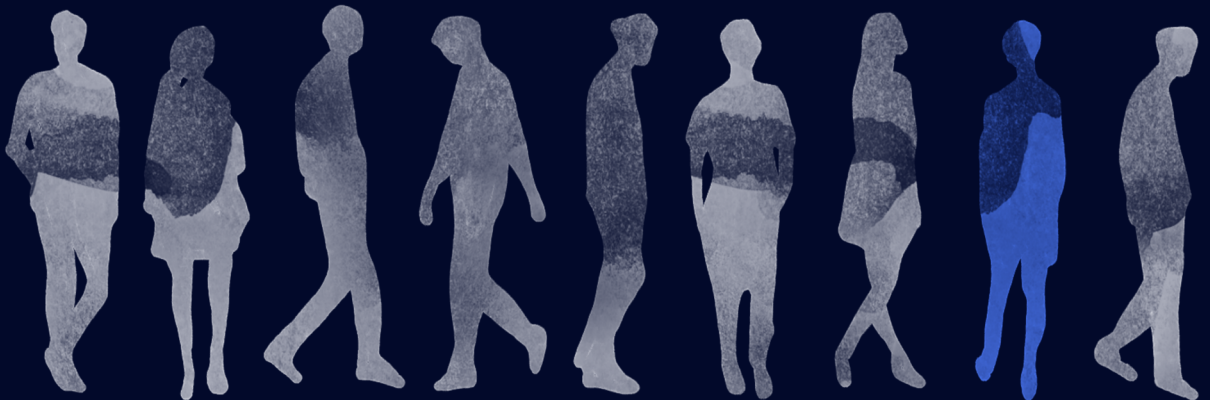


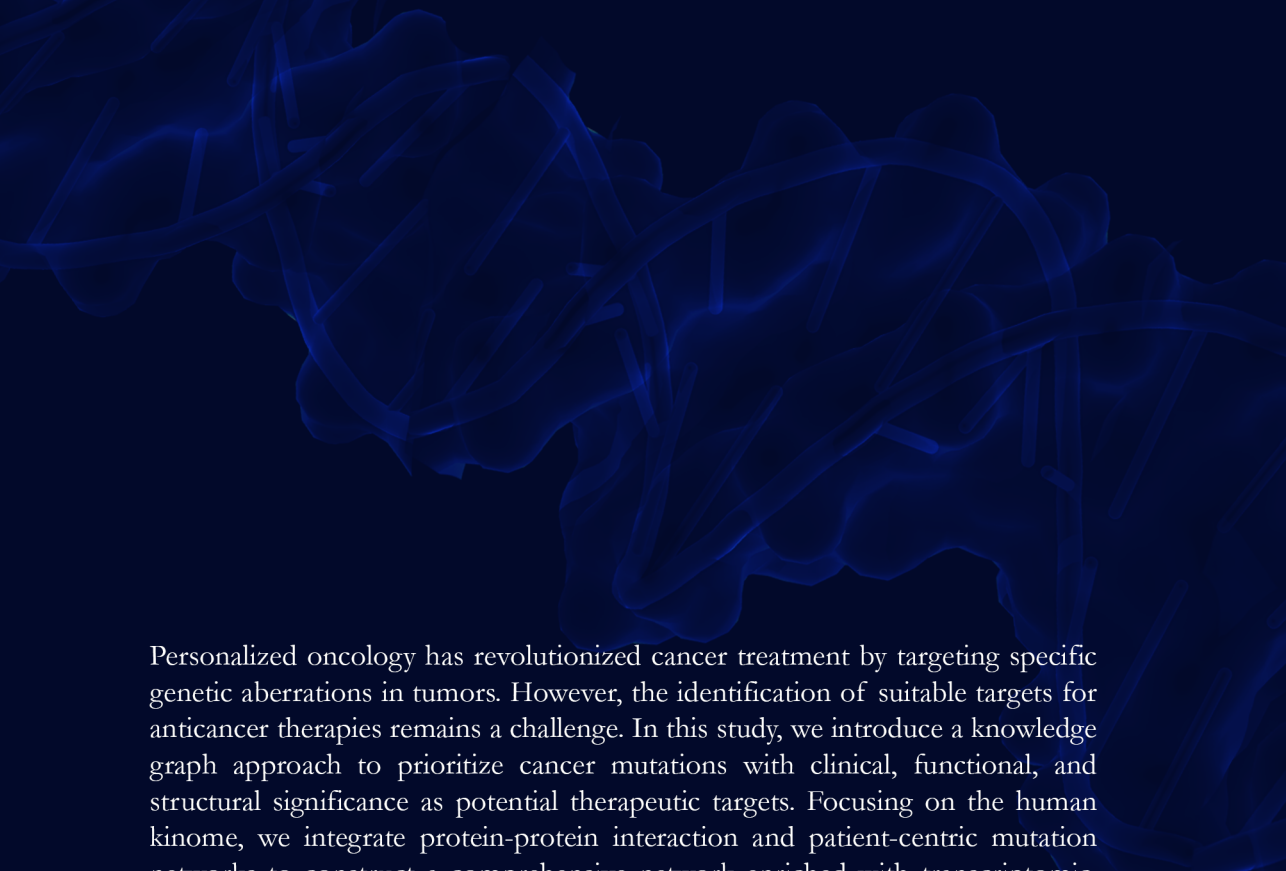
Chapter 8

Connecting the dots: A patient-centric knowledge graph approach to prioritize mutants for selective anticancer targeting

Marina Gorostiola González, Adriaan P. IJzerman, Laura H. Heitman,
Gerard J.P. van Westen

Adapted from *BioRxiv* doi: <https://doi.org/10.1101/2024.09.29.615658> (2024)



An abstract graphic in the top half of the page, consisting of numerous overlapping, translucent blue lines and shapes that form a complex, interconnected network, resembling a protein-protein interaction map or a data visualization of a knowledge graph. The lines vary in opacity and thickness, creating a sense of depth and complexity.

Personalized oncology has revolutionized cancer treatment by targeting specific genetic aberrations in tumors. However, the identification of suitable targets for anticancer therapies remains a challenge. In this study, we introduce a knowledge graph approach to prioritize cancer mutations with clinical, functional, and structural significance as potential therapeutic targets. Focusing on the human kinome, we integrate protein-protein interaction and patient-centric mutation networks to construct a comprehensive network enriched with transcriptomic, structural, and drug response data, together covering five layers of information. Moreover, we make the constructed knowledge graph publicly available, along with a plethora of scripts to facilitate further annotation and expansion of the network. Interactive visualization resources are also provided, ensuring accessibility for researchers regardless of computational expertise and enabling detailed analysis by cancer type and individual layers of information. This comprehensive resource has the potential to identify relevant mutations for targeted therapeutic interventions, thereby advancing personalized oncology and improving patient outcomes.



Introduction

Therapy selectivity, defined as the ability of a drug to bind specifically to its intended target, is a critical aspect of the drug development pipeline. Many therapies fail in clinical trials due to side effects caused by off-target binding – or, in other words, lack of selectivity¹. While the treatment of most diseases primarily requires target selectivity, other features may be crucial, such as tissue selectivity or, in the case of anticancer treatments, cancer-cell selectivity^{2,3}. Target selectivity is related to the pharmacological properties of the drug, whereas tissue selectivity is associated with its pharmacokinetic properties⁴. Cancer-cell-selectivity involves specifically attacking targets predominantly present in cancer cells while sparing healthy cells⁵. In combination, the optimization of these selectivity features is crucial to enabling efficient and safe therapies^{4,6}.

The advent of targeted anticancer therapies, also referred to as personalized or precision oncology, represents a significant shift in cancer treatment strategies and is based on the concept of therapy selectivity⁷. Compared to traditional cancer treatments such as chemotherapy and radiation therapy, targeted therapies leverage the unique genetic makeup of tumors to selectively target cancer cells⁷. Some of the characteristics of tumors that make this possible include the differential expression and the mutation of certain targets compared to their counterparts in healthy cells^{8,9}. Protein kinases are predominantly used as personalized anticancer targets given their high relevance in cancer signaling and aberrant genetic landscape across cancer types^{10,11}. In particular, mutated kinases may present functional or structural differences that distinguish them from their otherwise highly conserved protein family and that can be selectively targeted^{12,13}.

Current anticancer kinase targets commonly overexpressed in tumor tissue include human epidermal growth factor receptor 2 (HER2) in breast cancer and fms-like tyrosine kinase 3 (FLT3) in acute myeloid leukemia⁸. Mutated anticancer kinase targets leveraged in the clinic exhibit distinguishing features, such as the fusion protein BCR-ABL present in most patients with chronic myeloid leukemia and resulting from the aberrant coupling of the genes of breakpoint cluster region protein (BCR) and tyrosine-protein kinase ABL1¹⁴. Other structurally distinguishing features of mutated anticancer targets that promote selectivity include altered orthosteric binding pocket conformations and novel allosteric binding pocket formation^{15,16}. These have also been leveraged in the clinic, as exemplified by the epidermal growth factor receptor 1 (EGFR) L858R activating mutation targeted by selective orthosteric small molecule tyrosine kinase inhibitors (TKI)¹⁷ and clinical candidates targeting an allosteric pocket selectively in phosphoinositide 3-kinase α (PI3K α) activating mutants¹⁸.

To make personalized oncology more accessible, the scientific community is actively engaged in expanding the range of druggable (mutated) targets¹⁹. However, several challenges arise regarding the criteria for identifying suitable candidates²⁰. As previously discussed, the target should exhibit distinct characteristics compared to its healthy tissue counterpart to improve selectivity. Additionally, the candidate must play a functionally significant role in cancer progression to enhance efficacy. Lastly, the target should be relevant to a large group of patients with the cancer subtype. To achieve this, it is

necessary to conduct various experiments that delve into potential candidates, covering structural, functional, and multi-omics analyses – a process that can be quite time- and cost-intensive²¹. In this context, different holistic computational strategies have emerged as particularly effective in compiling heterogeneous data types and prioritizing target candidates^{22–25}.

Knowledge graphs, or complex networks, stand out as significant computational tools employed in cancer research due to their high versatility and interpretability²⁶. By adopting graph-based data representations, these methods facilitate the storage and comprehensive analysis of diverse data entities (nodes) and their interrelations (edges). This functionality not only enables explicit knowledge retrieval but also facilitates the exploration and prediction of implicit knowledge by applying complex network algorithm analyses or deep learning approaches²⁷. Some of the most fundamental graph-data representations in cancer research are gene regulatory networks and protein-protein interaction (PPI) networks, which represent causal or physical associations between entities of the same data type²⁸. The nodes in these networks can be enriched with other types of data, such as clinical-relevant genomic data²⁹, transcriptomics^{30–32}, multi-omics³³, disease-associated scores³², or inhibitor profiling³², and also combined with other networks^{34,35}. Alternatively, knowledge graphs can be constructed with heterogeneous data nodes representing entities other than genes or proteins, such as genetic variants³⁶, diseases^{36–39}, phenotypes^{38,40,41}, symptoms³⁹, treatments⁴¹, drugs^{36,37,40,41}, and risk and prevention factors³⁹. These graphs have a broad range of applications, spanning from cancer diagnosis and subtype classification³⁵ to prevention⁴¹ and treatment planning⁴⁰. Among these applications, there are also various tasks related to oncological drug discovery such as pathogenesis analysis³⁸, mutant driver³⁴ and resistance²⁹ prediction, biomarker³⁰ and target³³ identification, drug repurposing³⁷, and drug sensitivity prediction^{23,31}. However, the multi-faceted nature of anticancer drug selectivity, which makes it an intriguing subject for study as a knowledge graph, remains largely unexplored in this context.

In this study, we introduce a knowledge graph approach for prioritizing cancer mutations with clinical, functional, and structural significance as potential targets for selective anticancer therapies. Due to limitations in data availability, our focus was on the human kinome, representing the complete set of protein kinases encoded in the human genome. This knowledge graph was constructed by integrating two distinct networks. Firstly, a pre-existing PPI network was used, linking protein-encoding genes through phosphorylation events, which are the main kinase signaling events⁴². Secondly, a patient-centric network was developed, connecting kinase somatic mutations based on their co-occurrence in cancer patients sourced from the Genomic Data Commons (GDC) database⁴³. To enhance clinical and functional relevance, gene nodes in the knowledge graph were annotated with transcriptomics data. Additionally, mutation nodes underwent structural and functional annotation through analyses from primary sources including the protein data bank (PDB)⁴⁴ and KLIFS⁴⁵. Finally, all nodes were enriched with bioactivity data from ChEMBL to evaluate the druggability and drug sensitivity of mutations⁴⁶. This comprehensive network serves as a valuable resource in cancer research, potentially facilitating the identification of relevant mutations for targeted therapeutic interventions.

Results and discussion

Overview of primary sources: clinical, structural, and functional relevance

A knowledge graph was constructed to enable the prioritization of kinase cancer mutations as selective targets for anticancer therapies. The perfect candidate was defined as a mutation of clinical relevance with the maximum potential to induce differential pharmacological effects compared to its wild-type counterpart and negligible potential to develop resistance mechanisms⁴⁷. Clinical relevance was defined by a mutation recurrence in multiple cancer patients, but also by the kinase overexpression in a particular cancer type, which is a common druggability and cancer-cell selectivity indicator in targeted anticancer therapies. The last two conditions were further linked to the mutation's structural and functional relevance. In this context, orthosteric ligand binding pocket mutations were considered structurally relevant due to their potential to modify the pocket's conformation, which can be exploited to promote selectivity. These mutations were also considered functionally relevant because targeting them has the potential to directly disrupt the protein's function. Moreover, additional resistance mutations in the pocket have the risk of disrupting the binding of endogenous substrates needed for their activation and have thus a higher resistance threshold. Allosteric pockets have also been targeted in the past to increase target selectivity in cancer cells¹⁵, but they were not directly considered here due to constraints locating them. Functional relevance was also characterized by the importance of the target kinase in the cellular phosphorylation network. Apart from being central to crosstalk in cancer, kinases with a central role are less likely to have their signaling network re-routed and are therefore less prone to developing resistance⁴⁸.

Data was therefore collected from primary sources across five layers of information to address the key questions guiding the selection of selective mutation target candidates with clinical, structural, and functional relevance (**Figure 8.1**). The information pertained to either mutations (top three layers in **Figure 8.1**) or their corresponding targets (bottom two layers in **Figure 8.1**). Kinase mutations were derived from cancer somatic mutation data from the NIH GDC dataset compiled in **Chapter 5**^{43,49} and their connections enabled the analysis of mutation co-occurrence in a patient and overall mutation recurrence (third layer). Mutations were further annotated with information regarding their location on the protein (second layer) and their pharmacological effect with respect to the wild-type protein (first layer). Structural location was determined through the analysis of data from a family-independent source (PDB)⁴⁴ and a kinome-specific source (KLIFS)⁴⁵. The analysis of PDB complexes enabled the annotation of the distance between the mutated residue and the ligand centroid as a proxy for the distance to the orthosteric binding site, as described in **Chapter 4**. Additionally, the aligned position in the kinase binding pocket was annotated from KLIFS. The pharmacological effect of mutations was annotated from the combined analysis of the differences between bioactivity distributions in mutant and wild-type targets in ChEMBL and the Papyrus dataset, as previously introduced in **Chapter 4**. Apart from the kinases with cancer-related mutations, targets were derived from the phosphorylation PPI network defined by Olow *et al.*⁴² and their connections supported the analysis of phosphorylation events between

kinases and their substrates (fourth layer). Genes encoding the protein targets were further annotated with their differential expression in different cancer types compared to normal tissue that was derived from the GDC dataset (fifth layer).

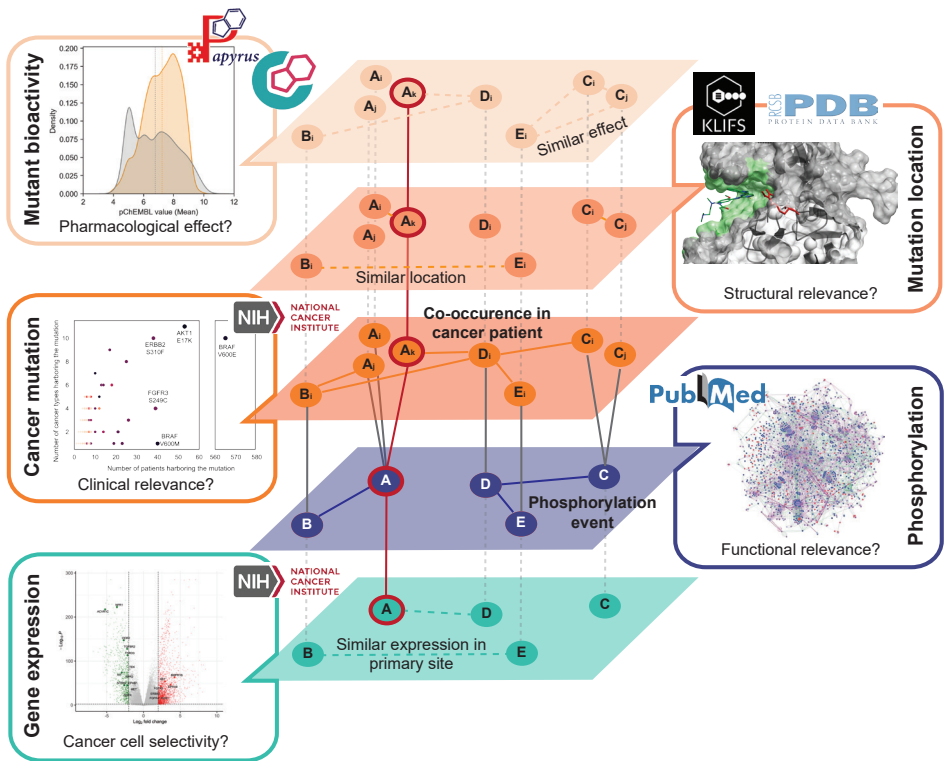


Figure 8.1. Overview of five layers of information contained in the knowledge graph. The top three layers correspond to somatic mutations, while the two bottom layers correspond to gene / protein targets. From top to bottom, the first layer represents the bioactivity differences triggered by mutations with respect to wild-type. The second layer represents the mutation’s structural location in the target protein. The third layer represents the occurrence of mutations in cancer patients. The fourth layer represents the phosphorylation events between targets. Finally, the fifth layer represents the differential expression of the target genes in tumor tissue in different cancer types (defined by the primary site where the tumor developed) compared to normal tissue. Each layer answers a distinct question about the mutation’s relevance as an anticancer target. The connections represented by lines between entities (nodes) within the third and fourth layers are used as edges in the knowledge graph, as well as the connections between nodes of those two layers. The connections within the rest of the layers, represented by dashed lines, are not kept in the knowledge graph but indicate how nodes relate based on the data represented in each layer. The red line connecting and highlighting nodes across the five layers exemplifies a potential candidate with information spawning across the five layers.

While knowledge graphs can incorporate a larger variety of information, such as Astra Zeneca’s BIKG which was constructed with 37 public and internal datasets⁵⁰, subsets of the graphs^{29,51} or more tailored representations³² are needed to answer specific research questions. For example, the CancerOmicsNet graph was created with the purpose of predicting therapeutic effects in various cancer cell lines. It incorporated layers

of information similar to those mentioned here, consolidated into a single graph using a phosphorylation PPI network³². In contrast, our approach involves integrating mutation-specific data that aligns with the specific objectives set for this graph. The data across these five layers enables an investigation into the clinical, structural, and functional implications of mutations and targets, while also tackling important considerations around druggability, like selectivity towards cancer cells and mutations. This information was incorporated into a knowledge graph and analyzed accordingly, as detailed in the following sections.

Knowledge graph architecture

Mutations and targets were connected in the knowledge graph by association edges and comprised the two types of nodes available. Mutation nodes were connected by edges representing co-occurrence in the same cancer patient. Target nodes – referred to as gene nodes from now on to denote target protein-coding genes – were connected by edges representing all phosphorylation events between kinases and their substrates (**Figure 8.2a**). The rest of the information collected from primary sources was collapsed from the five layers of information into one and stored as – mutation or gene – node attributes (**Figure 8.2b**). Edges representing cancer patient co-occurrence were also annotated with attributes representing the corresponding patient and cancer type, which allowed analysis per cancer type in subsequent sections. This simple graph architecture was chosen to keep mutations and their corresponding protein-coding genes as the central elements. The knowledge graph was constructed in NetworkX⁵² allowing multiple connections between the same two nodes. In the final kinome graph, this amounted to 78,782 nodes and 6,515,059 edges. Node entities and their relationships were determined manually here to maximize accuracy. However, other options are becoming available to extract them automatically, such as using large language models⁵³, in turn facilitating ontology mapping³⁹. These novel approaches are particularly relevant and promising in large knowledge graphs comprising numerous primary sources, although several challenges still need to be addressed, particularly regarding precision and biases in the data extracted⁵⁴.

The distribution of types and subtypes of nodes and edges is presented in **Table 8.1**. Gene nodes amounted to 1,571, of which 667 were kinases and 904 substrates. Of note, the original phosphorylation PPI network contained 774 kinase nodes but only 625 were kept after applying the filter for proteins with kinase activity as defined in the methods section. Additionally, 42 genes coding for proteins with kinase activity that were not included in the original phosphorylation PPI network were added to the graph due to the presence of mutations in cancer patients. Substrate types were further investigated, particularly within the context of membrane proteins. Specifically, 17 nodes were identified as G protein-coupled receptors (GPCRs), while 14 were categorized as solute carriers (SLCs). Additionally, 130 other substrates were found to be primarily localized to the plasma membrane. Moreover, out of the kinase nodes, 65 were also receptors. The majority of nodes (77,009) represented cancer mutations. Out of these, only 34 were associated with bioactivity data from both ChEMBL and the Papyrus dataset. Consequently,

an additional 202 kinase mutation nodes with bioactivity data were incorporated into the graph to allow the interpolation of mutation characteristics affecting bioactivity. In an expanded version of the graph aimed at drug discovery, other node types could include small molecules^{36,37,40} instead of summarizing the effects of mutations over all tested drugs. This could result in an additional layer of information where edges represent similarity between small molecules⁵⁵. An alternative would be to report the effect of individual kinase inhibitors as individual attributes in each kinase node³². However, other primary sources for mutation-driven bioactivity changes should be considered before implementing any of these expansions⁵⁶.

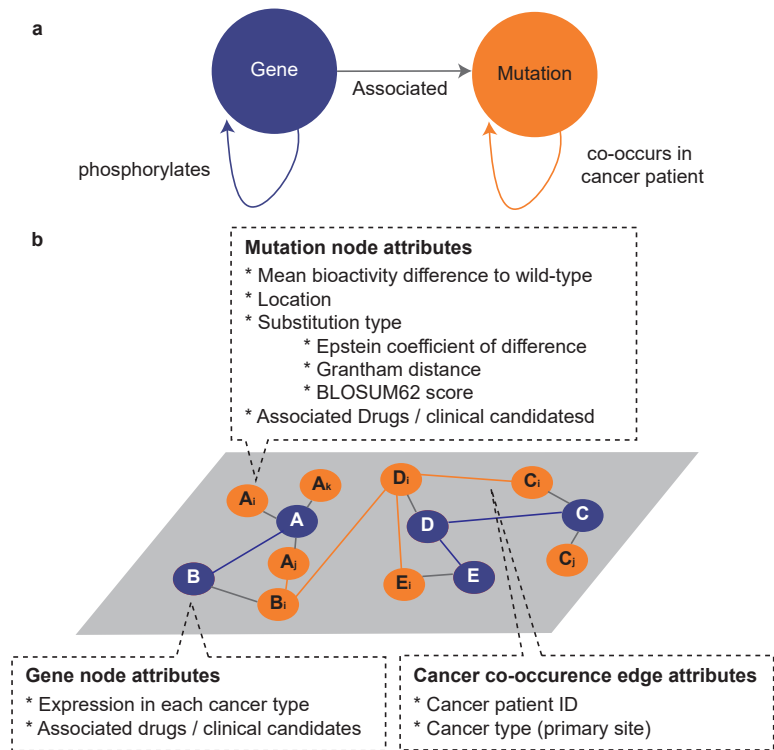


Figure 8.2. Knowledge graph architecture. **a)** Schematic representation of the two types of nodes in the graph (circles) and their edges (arrows). **b)** Example graph representation with attributes linked to mutation and gene nodes, as well as to cancer co-occurrence edges. Mutation nodes are represented in orange and gene nodes in blue. Edges connecting different types of nodes correspond to the edges described in a).

Filtering nodes from the original phosphorylation network resulted in a reduction of phosphorylation edges between gene nodes from the original count of 5,963 to 5,759. Cancer mutation nodes were connected to their respective genes via one or multiple edges, indicating the frequency of the mutation within that gene across patients. Cancer mutations were collected from 8,518 patients across 48 cancer types. As a result, the number of edges between cancer mutations and genes exceeded the count of cancer mutation nodes by 10,239. Conversely, the number of edges connecting genes and non-cancer mutations equaled the count of non-cancer mutation nodes. Over 6.4 million edges

represented the co-occurrence of two mutations in the same cancer patient, highlighting a high mutation burden in kinases. The distribution varied widely, with some patients having just one edge representing one mutation and others having up to 672,220 edges representing up to 1,159 unique mutations. The median number of co-occurring edges per patient was six, indicating most patients had few kinase mutations, while a few had a large number of mutations.

Table 8.1. Distribution of node and edge types and subtypes across the kinome knowledge graph.

Entity	Type	Subtype	Number of entities
Nodes	Gene	Kinase	667
		Substrate	904
	Mutation	Cancer	77,009
		Other (ChEMBL + Papyrus)	202
Edges	Gene - Gene	Phosphorylation	5,759
	Gene - Mutation	Cancer	87,248
		Other (ChEMBL + Papyrus)	202
	Mutation - Mutation	Cancer patient co-occurrence	6,421,798
		Other (ChEMBL + Papyrus multiple substitutions)	52

The analysis of node degrees in the knowledge graph confirmed this irregular pattern, highlighting the presence of a few nodes with exceptionally high degrees and a much larger number of nodes with lower degrees (**Figure 8.3**). A node degree represents the number of edges linking it to other nodes. As anticipated, recurrent cancer mutations and their corresponding genes exhibit high degrees within the knowledge graph. PIK3CA R88Q and BRAF V600E are the two mutations with the highest degree (10,275 and 6,219, respectively) and were present in 68 and 565 patients respectively (**Supplementary Table 8.1, 8.2**). Interestingly, other PIK3CA mutations with a higher occurrence frequency in cancer patients (PIK3CA E545K and H1047R present in 258 and 234 patients, respectively) showed comparatively lower node degrees (2,484 and 2,439, respectively). These results emphasize the importance of mutation co-occurrence and tumor mutation burden (TMB), which has recently been highlighted as a tumor biomarker⁵⁷. Accordingly, genes linked to TMB, such as TTN and OBSCN⁵⁸, were also highlighted in the knowledge graph based on their node degree. In a similar manner, mutations not highly recurrent across cancer patients but present in patients with a high TMB across cancer types (**Supplementary Table 8.3**) showed a high node degree, three of them even being present in natural variance (**Supplementary Table 8.2**). Mutation co-occurrence in patients has also been highlighted in other graph approaches independent of mutation recurrence⁵⁹. However, in order to pinpoint exclusively functionally relevant mutations, it might be appropriate to filter frequent mutations *a priori*⁶⁰. From a graph architectural point of view, the emphasis on individual patients that is central to our approach has also been reported in other patient-centric graphs. The most common approaches also consider genes in PPI networks that are enriched with multi-omics data^{61,62}. However, other graph architectures are possible where nodes represent individual patients with different multi-omics attributes⁶³. Incorporating patient nodes into the

existing graph and linking them to their corresponding mutations and genes could aid in filtering out patients with high TMB and mutations with low recurrence rates, resulting in a more refined graph. As elaborated in the subsequent sections, addressing data sparsity within the graph is a significant bottleneck for analysis, and constructing focused subgraphs can sometimes mitigate this issue.

