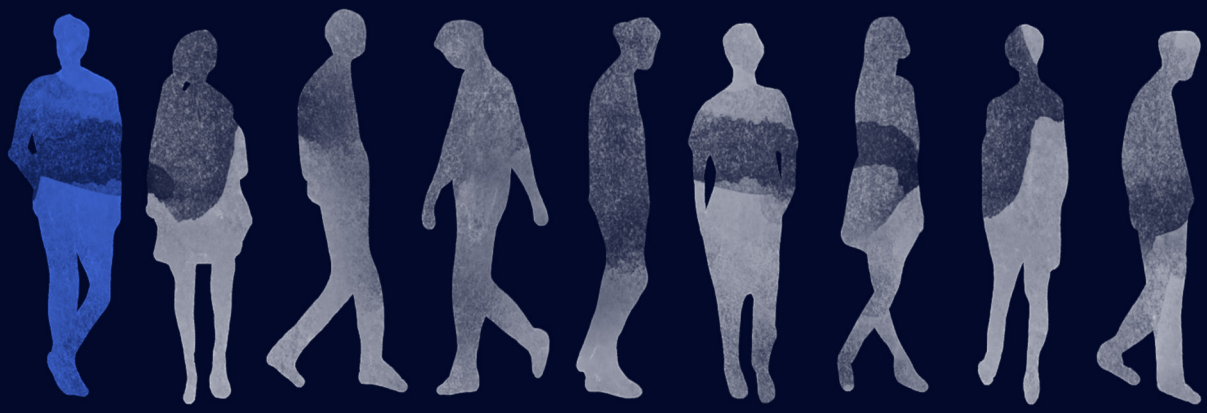
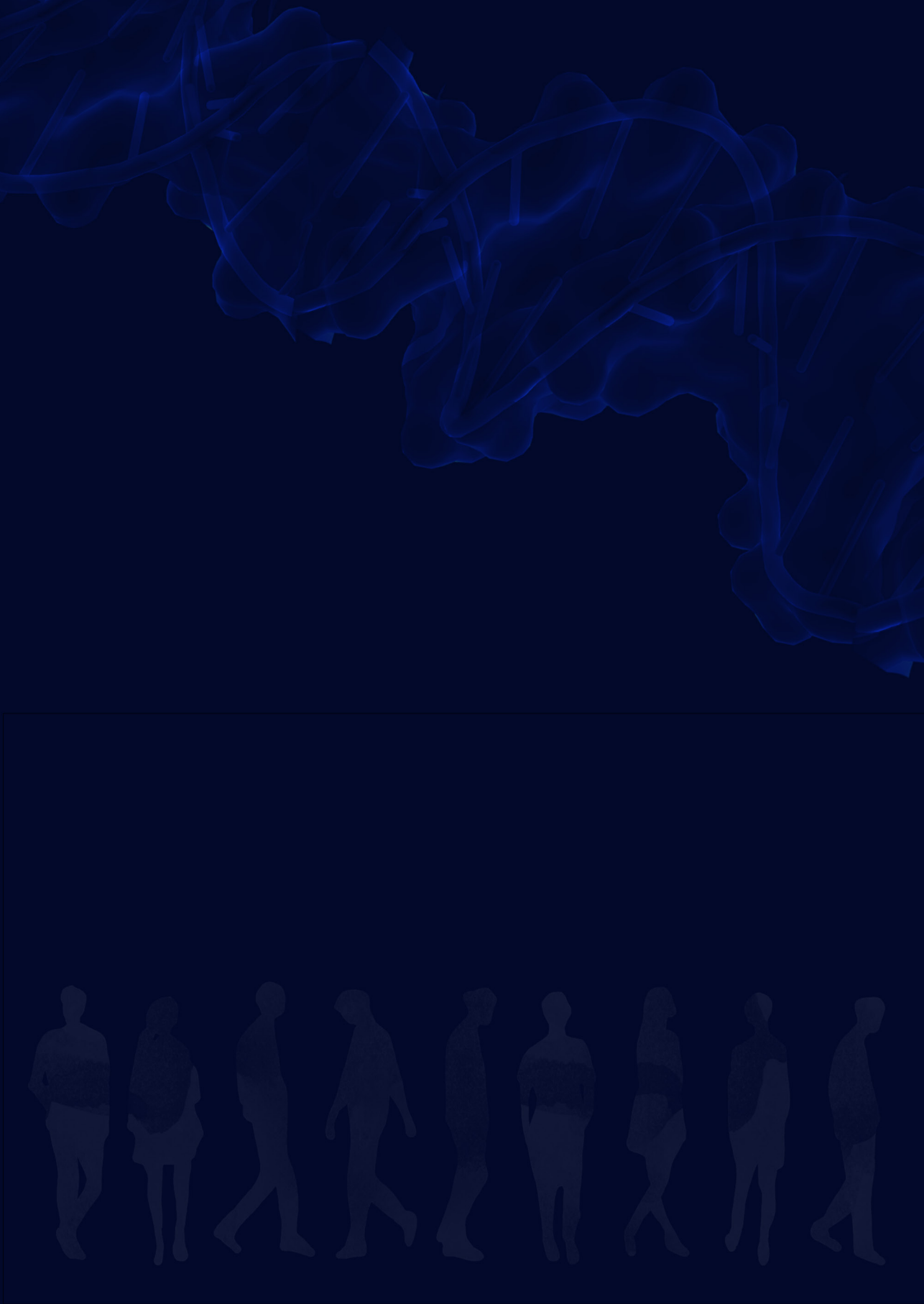
Getting personal: advancing personalized oncology through computational analysis of membrane proteins

