Universiteit
Leiden
The Netherlands

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

Gorostiola González, M.

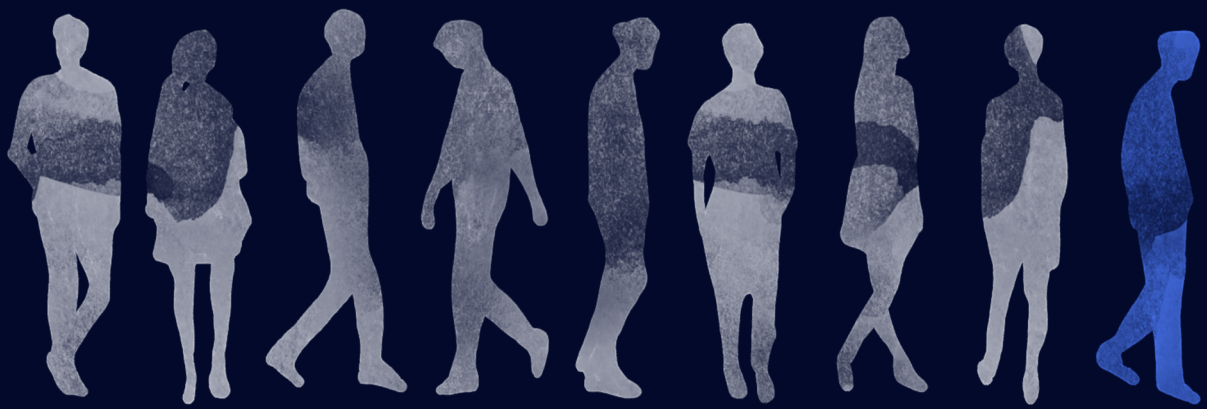
Citation
Gorostiola González, M. (2025, January 24). *Getting personal: advancing personalized oncology through computational analysis of membrane proteins*. Retrieved from https://hdl.handle.net/1887/4093962

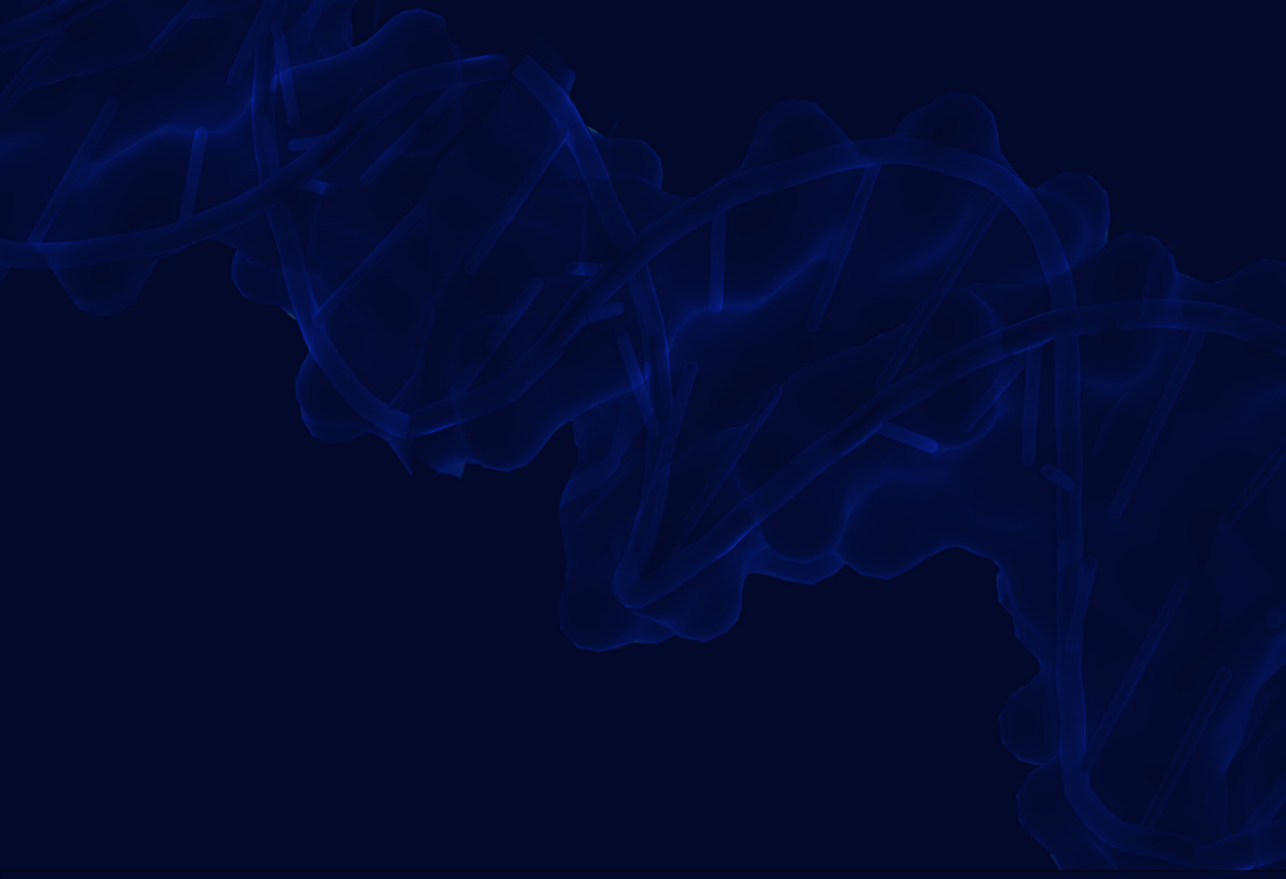
Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4093962