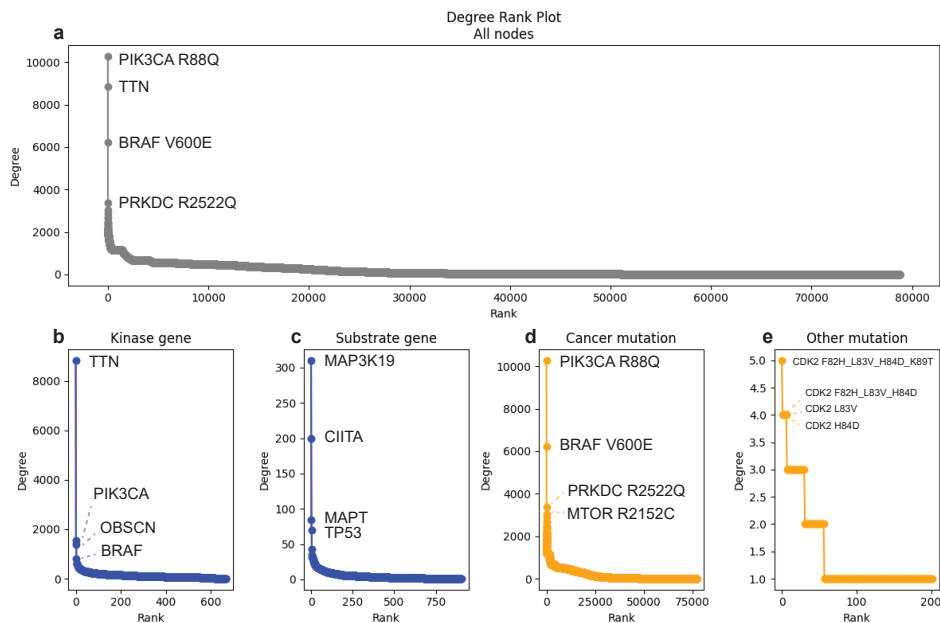


Figure 8.3. Node degree rank analysis in the kinome knowledge graph. Nodes are ranked based on their degree, which is calculated as the number of edges connecting the node to other nodes in the graph. The degree rank analysis is calculated for all nodes in the graph (a) as well as for each node subtype independently: kinase (b) and substrate (c) gene nodes, and cancer (d) and other (e) mutations. The top four ranked nodes in each case are labeled accordingly.

Attribute annotation sparsity across the graph

Nodes in the graph representing mutations and genes were annotated with attributes covering the information across the five layers of information previously described in **Figure 8.1**. Node attributes are crucial for graph analysis, as they can further denote the embeddings or labels for predictions in machine learning applications. In particular, the number of approved drugs and bioactivity changes were key mutation attributes that could be used to classify mutations for targeted therapy. However, the annotation density (this is, how many nodes have non-null values for a particular attribute) of these two and many other node attributes was rather low (as depicted in **Figure 8.4** for mutation nodes, and in **Supplementary Figure 8.1** for gene nodes).

Of the 1,571 gene nodes in the graph, 351 (22.34%) have at least one drug in any phase of development directly linked on ChEMBL as its target (**Supplementary Figure 8.1**).

However, only three kinase mutations are listed as the target for drugs approved or in development (**Figure 8.4f-h**). These are BRAF V600E and EGFR L858R activating mutations and EGFR T790M acquired resistance mutation, all previously linked to cancer. Apart from being too low to generate labels for classification tasks, these annotations are a direct consequence of the non-triviality of the variant annotation pipeline in bioactivity databases that we described in **Chapter 4**. For example, EGFR L858R does not have any approved drug directly linked to the mutation in the mechanism of action, probably due to the fact that most approved first-generation EGFR inhibitors were developed to target selectively either the activating deletion in exon 19 or the activating mutation L858R, and deletions are not fully curated in ChEMBL as of yet – although this is work in progress. Similarly, only 215 (0.28%) mutation nodes have a bioactivity change annotation (**Figure 8.4e**), which is a very small number for further machine learning analysis. These numbers could be increased with data from additional primary sources for mutant-induced bioactivity changes, such as the mutation-induced drug resistance database (MdrDB)⁵⁶.

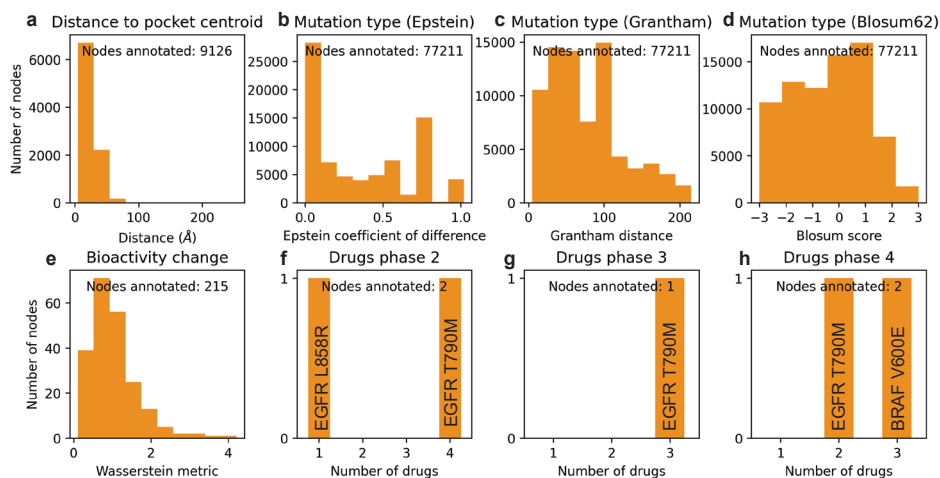


Figure 8.4. Density and distribution of mutation node attributes in the kinome knowledge graph. For each of the attributes, the number of mutation nodes with non-null values for the particular attribute is depicted on the y-axis. Moreover, the graphs represent the distribution of the attribute values across the mutation nodes on the x-axis, in the form of histograms or bar plots, depending on the density of each attribute. Distribution is represented as histograms for the distance to the pocket centroid calculated from PDB complexes (**a**), the mutation type as determined by the Epstein coefficient of difference (**b**), the mutation type as described by the Grantham distance (**c**), the evolutionary probability of the mutation type as described by the Blosom score in the Blosom62 matrix (**d**), and the bioactivity change represented by the Wasserstein distance between the bioactivity distribution for the mutation and the wild-type protein found in ChEMBL (**e**). Bar plots represent the number of drugs in different phases of development according to ChEMBL labeling: clinical phases 2-3 (**f-g**), and approved drugs (**h**). Pre-clinical candidates (phase 0) and drugs in clinical phase 1 are not included because they were not annotated in any mutation nodes. The mutations represented in the bar plots are labeled for reference.

Mutation attributes representing the characteristics of the amino acid substitution that could be used for the node embeddings were less sparsely represented (**Figure 8.4a-d**). Mapping of the three metrics representing mutation types (Epstein coefficient of

difference⁶⁴, Grantham distance⁶⁵, and Blosum62 score⁶⁶) could be done for all 100% of the mutations. These three metrics were selected to cover different aspects defining the amino acid substitutions, namely physicochemical properties (Epstein coefficient of difference, which is directional and represents the size and polarity difference between the wild-type and mutated amino acid; and Grantham distance, which is non-directional and calculates said difference based on atomic composition and molecular volume on top of polarity), and evolutionary conservation (Blosum score, which defines the evolutionary probability of a substitution relative to random probability, calculated for proteins clustered at 62% sequence similarity). Of note, 57% of the mutations had an Epstein coefficient of difference lower than 0.4, and 74% a Grantham distance below 100, meaning that the majority of substitutions were rather conservative. Simultaneously, 54% of the mutations were likely to happen by chance, as represented by a Blosum score of zero or higher, and in fact, 2,759 (3.57%) of all the mutations were also found to occur as natural variance in the 1000 Genomes dataset previously compiled in **Chapter 5**. The complementarity of the three amino acid substitution metrics was illustrated by the variations observed in their distributions (**Figure 8.4b-d**) while still aligning with clinical significance (**Figure 8.5**). From the 19 most recurrent mutations in cancer patients (occurring in more than 20 patients), six mutations (BRAF V600E, FGFR3 S249C, ERBB2 S310F, FGFR2 S252W, EGFR L858R, and PIK3CA C420R) were identified as disruptive based on an Epstein coefficient of difference over 0.4 and a Grantham distance greater than 100. Additionally, these mutations were determined to be less likely to occur by random chance based on a negative Blosum score. The most common mutation, BRAF V600E, occurring in 565 patients, had an Epstein coefficient of difference of 1.0 (**Figure 8.5a**), a Grantham distance of 121 (**Figure 8.5b**), and a Blosum score of -2 (**Figure 8.5c**). It is key to note, however, that three oncogenic PIK3CA mutations highly recurrent in breast cancer (E545K, H1047R, and E542K⁶⁷) would not be captured by any of these three metrics, while the three mutations with associated drugs under development (BRAF V600E, EGFR L858R, EGFR T790M) would. Therefore, although these annotations are useful to get a general understanding of the potential effect of an amino acid substitution, more advanced amino acid embeddings may be preferred for machine learning applications⁶⁸.

Attributes representing the structural location of the mutation in the protein were mapped to 9,126 mutations (11.82%) occurring in 161 genes when calculated from PDB complexes (**Figure 8.4**), and 3,854 mutations (4.99%) occurring in 295 genes were annotated with a KLIFS pocket position (**Supplementary Figure 8.2a**). Interestingly, only 1,890 mutations (2.44%) occurring in 138 kinases had a double structural annotation (from PDB and KLIFS), which highlighted the complementarity of these approaches. The annotations extracted from KLIFS covered a larger number of kinases since they also included pocket annotations for kinases with only apo structures available in the PDB, such as TTN. An additional advantage of the KLIFS annotation is that it can be directly linked to its structural relevance – functionally and pharmacologically. The PDB annotations, on the other hand, enabled the annotation of mutations in an additional 22 genes for which there was no data available in KLIFS. More importantly, they enabled the annotation of mutations outside of the ATP binding pocket, providing a more holistic view of the structural location of the mutations, which is very relevant

for drug response prediction⁶⁹. In fact, while secondary resistance mutations tend to be part of the ATP binding site, many activating mutations crucial for cancer development are outside of the KLIFS-defined binding pocket. For example, from the 25 most frequent cancer mutations shown in **Supplementary Table 8.1** and **Figure 8.5**, only EGFR L858R is part of the KLIFS binding site. This is also exemplified by the discrepancy identified between the number of mutations within the KLIFS binding pocket and the maximum mutation frequency at those positions (**Supplementary Figure 8.2a,c**). While positions c.l.69 and c.l.74 in the kinase catalytic loop were the most frequently mutated overall with 91 and 88 individual mutations respectively, the two most frequent individual mutations in cancer patients within the pocket were EGFR L858R in the activation loop (a.l.84) and ERBB2 V842I in the β -sheet VI (VI.67), occurring in 23 and 17 patients, respectively. It is worth noting that none of the seven binding pocket positions with mutations occurring in 10 or more patients were highly conserved, as reported by KLIFS. This highlights the need for cancer cells to not fully disrupt kinase function⁷⁰. The analysis of mutations annotated with both structural sources allowed for the correlation of the KLIFS pocket positions with precise distances to the ligand centroid. This demonstrated variability in the distances that is consistent with varying ligand sizes and binding modes observed among kinases (**Supplementary Figure 8.3**). However, this analysis also highlighted the presence of outliers representing measurement errors, which may arise from inconsistencies in sequence numbering and should be corrected in the pipeline. For kinases, one solution could be to utilize the KLIFS-curated structures, although this approach is specific to the kinome and cannot be expanded to other protein families. Other solutions are possible to structurally annotate cancer mutations and incorporate this information into PPI networks, but the existence of incomplete and incorrect structures in the PDB is a constant bottleneck⁷¹.

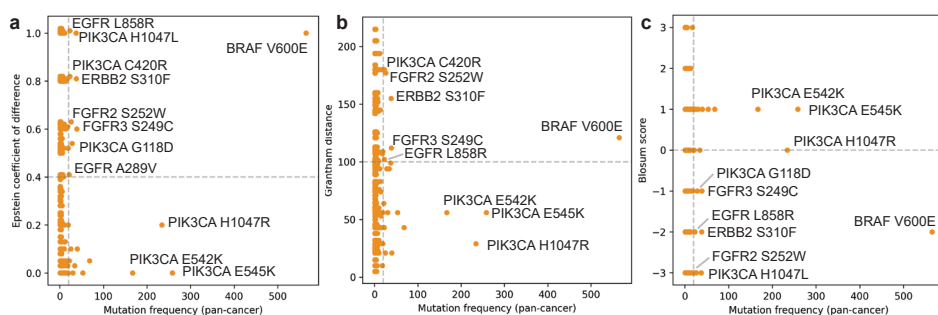


Figure 8.5. Correlation between cancer mutation frequency and three metrics describing the amino acid substitution in the kinome knowledge graph: Epstein coefficient of difference (**a**), Grantham distance (**b**), and Blosom score (**c**). Mutations occurring in more than 100 patients pan-cancer are labeled for reference. In **a**), an Epstein coefficient of difference of 0.4 is taken as an arbitrary threshold to distinguish between conservative (<0.4) and disruptive substitutions (>0.4), which are labeled for reference if they occur in more than 20 patients pan-cancer. In **b**), a Grantham distance of 100 is taken as an arbitrary threshold to distinguish between conservative (<100) and disruptive substitutions (>100), which are labeled for reference if they occur in more than 20 patients pan-cancer. In **c**), substitutions with an alignment happening less often than random chance as collected in the Blosom62 matrix (Blosom score < 0) are labeled for reference if they occur in more than 20 patients pan-cancer.

In general, the sparsity of knowledge graphs is a significant bottleneck for machine learning applications as it compromises the quality of the embedding methods used *a posteriori*⁷². However, sparsity in node attributes may not be a problem if the bias is in sync with the graph's objective and can be utilized for additional analysis. For example, some applications use customized attention mechanisms to prioritize nodes with more information for efficient data propagation to their neighbours²³. Similar approaches could be used in the knowledge graph developed here to predict bioactivity differences or approved drug development for mutations using node attribute competition tasks, where node attributes are inferred from the collective data in the graph⁷³. While some of the methods used for node attribute competition tasks are still robust with high node attribute sparsity (up to 80%)^{73,74}, a minimum might still be required. It is important to note that some knowledge graphs are not necessarily developed with the aim to apply additional (machine learning) analyses but with the intention to be a resource that can be interactively explored to gain a holistic view of a certain problem^{36,39}. In the last section, we explore the development and analysis of different subsets of the kinome knowledge graph, which can be a strategy both to decrease node attribute sparsity with the aim to apply machine learning analyses, as well as to create biologically relevant subgraphs that can be interactively explored for data extraction.

Subgraph analysis and interactive exploration: A case study for RTKs

The architecture of the knowledge graph supported the construction of biologically relevant subgraphs based on node and edge attributes. Although the Python package developed to build and analyze the graph supports filtering the graph for any attribute of interest, two main types of subgraphs were pre-defined and explored in this section. To illustrate these analysis options, a smaller graph was created specifically focusing on receptor tyrosine kinases (RTKs) and their substrates, mimicking the structure of the kinome graph (**Supplementary Tables 8.4-8.6** and **Supplementary Figures 8.4-8.8**). The smaller graph was built to facilitate analysis and interactive visualization, but it also served as a way to increase node attribute density while maintaining biological relevance. Compared to the kinome graph, the RTK graph contained approximately a seventh of the original nodes (11,989 nodes instead of 77,211) and almost 50 times fewer edges (141,311 instead of 6,515,059). All mutation node attributes considered in the previous section remained sparse, but in all cases the percentage of mutation nodes annotated increased. For example, the percentage of mutation nodes annotated with bioactivity data increased from 0.28% to 0.77%, and the number of nodes annotated with distance data increased from 11.82% to 15.31% for PDB-annotated nodes and from 4.99% to 7.17% for KLIFS-annotated nodes. While modest, this rise in attribute density is consistent with the significant emphasis on cancer research related to RTKs, as detailed in **Chapter 3**. The interactive visualization module enabled, among other views, the visualization of the phosphorylation edges in the RTK graph between the selected kinases and their substrates (dark and light blue, respectively in **Figure 8.6a**, where node size is proportional to the total number of drugs under development or approved for that node). To improve visualization, different node and edge types and subtypes can be left in the background, as was done in **Figure 8.6a** for mutation nodes and their edges. Of

note, the RTK graph still contains some kinases that are not receptors because they are substrates of RTKs, such as JAK2 which is phosphorylated by EGFR. This can be easily explored by zooming into a particular node, as shown in **Figure 8.6b** for EGFR and all nodes connected to it. The zoomed-in view also enables the interactive exploration of node attributes, on the right-side panel, which for a gene node includes among others expression values for different cancer types.

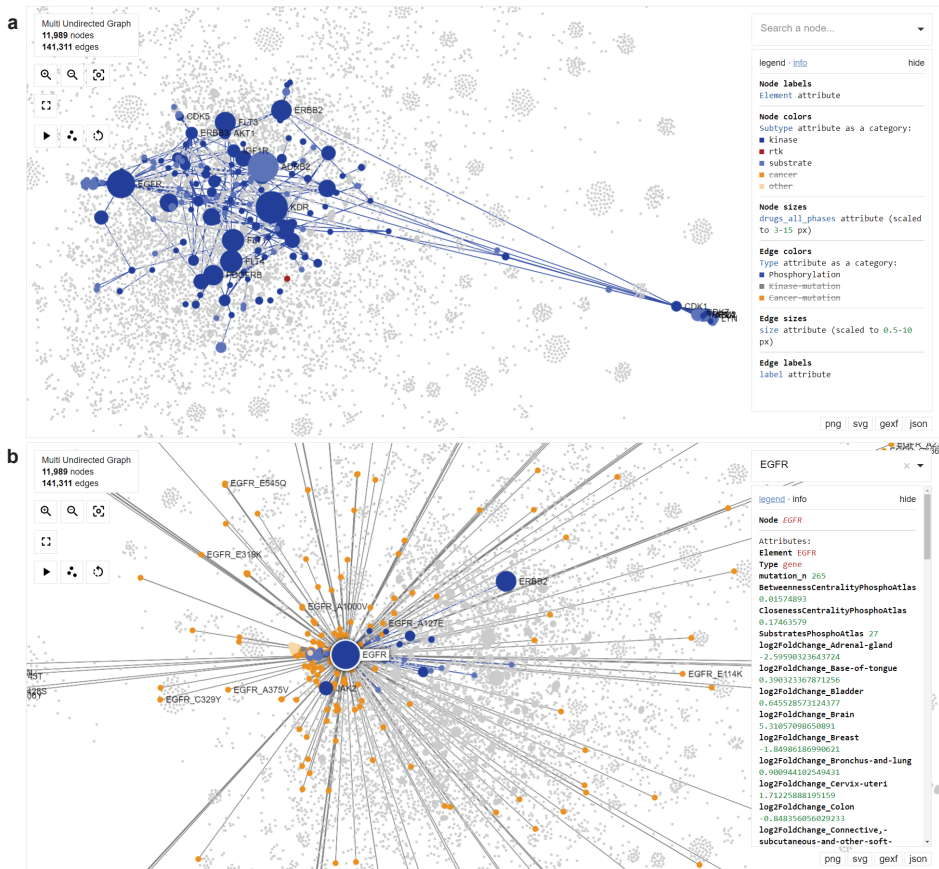


Figure 8.6. Interactive visualization example for the receptor tyrosine kinase knowledge graph. **a)** Visualization of the phosphorylation events between kinases (dark blue nodes) and their substrates (light blue nodes). Every highlighted edge represents a phosphorylation event. The visualization module, powered by IPysigma²⁵, includes a legend panel on the right side that enables search options and filtering and describes the attributes used for node color and size. To improve visualization, all mutation nodes and edges are hidden here. **b)** Example visualization when clicking on a particular node, in this case, EGFR. All edges connecting the node of interest and the paired nodes are highlighted, while the rest of the graph is kept in the background. When a node is selected, all the attributes associated with it are displayed on the right-side panel.

The first of the two pre-defined analysis options supports the construction and exploration of individual “layer” subgraphs that are based on the type of edges. This enables the division of the knowledge graph into three subgraphs: a purely cancer patient-centric

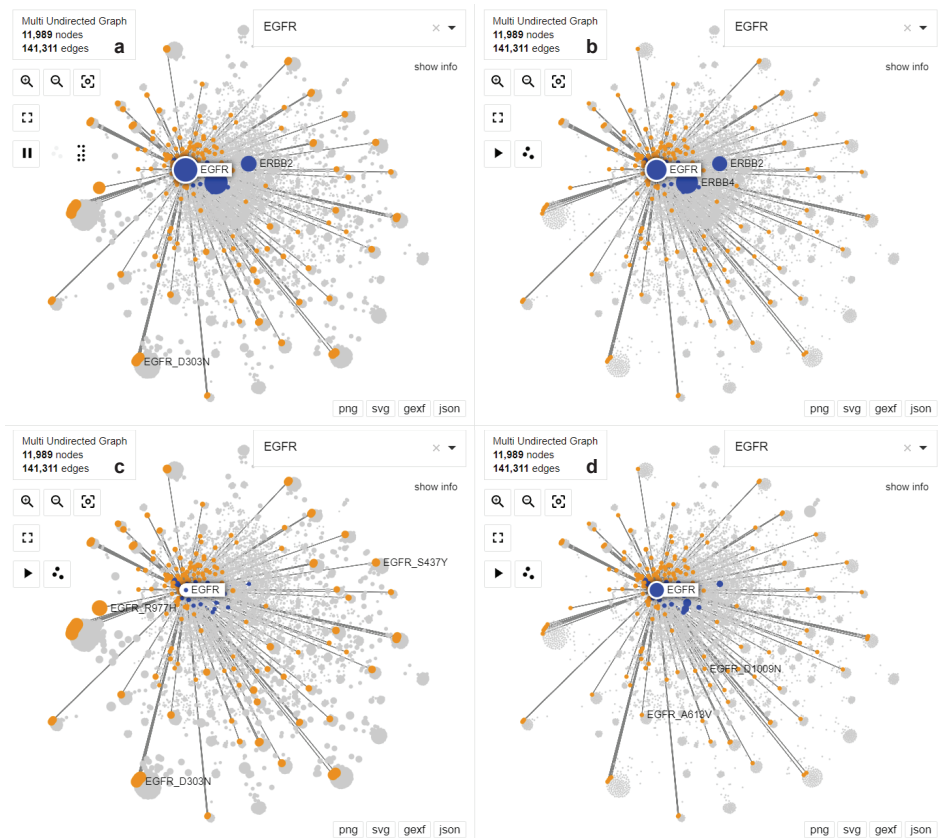


Figure 8.7. Node degree comparison across layers in the receptor tyrosine kinase knowledge graph, with a focus on nodes connected to EGFR for visualization purposes, as done in Figure 8.6b. Gene nodes are represented in blue and mutation nodes are represented in orange. Nodes that are not connected to EGFR are kept in the background and represented in grey. Node size in each panel is determined by the node degree calculated from the whole graph (a) or one of the three pre-defined analysis layers: kinase-mutation layer (b), cancer-mutation co-occurrence layer (c), or phosphorylation layer (d).

network linked by mutation co-occurrence, an association network between kinases and their mutations, and a phosphorylation PPI network. This division is important because it allows for the calculation of graph metrics for each layer independently. This, in turn, enables the characterization of the importance of specific nodes at various levels. By analyzing EGFR across multiple layers, it was possible to determine which layers influenced the final degree metric in the graph (Figure 8.7a). In this particular case, a high degree in gene nodes mostly arose from the kinase-mutation layer (Figure 8.7b) since there was an additional edge for each patient carrying a specific mutation. However, this high node degree could also result from a high degree in the phosphorylation layer (Figure 8.7d). A high degree in mutation nodes, however, mostly resulted from mutation co-occurrence in the same patient (Figure 8.7c). While the results were consistent with the anticipated graph architecture, conducting an analysis of each layer reveals that EGFR exhibits a high degree of connectivity not only due to its prevalence

in mutations across various patients but also because of its multiple interaction partners within the phosphorylation network. This analysis can also be extended to other metrics, for example, betweenness centrality (**Supplementary Figure 8.9**), which represents the influence a node has on the flow of information in the graph. This metric can in turn help determine different community clusters within the graph, as well as key nodes and edges relevant to connections between these communities.

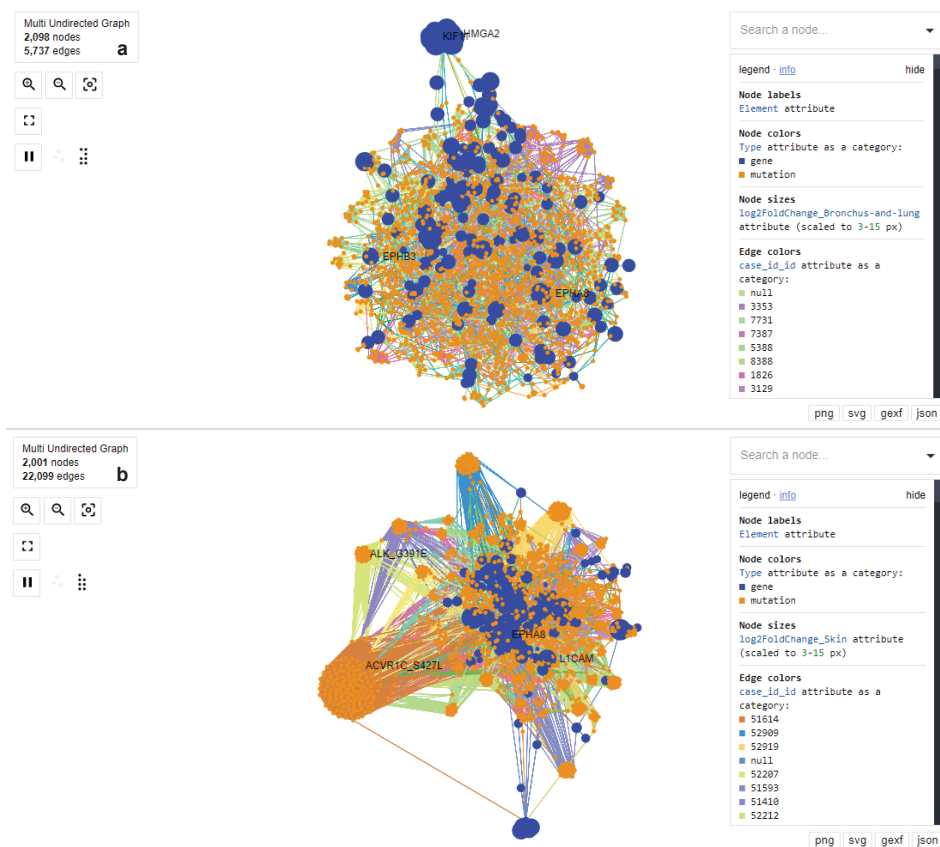


Figure 8.8. Interactive visualization of cancer type subgraphs derived from the receptor tyrosine kinase knowledge graph for the two most populated cancer types: bronchus and lung (**a**, 723 patients) and skin (**b**, 356 patients). Gene nodes are represented in blue and mutation nodes are represented in orange. Node size represents the differential expression (Log2 fold change) of genes in that cancer type tumor tissue compared to normal tissue. Each edge color represents a different patient.