Gorostiola González, M.

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 in[6]. 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 relapse[7–9].

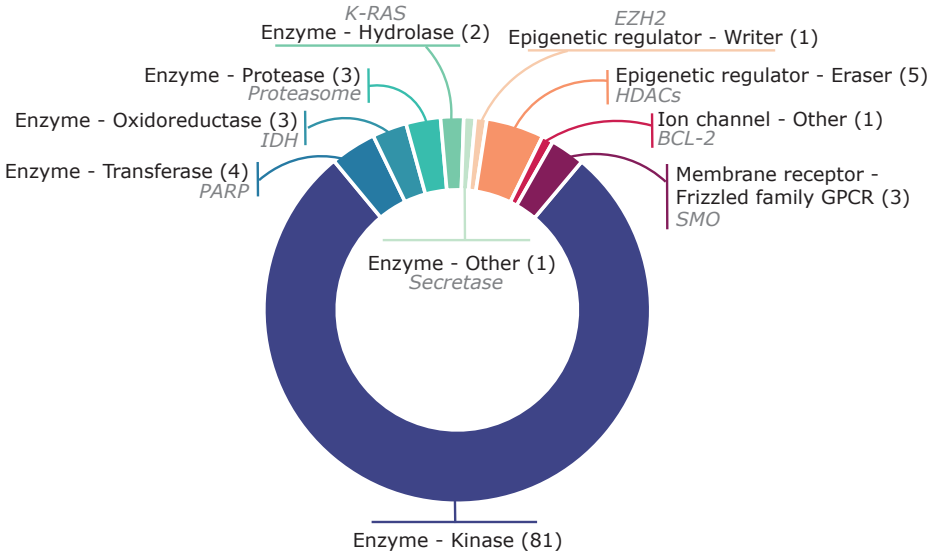


Figure 1.1. FDA-approved small molecule anticancer targeted agents from 2001 to 2023. The 104 approved drugs are distributed per target family according to the ChEMBL L1/L2 classification.

Personalized oncology comprises several therapeutical strategies that can be used when the patient meets certain specific profiling criteria[7]. 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 stage[8]. 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 vaccines[8]. 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 thesis[11]. Since the approval in 2001 of the first anticancer-targeted small molecule, imatinib, 104 small molecules have been approved for anticancer treatment[10]. However, while the eligibility of patients for targeted therapies is increasing, it was still estimated to be less than 15% in 2020[13].