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

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 treatments[1]. Unfortunately, successful personalized targeted therapies are currently reaching too few patients, and the drug discovery pipeline is costly and slow[2]. Computational tools are crucial to accelerate the rate at which novel drugs make it to the market[3]. 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 haystack[4].

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 proteins[5–7]. 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 structures[3]. 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 database[12]. 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 data[13]. To this end, knowledge graphs are good representations to maximize the use of heterogeneous data[14], 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 community[15]. 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 families[21]. 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 practitioners[26]. 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 approaches[27], 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 candidacy[32]. 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, respectively[33]. As an additional filtering step, several models could be added to the pipeline to test *a priori* the potential pathogenicity of specific missense mutations[34]. 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 interventions[38]. 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 domains[4]. 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 analyzed[43]. 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 therapies[44–46]. 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 2016[47], and the European Union announced Europe's Beating Cancer Plan in 2021[48]. 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

capabilities[3], 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 goal[49]. 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 ownership[50], it could lead to a significant impact. Private-public funding will kickstart in the short term higher accessibility to personalized therapy clinical trials[32]. 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 project[19], 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 medicine[58]. 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 testing[59]. 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 interest[60].

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 action[61]. 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. PANACEA[62] 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 role[65]. Finally, these analyses should also extend to the design of clinical trials to ensure efficient patient treatment and optimize the likelihood of novel drug approval[32].

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 prioritized[66]. 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 subpopulations. These considerations must be integrated at both the data collection and computational model-building levels[67]. Additionally, building trust in the centralized storage of data, implementing proper blinding of the data for research[68], and enhancing trust among clinical practitioners in computational applications will be significant challenges[69]. 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

1. Lassen, U. N. *et al.* Precision oncology: a clinical and patient perspective. *Future Oncol.* **17**, 3995–4009 (2021).
2. 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).
3. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
4. 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).
5. Blum, A., Wang, P. & Zenklusen, J. C. SnapShot: TCGA-Analyzed Tumors. *Cell* **173**, 530 (2018).
6. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
7. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
8. 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).
9. Sharp, A. K. *et al.* Biophysical insights into OR2T7: Investigation of a potential prognostic marker for glioblastoma. *Biophys. J.* **121**, 3706–3718 (2022).
10. 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).
11. Chang, H. *et al.* Machine Learning-Based Virtual Screening and Identification of the Fourth-Generation EGFR Inhibitors. *ACS Omega* **9**, 2314–2324 (2024).
12. Rodriguez-Espigares, I. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020).
13. 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).
14. 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).
15. Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol. Brain* **13**, 24 (2020).
16. Béquignon, O. J. M. *et al.* Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J. Cheminformatics* **15**, 3 (2023).
17. Austin, B. K., Firooz, A., Valafar, H. & Blenda, A. V. An Updated Overview of Existing Cancer Databases and Identified Needs. *Biology* **12**, 1152 (2023).
18. Tanoli, Z. *et al.* Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief. Bioinform.* **22**, 1656–1678 (2020).
19. 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).
20. Isberg, V. *et al.* GPCRdb: An information system for G protein-coupled receptors. *Nucleic Acids Res.* **44**, D356–D364 (2016).
21. 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).
22. 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).
23. 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).
24. 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.
25. 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).
26. Price, W. N. Big Data and Black-Box Medical Algorithms. *Sci. Transl. Med.* **10**, eaao5333 (2018).
27. Tiddi, I. & Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **302**, 103627 (2022).
28. Renaux, A. *et al.* A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC Bioinformatics* **24**, 324 (2023).
29. 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. *AMIA 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. Schaduangrat, 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. Piñeiro-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).