The second pre-defined analysis module allowed individual subgraphs to be easily constructed for each cancer type by filtering the cancer-related edges corresponding to patients with a particular cancer type. These subgraphs also included the phosphorylation events pertinent to cancer-type filtered nodes. The RTK graph covered 44 cancer types, of which the six most populated cancer types were bronchus and lung (723 patients), skin (356 patients), brain (312 patients), corpus uteri (302 patients), bladder (289 patients), and breast (273 patients). The analysis of cancer type subgraphs enabled the

distinction between recurrent and patient-specific mutations, since the former aggregate in the central part of the graph while the latter form clusters in the graph periphery. This phenomenon was clearly distinctive between the bronchus and lung subgraph (**Figure 8.8a**) and the skin subgraph (**Figure 8.8b**), where edges were colored to represent different patients. Most of the RTK mutations occurring in lung cancer patients were recurrent across patients, while several skin cancer patients harbored mutations that were specific for that patient. Similarly to the full graph, different node attributes can be used to interactively explore the graph. For example, the use of cancer type-specific gene expression (Log2 fold change) illustrated in **Figure 8.8** can aid in distinguishing genes with a higher relevance in that particular biological context. In fact, the construction of individual graphs for different genetic makeups is a common strategy when trying to prioritize cancer-related mutations^{76,77}. Additionally, each subgraph can be analyzed independently and the results can be combined and displayed in the main graph, as done with the layer analysis. This further enables the investigation of cancer type-specific relationships in a pan-cancer context (**Supplementary Figure 8.10** for degree analysis and **Supplementary Figure 8.11** for differential expression analysis), which tends to be the preferred strategy in personalized anticancer therapies⁷⁸.

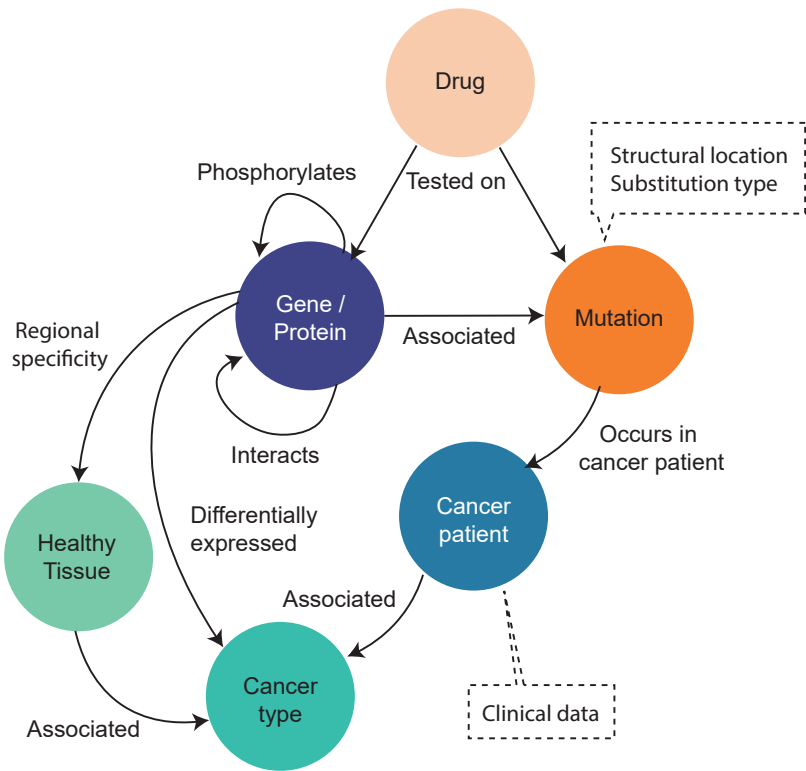


Figure 8.9. Proposed alternative knowledge graph architecture for the kinome patient-centric knowledge graph. Each circle represents a different node type and the arrows represent the edges between nodes. Dashed boxes include attributes that could be associated with specific types of nodes to increase information density.

Conclusions

A cancer patient-centric knowledge graph was constructed to aid in identifying mutations with advantageous characteristics to be targeted selectively by small molecules, hence reducing the side effects of anticancer therapies while maintaining high efficacy. Known oncogenic and targeted genes and mutations stand out from this exercise based on several of the node attributes collected across five layers of information, highlighting the potential of this graph in oncology drug discovery. Although initially tailored for the kinome, this framework can readily be adapted to other protein families by modifying kinome-specific primary sources. While the current graph facilitates interactive exploration and basic analyses, its limited node attribute density poses challenges for advanced machine learning applications following GNN embedding. To address this constraint, future exploration of alternative graph topologies, including the proposed complex topology in **Figure 8.9**, is recommended. This study underscores both the synergistic potential of integrating diverse data types and the critical need for expanded data availability to enhance predictive capabilities. Broader testing and reporting in publicly available databases will therefore be pivotal in advancing the value of knowledge graphs in personalized oncology applications.

Materials and Methods

Data collection from primary sources

Cancer-specific data was collected from the SQL implementation of the Genomic Data Commons (GDC) database version 22.0 previously described in **Chapter 5** and publicly available^{24,49}. This included cancer patient IDs (*case_id_id*) linked to their tumor's primary site where the cancer started developing. Through the manuscript, we refer to primary sites as cancer types for simplicity. Somatic mutations leading to amino acid substitutions were also collected from this dataset and linked to the patient in which they occurred. Finally, results from the differential expression analysis in cancer vs. normal samples conducted per cancer type in this dataset were also extracted for all available genes. All GDC-derived data was extracted with HUGO Gene Nomenclature Committee (HGNC) gene symbols. For reference, mutations were also annotated based on their inclusion in the natural variance dataset 1000 Genomes, previously described in **Chapter 5**.

A PPI phosphorylation network of the kinome was collected from the work of Olow *et al.*⁴² Nodes in the PPI network represent protein-coding genes of kinases and their substrates, identified by their HGNC symbols. The network was reconstructed using the edges file, and the nodes file was used to get node attributes, in particular the node subtype ("kinase", "substrate", or "both"). Nodes with subtype "both" were annotated as "kinase".

Bioactivity differences between mutants and wild-type proteins were computed as described in **Chapter 4** for the dataset of mutants annotated in ChEMBL 31 and the Papyrus dataset (version 5.5). In particular, the Wasserstein distance was calculated

between the bioactivity distributions of all annotated mutants and the wild-type protein. ChEMBL 31 was also used to extract drugs in all phases of development (0: preclinical, 1-3: clinical, 4: approved) linked via their mechanism of action to particular proteins and mutations. All proteins were represented by their UniProt accession codes.

The structural location of mutations was assessed twofold. The first approach was protein family-independent and entailed the calculation of the average distance between the mutated residues' centroid and the co-crystallized ligand's centroid across available structures for the protein in the PDB. All proteins were represented by their UniProt accession codes. This method was explained in detail in **Chapter 4**. The second approach was kinome-specific and entailed querying the KLIFS database⁴⁵. Through the API, the protein-coding gene's HGNC symbol was linked to all available structures in the database. These structures were used to query the 85-residue KLIFS binding pocket and the residues forming it. Finally, a consensus KLIFS binding pocket was defined for each queried gene by selecting the most representative residue numbers for each aligned position in the pocket.

Amino acid substitutions were annotated with their corresponding Epstein coefficient of difference⁶⁴ and Grantham's distance⁶⁵, as defined in **Chapter 4**. The Epstein coefficient of difference is directional and was therefore annotated for each individual substitution. Grantham's distance, on the other hand, is non-directional and was therefore linked to each absolute amino acid change independently of its direction. The Blosum62 score⁶⁶ was also included as a metric to define the likelihood of an amino acid substitution to happen more or less frequently than random change, based on evolutionary conservation.

Ontology mapping and protein family annotation

The complete dataset from The Human Protein Atlas resource⁷⁹ was downloaded to facilitate protein family filtering and ontology mapping between proteins (UniProt accession codes) and genes (HGNC gene symbols). The downloaded tab-separated file contained a subset of the data from version 23.0 of the resource corresponding to the fields available in the data portal. Kinases were annotated by selecting entries containing the term "Kinase" in their *Molecular function* field. Receptor tyrosine kinases (RTKs) were annotated as kinases additionally containing the term "Receptor" in their *Molecular function* field. Membrane proteins were annotated as those containing the term "Plasma membrane" in their *Subcellular main location* field. Substrates in the PPI phosphorylation network were further annotated as members of two membrane protein families of interest, G protein-coupled receptors (GPCRs) and solute carriers (SLCs). The former were filtered when the field *Protein class* contained the term "G-protein coupled". The latter was defined when the field *Gene* started with "SLC".

Graph building and interactive exploration

The knowledge graph constructed contained gene and mutation nodes. Gene nodes

were all nodes from the PPI phosphorylation network, including both kinases and substrates and kinases with somatic mutations in the GDC dataset. Edges between gene nodes were directly derived from the PPI network. Mutation nodes were kinase somatic mutations in the GDC dataset and kinase mutations with bioactivity annotations obtained from the dataset constructed in **Chapter 4**. Edges between mutations represented co-occurrence in the same patient of the GDC dataset. Mutation and gene nodes were connected by edges representing the association between mutations happening in a specific gene. Cancer mutations were additionally linked to their gene with an edge representing the patient where the mutation occurs. Gene nodes (both kinases and substrates) were annotated with differential expression data from GDC for all available cancer types. Mutation nodes were annotated with bioactivity and structural data when available. Structural data was annotated based on the mutation residue. All mutations were annotated with the Epstein coefficient of difference, based on the amino acid substitution. All nodes (genes and mutations) were annotated with the number of drugs in different phases of development that were associated with the protein encoded by the gene or the mutation in their mechanism of action. Gene nodes were further annotated as part of certain protein families of interest (membrane proteins, RTKs, GPCRs, SLCs). These data were saved in nodes and edges files.

NetworkX⁵² was used to build a multigraph in Python. The graph was built from the edges file and the nodes were annotated with their attributes using the nodes file. A package was constructed in Python to enable modular build, storage, and updates of the graph. To this end, a graph metadata file is created every time a graph is initialized with a new combination of graph name and edges/nodes files. The package also contains modules to facilitate visual interactive exploration of the data in the knowledge graph based on the graph visualization python package IPySigma⁷⁵.

Network analysis

Network analysis algorithms implemented in NetworkX were used to explore the knowledge graph and pinpoint relevant nodes. The network node metrics calculated were degree, degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and information centrality. The Python package built for this project enabled the calculation of these metrics for the complete network but also for subsets of it. Precomputed subsets included those containing only one type of edge: cancer co-occurrence, phosphorylation, and gene-mutation association. This distinction enabled the identification of key nodes in each of those layers of information. Additionally, each cancer type was analyzed independently by computing network subsets containing only edges corresponding to patients in a specific cancer type and the phosphorylation edges linking the subset genes. The package analysis modules not only enable the separate examination of these and other subsets but also facilitate the integration of the subset analysis results into the entire knowledge graph. The computed metrics can be further explored interactively.

References

1. Naga, D., Muster, W., Musvasva, E. & Ecker, G. F. Off-targetP ML: an open source machine learning framework for off-target panel safety assessment of small molecules. *Journal of Cheminformatics* **14**, 27 (2022).
2. Kalsekar, I., Koehler, J. & Mulvaney, J. Impact of ACE inhibitors on mortality and morbidity in patients with AMI: Does tissue selectivity matter? *Value in Health* **14**, 184–191 (2011).
3. Montoya, S. *et al.* Targeted Therapies in Cancer: To Be or Not to Be, Selective. *Biomedicines* **9**, 1591 (2021).
4. Vlot, A. H. C. *et al.* Target and Tissue Selectivity Prediction by Integrated Mechanistic Pharmacokinetic-Target Binding and Quantitative Structure Activity Modeling. *AAPS J* **20**, 11 (2017).
5. Blagosklonny, M. V. Selective protection of normal cells from chemotherapy, while killing drug-resistant cancer cells. *Oncotarget* **14**, 193–206 (2023).
6. Zhao, Z., Ukidve, A., Kim, J. & Mitragotri, S. Targeting Strategies for Tissue-Specific Drug Delivery. *Cell* **181**, 151–167 (2020).
7. Lassen, U. N. *et al.* Precision oncology: a clinical and patient perspective. *Future Oncology* **17**, 3995–4009 (2021).
8. Pessoa, J., Martins, M., Casimiro, S., Pérez-Plasencia, C. & Shoshan-Barmatz, V. Editorial: Altered Expression of Proteins in Cancer: Function and Potential Therapeutic Targets. *Front Oncol* **12**, 949139 (2022).
9. Waarts, M. R., Stonestrom, A. J., Park, Y. C. & Levine, R. L. Targeting mutations in cancer. *J Clin Invest* **132**, e154943 (2022).
10. Bhullar, K. S. *et al.* Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular Cancer* **17**, 48 (2018).
11. Liu, G., Chen, T., Zhang, X., Ma, X. & Shi, H. Small molecule inhibitors targeting the cancers. *MedComm* **3**, e181 (2022).
12. Liu, X. *et al.* Cryo-EM structures of cancer-specific helical and kinase domain mutations of PI3K α . *Proceedings of the National Academy of Sciences* **119**, e2215621119 (2022).
13. Dixit, A. *et al.* Sequence and Structure Signatures of Cancer Mutation Hotspots in Protein Kinases. *PLOS ONE* **4**, e7485 (2009).
14. Kong, Y. *et al.* Small Molecule Inhibitors as Therapeutic Agents Targeting Oncogenic Fusion Proteins: Current Status and Clinical. *Molecules* **28**, 4672 (2023).
15. Miller, M. S. *et al.* Identification of allosteric binding sites for PI3K α oncogenic mutant specific inhibitor design. *Bioorganic & Medicinal Chemistry* **25**, 1481–1486 (2017).
16. Kim, P., Zhao, J., Lu, P. & Zhao, Z. mutLBSgeneDB: mutated ligand binding site gene DataBase. *Nucleic Acids Res* **45**, D256–D263 (2017).
17. Zubair, T. & Bandyopadhyay, D. Small Molecule EGFR Inhibitors as Anti-Cancer Agents: Discovery, Mechanisms of Action, and Opportunities. *International Journal of Molecular Sciences* **24**, 2651 (2023).
18. Varkaris, A. *et al.* Allosteric PI3K α Inhibition Overcomes On-target Resistance to Orthosteric Inhibitors Mediated by Secondary PIK3CA Mutations. *Cancer Discovery* **14**, 227–239 (2024).
19. Dupont, C. A., Riegel, K., Pampaiah, M., Juhl, H. & Rajalingam, K. Druggable genome and precision medicine in cancer: current challenges. *The FEBS Journal* **288**, 6142–6158 (2021).
20. Radoux, C. J., Vianello, F., McCreig, J., Desai, N. & Bradley, A. R. The druggable genome: Twenty years later. *Front. Bioinform.* **2**, 958378 (2022).
21. Gorostiola González, M., Janssen, A. P. A., IJzerman, A. P., Heitman, L. H. & van Westen, G. J. P. Oncological drug discovery: AI meets structure-based computational research. *Drug Discovery Today* **27**, 1661–1670 (2022).
22. Pacini, C. *et al.* A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* **42**, 301–316 (2024).
23. Singha, M. *et al.* Unlocking the Potential of Kinase Targets in Cancer: Insights from CancerOmicsNet, an AI-Driven Approach to Drug Response Prediction in Cancer. *Cancers (Basel)* **15**, 4050 (2023).
24. Bongers, B. J. *et al.* Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors. *Scientific Reports* **12**, 21534 (2022).
25. Zhao, J., Cheng, F., Wang, Y., Arteaga, C. L. & Zhao, Z. Systematic Prioritization of Druggable Mutations in ~5000 Genomes Across 16 Cancer Types Using a Structural Genomics-based Approach. *Molecular & Cellular Proteomics* **15**, 642–656 (2016).
26. Silva, M. C., Eugénio, P., Faria, D. & Pesquita, C. Ontologies and Knowledge Graphs in Oncology Research. *Cancers (Basel)* **14**, 1906 (2022).
27. Zhang, W., Chien, J., Yong, J. & Kuang, R. Network-based machine learning and graph theory algorithms for precision oncology. *npj Precision Oncology* **1**, 25 (2017).
28. Cesareni, G., Sacco, F. & Perfetto, L. Assembling Disease Networks From Causal Interaction Resources. *Front. Genet.* **12**, 694468 (2021).
29. Gogleva, A. *et al.* Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat Commun* **13**, 1667 (2022).

30. Wang, Y. & Liu, Z.-P. Identifying biomarkers for breast cancer by gene regulatory network rewiring. *BMC Bioinformatics* **22**, 308 (2022).
31. Zhao, L. *et al.* Biological knowledge graph-guided investigation of immune therapy response in cancer with graph neural network. *Briefings in Bioinformatics* **24**, bbad023 (2023).
32. Pu, L. *et al.* An integrated network representation of multiple cancer-specific data for graph-based machine learning. *npj Syst Biol Appl* **8**, 1–8 (2022).
33. Niu, R., Guo, Y. & Shang, X. GLIMS: A two-stage gradual-learning method for cancer genes prediction using multi-omics data and co-splicing network. *iScience* **27**, 109387 (2024).
34. Hatano, N., Kamada, M., Kojima, R. & Okuno, Y. Network-based prediction approach for cancer-specific driver missense mutations using a graph neural network. *BMC Bioinformatics* **24**, 383 (2023).
35. Li, B. & Nabavi, S. A multimodal graph neural network framework for cancer molecular subtype classification. *BMC Bioinformatics* **25**, 27 (2024).
36. Quan, X., Cai, W., Xi, C., Wang, C. & Yan, L. AIMedGraph: a comprehensive multi-relational knowledge graph for precision medicine. *Database* **2023**, baad006 (2023).
37. Bang, D., Lim, S., Lee, S. & Kim, S. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nat Commun* **14**, 3570 (2023).
38. Renaux, A. *et al.* A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC Bioinformatics* **24**, 324 (2023).
39. Maghawry, N., Ghoniemy, S., Shaaban, E. & Emara, K. An Automatic Generation of Heterogeneous Knowledge Graph for Global Disease Support: A Demonstration of a Cancer Use Case. *Big Data and Cognitive Computing* **7**, 21 (2023).
40. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci Data* **10**, 67 (2023).
41. Jin, S., Liang, H., Zhang, W. & Li, H. Knowledge Graph for Breast Cancer Prevention and Treatment: Literature-Based Data Analysis Study. *JMIR Medical Informatics* **12**, e52210 (2024).
42. Olow, A. *et al.* An atlas of the human kinome reveals the mutational landscape underlying dysregulated phosphorylation cascades in cancer. *Cancer Research* **76**, 1733–1745 (2016).
43. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
44. Burley, S. K. *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* **47**, D464–D474 (2019).
45. Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. P. & Kooistra, A. J. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res* **49**, D562–D569 (2021).
46. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
47. Danesi, R. *et al.* Druggable targets meet oncogenic drivers: opportunities and limitations of target-based classification of tumors and the role of Molecular Tumor Boards. *ESMO Open* **6**, 100040 (2021).
48. Yang, Y., Li, S., Wang, Y., Zhao, Y. & Li, Q. Protein tyrosine kinase inhibitor resistance in malignant tumors: molecular mechanisms and future perspective. *Sig Transduct Target Ther* **7**, 1–36 (2022).
49. Bongers, B. *et al.* Data underlying the article: Pan-cancer in silico analysis of somatic mutations in G-protein coupled receptors: The effect of evolutionary conservation and natural variance. Available at <https://doi.org/10.4121/15022410.V1> (2021).
50. Geleta, D. *et al.* Biological Insights Knowledge Graph: an integrated knowledge graph to support drug development. Preprint at *BioRxiv* <https://doi.org/10.1101/2021.10.28.466262> (2021).
51. Rozemberczki, B. *et al.* MOOMIN: Deep Molecular Omics Network for Anti-Cancer Drug Combination Therapy. Preprint at *ArXiv* <https://doi.org/10.48550/arXiv.2110.15087> (2022).
52. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in *Proceedings of the 7th Python in Science Conference* (2008).
53. Karim, M. R. *et al.* From Large Language Models to Knowledge Graphs for Biomarker Discovery in Cancer. Preprint at *ArXiv* <https://doi.org/10.48550/arXiv.2310.08365> (2023).
54. Pan, J. Z. *et al.* Large Language Models and Knowledge Graphs: Opportunities and Challenges. Preprint at *ArXiv* <http://arxiv.org/abs/2308.06374> (2023).
55. Zhang, N. *et al.* Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLoS Computational Biology* **11**, e1004498 (2015).
56. Yang, Z. *et al.* A mutation-induced drug resistance database (MdrDB). *Commun Chem* **6**, 1–9 (2023).
57. Sha, D. *et al.* Tumor Mutational Burden (TMB) as a Predictive Biomarker in Solid Tumors. *Cancer Discov* **10**, 1808–1825 (2020).
58. Oh, J.-H. *et al.* Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *npj Genom. Med.* **5**, 1–11 (2020).
59. Ibáñez, M. *et al.* The Mutational Landscape of Acute Promyelocytic Leukemia Reveals an Interacting Network of Co-Occurrences and

- Recurrent Mutations. *PLOS ONE* **11**, e0148346 (2016).
60. Zhong, X., Yang, H., Zhao, S., Shyr, Y. & Li, B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC Genomics* **16**, S7 (2015).
 61. Li, B., Wang, T. & Nabavi, S. Cancer molecular subtype classification by graph convolutional networks on multi-omics data. in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* 1–9 (2021)
 62. Guo, H., Lv, X., Li, Y. & Li, M. Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification. *Brief Funct Genomics* **22**, 463–474 (2023).
 63. Tanvir, R. B., Islam, M. M., Sobhan, M., Luo, D. & Mondal, A. M. MOGAT: A Multi-Omics Integration Framework Using Graph Attention Networks for Cancer Subtype Prediction. *International Journal of Molecular Sciences* **25**, 2788 (2024).
 64. Epstein, C. J. Non-randomness of Amino-acid Changes in the Evolution of Homologous Proteins. *Nature* **215**, 355–359 (1967).
 65. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
 66. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915–10919 (1992).
 67. Martínez-Sáez, O. *et al.* Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Research* **22**, 45 (2020).
 68. Boyer, S., Money-Kyrle, S. & Bent, O. Predicting protein stability changes under multiple amino acid substitutions using equivariant graph neural networks. Preprint at *ArXiv* <http://arxiv.org/abs/2305.19801> (2023).
 69. Robichaux, J. P. *et al.* Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* **597**, 732–737 (2021).
 70. Van Linden, O. P. J., Kooistra, A. J., Leurs, R., De Esch, I. J. P. & De Graaf, C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space. *Journal of Medicinal Chemistry* **57**, 249–277 (2013).
 71. Engin, H. B., Kreisberg, J. F. & Carter, H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLOS ONE* **11**, e0152929 (2016).
 72. Pujara, J., Augustine, E. & Getoor, L. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* 1751–1756 (2017).
 73. Xu, J., Zhang, W., Duan, Q. & Li, S. HOAP: Node attribute completion of knowledge graph based on high-order neighbor attribute propagation. Preprint at *Research Square* <https://doi.org/10.21203/rs.3.rs-1937079/v1> (2022).
 74. Guo, D., Chu, Z. & Li, S. Fair Attribute Completion on Graph with Missing Attributes. Preprint at *ArXiv* <http://arxiv.org/abs/2302.12977> (2023).
 75. Plique, G. ipysigma. Available at Zenodo <https://doi.org/10.5281/zenodo.7521476> (2023).
 76. Boutros, A. *et al.* Activity and safety of first-line treatments for advanced melanoma: A network meta-analysis. *European Journal of Cancer* **188**, 64–79 (2023).
 77. Bosdriesz, E. *et al.* Identifying mutant-specific multi-drug combinations using comparative network reconstruction. *iScience* **25**, 104760 (2022).
 78. Hu, C. *et al.* Optimizing drug combination and mechanism analysis based on risk pathway crosstalk in pan cancer. *Sci Data* **11**, 74 (2024).
 79. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

Supplementary Information

Supplementary Table 8.1. Top 25 cancer mutation nodes with the highest mutation frequency (pan-cancer) in the kinome knowledge graph.

Cancer mutation	Frequency
BRAF_V600E	565
PIK3CA_E545K	258
PIK3CA_H1047R	234
PIK3CA_E542K	167
PIK3CA_R88Q	68
AKT1_E17K	53
BRAF_V600M	40
FGFR3_S249C	39
ERBB2_S310F	38
PIK3CA_H1047L	37
PIK3CA_N345K	34
PIK3CA_E726K	30
PIK3CA_G118D	28
FGFR2_S252W	26
ERBB3_V104M	25
EGFR_L858R	23
PIK3CA_C420R	23
PIK3CA_Q546R	22
EGFR_A289V	21
PIK3CA_E453K	19
EGFR_G598V	19
PIK3CA_R108H	19
PIK3CA_E545A	18
PIK3CA_M1043I	18
MAPK1_E322K	18

Supplementary Table 8.2. Top 25 ranked cancer mutation nodes in the kinome knowledge graph according to their degree. For reference, the mutation frequency across all cancer patients analyzed is reported, and the rank that would correspond to the cancer mutation if it was calculated based on the mutation frequency, if this rank is 1-25 (otherwise reported as >25). Additionally, the number of unique mutations reported for the gene is reported. *Mutations present in the natural variance dataset 1000 Genomes.

Cancer mutation	Degree	Rank	Mutation frequency	Frequency rank	Gene cancer mutations
PIK3CA_R88Q	10275	1	68	5	268
BRAF_V600E	6219	2	565	1	132
PRKDC_R2522Q	3376	3	11	>25	521
MTOR_R2152C	3042	4	4	>25	345
NEK3_S284L*	2898	5	6	>25	52
PAK5_E144K	2705	6	7	>25	203
GCK_A2V	2661	7	4	>25	79
PIK3CA_E545K	2484	8	258	2	268
PDK2_A259V	2483	9	3	>25	40
PIK3CA_H1047R	2439	10	234	3	268
TTN_R2506Q	2409	11	8	>25	7791
CAMK1D_S360L	2352	12	6	>25	74
TEK_S599L	2321	13	4	>25	185
DCAF1_R855Q	2262	14	5	>25	153
MAP3K15_R493W	2191	15	7	>25	199
TTN_D19391N	2175	16	8	>25	7791
ROCK1_R1012Q	2159	17	5	>25	186
SMG1_R803H	2127	18	3	>25	386
HIPK1_R875H	2123	19	5	>25	148
ROCK1_R590Q	2118	20	5	>25	186
STK3_S344L	2089	21	5	>25	963
ROCK2_R339Q	2071	22	4	>25	153
TTN_R33466C*	2064	23	4	>25	7791
DGKB_R685Q*	2059	24	3	>25	217
IP6K1_R329H	2054	25	4	>25	45

Supplementary Table 8.3. Top 25 most frequently mutated proteins per cancer type (as defined by primary site). Mutation frequency is calculated as the sum of all mutations in that protein-cancer type pair.

Protein	Cancer type	Mutation frequency
TTN	Skin	1,931
TTN	Corpus uteri	1,808
TTN	Bronchus and lung	1,402
TTN	Colon	581
TTN	Stomach	553
PIK3CA	Corpus uteri	367
PIK3CA	Breast	357
OBSCN	Corpus uteri	334
TTN	Bladder	328
TTN	Brain	321
BRAF	Thyroid gland	290
BRAF	Skin	290
TTN	Breast	277
TTN	Cervix uteri	207
TTN	Ovary	202
OBSCN	Skin	195
TTN	Rectum	192
SMG1	Corpus uteri	173
LRRK2	Corpus uteri	170
TAF1	Corpus uteri	168
PRKDC	Corpus uteri	161
MYO3A	Corpus uteri	157
OBSCN	Bronchus and lung	150
OBSCN	Colon	147
EGFR	Brain	146

Supplementary Table 8.4. Distribution of node and edge types and subtypes across the receptor tyrosine kinase knowledge graph.