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 limited[17]. 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 therapeutics[19]. 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 diseases[20]. 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 mutation[28]. 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 selectivity[29]. 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 molecules[30].

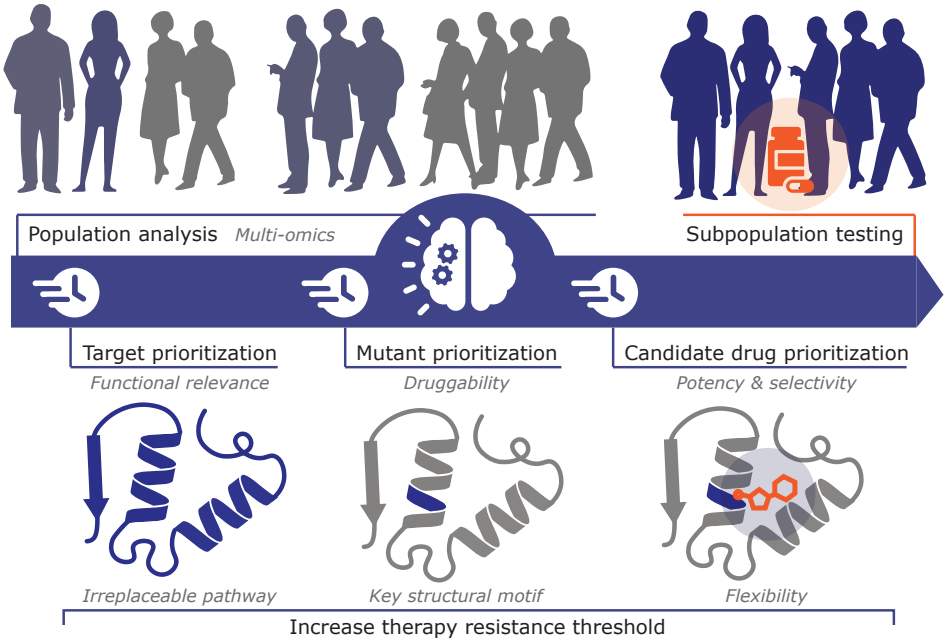


Figure 1.2. Three levels of computationally driven prioritization to accelerate personalized small molecule hit identification.

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 subpopulations[31]. 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 AlphaFold[34], 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 complexes[33].

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 targets[21]. 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 proteins[35].

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 candidates[39]. The role of certain protein membrane families in cancer, particularly receptor tyrosine kinases (RTKs) has been extensively highlighted[40]. 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 patients[40].

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 Smoothened (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 conservation[44]. 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 alike[47]. Subsequently, GPCRs are gaining interest as anticancer targets, with some inhibitors in clinical trials particularly as immunotherapy[48]. 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 community[49]. Among other substrates, SLCs transport metabolites, neurotransmitters, amino acids and ions, and their expression is dysregulated in several cancer types[50].

