Universiteit
Leiden
The Netherlands

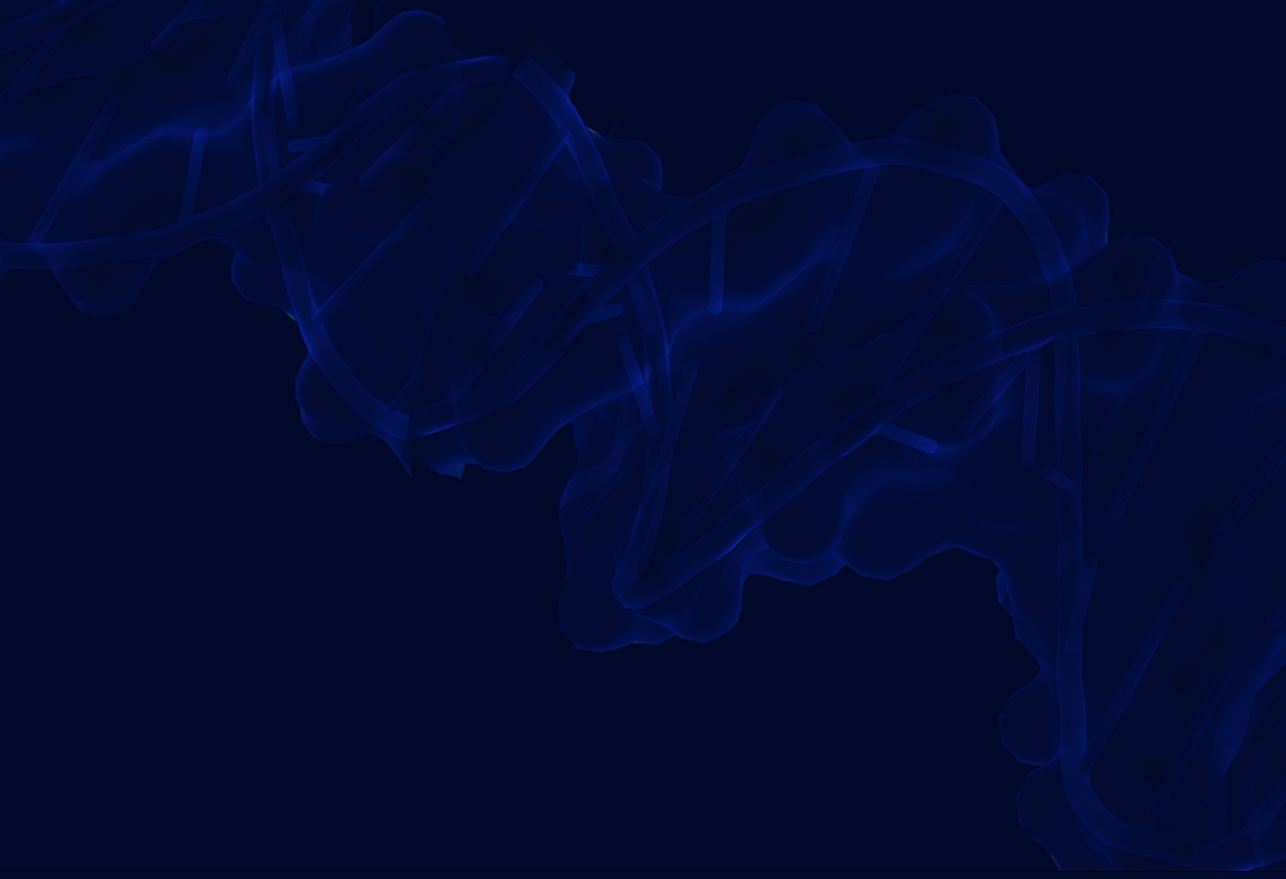
# Getting personal: advancing personalized oncology through computational analysis of membrane proteins

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).



## 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.