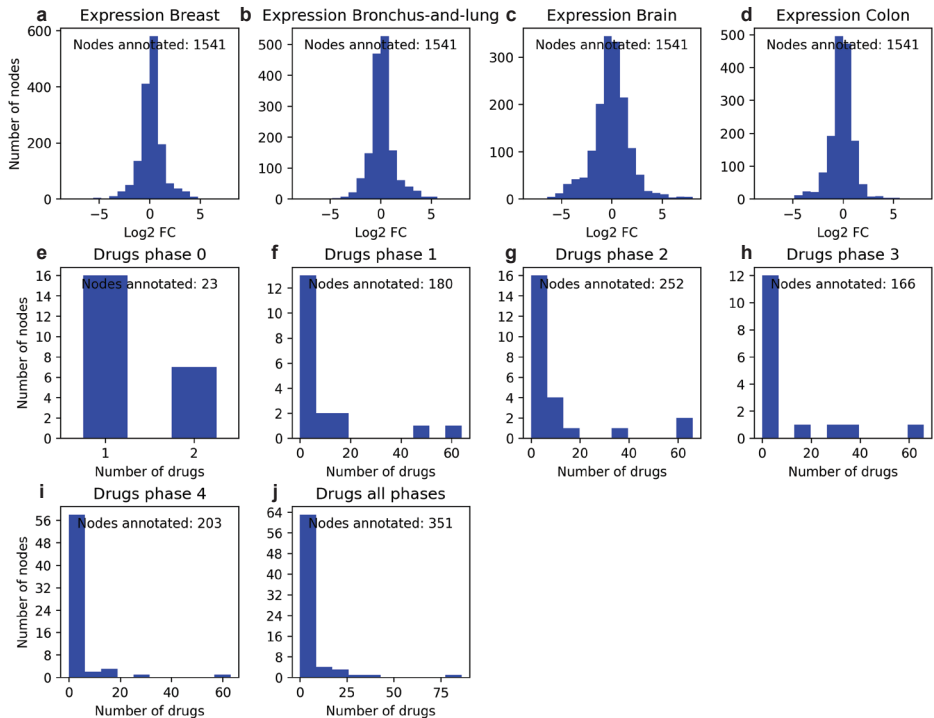
Entity	Type	Subtype	Number of entities
Nodes	Gene	Kinase	110
		Receptor kinase (not in PPI as kinase)	1
		Substrate	142
	Mutation	Cancer	11,660
		Other (ChEMBL + Papyrus)	76
Edges	Gene - Gene	Phosphorylation	673
	Gene - Mutation	Cancer	13,353
		Other (ChEMBL + Papyrus)	76
	Mutation - Mutation	Cancer patient co-occurrence	127,187
		Other (ChEMBL + Papyrus multiple substitutions)	22

Supplementary Table 8.5. Top 25 cancer mutation nodes with the highest mutation frequency (pan-cancer) in the receptor tyrosine kinase knowledge graph.

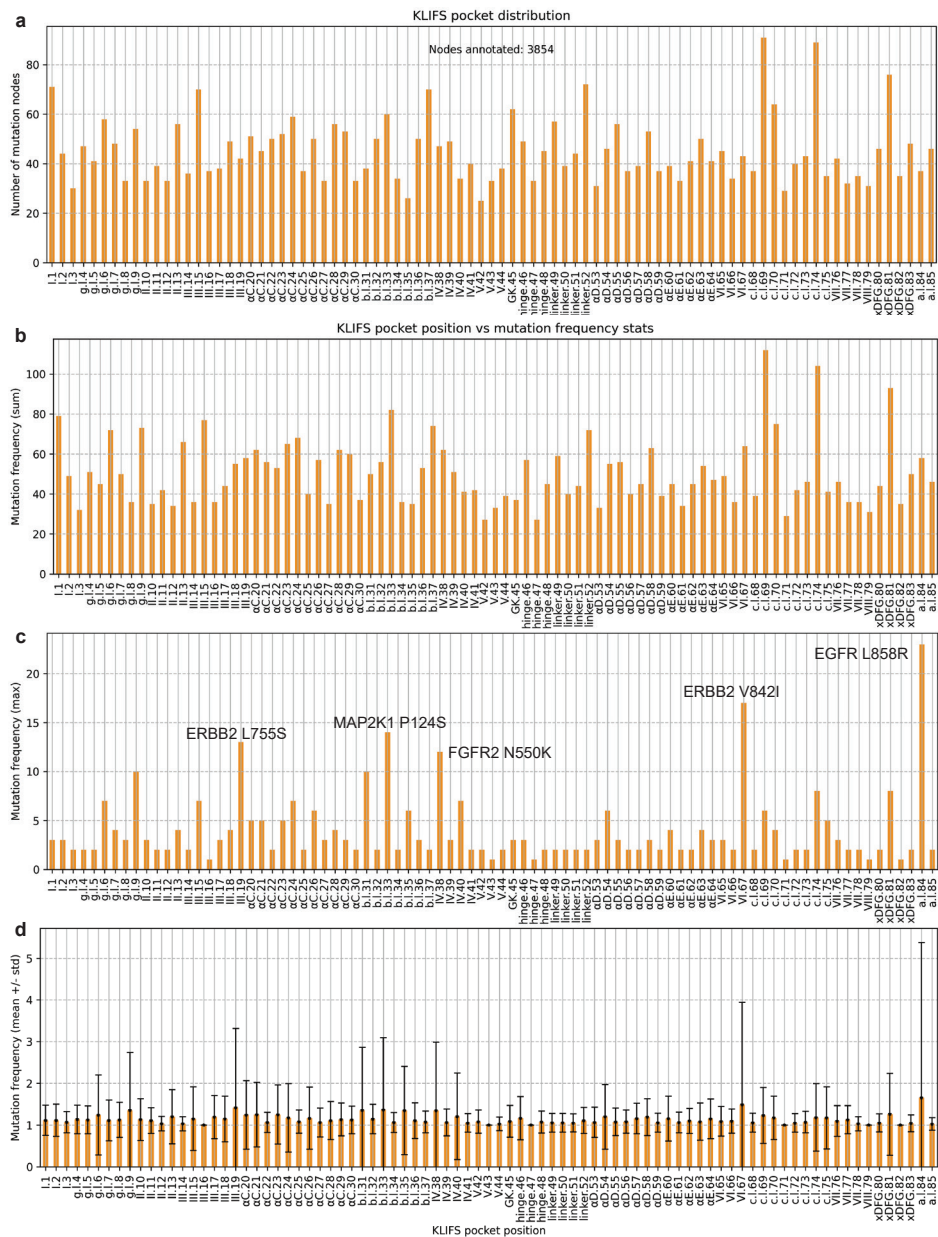
Cancer mutation	Frequency
FGFR3_S249C	39
ERBB2_S310F	38
FGFR2_S252W	26
ERBB3_V104M	25
EGFR_L858R	23
EGFR_A289V	21
EGFR_G598V	19
ERBB2_V842I	17
ERBB2_R678Q	14
ERBB2_L755S	13
FGFR2_N550K	12
FGFR3_Y375C	10
ERBB2_V777L	10
EPHA6_R268C	8
KIT_D816V	8
FLT3_D835Y	8
KDR_R1032Q	8
EGFR_L62R	7
EGFR_R222C	7
FGFR2_C383R	7
MUSK_R854Q	7
ERBB4_R711C	7
EGFR_L861Q	6
KIT_K642E	6
FGFR1_N577K	6

Supplementary Table 8.6. Top 25 ranked cancer mutation nodes in the receptor tyrosine kinase knowledge graph according to their degree. For reference, the mutation frequency across all cancer patients analyzed is reported, and the rank that would correspond to the cancer mutation if it was calculated based on the mutation frequency, if this rank is 1-25 (otherwise reported as >25). Additionally, the number of unique mutations reported for the gene is reported. *Mutations present in the natural variance dataset 1000 Genomes.

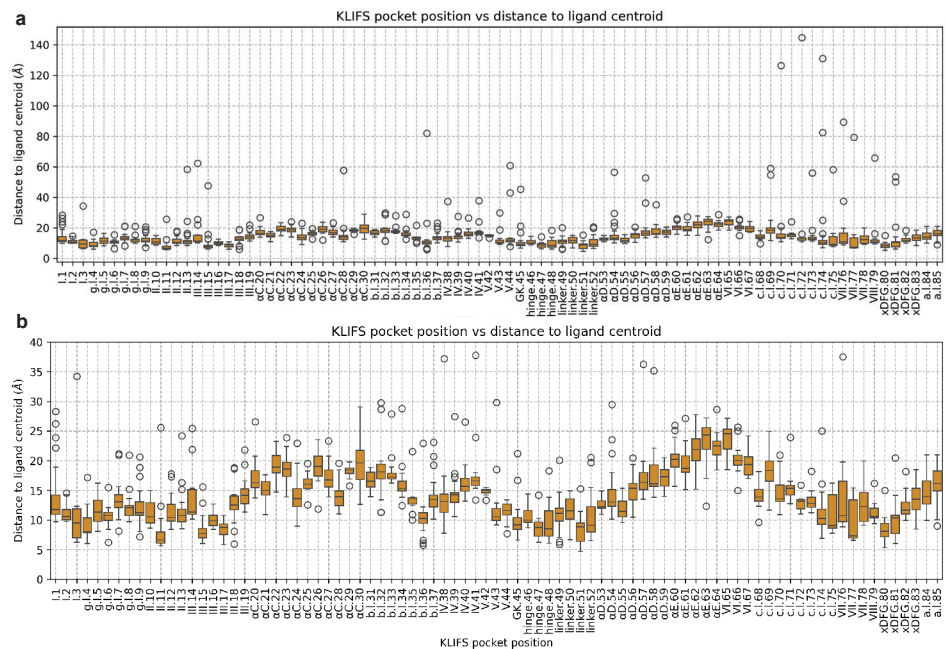
Cancer mutation	Degree	Rank	Mutation frequency	Frequency rank	Gene cancer mutations
TEK_S599L	391	1	4	>25	185
EPHA10_D881N	305	2	2	>25	140
KDR_R1032Q	300	3	8	17	331
KIT_R888Q	274	4	4	>25	214
EPHA6_D243N	263	5	6	>25	336
PDGFRA_E156D	252	6	2	>25	293
DDR1_D714N	246	7	2	>25	112
ERBB3_R916Q*	237	8	2	>25	229
FGFR2_R165W	232	9	2	>25	166
EPHA4_R745H	232	10	2	>25	192
CSF1R_D565N	229	11	2	>25	138
INSR_R924Q	226	12	2	>25	201
EGFR_R977H	217	13	2	>25	265
EPHA6_R788C	209	14	2	>25	336
ACVR1C_R245Q	207	15	3	>25	91
EPHA2_E523K	203	16	2	>25	171
FLT1_E144K*	201	17	3	>25	264
PDGFRA_K196N	200	18	2	>25	293
EPHA8_A685T	199	19	2	>25	185
MET_R412C	198	20	2	>25	214
RYK_R563Q	195	21	4	>25	64
MET_L982M	195	22	2	>25	214
EPHB1_R743W	194	23	2	>25	286
EPHA1_R261W	194	24	3	>25	133
MUSK_R572K	192	25	2	>25	174



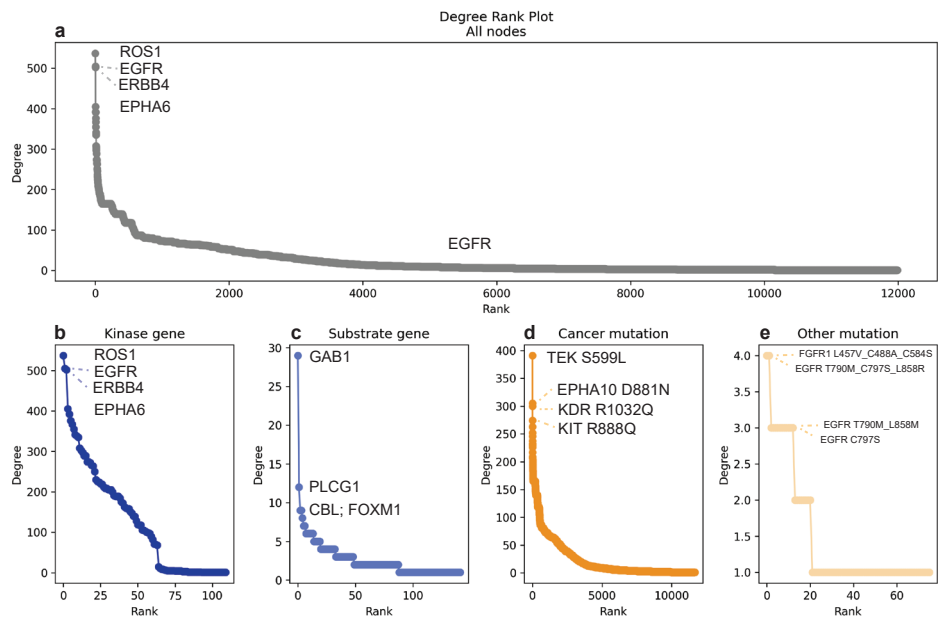
Supplementary Figure 8.1. Density and distribution of gene node attributes. For each of the attributes, the number of gene nodes with non-null values for the particular attribute is depicted on the y-axis. Moreover, the graphs represent the distribution of the attribute values across the gene nodes on the x-axis, in the form of histograms or bar plots, depending on the density of each attribute. Four cancer types are selected as an example to show the distribution of differential expression Log2 fold change (Log2FC) between the tumor tissue and normal tissue (a-d). The rest of the graphs represent the number of drugs in different phases of development according to ChEMBL labeling: pre-clinical “0” phase (e), clinical phases 1-3 (f-h), and approved drugs (i). The total number of drugs in any phase is represented in (j).



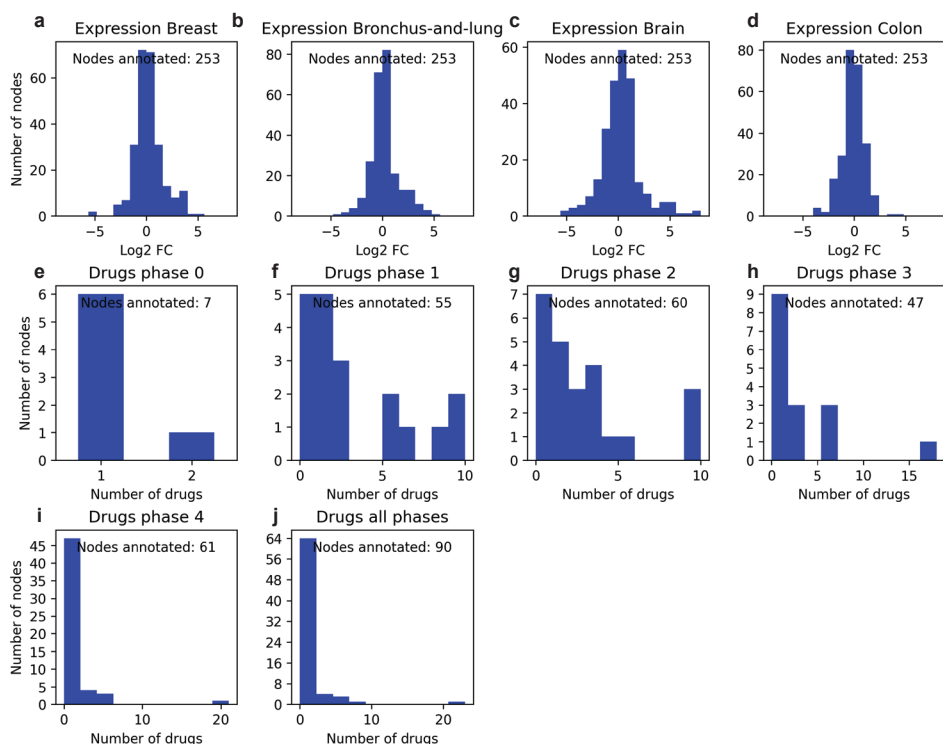
Supplementary Figure 8.2. Density and distribution of KLIFS structural attribute annotations in mutation nodes. **a)** Number of mutation nodes with an annotation for each particular position of the 85-consensus kinase pocket defined by KLIFS. **b-d)** Cancer mutation frequency statistics for each pocket position: sum of mutation frequency in each position (**b**), maximum mutation frequency reported for each position, with the mutations with the top five frequencies labeled (**c**), and mean +/- standard deviation of mutation frequency reported for each pocket position (**d**).



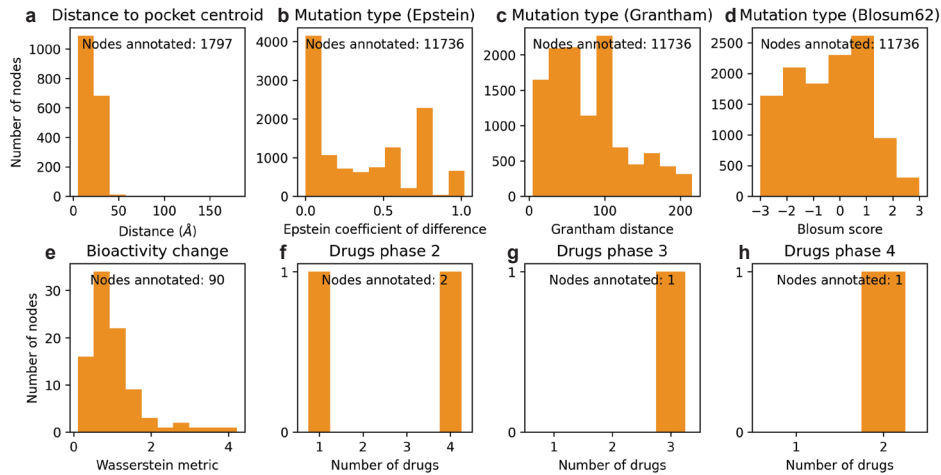
Supplementary Figure 8.3. Average distance to ligand centroid for every KLIFS pocket position calculated for each mutation in that position from available PDB complexes. One distance value is recorded per available mutation in the kinome graph. **a)** Complete distribution. **b)** Distribution of values in the range 0-40 Å.



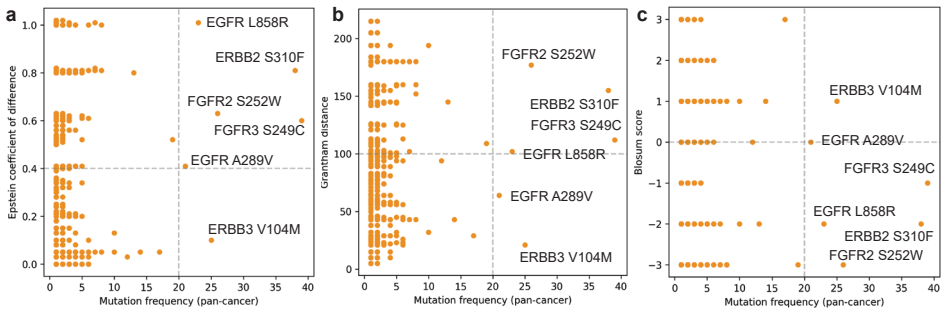
Supplementary Figure 8.4. Node degree rank analysis for the receptor tyrosine kinase knowledge graph. Nodes are ranked based on their degree, which is calculated as the number of edges connecting the node to other nodes in the graph. The degree rank analysis is calculated for all nodes in the graph (a) as well as for each node subtype independently: kinase (b) and substrate (c) gene nodes, and cancer (d) and other (e) mutations. The top four ranked nodes in each case are labeled accordingly.



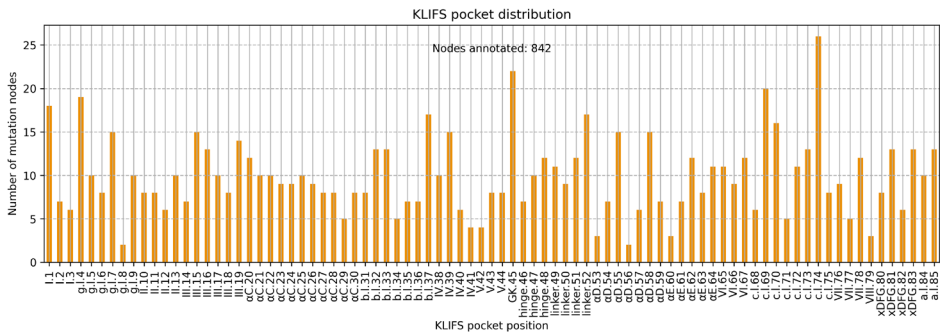
Supplementary Figure 8.5. Density and distribution of gene node attributes in the receptor tyrosine kinase knowledge graph. For each of the attributes, the number of gene nodes with non-null values for the particular attribute is depicted on the y-axis. Moreover, the graphs represent the distribution of the attribute values across the gene nodes on the x-axis, in the form of histograms or bar plots, depending on the density of each attribute. Four cancer types are selected as an example to show the distribution of differential expression Log2 fold change (Log2FC) between the tumor tissue and normal tissue (**a-d**). The rest of the graphs represent the number of drugs in different phases of development according to ChEMBL labeling: pre-clinical “0” phase (**e**), clinical phases 1-3 (**f-h**), and approved drugs (**i**). The total number of drugs in any phase is represented in (**j**).



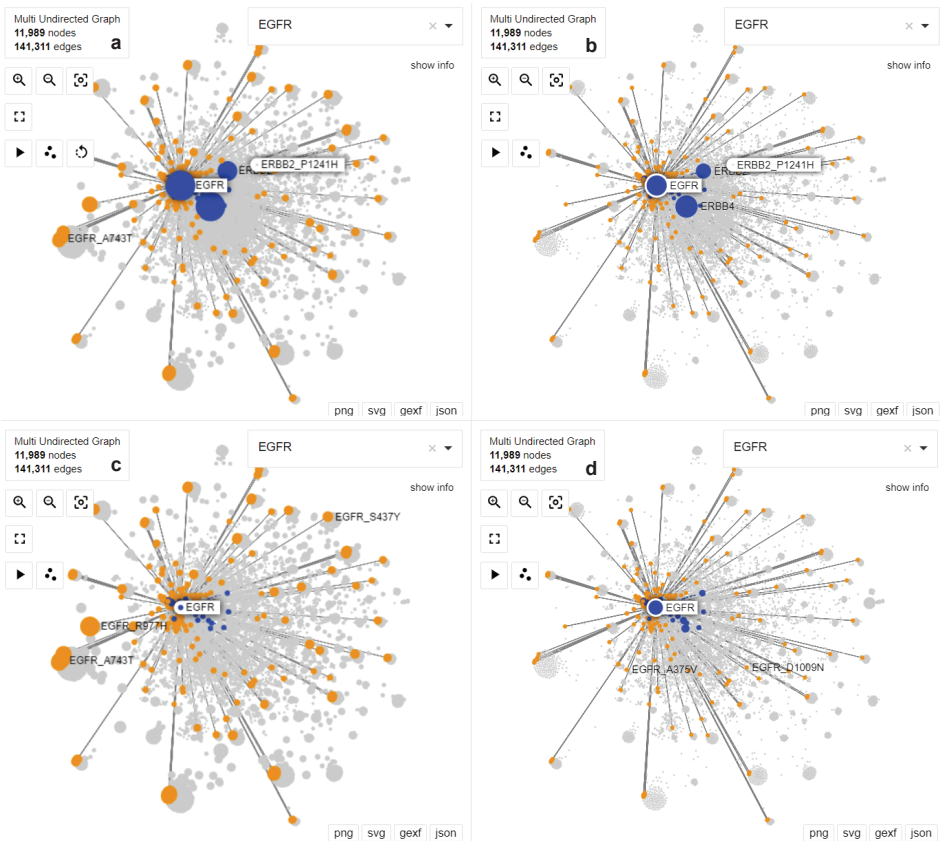
Supplementary Figure 8.6. Density and distribution of mutation node attributes in the receptor tyrosine kinase knowledge graph. For each of the attributes, the number of mutation nodes with non-null values for the particular attribute is depicted on the y-axis. Moreover, the graphs represent the distribution of the attribute values across the mutation nodes on the x-axis, in the form of histograms or bar plots, depending on the density of each attribute. Distribution is represented as histograms for the distance to the pocket centroid calculated from PDB complexes (a), the mutation type as determined by the Epstein coefficient of difference (b), the mutation type as described by the Grantham distance (c), the evolutionary probability of the mutation type as described by the Blosum score in the Blosum62 matrix (d), and the bioactivity change represented by the Wasserstein distance between the bioactivity distribution for the mutation and the wild-type protein found in ChEMBL (e). Bar plots represent the number of drugs in different phases of development according to ChEMBL labeling: clinical phases 2-3 (f-g), and approved drugs (h). Pre-clinical candidates (phase 0) and drugs in clinical phase 1 are not included because they were not annotated in any mutation nodes.



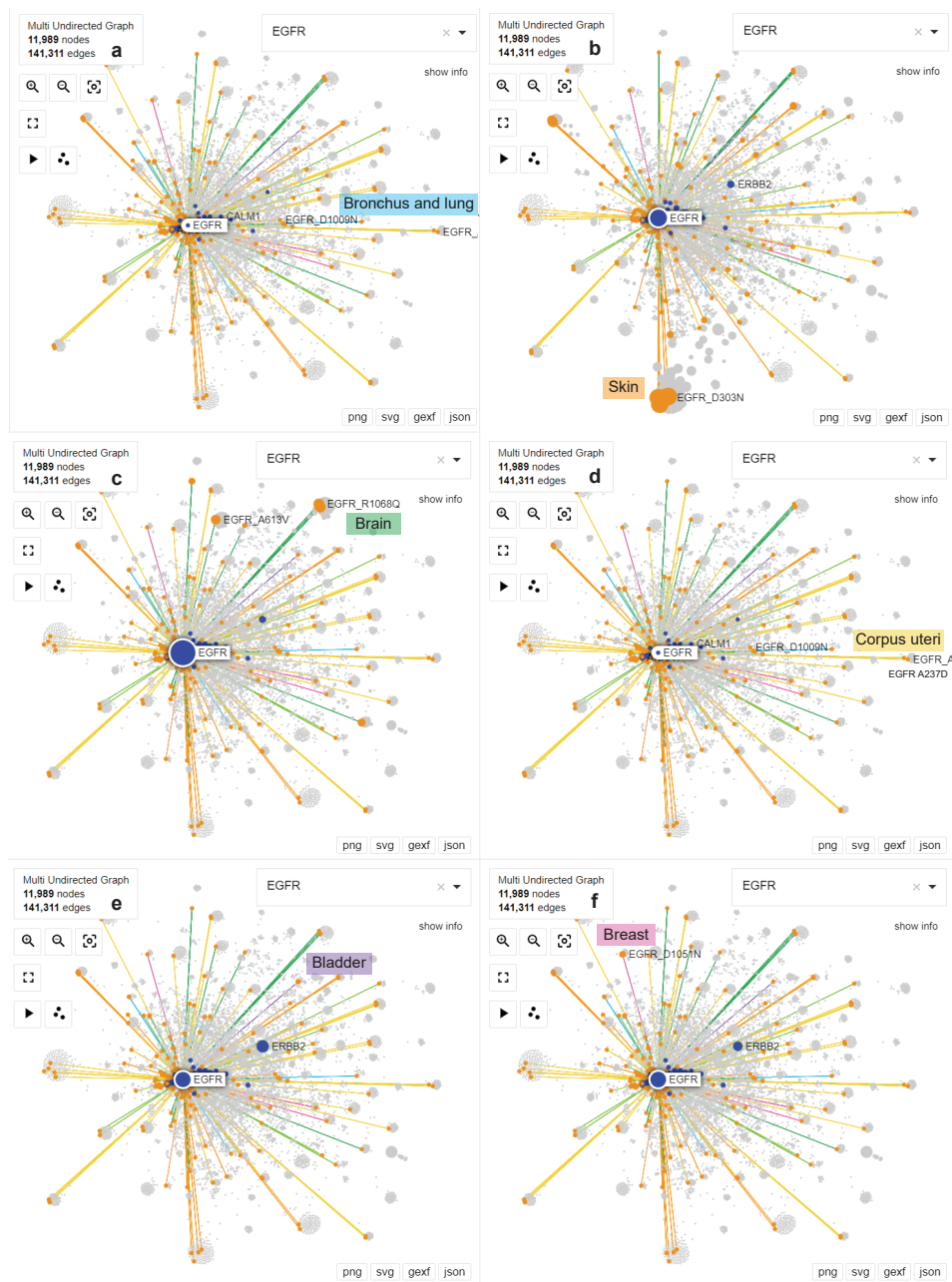
Supplementary Figure 8.7. Correlation between cancer mutation frequency and three metrics describing the amino acid substitution in the kinome knowledge graph: Epstein coefficient of difference (a), Grantham distance (b), and Blosum score (c). Mutations occurring in more than 20 patients pan-cancer are labeled for reference. In a), an Epstein coefficient of difference of 0.4 is taken as an arbitrary threshold to distinguish between conservative (<0.4) and disruptive substitutions (>0.4). In b), a Grantham distance of 100 is taken as an arbitrary threshold to distinguish between conservative (<100) and disruptive substitutions (>100). In c), substitutions with an alignment happening less often than random chance as collected in the Blosum62 matrix are represented by a Blosum score < 0.