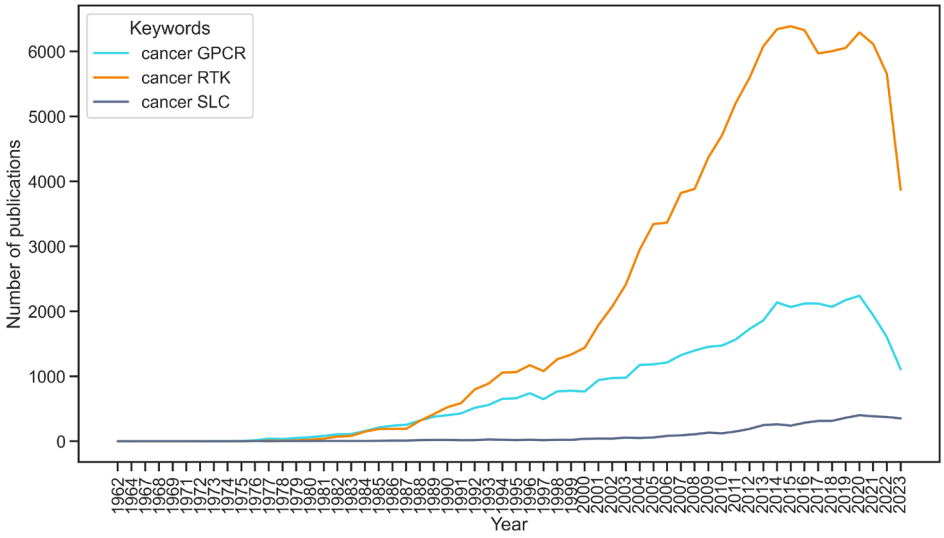


Figure 1.3. Number of publications retrieved from PubMed with the combination of keywords "cancer" and three membrane protein families: RTKs, GPCRs, and SLCs. Data was retrieved in November 2023, therefore the number of publications related to years 2020-2023 shows a drop corresponding to publication 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 Commons[51] 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

1. 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).
2. Debela, D. T. *et al.* New approaches and procedures for cancer treatment: Current perspectives. *SAGE Open Med.* **9**, 1–10 (2021).
3. Hashim, D. *et al.* The global decrease in cancer mortality: trends and disparities. *Ann. Oncol.* **27**, 926–933 (2016).
4. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
5. Wang, Y. & Feng, W. Cancer-related psychosocial challenges. *Gen. Psychiatry* **35**, e100871 (2022).
6. 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).
7. 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).
8. Krzyszczyk, P. *et al.* The growing role of precision and personalized medicine for cancer treatment. *Technology* **6**, 79–100 (2018).
9. Lassen, U. N. *et al.* Precision oncology: a clinical and patient perspective. *Future Oncol.* **17**, 3995–4009 (2021).
10. Zhong, L. *et al.* Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct. Target. Ther.* **6**, 1–48 (2021).
11. Min, H.-Y. & Lee, H.-Y. Molecular targeted therapy for anticancer treatment. *Exp. Mol. Med.* **54**, 1670–1694 (2022).
12. Waarts, M. R., Stonestrom, A. J., Park, Y. C. & Levine, R. L. Targeting mutations in cancer. *J. Clin. Invest.* **132**, e154943 (2022).
13. 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).
14. Mullard, A. 2021 FDA approvals. *Nat. Rev. Drug Discov.* **21**, 83–88 (2022).
15. Mullard, A. 2022 FDA approvals. *Nat. Rev. Drug Discov.* **22**, 83–88 (2023).
16. Mullard, A. 2023 FDA approvals. *Nat. Rev. Drug Discov.* **88**, 88–95 (2024).
17. O'Dwyer, P. J. *et al.* The NCI-MATCH trial: lessons for precision oncology. *Nat. Med.* **29**, 1349–1357 (2023).
18. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2020 update. *Pharmacol. Res.* **152**, 104609 (2020).
19. 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).
20. Rahman, M. M. *et al.* Emerging Promise of Computational Techniques in Anti-Cancer Research: At a Glance. *Bioengineering* **9**, 335 (2022).
21. 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).
22. Reyna, M. A. *et al.* Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **11**, 729 (2020).
23. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
24. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
25. Dinalankara, W. *et al.* Digitizing omics profiles by divergence from a baseline. *Proc. Natl. Acad. Sci.* **115**, 4545–4552 (2018).
26. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385 (2018).
27. Chitrala, 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).
28. 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).
29. 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).
30. Robichaux, J. P. *et al.* Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* **597**, 732–737 (2021).
31. Arjmand, B. *et al.* Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front. Genet.* **13**, 824451 (2022).
32. Qureshi, R. *et al.* AI in drug discovery and its clinical relevance. *Heliyon* **9**, e17575 (2023).
33. 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. & Haueis, 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).