Supplementary Figure 8.8. Number of mutation nodes in the receptor tyrosine kinase graph with an annotation for each particular position of the 85-consensus kinase pocket defined by KLIFS.

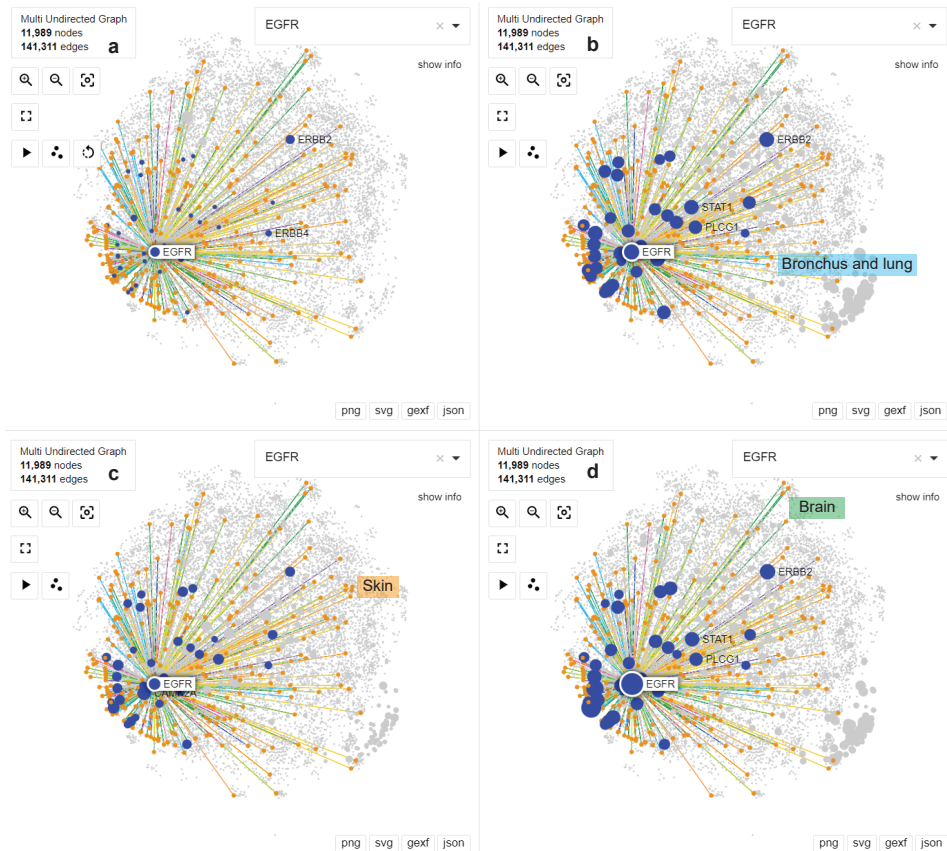


Supplementary Figure 8.9. Node betweenness centrality comparison across layers in the receptor tyrosine kinase knowledge graph, with a focus on nodes connected to EGFR for visualization purposes. Gene nodes are represented in blue and mutation nodes are represented in orange. Nodes that are not connected to EGFR are kept in the background and represented in grey. Node size in each panel is determined by the node degree calculated from the whole graph (a) or one of the three pre-defined analysis layers: kinase-mutation layer (b), cancer-mutation co-occurrence layer (c), or phosphorylation layer (d).



Supplementary Figure 8.10. Node degree comparison across the six most populated cancer types in the receptor tyrosine kinase knowledge graph, with a focus on nodes connected to EGFR for visualization purposes. Gene nodes are represented in blue and mutation nodes are represented in orange. Nodes that are not connected to EGFR are kept in the background and represented in grey. Each edge color represents a different cancer type. Node size in each panel is determined by the node degree calculated from the subgraphs for six cancer types: bronchus and lung (a, 723 patients – 56 with EGFR ▶

► mutations. Represented by blue edges), skin (**b**, 356 patients – 29 with EGFR mutations. Represented by orange edges), brain (**c**, 312 patients – 127 with EGFR mutations. Represented by green edges), corpus uteri (**d**, 302 patients – 29 with EGFR mutations. Represented by yellow edges), bladder (**e**, 289 patients – 7 with EGFR mutations. Represented by purple edges), or breast (**f**, 273 patients – 13 with EGFR mutations. Represented by pink edges).

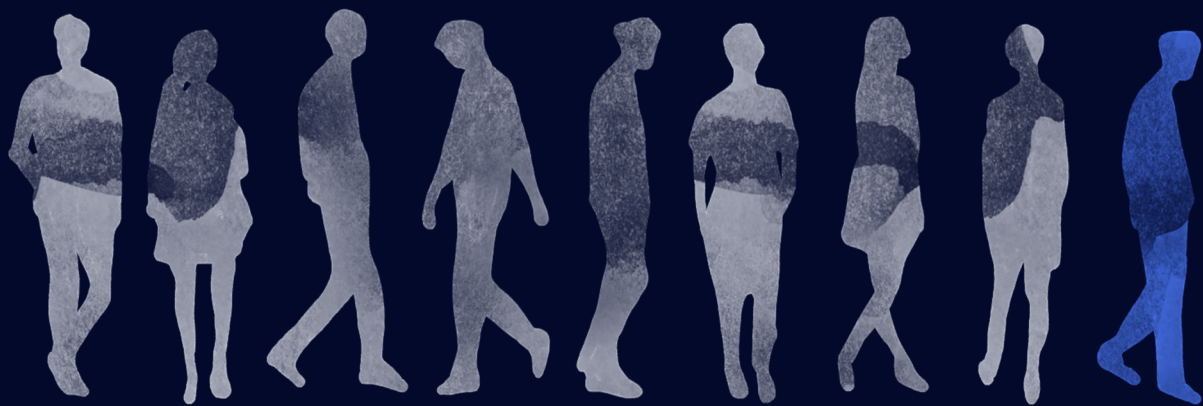


Supplementary Figure 8.11. Comparative visualization of the receptor tyrosine kinase knowledge graph with node sizes representing different attributes. The focus is on nodes connected to EGFR for visualization purposes. Gene nodes are represented in blue and mutation nodes are represented in orange. Nodes that are not connected to EGFR are kept in the background and represented in grey. Each edge color represents a different cancer type. Node size represents the number of approved drugs (**a**) or the differential expression (Log2 fold change) in tumor tissue compared to healthy tissue in the three most populated cancer types: bronchus and lung (**b**, 723 patients – 56 with EGFR mutations. Represented by blue edges), skin (**c**, 356 patients – 29 with EGFR mutations. Represented by orange edges), and brain (**d**, 312 patients – 127 with EGFR mutations. Represented by green edges).



Chapter 9

General conclusions and future perspectives



Conclusions from this thesis

In a world witnessing a rising prevalence of cancer, personalized oncology stands out as a beacon of hope for more effective and safer treatments¹. Unfortunately, successful personalized targeted therapies are currently reaching too few patients, and the drug discovery pipeline is costly and slow². Computational tools are crucial to accelerate the rate at which novel drugs make it to the market³. Applied to personalized oncology, they can be a key instrument to expand beyond the state-of-the-art anticancer protein targets, but also to pinpoint druggable genetic alterations and screen large molecular libraries in order to find the needle in the haystack⁴.

Computational statistical analyses have proven successful in the past as a means to investigate large amounts of omics data that have led to the prioritization of the currently targeted anticancer proteins⁵⁻⁷. Other computational drug discovery strategies have been implemented for these and related proteins to assist the drug discovery pipeline, as highlighted in **Chapter 2**. Currently, these methods lack evaluation on understudied protein families in cancer research. However, this is precisely where they could contribute to expanding the pool of anticancer targets, thus increasing patient eligibility. Therefore, in this thesis, the computational efforts were focused on the development of pipelines that can be applied to prioritize anticancer targets from underexplored families. In particular, the focus lay on membrane proteins such as GPCRs and SLCs, which in **Chapter 3** I highlight as potential targets for anticancer therapies with clear experimental hurdles.

Through the work developed in this thesis, I demonstrated that back-to-back computational pipelines can be designed to accelerate the development of personalized treatments targeting membrane proteins. Firstly, targets of a particular family can be prioritized based on somatic mutation enrichment in cancer patients across functionally relevant motifs, as was done in **Chapter 5** for GPCRs. Secondly, the effect of cancer-related mutations on prioritized targets can be studied to assess their druggability with structure-based (SB) methods, as was showcased in **Chapter 6** for glutamate transporter EAAT1 and in the literature for GPCRs^{8,9}. Finally, a selection of prioritized mutants that show differential dynamic effects compared to the wild-type version of the protein can be screened against a large virtual library of candidate drugs. To this end, I proposed the development of mutant-aware virtual screening methods, as shown in **Chapter 7** for protein descriptors that maximize the dynamic differences in mutant targets to achieve potent and selective targeted therapies. Yet, these applications encountered a multitude of challenges that combined the inherent hurdles of computational drug discovery methods with those of cancer and membrane protein research.

One of the main challenges in computational drug discovery is data availability. Data-driven approaches such as machine learning (ML) and other statistical methods are highly dependent on data quantity and quality. SB methods are dependent on the availability of resolved protein structures³. The additional focus on membrane proteins provides an extra strain on data availability, as I hypothesized in **Chapter 3** for all types of data and confirmed in **Chapter 4** for mutant bioactivity data. In Chapter 4 it was observed that established anticancer targets, such as EGFR and BRAF, harbor the most mutant

bioactivity data in ChEMBL and that this data concentrates on a few clinically relevant variants. In turn, this meant that models to predict mutant bioactivity data were only predictable for known targets. Indeed, the very limited availability of mutant bioactivity data for GPCRs did not allow the construction of mutant PCM models in **Chapter 7**, thus confirming the negative effect of this bias. In contrast, there are many bioactivity models in the literature for established anticancer targets^{10,11}. Structural data availability also played a big role in this chapter, where the GPCRs analyzed were selected based on the availability of pre-computed molecular dynamics (MD) simulations on an open-source database¹². Moreover, the availability of structural data is a limiting factor in all steps where SB methods are used, such as in **Chapter 6**. In some cases, however, the lack of one type of data can be compensated by another for the same protein due to the high correlation between data types, for example, different omics and imaging data¹³. To this end, knowledge graphs are good representations to maximize the use of heterogeneous data¹⁴, which can be deployed in protein families where several members are known anticancer targets, as demonstrated in **Chapter 8** for RTKs.

It is crucial not only to recognize the importance of data but also to ensure its accessibility and reusability within the community¹⁵. Promisingly, there is a commendable initiative within the scientific community to develop open-source databases and datasets for cancer research and drug discovery that facilitate easy exploration, both manually and computationally^{16–19}. As a bonus point, even if created for other purposes, these databases can be repurposed for anticancer research. For example, in **Chapter 7** I was able to reuse mutagenesis data and compute mutant MD simulations from publicly available resources for GPCRs^{12,20}. Tools based on AlphaFold have been developed for similar applications, but they lack expert knowledge on particular protein families²¹. Therefore, it is advantageous if the protein family under investigation has been studied for therapeutic purposes other than cancer research. This ensures the availability of open-source resources, as seen with GPCRs in comparison to SLCs. Recognizing the importance of open data, I contributed two datasets to the community to further facilitate personalized oncology research. Firstly, in **Chapter 4** I developed a mutant-aware dataset extracted from ChEMBL and Papyrus ready for bioactivity modeling. Of note, the pipeline employed to develop this dataset will be integrated into ChEMBL to improve the database's variant annotation pipeline in the future. Secondly, a GDC database SQL implementation was developed in **Chapter 5** and used in all chapters of this thesis. This SQL dataset was crucial for computational multi-omics analysis of combined cancer projects in this thesis. The community has also taken note of its importance, with over 820 dataset downloads at the time of writing since its publication in October 2021.

Given the high complexity of cancer, the combination of data-driven and structural approaches is a promising strategy to cover as many disease-related factors as possible, as I summarized in **Chapter 2**. However, this combination introduces its own set of additional challenges. It is important to keep in mind that errors are inevitable in computational drug discovery, both related to data and methodologies^{22,23}. Therefore, while stacking multiple computational methods can be beneficial, it introduces a distinct risk of accumulating uncertainties. This concern potentially surfaced in **Chapter 7**, where I devised MD-based protein descriptors for modeling applications, termed 3DDPDs. The

MD-based descriptors outperformed all other protein descriptors they were compared to, particularly in more challenging validation strategies. However, the outcomes derived from MD simulations, notably, exhibit a high degree of stochasticity, as evidenced in **Chapter 6** for multiple replicates for EAAT1. Consequently, the incorporation of uncertainty measures or replicates becomes highly pertinent, which was not implemented in **Chapter 7**. Therefore, it is crucial to subject these combined approaches to testing in diverse scenarios and to institute a rigorous validation process, encompassing benchmark strategies and estimations for predicting uncertainties^{24,25}. Although the fully integrated AI-structural pipelines, as exemplified in **Chapter 7**, hold significant promise, the sequential pipelines possess the advantage of validation at different stages thus reducing the risk of uncertainty accumulation.

The interpretability of models is pivotal for the incorporation of computational approaches into the drug discovery and clinical pipeline. Models perceived as “black boxes” that produce valuable results that cannot be linked back to the underlying data are not well received by clinical practitioners²⁶. While SB methods are highly interpretable, ML models have higher risks of becoming “black boxes”. In **Chapter 7**, I address this challenge by crafting dynamic descriptors that can be traced back to specific amino acids in the structure of the protein. Consequently, if certain features from these descriptors emerge as crucial for the model, it allows us to hypothesize that variations in protein dynamics at these specific locations contribute to differences in bioactivity. However, in terms of interpretability, knowledge graphs are considered one of the most comprehensive computational approaches²⁷, as the one described in **Chapter 8**. In this framework, all the links between data types are defined, enabling the users to navigate and identify the most relevant connections. Integrating “black box” deep learning algorithms on top of the graph, which extract predicted links or significant nodes, still provides the users with the graph itself for reference, aiding in understanding the rationale behind the established connections. These reasons explain the current and future extensive applicability of knowledge graphs in the context of (oncological) drug discovery^{28–31}.

On top of being interpretable, the outcomes generated from the computational pipeline should consistently align with clinical relevance. Specifically, potential anticancer targets and genetic alterations ought to apply to a sufficiently substantial subpopulation, warranting further investigation toward clinical candidacy³². However, it is difficult to fully assess this relevance. For example, in **Chapter 6** and **Chapter 7**, I selected several mutations present in cancer patients in EAAT1 and GPCRs, respectively, for analysis. I compared these mutations to natural variance to confirm that they are cancer-specific. Nevertheless, these mutations occurred only in one or two patients across various cancer types (pan-cancer). To provide context, mutations associated with approved anticancer-targeted therapies, such as EGFR L858R or BRAF V600E, are observed in a higher number of patients in the GDC dataset - specifically 56 and 621, respectively³³. As an additional filtering step, several models could be added to the pipeline to test *a priori* the potential pathogenicity of specific missense mutations³⁴. However, even mutations with a low frequency that are not necessarily cancer drivers can confer an advantage for survival or selectivity in anticancer therapies^{35–37}. Similarly, low-frequency mutations in conserved positions across protein families as identified in **Chapter 5** for GPCRs

could be proposed as therapeutical targets for poly-pharmacological interventions³⁸. Furthermore, these mutations may be linked to differential expression or other (epi)genetic alterations, rendering them promising targets for further investigation^{39,40}. Finally, it is essential to recognize the significance of methods, such as the ones I have developed in this thesis, due to their broad flexibility and thus applicability. These methods lay the groundwork for assessing membrane protein somatic mutations that may be deemed of higher clinical relevance in the future.

The road to clinical relevance is paved by reproducibility and experimental validation. While computational approaches play a key role in generating hypotheses to enhance the success rate throughout the pipeline, experimental testing is indispensable for their validation^{41,42}. Indeed, progress in cancer biology and medicinal chemistry is equally significant alongside advancements in computational drug discovery. This synergy is crucial for enabling personalized oncology, emphasizing the substantial collaboration among these three domains⁴. I exemplified this synergy in **Chapter 6**, where a combined *in silico* and *in vitro* approach was used to evaluate the effect of cancer-related mutations in the EAAT1 glutamate transporter. It is important to realize, though, that one biological experiment is not always enough due to the high complexity of the systems being analyzed⁴³. In this sense, the computational pipelines themselves can be modified to prioritize targets with a better chance to be further validated computationally or experimentally in one or several experiments, as it was demonstrated in **Chapter 5**. Here, multi-objective optimization was used to highlight GPCRs as potential anticancer targets based on a high enrichment of mutations in functionally relevant conserved domains in cancer patients compared to natural variance. However, the optimization algorithm allowed the introduction of additional practical objectives that helped bring forward GPCRs with better chances to be followed up experimentally based on the availability of in-house assays.

As a final note, the methods presented in this thesis were developed with the aim of broad applicability across various targets and protein families. However, substantial optimization is essential to achieve true target-agnostic capability. As previously discussed, certain protein families may currently lack sufficient data for implementing specific steps outlined in this thesis. Nevertheless, it is vital to recognize these challenges while being mindful of the potential for expansion and improvement.

Future perspectives

The dedication of the scientific community to progress towards improved anticancer therapies is evident, as reflected by the majority of approved drugs over the past decade consistently being targeted anticancer therapies⁴⁴⁻⁴⁶. What is even more important, governments and funding organizations recognize the massive burden of cancer in our society and are putting strategies in place to fight it. In the USA, the Cancer Moonshot program was launched in 2016⁴⁷, and the European Union announced Europe's Beating Cancer Plan in 2021⁴⁸. Increased funding holds the potential for significant impact. Promisingly, the main challenges highlighted in this thesis are expected to be addressed in the coming years due to the growing availability of data and enhanced computational

capabilities³, which will precipitate broader applicability and expansion to understudied protein families. Nevertheless, the impracticality of exploring every potential target and mutation in the genome remains, as it could clutter scientific literature and dilute the impact of individual applications. Hence, a clear and focused approach is essential.

The COVID-19 pandemic has demonstrated the remarkable achievements possible when the scientific community collaborates towards a shared goal⁴⁹. Similarly, the cancer pandemic deserves a unified effort. In this context, international bodies could play a pivotal role by assigning quotas to pharmaceutical companies and academic institutions, ensuring a coordinated and complementary allocation of resources towards cancer research. Although such distribution would not be short of challenges regarding funding and IP ownership⁵⁰, it could lead to a significant impact. Private-public funding will kickstart in the short term higher accessibility to personalized therapy clinical trials³². In the long term, a better understanding of the disease will lead to more accurate treatment plans that will reduce the immense economic burden of cancer, estimated to be 100 billion € annually in the EU^{48,51}. Subsequently, the cost reduction resulting from improved personalized oncology treatments will offset the additional expenses incurred in research. I propose that computational tools will play a crucial role in defining and streamlining the various steps required for accelerated and impactful outcomes. These computational pipelines should particularly focus on:

1. *Design and implementation of machine-readable open-source cancer databases*

Whole Genome Sequencing (WGS) projects such as The Cancer Genome Atlas (TCGA)⁵², The Pan-Cancer Analysis of Whole Genomes (PCAWG)⁵³, and more recently The International Cancer Genome Consortium (ICGC)⁵⁴ and the 100,000 Cancer Genomes project¹⁹, play a pivotal role in analyzing the heterogeneity and complexity of cancer. Raw sequencing data from these projects is often available for download from data repositories. Additionally, many of these projects have developed intuitive web-based interfaces that allow exploration of the analyzed results. However, bulk downloads of analyzed results – e.g. somatic mutations, differentially expressed genes/proteins – are rarely available. Furthermore, the data is dispersed across various data portals, leading to considerable variations in analysis pipelines and the format of the contained data. As a consequence, these limitations impose constraints on the possibility of performing analyses across the totality of the data accumulated across patients and data types, making it accessible primarily to bioinformatics experts or limiting it to the scope of very focused and smaller datasets. In this context, the development of centralized computational pipelines could ensure consistency in multi-omics data processing and analysis. These efforts could be supported by the use of large language models, such as ChatGPT, which are already showing potential in biological applications^{55,56}. Furthermore, the use of centralized data collection and relational database storage systems as the one presented in Chapter 5 would facilitate data collection across hospitals and data sharing and reusability among researchers.

2. *Identification of key biomarkers for diagnosis and personalized treatment*

Expanding on the work of this thesis, I anticipate that the holistic analysis of multi-omics, bioactivity, and structural data will be key to pinpointing the biomarkers that define subpopulations of cancer patients and the targets that make good candidates for diagnosis and selective targeting. Knowledge-based approaches as presented in Chapter 8 expanded to all protein families, like canSAR.ai (more focused on protein-ligand interaction)⁵⁷, or BOCK (more focused on multi-omics information)²⁸, are a good starting point. A gold standard model would integrate multi-omics data with clinical biomarkers and protein-ligand interaction, as it has already been proposed for non-oncological personalized medicine⁵⁸. To amplify the impact of the results, these analyses should be seamlessly integrated with experimental validation. Access to experimental methods that are cost-effective and easier to set up should be facilitated across computational labs. This would enhance high-throughput screening, allowing for the assessment of model accuracy before engaging in virtual screening of a subset of compounds. Promising approaches to this end are platforms that allow the automation of chemical synthesis and testing⁵⁹. On top of assessing prediction accuracy, the implementation of these platforms would allow scientists to engage in active learning, which can be used in computational drug discovery to better screen the chemical space of interest⁶⁰.

3. *Prediction of optimal treatment strategies*

The high cost and personal burden associated with cancer largely stem from the challenging decision-making process for determining the optimal treatment strategy. Oncologists face difficult choices when devising a treatment plan, often requiring multiple rounds of treatment before identifying an effective course of action⁶¹. Computational approaches have the potential to provide significant benefits by integrating all clinical data associated with biomarkers that need testing in a patient. A program based on holistic analyses, taking the patient's omics data as input, can serve as a valuable tool for streamlining and enhancing the decision-making process in clinical settings. PANACEA⁶² and PanDrugs^{63,64} are examples recently developed in this direction. The former employs a knowledge graph coupled with a distance-based method to prioritize treatments based only on genomic data. The latter uses a double-scoring scheme, where both a drug score and a gene score are calculated based on the patient's multi-omics data input. Future implementations should aim to merge features from both approaches, incorporating multi-omics data and adopting a more holistic perspective to address the problem comprehensively. This approach would consider all potential treatment options, not only targeted small molecules but also innovative approaches such as cancer vaccines and immunotherapy – where membrane proteins such as GPCRs already play a crucial role⁶⁵. Finally, these analyses should also extend to the design of clinical trials to ensure efficient patient treatment and optimize the likelihood of novel drug approval³².

4. *Prioritization of the main research gaps*

Ultimately, the centralized and organized storage of cancer-related data, as proposed in (1), would not only streamline the identification of potential biomarkers, targets (2), and treatment strategies (3). The analysis of these datasets could also be coupled with uncertainty estimates to precisely identify research areas where projects and data generation should be prioritized⁶⁶. In order for this system to be implemented in the future, several challenges would need to be addressed. One of the primary concerns to address will be the reduction of inequality, focusing on ensuring universal accessibility. It is crucial not only to make these advancements accessible to everyone but equally important not to overlook patients in small sub-populations. These considerations must be integrated at both the data collection and computational model-building levels⁶⁷. Additionally, building trust in the centralized storage of data, implementing proper blinding of the data for research⁶⁸, and enhancing trust among clinical practitioners in computational applications will be significant challenges⁶⁹. Several discussions will be required to determine appropriate centralization systems at different levels that comply with patient privacy standards. In this regard, global systems are likely to present more complications compared to national or supranational entities with established shared policies and funding mechanisms, like the European Union. Hopefully, governing entities will be able to recognize the importance of the problem at hand and set differences aside to work together towards a common goal.

Final remarks

This thesis emphasizes the importance of using AI and structure-based methods to efficiently explore novel personalized oncology treatments with increased efficacy and decreased side effects. This is done by defining three levels where anticancer targets, genetic alterations, and potential drugs are prioritized, respectively. The methods outlined in this thesis were developed with a focus on membrane proteins as a proxy for underexplored proteins in cancer research but with the goal of being broadly applicable across different targets and protein families. While tailoring each new application to its specific requirements is necessary, having a diverse range of approaches to choose from enhances the likelihood of developing the most suitable pipeline. This is vital in the quest to find effective and safe medicines for all cancer patients.

References

- Lassen, U. N. *et al.* Precision oncology: a clinical and patient perspective. *Future Oncol.* **17**, 3995–4009 (2021).
- Haslam, A., Kim, M. S. & Prasad, V. Updated estimates of eligibility for and response to genome-targeted oncology drugs among US cancer patients, 2006–2020. *Ann. Oncol.* **32**, 926–932 (2021).
- Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
- Stuart, D. D. *et al.* Precision Oncology Comes of Age: Designing Best-in-Class Small Molecules by Integrating Two Decades of Advances in Chemistry, Target Biology, and Data Science. *Cancer Discov.* **13**, 2131–2149 (2023).
- Blum, A., Wang, P. & Zenklusen, J. C. SnapShot: TCGA-Analyzed Tumors. *Cell* **173**, 530 (2018).
- Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
- Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- Muthiah, I., Rajendran, K., Dhanaraj, P. & Vallinayagam, S. In silico structure prediction, molecular docking and dynamic simulation studies on G Protein-Coupled Receptor 116: a novel insight into breast cancer therapy. *J. Biomol. Struct. Dyn.* **39**, 4807–4815 (2021).
- Sharp, A. K. *et al.* Biophysical insights into OR2T7: Investigation of a potential prognostic marker for glioblastoma. *Biophys. J.* **121**, 3706–3718 (2022).
- Srisongkram, T. & Weerapreeyakul, N. Drug Repurposing against KRAS Mutant G12C: A Machine Learning, Molecular Docking, and Molecular Dynamics Study. *Int. J. Mol. Sci.* **24**, 669 (2023).
- Chang, H. *et al.* Machine Learning-Based Virtual Screening and Identification of the Fourth-Generation EGFR Inhibitors. *ACS Omega* **9**, 2314–2324 (2024).
- Rodríguez-Espigares, I. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020).
- Kong, J. *et al.* Integrative, Multimodal Analysis of Glioblastoma Using TCGA Molecular Data, Pathology Images, and Clinical Outcomes. *IEEE Trans. Biomed. Eng.* **58**, 3469–3474 (2011).
- Wilcke, X., Bloem, P. & de Boer, V. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Sci.* **1**, 39–57 (2017).
- Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol. Brain* **13**, 24 (2020).
- Béguignon, O. J. M. *et al.* Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J. Cheminformatics* **15**, 3 (2023).
- Austin, B. K., Firooz, A., Valafar, H. & Blenda, A. V. An Updated Overview of Existing Cancer Databases and Identified Needs. *Biology* **12**, 1152 (2023).
- Tanoli, Z. *et al.* Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief. Bioinform.* **22**, 1656–1678 (2020).
- Sosinsky, A. *et al.* Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat. Med.* **30**, 279–289 (2024).
- Isberg, V. *et al.* GPCRdb: An information system for G protein-coupled receptors. *Nucleic Acids Res.* **44**, D356–D364 (2016).
- Zheng, L. *et al.* MoDAFold: a strategy for predicting the structure of missense mutant protein based on AlphaFold2 and molecular dynamics. *Brief. Bioinform.* **25**, bbae006 (2024).
- Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* **26**, 1040–1052 (2021).
- Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discov. Today* **26**, 511–524 (2021).
- Walters, P. We Need Better Benchmarks for Machine Learning in Drug Discovery. Available at <http://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html> (2023). Accessed 2023-12-08.
- Rasmussen, M. H., Duan, C., Kulik, H. J. & Jensen, J. H. Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets. Preprint at *ChemRxiv* <https://doi.org/10.26434/chemrxiv-2023-w93dm> (2023).
- Price, W. N. Big Data and Black-Box Medical Algorithms. *Sci. Transl. Med.* **10**, eaao5333 (2018).
- Tiddi, I. & Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **302**, 103627 (2022).
- Renaux, A. *et al.* A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC Bioinformatics* **24**, 324 (2023).
- Hatano, N., Kamada, M., Kojima, R. & Okuno, Y. Network-based prediction approach for cancer-specific driver missense mutations using a graph neural network. *BMC Bioinformatics* **24**, 383 (2023).

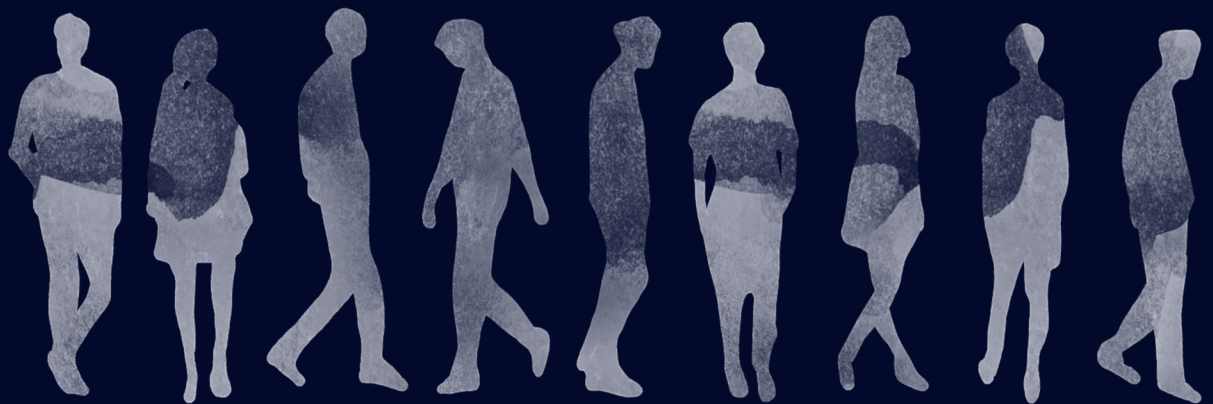
30. Bang, D., Lim, S., Lee, S. & Kim, S. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nat. Commun.* **14**, 3570 (2023).
31. Gogleva, A. *et al.* Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat. Commun.* **13**, 1667 (2022).
32. Fountzilas, E., Tsimberidou, A. M., Vo, H. H. & Kurzrock, R. Clinical trial design in the era of precision medicine. *Genome Med.* **14**, 101 (2022).
33. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
34. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
35. Lusito, E. *et al.* Unraveling the role of low-frequency mutated genes in breast cancer. *Bioinformatics* **35**, 36–46 (2019).
36. Klebanov, N. *et al.* Burden of unique and low prevalence somatic mutations correlates with cancer survival. *Sci. Rep.* **9**, 4848 (2019).
37. Monticelli, M. *et al.* Passenger mutations as a target for the personalized therapy of cancer. Preprint at *PeerJ Preprints* <https://doi.org/10.7287/peerj.preprints.27338v1> (2018).
38. Jones, D. *et al.* Polypharmacology Within the Full Kinome: a Machine Learning Approach. *AMLA Jt. Summits Transl. Sci. Proc.* **2017**, 98–107 (2018).
39. Masica, D. L. & Karchin, R. Correlation of Somatic Mutation and Expression Identifies Genes Important in Human Glioblastoma Progression and Survival. *Cancer Res.* **71**, 4550–4561 (2011).
40. Jiang, L., Yu, H. & Guo, Y. Modeling the relationship between gene expression and mutational signature. *Quant. Biol.* **11**, 31–43 (2023).
41. Schaduengrat, N. *et al.* Towards reproducible computational drug discovery. *J. Cheminformatics* **12**, 9 (2020).
42. Li, H. *et al.* Computational drug development for membrane protein targets. *Nat. Biotechnol.* **42**, 229–242 (2024).
43. Dang, C. V. Reproducibility in Cancer Biology: Mixed outcomes for computational predictions. *eLife* **6**, e22661 (2017).
44. Mullard, A. 2021 FDA approvals. *Nat. Rev. Drug Discov.* **21**, 83–88 (2022).
45. Mullard, A. 2022 FDA approvals. *Nat. Rev. Drug Discov.* **22**, 83–88 (2023).
46. Mullard, A. 2023 FDA approvals. *Nat. Rev. Drug Discov.* **88**, 88–95 (2024).
47. Public Law 114 - 255—114th Congress (2015–2016): 21st Century Cures Act (2016, December 13).
48. Commission, ‘Communication from the Commission to the European Parliament and the Council: Europe’s beating cancer plan’ COM(2021)44.
49. Saag, M. Wonder of wonders, miracle of miracles: the unprecedented speed of COVID-19 science. *Physiol. Rev.* **102**, 1569–1577 (2022).
50. COVID has shown the power of science–industry collaboration. *Nature* **594**, 302–302 (2021).
51. Vellekoop, H. *et al.* The Net Benefit of Personalized Medicine: A Systematic Literature Review and Regression Analysis. *Value Health* **25**, 1428–1438 (2022).
52. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68 (2015).
53. Goldman, M. J. *et al.* A user guide for the online exploration and visualization of PCAWG data. *Nat. Commun.* **11**, 3400 (2020).
54. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
55. Joachimiak, M. P., Caufield, J. H., Harris, N. L., Kim, H. & Mungall, C. J. Gene Set Summarization using Large Language Models. Preprint at *ArXiv* <https://doi.org/10.48550/arXiv.2305.13338> (2023).
56. Caufield, J. H. *et al.* Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. Preprint at *ArXiv* <https://doi.org/10.48550/arXiv.2304.02711> (2023).
57. di Micco, P. *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.* **51**, D1212–D1219 (2023).
58. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci. Data* **10**, 67 (2023).
59. Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H. & Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **40**, 592–604 (2019).
60. Khalak, Y., Tresadern, G., Hahn, D. F., de Groot, B. L. & Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J. Chem. Theory Comput.* **18**, 6259–6270 (2022).
61. Glatzer, M., Panje, C. M., Sirén, C., Cihoric, N. & Putora, P. M. Decision Making Criteria in Oncology. *Oncology* **98**, 370–378 (2018).
62. Ulgen, E., Ozisik, O. & Sezerman, O. U. PANACEA: network-based methods for pharmacotherapy prioritization in personalized oncology. *Bioinformatics* **39**, btad022 (2023).
63. Piñero-Yáñez, E. *et al.* PanDrugs: A novel method to prioritize anticancer drug treatments according

- to individual genomic data. *Genome Med.* **10**, 41 (2018).
64. Jiménez-Santos, M. J. *et al.* PanDrugs2: prioritizing cancer therapies using integrated individual multi-omics data. *Nucleic Acids Res.* **51**, W411–W418 (2023).
 65. Fan, T. *et al.* Therapeutic cancer vaccines: advancements, challenges, and prospects. *Signal Transduct. Target. Ther.* **8**, 1–23 (2023).
 66. Zang, X. *et al.* Prioritizing additional data collection to reduce decision uncertainty in the HIV/AIDS response in 6 US cities: a value of information analysis. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* **23**, 1534–1542 (2020).
 67. Cobanaj, M. *et al.* Advancing equitable and personalized cancer care: Novel applications and priorities of artificial intelligence for fairness and inclusivity in the patient care workflow. *Eur. J. Cancer* **198**, 113504 (2024).
 68. Broekstra, R., Aris-Meijer, J., Maeckelberghe, E., Stolk, R. & Otten, S. Trust in Centralized Large-Scale Data Repository: A Qualitative Analysis. *J. Empir. Res. Hum. Res. Ethics* **15**, 365–378 (2020).
 69. Steerling, E., Siira, E., Nilsen, P., Svedberg, P. & Nygren, J. Implementing AI in healthcare—the relevance of trust: a scoping review. *Front. Health Serv.* **3**, 1211150 (2023).



Appendix **A**

GDC SQL implementation v22.0
(release 16th January 2020)
Quick start guide



What is the GDC?

The NCI Genome Data Commons (GDC)¹ is a publicly available cancer knowledge network to provide the cancer research community with a harmonized and curated data service for genome/transcriptome sequence data and standardized analyses for derived data from different cancer studies. The GDC is currently the official repository of data from the TCGA project² and other more recent and ongoing whole genome cancer sequencing projects such as the TARGET and CGCI projects^{3,4}.

GDC conventional data channels versus GDC SQL local implementation

The data provided by the GDC is available through three different channels: a data portal, a data transfer tool, and an API. From the data portal, part of the data (e.g. cases, genes, mutations, clinical data) is accessible for visualization and analysis, with a number of tools available for this purpose (i.e. Oncogrid, survival plots, cohort comparison). The data portal also allows exploration of the repository, where the data is stored in different file formats. The data comprised in those files, however, is only accessible upon download, which can be done through the data portal (for a small number of files), or through the data transfer tool (from the command line, for a larger amount of files), providing a data manifest previously generated in the data portal. Furthermore, there is an API that grants access to all of the data available on the data portal through different endpoints, as well as to the generation of data manifests for data downloads. Although the conventional GDC data channels are extremely useful for data visualization and retrieval of specific - limited - queries, it is not the most appropriate tool for big data analysis, since the links between different data types are sometimes unclear, and not all the data types are available from the same channels. Moreover, the GDC repository is updated every 2-3 months with new entries, and the conventional data channels only allow data retrieval from the most recent release, which can be prejudicial for projects running for a longer time. Due to these factors, I made the decision to develop a local SQL implementation for GDC using data acquired from all conventional sources. This implementation aims to streamline access to all data from a specific release. The data model that I formulated was carefully crafted to facilitate large-scale data queries and incorporates relevant data types essential for cancer research in both my thesis and related collaborations. Several tests showed that the data contained in the SQL local implementation is almost the same as that of the data portal (for the data types available in the data portal), although some minor differences are found due to errors in the database. The conventional channels, however, are still very useful tools for data visualization and analysis, but the release version needs to always be taken into account.

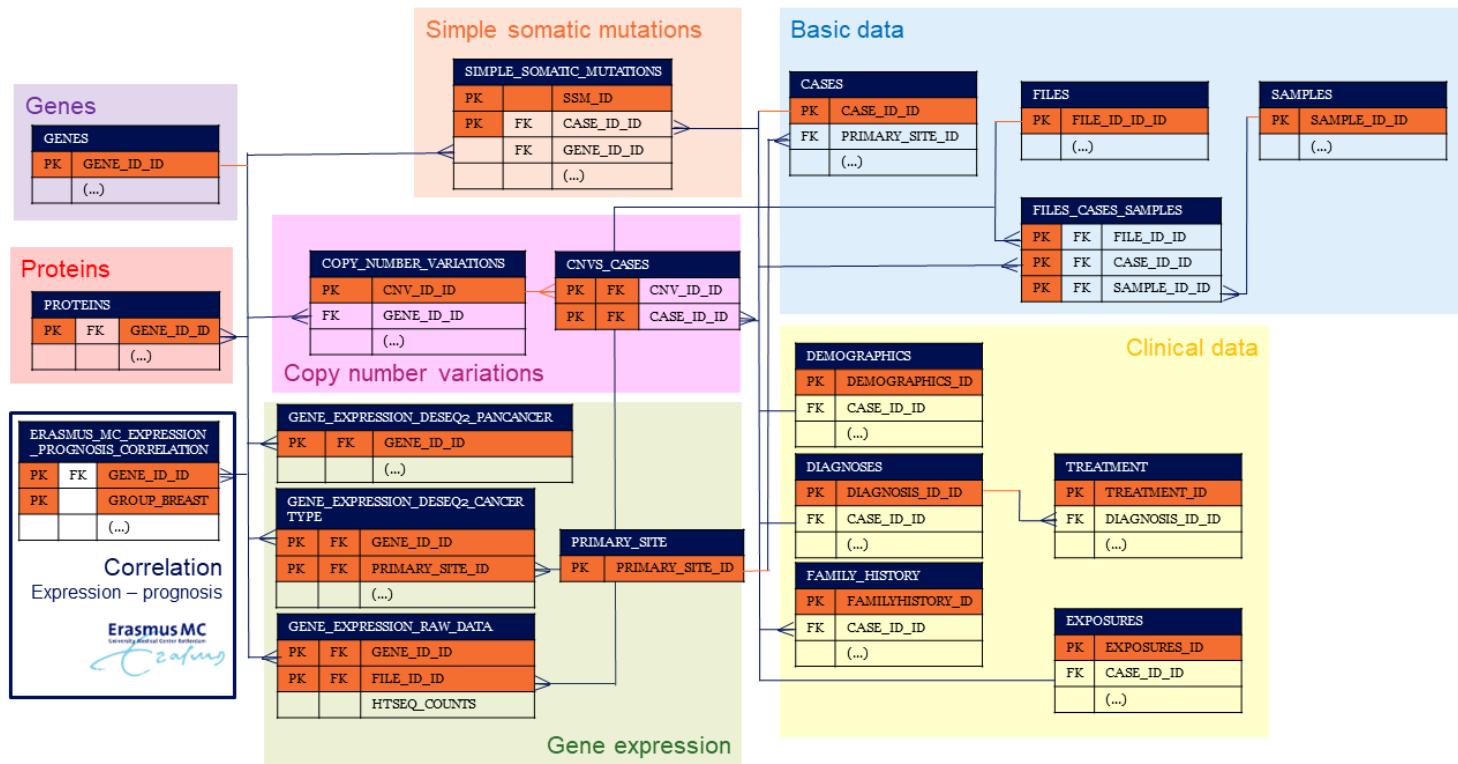


Figure A.1. The basic architecture of the GDC SQL local implementation. Only primary and foreign keys are depicted in the diagram, as well as the connections between them.

The GDC SQL local implementation's basic structure

The SQL local implementation features 19 tables organized into eight fields connected by a network of primary (PK) and foreign (FK) keys to optimize storage and query processing, as shown in **Figure A.1**. Unique numerical values are used for all PKs, and FKs reference PKs in parent tables. There is only one exception to this rule, explained in more detail in the section *The connection between cases, samples, and files*. Some tables (*files_cases_samples*, *cnvs_cases*, and *primary_site*) serve mainly as connection tables and lack additional properties. The full database schema, including all properties, can be found in the associated online repository for **Chapter 5**⁵.

Description of the data fields and their source

The seven data fields in **Figure A.1** depict diverse data types gathered from GDC conventional data channels.

a) Basic data

The tables in this field contain basic data properties for cases, samples, and files. “Cases” is the term used in GDC for patients. The connection between cases, samples, and files is crucial for analyzing the data in the database. The data was obtained through API queries to cases and files endpoints. More detailed relationships and their implications are discussed in *The connection between cases, samples, and files* section.

b) Clinical data

Some of the patients in the GDC have associated clinical data, depending on the cancer project. For those cases, five different tables are available (demographics, family history, exposures, diagnoses, and treatments). The data contained in these tables was obtained by querying the API's cases endpoint.

c) Simple somatic mutations

This table contains all the data associated with genomic sequencing. The data was obtained by querying the API's ssms endpoint and filtered as in the data portal to only keep the canonical transcript's data when several transcripts were available.

d) Copy number variations

Most data is available in the data portal, but copy number variation data can be found only through the API's cnvs endpoint.

e) Gene expression

This field includes raw transcriptomic data (RNA seq HTSeq counts) and analyzed gene expression annotations, making it one of the most challenging fields due to the lack of availability in the data portal and API. I obtained and analyzed the files using

a specific pipeline detailed in the *Analysis of gene expression data* section.

f) Genes

This is a field that provides gene information for different tables. The data for this field was obtained from the API's ssms endpoint, and extracted from the Simple somatic mutations table to be able to provide information to a larger number of tables.

g) Proteins

Similarly to the field *Genes*, this field provides protein information for different tables.

h) Erasmus MC expression-progression correlation

This field provides information derived from an Erasmus MC (CC. J.W.M. Martens) correlation analysis between breast cancer patient's gene expression data and their cancer progression profiles.

Guide to the most useful data properties

All properties obtained from the GDC API conserve their original names, except for the PKs and FKs, which were manually created to give numerical references. Therefore, the description for some of the properties is available in the GDC data dictionary online. In general, the description of the properties is intuitive. The properties in the tables corresponding to the *Gene expression* field are derived from differential expression analysis and are further detailed in the section *Analysis of gene expression data*. The tables *files_cases_samples*, *cns_cases*, and *primary_site* are mainly connection tables, therefore they do not contain useful properties for other purposes than linking tables. The most useful properties in the most relevant tables are described in **Tables A.1-A.11** with definitions based on the GDC data dictionary (https://docs.gdc.cancer.gov/Data_Dictionary/viewer).

Table A.1. Description of the most useful properties in table cases.

Table: cases	
Property	Meaning
primary_site	Primary site or the general location of the cancer, as categorized by the World Health Organization (WHO). Can be used as a replacement for cancer type. E.g. Adrenal gland, Breast, Bronchus and lung.
disease_type	Type of malignant disease as categorized by the World Health Organization's (WHO) International Classification of Diseases for Oncology (ICD-O). E.g. Blood Vessel Tumors, Mesothelial Neoplasms.

Table A.2. Description of the most useful properties in table samples.

Table: samples	
Property	Meaning
sample_type	Origin of a biological sample utilized in a laboratory analysis. E.g. Blood Derived Cancer - Bone Marrow, Primary Tumor, Metastatic, Solid Tissue Normal, RNA, Slides
tissue_type	Type of tissue based on its disease status or proximity to tumor tissue. E.g. Tumor, Normal, Peritumoral *NOTE: even though this would be the perfect property to define whether a sample is derived from a tumor or normal tissue, it is often unknown or not described, so for that purpose is better to use sample_type

Table A.3. Description of the most useful properties in table files.

Table: files	
Property	Meaning
data_category	Category of data included in a file. E.g. Simple Nucleotide Variation, Clinical
data_type	Detailed data type included in a file. E.g. Gene Expression Quantification, Slide Image
experimental_strategy	Experimental strategies employed for molecular characterization of the cancer. E.g. WGS, miRNA-Seq
workflow_type	Bioinformatic workflow used for analysis of the data. E.g. DNACopy, HTSeq - Counts, GENIE Simple Somatic Mutation

Table A.4. Description of the most useful properties in table diagnoses.

Table: diagnoses	
Property	Meaning
age_at_diagnosis	Age at the time of diagnosis as number of days since birth.
last_known_disease_status	Last known condition of an individual's disease. E.g. Tumor free, Distant met recurrence/progression
progression_or_recurrence	Yes/No/Unknown indicator to identify whether a patient has had a new tumor after initial treatment.
ajcc_clinical_stage	Stage group determined from clinical information on the tumor, regional node, and metastases to group patients with similar prognosis for cancer. E.g. Stage 0, Stage IA2
ajcc_pathologic_stage	Spread of the disease through the body based on cancer staging using AJCC criteria.
igcccg_stage	Staging according to the International Germ Cell Cancer Collaborative Group (IGCCCG), used to further classify metastatic testicular tumors. E.g. Good Prognosis, Poor Prognosis

Table A.5. Description of the most useful properties in table treatments.

Table: treatments	
Property	Meaning
therapeutic_agents	Individual agent(s) used in treatment. E.g. 10-Deacetylaxol
treatment_effect	Effect a treatment had on the tumor. E.g. complete Necrosis (No Viable Tumor), No Necrosis
treatment_type	Type of treatment used. E.g. Chemotherapy, Immunotherapy (Including Vaccines)

Table A.6. Description of the most useful properties in table simple_somatic_mutations.

Table: simple_somatic_mutations	
Property	Meaning
mutation_type	General type of mutation. E.g. Substitution, Deletion, Insertion
mutation_subtype	Detailed subtype of mutation. E.g. Missense, Frameshift, Stop Gained, Intron
gene_id	Ensembl gene id
aa_change	Amino acid change in the protein affected by a mutation in a protein-coding gene. E.g. V600E, K15Rfs*5, empty (for deletions)
sift_impact	Sorting Intolerant From Tolerant (SIFT [®]) predicted category respect to the likelihood of a phenotypic effect upon mutation: <ul style="list-style-type: none">• tolerated: Not likely• tolerated_low_confidence: More likely than tolerated• deleterious: Likely• deleterious_low_confidence: Less likely than deleterious
vep_impact	Ensembl Variant Effect Predictor (VEP [®]) predicted category respect to the extent of the impact on protein function upon mutation : <ul style="list-style-type: none">• HIGH (H): Disruptive impact on the protein, e.g. truncation, loss of function• MODERATE (M): Non-disruptive but might change protein effectiveness• LOW (L): Mostly harmless• MODIFIER (MO): Non-coding variants or variants affecting non-coding genes, therefore the impact is difficult to predict
polyphen_impact	Polymorphism Phenotyping (Polyphen [®]) predicted category respect to the possibility to affect protein structure or function: <ul style="list-style-type: none">• probably damaging (PR): Highly possible• possibly damaging (PO): Possible• benign (BE): Not likely• unknown (UN): Difficult to make a prediction

Table A.7. Description of the most useful properties in table `copy_number_variations`.

Table: <code>copy_number_variations</code>	
Property	Meaning
<code>gene_id</code>	Ensembl gene id
<code>cnv_change</code>	Copy number estimation based on the GDC <i>Copy Number Variation Analysis Pipeline</i> . Three categories are defined based on the focal CNV values: <ul style="list-style-type: none">loss (-1): focal CNV values smaller than -0.3gain (+1): focal CNV values larger than 0.3neutral (0): focal CNV values between -0.3 and 0.3

Table A.8. Description of the most useful properties in table `genes`.

Table: <code>genes</code>	
Property	Meaning
<code>gene_id</code>	Ensembl gene id
<code>symbol</code>	HGNC symbol for the gene analyzed

Table A.9. Description of the most useful properties in tables `gene_expression_deseq2_pancancer` and `gene_expression_deseq2_cancertype`.

Table: <code>gene_expression_deseq2_pancancer</code> / <code>gene_expression_deseq2_cancertype</code>	
Property	Meaning
<code>gene_id</code>	Ensembl gene id
<code>expression_status</code>	Gene expression estimation upon differential expression analysis of tumor vs. normal samples with DESeq2 as detailed in section <i>Analysis of gene expression data</i> . <ul style="list-style-type: none">Genes with <code>log2_fold_change</code> values larger than 2 and <code>p_value</code> values lower than 0.05 are categorized as “overexpressed”Genes with <code>log2_fold_change</code> values lower than -2 and <code>p_value</code> values lower than 0.05 are categorized as “underexpressed”Genes with <code>log2_fold_change</code> values between -2 and 2 and <code>p_value</code> values lower than 0.05 are categorized as “neutral”Genes with <code>p_value</code> values higher than 0.05 are categorized as “not significant”

Table A.10. Description of the most useful properties in table `proteins`.

Table: <code>proteins</code>	
Property	Meaning
<code>gene_id</code>	Ensembl gene id
<code>SwissProt</code>	UniProt accession code for the protein corresponding to the gene analyzed

Table A.11. Description of the most useful properties in erasmus_mc_expression_prognosis_correlation table.

Table: erasmus_mc_expression_prognosis_correlation	
Property	Meaning
gene_id	Ensembl gene id
group_breast	Group to which the breast cancer patients belong to: <ul style="list-style-type: none">• ERpos: ER positive• ERneg: ER negative• TN: triple negative (negative for ER, PR and ERBB2)• ALL: all patients
expression_prog_corr	The existence or not of correlation between gene expression and negative prognosis (rapid progression leading to metastasis): <ul style="list-style-type: none">• “True” for hr_value > 1 and p_value < 0.05• “False” for hr_value < 1 and pvalue < 0.05• “ns” or not statistically significant for p_value > 0.05 and any hr_value

The connection between cases, samples, and files

The basic information field is crucial to understand the patient’s data. All the rest of the fields are connected to this one, either by the *case_id_id* or the *file_id_id*. The cases table provides important information, like the primary site of the tumor, while the samples table provides information about the type of sample (e.g. normal, tumor). Even though the data directly connected to *case_id_id* (e.g. simple somatic mutations or clinical data) cannot be directly connected to a specific sample, the information provided by that link can be still very useful.

It is important to note, however, that the relationship between cases, files, and samples is complicated. A single case can be associated to different files and to different samples. At the same time, a file can be created with data from different samples, even from different cases. This can be confusing when the same case has samples of different types (**Figure A.2**). All these possible scenarios need to be considered when querying the database.

Moreover, these relationships can be retrieved from the GDC API through two different ways: a) using the cases endpoint, the cases - samples, and the cases - files relationships can be retrieved, and b) using the files endpoint, the files - cases - samples triple relationship can be retrieved. Here, I used both ways to obtain data for the *files_cases_samples* table. This means that some file-sample relationships are not defined. In order to be able to add the additional cases - samples and cases -files relationships, they were linked to an empty value in the files and samples tables, respectively.

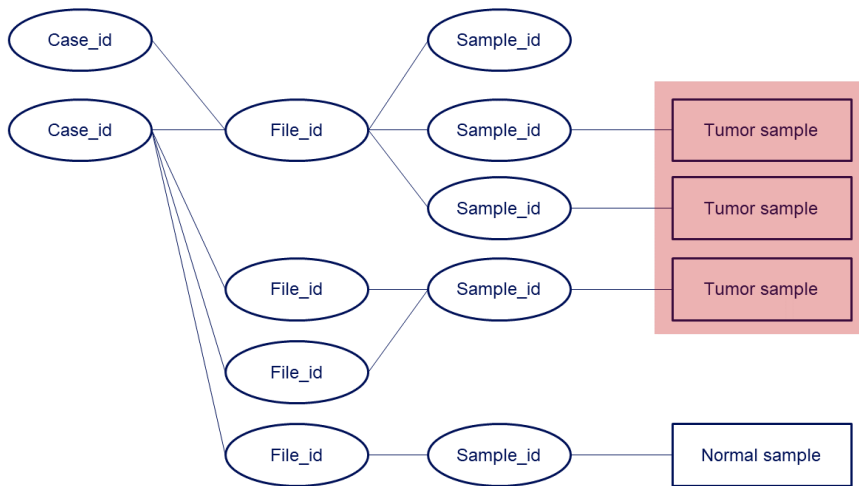


Figure A.2. Relationships between cases, files, and samples available in the GDC.

Analysis of gene expression data

Even though transcriptomic data is available in files from the GDC data transfer tool, these need further analysis in order to be properly interpreted. From files with *work-flow_type* equal to “HTSeq - Counts”, I performed differential expression analyses with DESeq2 in order to assess the over- and under-expression of genes in tumor vs. normal samples. I made use of the files-cases-samples relationships in order to define the origin of the different RNA sequencing files. I performed two types of analyses: a) pan-cancer differential expression analysis and b) per cancer type differential expression analysis. The cancer type was defined based on the *primary_site* property. The potential batch effect introduced by samples from different projects was accounted for by introducing it as the covariate in the analysis. The DESeq2 analysis was performed using the Leiden University supercomputer facilities (ALICE), and the results were uploaded to the GDC SQL local implementation, together with an interpretation of the results (property *expression_status*). Moreover, in the GDC SQL local implementation, there is a *gene_expression_raw_data* table, where the raw counts from the HTSeq - Counts files are included, in order to be able to perform differential expression analyses a posteriori from raw data on custom cohorts.

A

Erasmus MC prognosis analysis

The data from Erasmus MC was provided by J.W.M. Martens for breast tumors and breast cell lines. Regarding the tumors, this is a cohort of their own data (n = 344) supplemented with publicly available samples that all run on the same chip type (867 samples in total). Clinically, the samples are similar as well, all are lymph-node negative and have not been adjuvantly treated (no chemo / hormonal therapy after surgery to

remove the primary tumor). They also know the metastasis-free survival (MFS) of these patients, and they then view the prognosis in all samples or separately for ER negatives, ER positives, and Triple negatives (negative for ER, PR, and ERBB2). With a Cox regression, they calculated a Hazard Ratio (“hr_value”) with a p-value. This was done on the expression data as a continuous value. A $HR > 1$ means a correlation exists between high expression and poor prognosis (short time between primary and metastasis).

References

1.

Grossman, R. L. *et al.* Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

2.

Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68 (2015).

3.

Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).

4.

Thomas, N. *et al.* Genetic subgroups inform on pathobiology in adult and pediatric Burkitt lymphoma. *Blood* **141**, 904–916 (2023).

5.

Bongers, B. *et al.* Data underlying the article: Pan-cancer in silico analysis of somatic mutations in G-protein coupled receptors: The effect of evolutionary conservation and natural variance. Available at <https://doi.org/10.4121/15022410.V1> (2021).

6.

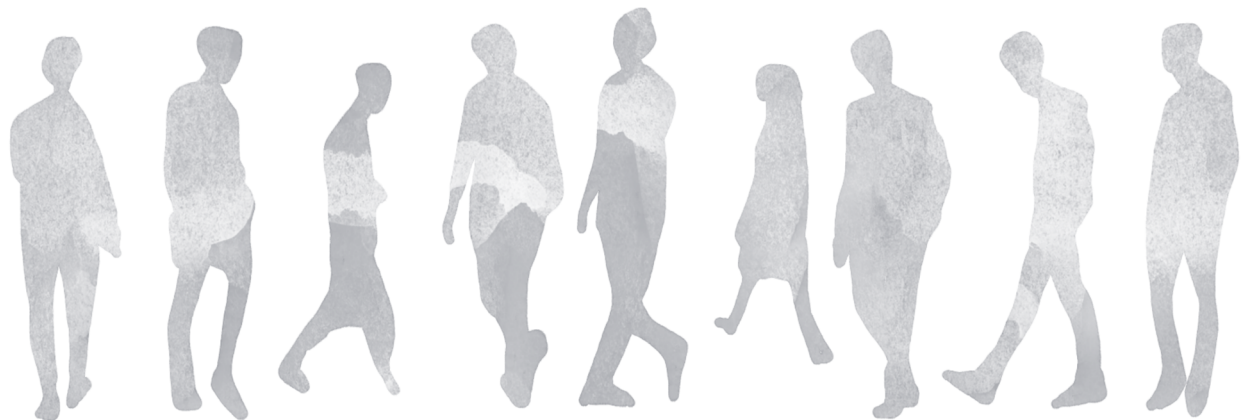
Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

7.

McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).

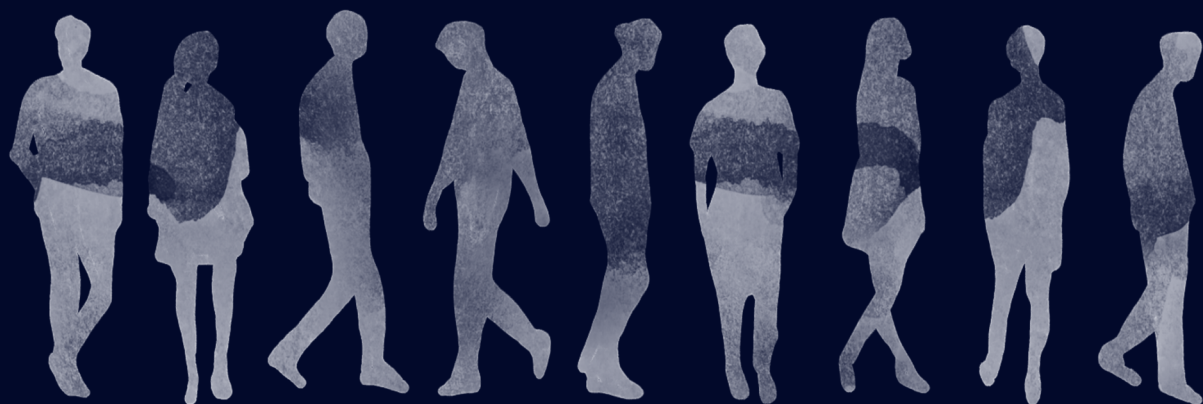
8.

Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).



Appendix **B**

Data and software availability



This thesis was created with the aim to promote FAIR (Findable, Accessible, Interoperable, and Reusable) data management principles and open source software development practices. To this end, whenever possible, data was obtained from public databases and repositories, and open-source software was used. Similarly, novel datasets and code derived from the practical chapters of this thesis (Chapters 4-8) were made available through public repositories:

Chapter 4

Data availability

The ChEMBL 31 data used in this chapter is available online (https://doi.org/10.6019/CHEMBL_database.31). The bioactivity data from the Papyrus dataset is available online (<https://doi.org/10.5281/zenodo.7373214>). All protein structures used in this chapter are available on the RCSB Protein Data Bank (<https://www.rcsb.org/>). The bioactivity-enhanced bioactivity dataset derived from this chapter is available on Zenodo (<https://doi.org/10.5281/zenodo.11236694>).

Software availability

The Python 3.10 code used to compile and analyze the data for this chapter is available on Zenodo (<https://doi.org/10.5281/zenodo.11236694>) and maintained on GitHub (https://github.com/CDDLeiden/chembl_variants).

Chapter 5

Data availability

The protein structures used in this chapter are available on the RCSB Protein Data Bank (<https://www.rcsb.org/>). The ChEMBL 27 data used in this chapter is available online (https://doi.org/10.6019/CHEMBL_database.27). The G protein-coupled receptor information derived from the GPCRdb database is available online (<https://gpcrdb.org/>). The GDC v22.0 SQL implementation and the compilation of the 1000 Genomes dataset, as well as all datasets for analysis derived from this chapter are available on the 4TU repository (<https://doi.org/10.4121/15022410>).

Software availability

The source code used to produce the results in this chapter was generated using the commercial software package Accelrys Pipeline Pilot 2018 version 18. All Pipeline Pilot protocols, as well as the Python 3.8 code used to generate the figures for this chapter, are available on the 4TU repository (<https://doi.org/10.4121/15022410>).

Chapter 6

Data availability

The GDC v22.0 SQL implementation and the compilation of the 1000 Genomes dataset are available in online repositories (see Chapter 5 *Data availability*). All protein structures used in this chapter are available on the RCSB Protein Data Bank (<https://www.rcsb.org/>). The input files needed to generate the molecular dynamics simulations in this chapter using Desmond, as well as the results from Monte Carlo mutagenesis and 4D docking are available on Zenodo (<https://doi.org/10.5281/zenodo.11236571>).

Chapter 6

Software availability

The commercial software ICM-Pro version 3.9-2c and open source Desmond version 2021.1 were used in this chapter. The analysis of the molecular dynamics simulations was done with PyMol version 2.5.2 and Python 3.8. All the projects and scripts are available on Zenodo (<https://doi.org/10.5281/zenodo.11236571>).

Chapter 7

Data availability

The bioactivity data used in this chapter was obtained from the Papyrus dataset and is available online (<https://doi.org/10.5281/zenodo.7373214>). The wild-type molecular simulations were obtained from the GPCRmd database and are available online (<https://submission.gpcrmd.org/home/>). The G protein-coupled receptor information derived from the GPCRdb database is available online (<https://gpcrdb.org/>). The input files needed to generate the mutant molecular dynamics simulations in this chapter using AceMD are available on Zenodo (<https://doi.org/10.5281/zenodo.7957235>).

Software availability

The Python 3.8 code used to generate and analyze the results in this chapter is available on Zenodo (<https://doi.org/10.5281/zenodo.8026883>) and maintained at GitHub (<https://github.com/CDDLeiden/3ddpd>).

Chapter 8

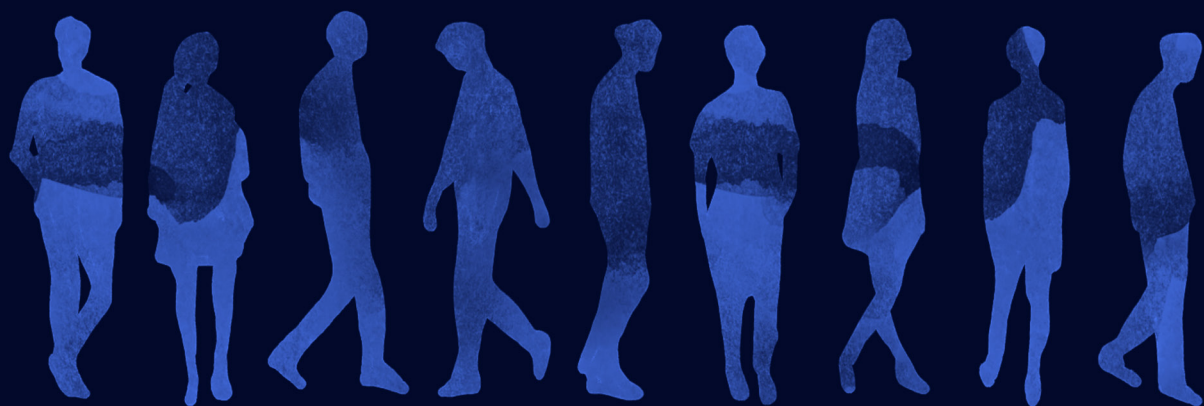
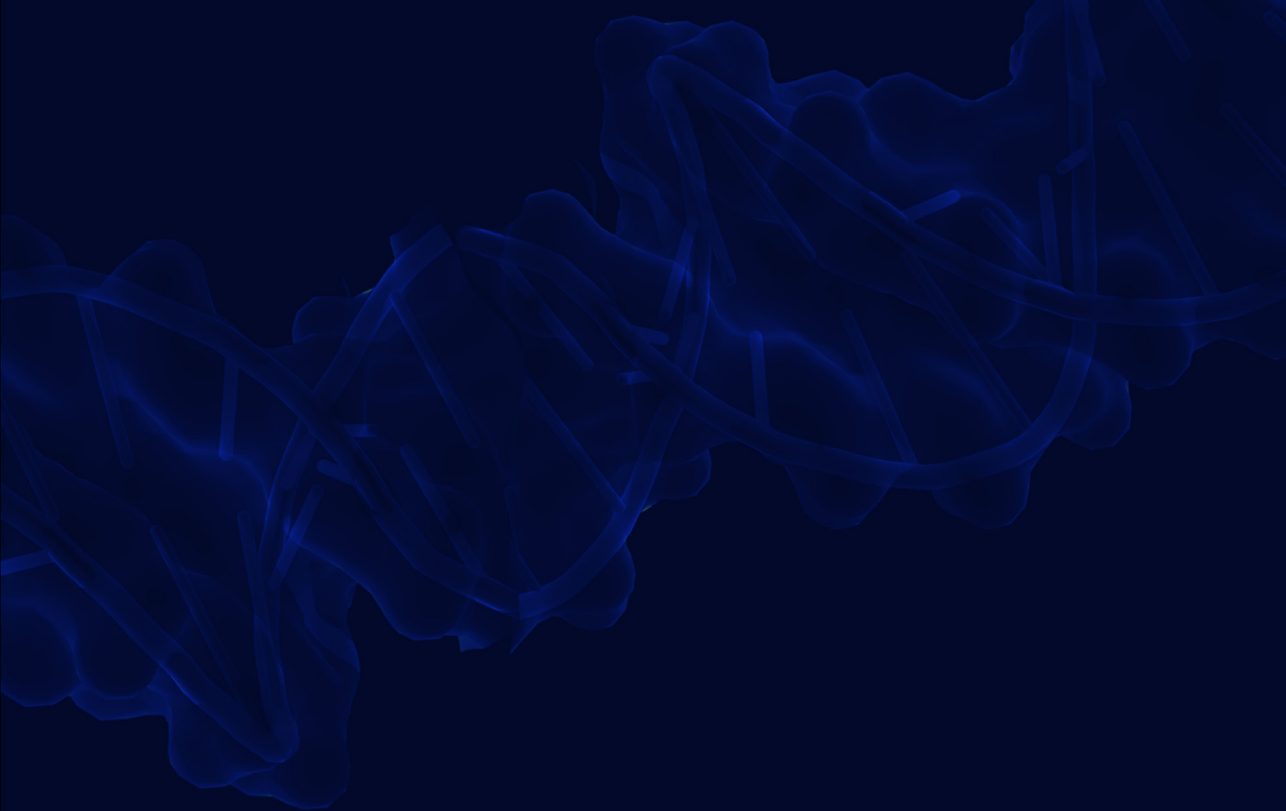
Data availability

The data used in this chapter was previously compiled and made available in previous chapters, except for the phosphorylation network, which is freely available online upon registration (<https://cancer.ucsf.edu/phosphoatlas>), and the kinome data from the KLIFS database, which is available online (<https://klifs.net/browse.php>). The input files needed to generate the kinome and receptor tyrosine kinase knowledge graphs, as well the pickle files to re-generate the graphs are available on Zenodo (<https://doi.org/10.5281/zenodo.11236776>). This repository also contains the HTML interactive visualization sessions described in the chapter.

Software availability

The Python 3.10 code used to compile and analyze the knowledge graphs is available on Zenodo (<https://doi.org/10.5281/zenodo.11236776>).





Summary

Cancer is considered the silent pandemic of the 21st century and the second leading cause of death worldwide. The significant heterogeneity of this disease, seen across various cancer types, individuals, and even tumor cells, makes it extremely challenging to treat effectively and safely in all patients. Personalized oncology has emerged as an efficient strategy to leverage the differences present in cancer for the selective targeting of tumor cells. This approach aims to reduce side effects while maintaining or enhancing therapeutic efficacy. However, the availability of personalized therapies is currently limited, leaving many cancer patients longing for more selective treatments. In this context, computational tools play a crucial role in exploring unresolved questions in cancer research and accelerating the discovery of new proteins that can be selectively targeted in anticancer therapies. One main advantage of using computational tools is the ability to investigate promising protein families that have been overlooked in cancer research due to experimental limitations or publication bias, such as membrane proteins. This thesis delves into the potential of computational tools in prioritizing novel targets, mutations, and drugs for use in personalized oncology, with a specific focus on membrane proteins.

This concept is first introduced in **Chapter 1**, where the three prioritization levels are linked to functional relevance, druggability, therapy potency, selectivity, and resistance. The main promises and challenges in personalized oncology are delineated in this chapter, followed by an overview of computational methods used in drug discovery that can be extrapolated to oncological research. In particular, it is introduced how these methods can be applied to the study of membrane proteins as promising yet experimentally challenging anticancer targets. These concepts are further expanded upon throughout this thesis.

Chapter 2 reviews the wide range of computational tools that can be applied in oncological drug discovery. The main focus of this chapter is on two main categories: artificial intelligence (AI) and structure-based (SB) methods. These two categories are outlined independently, but the increased potential of their combination is highlighted, especially in the context of cancer research, which requires multidisciplinary solutions. By reviewing a selection of combined applications in cancer-related targets, the reader gains an understanding of the potential of the methodologies developed and applied throughout this thesis.

The applications discussed in Chapter 2 primarily focus on established anticancer targets, in particular soluble protein kinases. However, broadening the range of anticancer targets is crucial for expanding access to personalized oncology treatments for a larger population. **Chapter 3** emphasizes membrane proteins as potential new anticancer targets that can be explored using computational tools to address the experimental challenges that make them less appealing to study compared to soluble proteins. This chapter also identifies the main challenges in computational drug discovery for membrane proteins, which are primarily related to data availability and publication bias. Within this context, three protein families with varying levels of representation in the literature are highlighted and examined: receptor tyrosine kinases (RTKs), G protein-coupled

receptors (GPCRs), and solute carriers (SLCs).

Chapter 4 further explores the differences in data availability between protein families and individual targets. This chapter highlights the strong correlation between publication bias and data present in publicly available databases, specifically bioactivity data related to mutant proteins or genetic variants in the ChEMBL database. This data, vital for oncological drug discovery, is significantly enriched on known anticancer targets and genetic variants, particularly kinases and RTKs. The chapter emphasizes the importance of this data in computational drug discovery through benchmarking variant-agnostic and variant-aware bioactivity models, which can be utilized in oncological drug discovery under appropriate circumstances. Additionally, the chapter offers data, tools, and recommendations to aid in the curation of high-quality variant-annotated datasets for bioactivity modeling.

Chapters 5-7 focus on developing various computational applications to address the three levels of prioritization introduced in Chapter 1. These applications utilize the methods introduced in Chapter 2 and are applied to the lesser explored membrane protein families introduced in Chapter 3.

Target prioritization for GPCRs is discussed in **Chapter 5**. This chapter evaluates the functional relevance of individual GPCRs in cancer by analyzing pan-cancer related mutations in comparison to natural variance. The results in this chapter and the subsequent ones are based on a cancer patient dataset created for this thesis from the Genomic Data Commons (GDC) database, which is made available for public use and is computationally friendly and version-stable. Mutations enriched in cancer and located in functionally-conserved motifs are considered priorities in identifying 52 GPCRs as potential anticancer targets using a multi-objective optimization approach. This approach also allows for the inclusion of practical objectives, such as additional computational and experimental resources, and can be further integrated with SB analyses for specific receptors of interest, including the methods described in Chapter 6.

Chapter 6 covers mutant prioritization for the glutamate transporter EAAT1, a member of the SLC family. Cancer-related mutations from the GDC dataset found near the orthosteric and allosteric binding pockets are computationally tested to assess their impact on protein conformation and function. Molecular dynamics (MD) simulations and docking experiments suggest that certain cancer-related mutations, specifically R479W, induce a conformational change that can be leveraged in personalized oncology. Additionally, this chapter demonstrates the translatability of computational findings to real-world applications through *in vitro* experimental validation of the mutations' effects on transporter function and response to pharmacological intervention.

Drug prioritization is discussed in **Chapter 7**, which introduces a method developed to enhance virtual screening of large libraries of candidate drugs for mutant GPCRs. This chapter presents the creation of novel 3D dynamic protein descriptors (3DDPDs) based on MD simulations to improve the representation of mutant proteins for proteochemometric bioactivity modeling. Results show that these novel descriptors outperform sequence-based descriptors in wild-type GPCR bioactivity modeling. However, evaluation

of their applicability in mutant GPCRs is pending due to data availability constraints discussed in Chapter 4.

The lessons learned from Chapters 4-7 culminate in **Chapter 8**, where a holistic approach is taken to integrate all types of data previously discussed. In this chapter, a patient-centric knowledge graph is developed with the goal of prioritizing mutated proteins for targeted therapy in cancer. This approach combines the structural and bioactivity data analyses from Chapter 4, as well as the cancer and natural variance data from Chapter 5. Additionally, it builds upon the concepts explored in Chapters 4-7 to prioritize targets and mutations that are functionally, structurally, and clinically relevant. Due to limitations in data availability, the focus of this chapter is primarily on kinases, specifically RTKs. However, like the preceding chapters, it is designed to be adaptable to any protein type if the necessary data becomes accessible in the future. Advanced modeling algorithms could also be utilized to enhance the knowledge graph as more data becomes available.

Finally, **Chapter 9** provides a summary of the conclusions drawn from the preceding chapters within the wider context of computational oncological drug discovery. Overall, the methods developed within this thesis expand the range of available tools for selecting novel targets, mutants, and drug candidates for personalized oncology applications. However, in order to achieve clinical significance, collaborative efforts must be maintained within the scientific community to focus on cancer research initiatives. Computational tools similar to those developed in this thesis can be extremely beneficial for tasks such as designing and implementing machine-readable open-source cancer databases, identifying key biomarkers for diagnosis and personalized treatment, predicting optimal treatment strategies, and prioritizing key research areas. Ultimately, it is only through collaborative and focused efforts that effective and safe treatments can be developed for all patients fighting cancer.

Samenvatting

Kanker wordt beschouwd als de stille pandemie van de 21e eeuw en is de tweede belangrijkste doodsoorzaak wereldwijd. De aanzienlijke heterogeniteit van de ziekte kanker, gezien in verschillende kankersoorten, individuen en zelfs tumorcellen, maakt het buitengewoon uitdagend om het effectief en veilig bij alle patiënten te behandelen. Gepersonaliseerde oncologie is naar voren gekomen als een efficiënte strategie om gebruik te maken van de verschillen die aanwezig zijn in kanker voor de selectieve targeting van tumorcellen. Deze aanpak streeft ernaar de bijwerkingen te verminderen terwijl de therapeutische effectiviteit behouden blijft of verbetert. Echter, de beschikbaarheid van gepersonaliseerde therapieën is momenteel beperkt, waardoor veel kankerpatiënten verlangen naar meer selectieve behandelingen. In deze context spelen computationele hulpmiddelen een cruciale rol bij het verkennen van onopgeloste vragen in kankeronderzoek en het versnellen van de ontdekking van nieuwe eiwitten die selectief getarget kunnen worden in antikankertherapieën. Een belangrijk voordeel van het gebruik van computationele hulpmiddelen is het vermogen om veelbelovende eiwitfamilies te onderzoeken die over het hoofd zijn gezien in kankeronderzoek door experimentele beperkingen of publicatiebias, zoals membraaneiwitten. Dit proefschrift onderzoekt het potentieel van computationele hulpmiddelen bij het voorrang geven aan nieuwe doeleiwitten, mutaties en medicijnen voor gebruik in gepersonaliseerde oncologie, met een specifieke focus op membraaneiwitten.

Dit concept wordt eerst geïntroduceerd in **Hoofdstuk 1**, waar de drie prioriteitsniveaus worden gekoppeld aan functionele relevantie, kans op genezing, therapie-effectiviteit, selectiviteit en resistentie. De belangrijkste beloften en uitdagingen in gepersonaliseerde oncologie worden in dit hoofdstuk uiteengezet, gevolgd door een overzicht van computationele methoden die worden gebruikt in geneesmiddelenonderzoek en te extrapoleren zijn naar oncologisch onderzoek. In het bijzonder wordt belicht hoe deze methoden kunnen worden toegepast op de studie van membraaneiwitten als veelbelovende maar experimenteel uitdagende antikankertargets. Deze concepten worden verder uitgewerkt in dit proefschrift.

Hoofdstuk 2 bespreekt de brede reeks computationele hulpmiddelen die kunnen worden toegepast in oncologische medicijnontdekking. In het bijzonder komen twee hoofdcategorieën aan bod: kunstmatige intelligentie (AI) en structuur-gebaseerde (SB) methoden. Deze twee categorieën worden onafhankelijk uiteengezet, maar het verhoogde potentieel van hun combinatie wordt benadrukt, vooral in de context van kankeronderzoek dat multidisciplinaire oplossingen vereist. Door een selectie van gecombineerde toepassingen in kankergerelateerde doelen te beoordelen, krijgt de lezer inzicht in het potentieel van de methodologieën die zijn ontwikkeld en toegepast in dit proefschrift.

De toepassingen die in Hoofdstuk 2 worden besproken, richten zich voornamelijk op gevestigde antikankertargets, met name oplosbare proteïne-kinasen. Het verbreden van het scala aan antikankertargets is echter cruciaal voor het uitbreiden van de toegang tot gepersonaliseerde oncologiebehandelingen voor een grotere populatie. **Hoofdstuk 3** benadrukt membraaneiwitten als potentiële nieuwe antikankertargets die kunnen

worden onderzocht met behulp van computationele hulpmiddelen om de experimentele uitdagingen aan te pakken die ze minder aantrekkelijk maken om te bestuderen in vergelijking met oplosbare eiwitten. Dit hoofdstuk identificeert ook de belangrijkste uitdagingen in computationele medicijnontdekking voor membraaneiwitten, die voornamelijk te maken hebben met de beschikbaarheid van gegevens en publicatiebias. In deze context worden drie eiwitfamilies met verschillende niveaus van vertegenwoordiging in de literatuur uitgelicht en onderzocht: receptor tyrosine kinasen (RTKs), G-eiwit gekoppelde receptoren (GPCRs) en solute carriers (SLCs).

Hoofdstuk 4 onderzoekt verder de verschillen in de beschikbaarheid van gegevens tussen eiwitfamilies en individuele doelen. Dit hoofdstuk benadrukt de sterke correlatie tussen publicatiebias en gegevens in openbaar beschikbare databases, specifiek bioactiviteitsgegevens gerelateerd aan gemuteerde eiwitten of genetische varianten in de ChEMBL database. Deze gegevens, die van vitaal belang zijn voor oncologische medicijnontdekking, zijn aanzienlijk verrijkt op bekende antikankertargets en genetische varianten, met name kinasen en RTKs. Het hoofdstuk benadrukt het belang van deze gegevens in computationele medicijnontdekking door benchmarking van variant-agnostische en variant-bewuste bioactiviteitsmodellen die kunnen worden gebruikt in oncologische medicijnontdekking onder geschikte omstandigheden. Daarnaast biedt het hoofdstuk gegevenshulpmiddelen en aanbevelingen om te helpen bij het samenstellen van hoogwaardige variant-geannoteerde datasets voor bioactiviteitsmodellering.

Hoofdstukken 5-7 richten zich op het ontwikkelen van verschillende computationele toepassingen om de drie niveaus van prioritering te behandelen die in Hoofdstuk 1 zijn geïntroduceerd. Deze toepassingen maken gebruik van de methoden die zijn vermeld in Hoofdstuk 2 en worden toegepast op de minder onderzochte membraaneiwitfamilies die zijn geïntroduceerd in Hoofdstuk 3.

Target-prioritering voor GPCRs wordt besproken in **Hoofdstuk 5**. Dit hoofdstuk evalueert de functionele relevantie van individuele GPCRs in kanker door pan-kanker gerelateerde mutaties te analyseren in vergelijking met natuurlijke variatie. De resultaten in dit hoofdstuk en de daaropvolgende zijn gebaseerd op een kankerpatiëntendataset die is gemaakt voor dit proefschrift uit de Genomic Data Commons (GDC) database, die beschikbaar is gesteld voor openbaar gebruik en computationeel vriendelijk en versie-stabiel is. Mutaties die verrijkt zijn in kanker en zich bevinden in functioneel bewaarde motieven worden beschouwd als prioriteiten bij het identificeren van 52 GPCRs als potentiële antikankertargets met behulp van een multi-objectieve optimalisatie aanpak. Deze aanpak maakt het ook mogelijk om praktische doelen, zoals aanvullende computationele en experimentele middelen, op te nemen en kan verder worden geïntegreerd met SB-analyses voor specifieke receptoren van interesse, inclusief de methoden die worden beschreven in Hoofdstuk 6.

Hoofdstuk 6 behandelt mutant-prioritering voor de glutamaat transporter EAAT1, een lid van de SLC-familie. Kankergerelateerde mutaties uit de GDC-dataset, gevonden in de buurt van de orthostere en allosterie bindingsplaats, worden computationeel getest om hun impact op eiwitconformatie en functie te beoordelen. Moleculaire dynamica (MD) simulaties en docking experimenten suggereren dat bepaalde kankergerelateerde

mutaties, met name R479W, een conformationele verandering veroorzaken die kan worden benut in gepersonaliseerde oncologie. Daarnaast demonstreert dit hoofdstuk de overdraagbaarheid van computationele bevindingen naar de echte wereld door middel van *in vitro* experimentele validatie van de effecten van de mutaties op transportfunctie en respons op farmacologische interventie.

Medicijnprioritering wordt besproken in **Hoofdstuk 7**, waarin een methode wordt geïntroduceerd om virtuele screening van grote bibliotheken van kandidaat-medicijnen voor mutante GPCRs te verbeteren. Dit hoofdstuk presenteert de creatie van nieuwe 3D dynamische eiwitdescriptoren (3DDPDs) op basis van MD simulaties om de representatie van gemuteerde eiwitten voor proteochemometrics bioactiviteitsmodellering te verbeteren. Resultaten tonen aan dat deze nieuwe descriptoren beter presteren dan sequentie-gebaseerde descriptoren in wild-type GPCR bioactiviteitsmodellering. Echter, evaluatie van hun toepasbaarheid in gemuteerde GPCRs wacht nog op verbetering van de beperkingen in gegevens beschikbaarheid die worden besproken in Hoofdstuk 4.

De lessen uit Hoofdstukken 4-7 culminereren in **Hoofdstuk 8**, waar een holistische benadering wordt gevolgd om alle eerder besproken gegevenssoorten te integreren. In dit hoofdstuk wordt een patiëntgerichte kennisgrafiek ontwikkeld met als doel het prioriteren van gemuteerde eiwitten voor gerichte therapie bij kanker. Deze benadering combineert de structurele en bioactiviteitsgegevens analyses uit Hoofdstuk 4 evenals de kanker- en natuurlijke variatiegegevens uit Hoofdstuk 5. Daarnaast bouwt het voort op de concepten die zijn verkend in Hoofdstukken 4-7 om doelen en mutaties te prioriteren die functioneel, structureel en klinisch relevant zijn. Vanwege beperkingen in gegevensbeschikbaarheid ligt de focus van dit hoofdstuk voornamelijk op kinasen, specifiek RTKs. Echter, net als de voorgaande hoofdstukken, is het ontworpen om toepasbaar te zijn in elk eiwittype als de noodzakelijke gegevens in de toekomst beschikbaar komen. Geavanceerde modellering algoritmen zouden ook kunnen worden gebruikt om de kennisgrafiek te verbeteren naarmate meer gegevens beschikbaar komen.

Ten slotte biedt **Hoofdstuk 9** een samenvatting van de conclusies die zijn getrokken uit de voorgaande hoofdstukken in de bredere context van computationele oncologische medicijnondekking. Over het algemeen breiden de methoden die in dit proefschrift zijn ontwikkeld het scala aan beschikbare hulpmiddelen uit voor het selecteren van nieuwe doeleiwitten, mutanten en medicijn kandidaten voor gepersonaliseerde oncologische toepassingen. Om echter klinische significantie te bereiken, moeten er verder samengewerkt worden binnen de wetenschappelijke gemeenschap om te focussen op kankeronderzoeksinitiatieven. Computationele hulpmiddelen zoals die in dit proefschrift ontwikkeld zijn kunnen uiterst nuttig zijn voor taken zoals het ontwerpen en implementeren van machine-leesbare open-source kankerdatabases, het identificeren van belangrijke biomarkers voor diagnose en gepersonaliseerde behandeling, het voorspellen van optimale behandelingsstrategieën en het prioriteren van belangrijke onderzoeksgebieden. Zo kunnen door samenwerking met andere wetenschappers effectieve en veilige behandelingen worden ontwikkeld voor alle patiënten die tegen kanker vechten.

Resumen

El cáncer se considera la pandemia silenciosa del siglo XXI y la segunda causa principal de muerte en todo el mundo. La heterogeneidad que lo caracteriza, la cual se manifiesta entre distintos tipos de cáncer, individuos e incluso células tumorales, hace que sea extremadamente difícil de tratar de manera efectiva y segura en todos los pacientes. La oncología personalizada ha surgido como una estrategia eficiente para aprovechar las diferencias presentes en el cáncer y así atacar específicamente las células tumorales. Este enfoque tiene como objetivo reducir los efectos secundarios mientras mantiene o mejora la eficacia terapéutica. Sin embargo, la disponibilidad de terapias personalizadas es actualmente limitada, lo que deja a muchos pacientes con cáncer deseando poder optar a tratamientos más selectivos o específicos. En este contexto, las herramientas computacionales juegan un papel crucial en la exploración de preguntas no resueltas en la investigación del cáncer y en la aceleración del descubrimiento de nuevas proteínas que pueden ser atacadas selectivamente en terapias anticancerígenas como dianas terapéuticas. Una ventaja principal del uso de herramientas computacionales es la capacidad de investigar familias de proteínas prometedoras que han sido pasadas por alto en la investigación del cáncer debido a limitaciones experimentales o de sesgo de publicación, como las proteínas de membrana. Esta tesis profundiza en el potencial de las herramientas computacionales en la priorización de nuevas dianas terapéuticas, mutaciones y candidatos a fármacos para su uso en la oncología personalizada, con un enfoque específico en las proteínas de membrana.

Este concepto se introduce primero en el **Capítulo 1**, donde se vinculan los tres niveles de priorización con la relevancia funcional, la capacidad de ser atacadas con medicamentos, la potencia de la terapia, su selectividad y la posibilidad de generar resistencias. En este capítulo se delinean las principales promesas y desafíos de la oncología personalizada, seguidos de una visión general de los métodos computacionales utilizados en el descubrimiento de fármacos que pueden extrapolarse a la investigación oncológica. En particular, se introduce cómo estos métodos pueden aplicarse al estudio de las proteínas de membrana como dianas terapéuticas anticancerígenas prometedoras pero experimentalmente complejas. Estos conceptos se detallan a lo largo de esta tesis.

El **Capítulo 2** revisa la amplia gama de herramientas computacionales que pueden aplicarse en el descubrimiento de fármacos oncológicos. El enfoque principal de este capítulo está en dos categorías: inteligencia artificial (IA) y métodos basados en la estructura (SB). Estas dos categorías se describen de manera independiente, pero se destaca el potencial sinérgico de su combinación, especialmente en el contexto de la investigación del cáncer, la cual requiere soluciones multidisciplinarias. Al revisar una selección de aplicaciones combinadas en cáncer, el lector obtiene una comprensión del potencial de las metodologías desarrolladas y aplicadas a lo largo de esta tesis.

Las aplicaciones discutidas en el Capítulo 2 se centran principalmente en dianas terapéuticas anticancerígenas ampliamente consolidadas, en particular las proteínas citosolubles. Sin embargo, ampliar el rango de dianas anticancerígenas es crucial para expandir el acceso a los tratamientos de oncología personalizada a una población más

amplia. El **Capítulo 3** enfatiza el uso de proteínas de membrana como posibles nuevas dianas anticancerígenas. Estas pueden ser exploradas usando herramientas computacionales para abordar los desafíos experimentales que las hacen menos atractivas de estudiar en comparación con las proteínas solubles. Este capítulo también identifica los principales desafíos en el descubrimiento computacional de fármacos para proteínas de membrana, que están principalmente relacionados con la disponibilidad de datos y el sesgo de publicación. En este contexto, se destacan y examinan tres familias de proteínas con diferentes niveles de representación en la literatura: receptores tirosina-cinasa (RTKs), receptores acoplados a proteínas G (GPCRs) y transportadores de solutos (SLCs).

El **Capítulo 4** explora en más detalle las diferencias en la disponibilidad de datos entre familias de proteínas y dianas a nivel individual. Este capítulo resalta la fuerte correlación entre el sesgo de publicación y los datos presentes en bases de datos públicas, específicamente los datos de bioactividad relacionados con proteínas mutantes o variantes genéticas en la base de datos ChEMBL. Estos datos, vitales para el descubrimiento de fármacos oncológicos, están mayoritariamente presentes en dianas anticancerígenas y variantes genéticas de relevancia establecida, en particular cinasas y RTKs. El capítulo enfatiza la importancia de estos datos en el descubrimiento computacional de fármacos a través de la evaluación comparativa de modelos de predicción de bioactividad teniendo o no en cuenta la variabilidad genética. Estos modelos pueden ser utilizados en el descubrimiento de fármacos oncológicos si se dan las circunstancias adecuadas. Además, el capítulo ofrece herramientas de datos y recomendaciones para ayudar en la preparación de conjuntos de datos con variantes de alta calidad para la predicción de bioactividad.

Los Capítulos 5-7 se centran en el desarrollo de varias aplicaciones computacionales para abordar los tres niveles de priorización introducidos en el Capítulo 1. Estas aplicaciones utilizan los métodos introducidos en el Capítulo 2 y se aplican a las familias de proteínas de membrana menos exploradas introducidas en el Capítulo 3.

La priorización de dianas terapéuticas se trata en el **Capítulo 5** con enfoque en la familia de GPCRs. Este capítulo evalúa la relevancia funcional de GPCRs en el cáncer mediante el análisis de mutaciones relacionadas con el cáncer en comparación con la variación natural. Los resultados en este capítulo y los siguientes se basan en un conjunto de datos de pacientes con cáncer creado para esta tesis a partir de la base de datos Genomic Data Commons (GDC), que está disponible para su uso público. Las mutaciones más comunes en cáncer y localizadas en motivos conservados funcionalmente se consideran objetivos prioritarios para identificar 52 GPCRs como posibles dianas anticancerígenas utilizando un enfoque de optimización multiobjetivo. Este enfoque también permite la inclusión de objetivos prácticos, como la disponibilidad de recursos computacionales y experimentales adicionales, y puede integrarse con análisis SB para receptores específicos de interés, como por ejemplo los métodos descritos en el Capítulo 6.

El **Capítulo 6** cubre la priorización de mutantes para el transportador de glutamato EAAT1, un miembro de la familia SLC. Las mutaciones relacionadas con el cáncer del conjunto de datos GDC encontradas cerca de los lugares de unión ortostéricos (sitio activo) y alostéricos se evalúan computacionalmente para valorar su impacto en la conformación y función de la proteína. Las simulaciones de dinámica molecular (MD) y los

experimentos de acoplamiento (docking) sugieren que ciertas mutaciones relacionadas con el cáncer, específicamente R479W, inducen un cambio conformacional que puede ser aprovechado en oncología personalizada. Además, este capítulo demuestra la transferibilidad de los hallazgos computacionales a aplicaciones del mundo real a través de la validación experimental *in vitro* de los efectos de estas mutaciones en la función del transportador y su respuesta a intervención farmacológica.

La priorización de moléculas como fármacos se desarrolla en el **Capítulo 7**, que introduce un método desarrollado para mejorar el cribado virtual de grandes bibliotecas de candidatos a medicamentos que ataquen variantes genéticas de GPCRs. Este capítulo presenta la creación de nuevos descriptores de proteínas que son 3D y dinámicos (3DDPDs), que se basan en simulaciones MD para mejorar la representación de proteínas mutantes en la modelización proteoquimométrica de bioactividad. Los resultados muestran que estos nuevos descriptores superan a los descriptores basados en secuencias proteicas en la modelización de bioactividad de GPCRs no mutados. Sin embargo, la evaluación de su aplicabilidad en GPCRs mutantes está pendiente debido a las limitaciones de disponibilidad de datos discutidas en el Capítulo 4.

Las lecciones de los Capítulos 4-7 encuentran su culminación en el **Capítulo 8**, donde se adopta un enfoque holístico para integrar todos los tipos de datos discutidos previamente. En este capítulo se desarrolla un grafo de conocimiento (knowledge graph) centrado en el paciente con el objetivo de priorizar proteínas mutadas para la terapia dirigida en el cáncer. Este enfoque combina los análisis de datos estructurales y de bioactividad del Capítulo 4, así como los datos de cáncer y variación natural del Capítulo 5. Además, se basa en los conceptos explorados en los Capítulos 4-7 para priorizar dianas terapéuticas y mutaciones que sean funcional, estructural y clínicamente relevantes. Debido a las limitaciones en la disponibilidad de datos, el enfoque de este capítulo se centra principalmente en cinasas, específicamente RTKs. Sin embargo, al igual que los capítulos anteriores, está diseñado para amoldarse a cualquier tipo de proteína si los datos necesarios están disponibles en el futuro. A medida que haya más datos disponibles, también se podrían utilizar algoritmos de modelización avanzados.

Finalmente, el **Capítulo 9** ofrece un resumen de las conclusiones extraídas de los capítulos anteriores en el contexto más amplio del descubrimiento computacional de fármacos oncológicos. En general, los métodos desarrollados en esta tesis amplían la gama de herramientas disponibles para seleccionar nuevas dianas terapéuticas, mutantes y candidatos a medicamentos para aplicaciones de oncología personalizada. Sin embargo, para lograr una significancia clínica, se deben mantener los esfuerzos de colaboración dentro de la comunidad científica para enfocar las iniciativas de investigación del cáncer. Las herramientas computacionales similares a las desarrolladas en esta tesis pueden ser extremadamente beneficiosas para tareas como diseñar e implementar bases de datos de cáncer de código abierto, identificar biomarcadores clave para el diagnóstico y tratamiento personalizado, predecir estrategias de tratamiento óptimas y priorizar áreas clave de investigación. En última instancia, solo a través de esfuerzos colaborativos y enfocados se pueden desarrollar tratamientos efectivos y seguros para todos los pacientes que luchan contra el cáncer.

List of publications

Part of this thesis

1. Gorostiola González, M., Janssen, A.P.A., IJzerman, A.P., Heitman, L.H. & van Westen, G.J.P. Oncological drug discovery: AI meets structure-based computational research. *Drug Discovery Today*, **27**, 1661–1670 (2022).
2. Gorostiola González, M., Rakers, P., Jespers, W., IJzerman, A.P., Heitman, L.H. & van Westen, G.J.P. Computational characterization of membrane proteins as anticancer targets: Current challenges and opportunities. *International Journal of Molecular Sciences*, **25**, 3698 (2024).
3. Gorostiola González, M.[†], Béquignon, O.J.M.[†], Manners, J.M., Zdrazil, B., Leach, A.R., IJzerman, A.P., Heitman, L.H. & van Westen, G.J.P. Excuse me, there is a mutant in my bioactivity soup! A comprehensive analysis of the genetic variability landscape of bioactivity databases and its effect on activity modelling. Preprint at *ChemRxiv* <https://doi.org/10.26434/chemrxiv-2024-kxlgm> (2024).
4. Bongers, B.J.[†], Gorostiola González, M.[†], Wang, X., van Vlijmen, H.W.T., Jespers, W., Gutiérrez-de-Terán, H., Ye, K., IJzerman, A.P., Heitman, L.H. & van Westen, G.J.P. Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors. *Scientific Reports*, **12**, 21534 (2022).
5. Gorostiola González, M.[†], Sijben, H.[†], Dall'Acqua, L., Liu, R., IJzerman, A.P., Heitman, L.H. & van Westen, G.J.P. Molecular insights into disease-associated glutamate transporter (EAAT1 / *SLC1A3*) variants using *in silico* and *in vitro* approaches. *Frontiers in Molecular Biosciences*, **10**, 3389 (2023).
6. Gorostiola González, M., van den Broek, R.L., Braun, T.G.M., Chatzopoulou, M., Jespers, W., IJzerman, A.P., Heitman, L.H. & van Westen, G.J.P. 3DDPDs: describing protein dynamics for proteochemometric bioactivity prediction. A case for (mutant) G protein-coupled receptors. *Journal of Cheminformatics* **15**, 74 (2023).
7. Gorostiola González, M., IJzerman, A.P. & van Westen, G.J.P. A patient-centric knowledge graph approach to prioritize mutants for selective anti-cancer targeting. Preprint at *BioRxiv* <https://doi.org/10.1101/2024.09.29.615658> (2024).

[†] These authors contributed equally

Other publications

8. Dilweg, M.A., Gorostiola González, M., de Ruiter, M.D., Meijboom, N.J., van Veldhoven, J.P.D., Liu, R., Jespers, W., van Westen, G.J.P., Heitman, L.H., IJzerman, A.P. & van der Es, D. Exploring novel dilazep derivatives as hENT1 inhibitors and potentially covalent molecular tools. *Purinergic Signalling*, online ahead of printing (2024).
9. van den Maagdenberg, H.W., Sicho, M., Alencar Araripe, D., Luukkonen, S.,

- Scoenmaker, L., Jespers, M., Béquignon, O.J.M., Gorostiola González, M., van den Broek, R.L., Bernatavicius, A., van Hasselt, J.G.C., van der Graaf, P.H. & van Westen, G.J.P. QSPRpred: a Flexible Open-Source Quantitative Structure-Property Relationship Modelling Tool. Preprint at *ChemRxiv* <https://doi.org/10.26434/chemrxiv-2024-m9989> (2024).
10. van Veggel, L., Mocking, T.A.M., Sijben, H.J., Liu, R., Gorostiola González, M., *et al.* Still in Search for an EAAT Activator: GT949 Does Not Activate EAAT2, nor EAAT3 in Impedance and Radioligand Uptake Assays. *ACS Chemical Neuroscience*, **15**, 1424-1431 (2024).
 11. Mullooney, M.W., Duncan, K.R., Elsayed, S.S. *et al.* Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery*, **22**, 895-916 (2023).
 12. den Hollander, L.S., Béquignon, O.J.M., Wang, X., van Wezel, K., Broekhuis, J., Gorostiola González, M., de Visser, K.E., IJzerman, A.P., van Westen, G.J.P. & Heitman, L.H. Impact of cancer-associated mutations in CC chemokine receptor 2 on receptor function and antagonism. *Biochemical Pharmacology*, **208**, 115399 (2023).
 13. Feng, C., Wang, X., Jespers, W., Liu, R., Zamarbide Losada, S.D., Gorostiola González, M., van Westen, G.J.P., Danen, E.H.J. & Heitman, L.H. Cancer-Associated Mutations of the Adenosine A2A Receptor Have Diverse Influences on Ligand Binding and Receptor Functions. *Molecules*, **27**, 4676 (2022).
 14. Kovacikova, K., Gorostiola González, M., Jones, R., Reguera, J., Gigante, A., Pérez-Pérez, M.J., Pürstinger, G., Moesslacher, J., Langer, T., Jeong, L.S., Delang, L., Neyts, J., Snijder, E.J., van Westen, G.J.P. & van Hemert M.J. Structural insights into the mechanisms of action of functionally distinct classes of Chikungunya virus nonstructural protein 1 inhibitors. *Antimicrobial Agents and Chemotherapy*, **65**, e02566-20 (2021).
 15. Burggraaff, L., Lenselink, E.B., Jespers, W., van Engelen, J., Bongers, B.J., Gorostiola González, M., Liu, R., Hoos, H.H., van Vlijmen, H.W.T., IJzerman, A.P. & van Westen, G.J.P. Successive statistical and structure-based modeling to identify chemically novel kinase inhibitors. *Journal of Chemical Information and Modeling*, **60**, 4283-4295 (2020).

Scientific communications

2024	Excuse me, there is a mutant in my bioactivity soup! A comprehensive analysis of the genetic variability landscape of bioactivity databases and its effect on activity modelling (<i>poster</i>). 24th EuroQSAR . Barcelona, Spain.
2023	Analysis of cancer mutations directed towards rational design of selective inhibitors for RTKs (<i>poster</i>). 9th Joint Sheffield Conference on Cheminformatics . Sheffield, UK. Assessment of cancer-related glutamate transporter (EAAT1 / <i>SLC1A3</i>) mutants using a combination of <i>in vitro</i> and <i>in silico</i> approaches (<i>poster</i>). ONCODE Annual Scientific Meeting . Amersfoort, the Netherlands. Accelerating personalized oncology: AI and structural methods in oncological drug discovery (<i>oral</i>). LACDR Spring Symposium . Leiden, the Netherlands.
2022	Pan-cancer analysis of somatic mutations in G protein-coupled receptors (<i>poster</i>). ULLA Summer School . Uppsala, Sweden. Describing protein dynamics for proteochemometric bioactivity prediction: 3DDPDs (<i>oral, selected</i>). 12th International Conference on Chemical Structures (ICCS) . Noordwijkerhout, the Netherlands. Assessment of cancer-related glutamate transporter (EAAT1 / <i>SLC1A3</i>) mutants using a combination of <i>in vitro</i> and <i>in silico</i> approaches (<i>poster</i>). FIGON Dutch Medicines Day & EUFEPS Annual meeting . Leiden, the Netherlands. Assessment of cancer-related glutamate transporter (EAAT1 / <i>SLC1A3</i>) mutants using a combination of <i>in vitro</i> and <i>in silico</i> approaches (<i>poster</i>). LACDR Spring Symposium . Leiden, the Netherlands.
2021	Pan-cancer analysis of somatic mutations in G protein-coupled receptors (<i>poster</i>). ONCODE Annual Scientific Meeting . Online. Pan-cancer analysis of somatic mutations in G protein-coupled receptors (<i>poster</i>). LACDR Spring Symposium . Online.
2020	Analysis of cancer mutations directed towards rational design of selective inhibitors for RTKs (<i>poster</i>). FIGON Dutch Medicines Day . Online. Analysis of cancer mutations directed towards rational design of selective inhibitors for RTKs (<i>poster</i>). LACDR Spring Symposium . Online. <i>Awarded poster prize.</i>

Curriculum Vitae

Marina Gorostiola González was born on November 20th, 1994 in Medina de Pomar, Spain. After graduating with honors in 2012 from the International Baccalaureate program at Cardenal López de Mendoza in Burgos, she started her university studies in the city of Salamanca, Spain. In 2017, she graduated with honors from the five-year BSc and MSc program in Pharmacy at the Faculty of Pharmacy of the University of Salamanca. In addition to her academic pursuits, Marina served as a class representative in the student union. During her studies, she explored different career options with curricular and extracurricular training in clinical pharmacy at the Valdecilla University Hospital in Santander, Spain, and at the Mount Carmel Hospital in Attard, Malta; community pharmacy at her mother's pharmacy in Burgos, Spain; and research at the Cancer Research Center, as well as the department of Pharmaceutical Sciences at the Faculty of Pharmacy in Salamanca, Spain. It was at this department that during her fifth year, Marina performed the first computational research project of her design supervised by Prof. dr. Maria José García Sánchez and Prof. dr. María Dolores Santos Buelga to investigate *in silico* drug absorption in celiac patients, which was later published in the journal *FarmaJournal*, edited by the University of Salamanca.

In September 2017, Marina started her MSc in Bio-Pharmaceutical Sciences specializing in computational drug discovery research at the University of Leiden, the Netherlands. She first performed a nine-month internship at the Division of Drug Discovery and Safety (now Division of Medicinal Chemistry), under the supervision of Prof. dr. Gerard van Westen and Dr. Lindsey Burggraaff in the computational drug discovery (CDD) group. This work entailed the computational prioritization of cancer-related kinase mutants to target selectively and inspired this thesis. During her time in the CDD group, Marina also participated in the “Multi-Targeting Drug” DREAM challenge, culminating in her first contribution to a computational drug discovery publication. As part of her MSc program, Marina wrote a literature review under the supervision of Dr. Joost Beltman and took several courses on Pharmaceutical bioinformatics from Uppsala University. Finally, she performed a six-month internship at the Molecular Modelling & Design Department in Galapagos NV, Belgium, under the supervision of Dr. Nicolas Triballeau and Dr. Bart Lenselink. This project focused on the development of a 3D machine learning-based scoring function for kinase-ligand interactions. In September 2019, Marina graduated *cum laude* from her MSc at Leiden University.

In December 2019, Marina started her PhD in computational drug discovery at the Division of Drug Discovery and Safety at Leiden University under the supervision of Prof. dr. Gerard van Westen, Prof. dr. Laura Heitman, and Prof. dr. Ad IJzerman. This PhD position was supported by the Oncode Institute, an independent research institute dedicated to cancer research in the Netherlands. This project focused on developing methods to characterize membrane proteins as anticancer targets using a combination of AI and structure-based computational tools. During her PhD research project, Marina presented her work at several national and international conferences both as poster and oral presentations. These included Oncode meetings and FIGON Dutch Medicine Days in the Netherlands, the International Conference on Chemical Structures (ICCS) in the

Netherlands in 2022, and the Joint Sheffield Conference on Cheminformatics in the UK in 2023. She also participated in the Lorentz Workshop in AI for natural product drug discovery in the Netherlands in 2021 and the ULLA summer school in Sweden in 2023. As part of her PhD, Marina also supervised several bachelor and master students in their research projects and initiated several national and international collaborations.

Marina continues her career in computational drug discovery as a researcher at Chemotargets in Barcelona, Spain.

Acknowledgments

Life is simply a series of random turns and events, and it is up to us to navigate our way through them effectively. I would not have reached this point without the support of the people who guided me to make the right decisions along the way, both scientifically and personally.

Thank you, Gerard, for giving me the space to start a career in computational drug discovery. One conversation with you was all it took. It just felt right.

I am extremely grateful for my promotion team. Challenging as it was to have three supervisors - Gerard, Laura, and Ad - you have significantly improved my scientific output and personal development. More importantly, you have taught me through your own experience the significance of collaborative work and the necessity to maintain a balanced personal-professional life. I cannot stress enough the exceptional job you are doing in promoting this and how crucial it is for young scientists to witness.

I would like to express my gratitude to all the colleagues and external collaborators who have played a significant role in both the development of this thesis and other related projects. Your contributions have been instrumental in enhancing my skills and fostering my scientific curiosity. A special thanks goes to Brandon, Olivier, Huub, and Willem for their invaluable input in this thesis.

A big shout-out to the students who performed their research under my supervision. Remco, Veerle, Thomas, Donald, Marit, Pepijn, and Magdalini, I hope I was able to inspire in you a love for science.

I would not have stayed sane during this PhD if it was not for the magnificent working environment fostered at the Medicinal Chemistry division. I have truly enjoyed scientific and social events alike. To all present and past colleagues: thank you for making it possible! Special mention to Brandon, Olivier, Majlen, Jara, Willem, Andrius, Sohvi, and David. I have no words to describe the magnificent people you are. Instrumental to my sanity were also all friends outside of University. Leiden became home thanks to you, and I will miss you deeply.

Mamá, papá, sois la razón y el propósito de esta tesis. Gracias por darme las alas para perseguir mis sueños y el apoyo incondicional para alcanzarlos. Todos mis éxitos son vuestros, siempre. *[Mom, Dad, you are the reason and purpose behind this thesis. Thank you for giving me the wings to pursue my dreams and for the unconditional support to achieve them. All of my successes are yours, always.]*

Finally, to my life partner, Konsta, a special thanks for enduring the ups and downs of my PhD journey. Your calm keeps me grounded and your confidence gives me the strength to step beyond my comfort zone. I could not have done this without you. Now, what's next?

