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

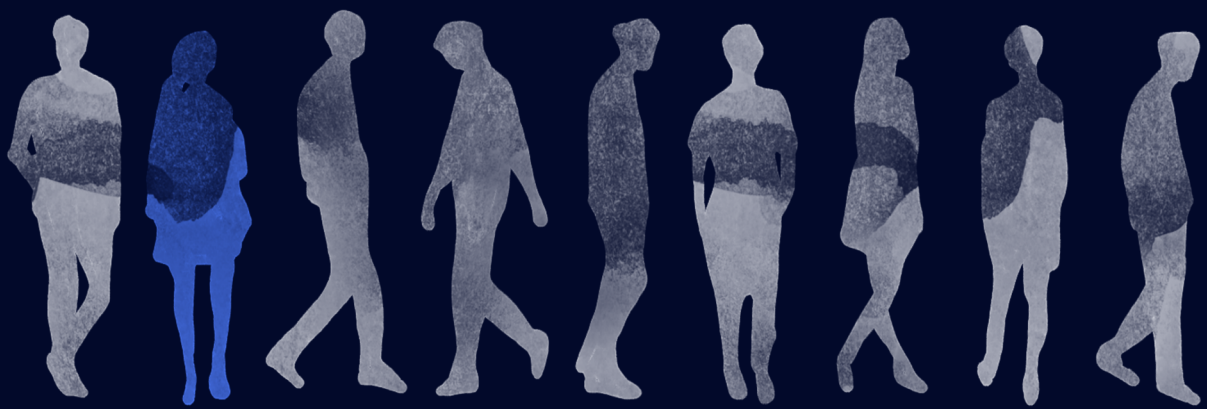
Gorostiola González, M.

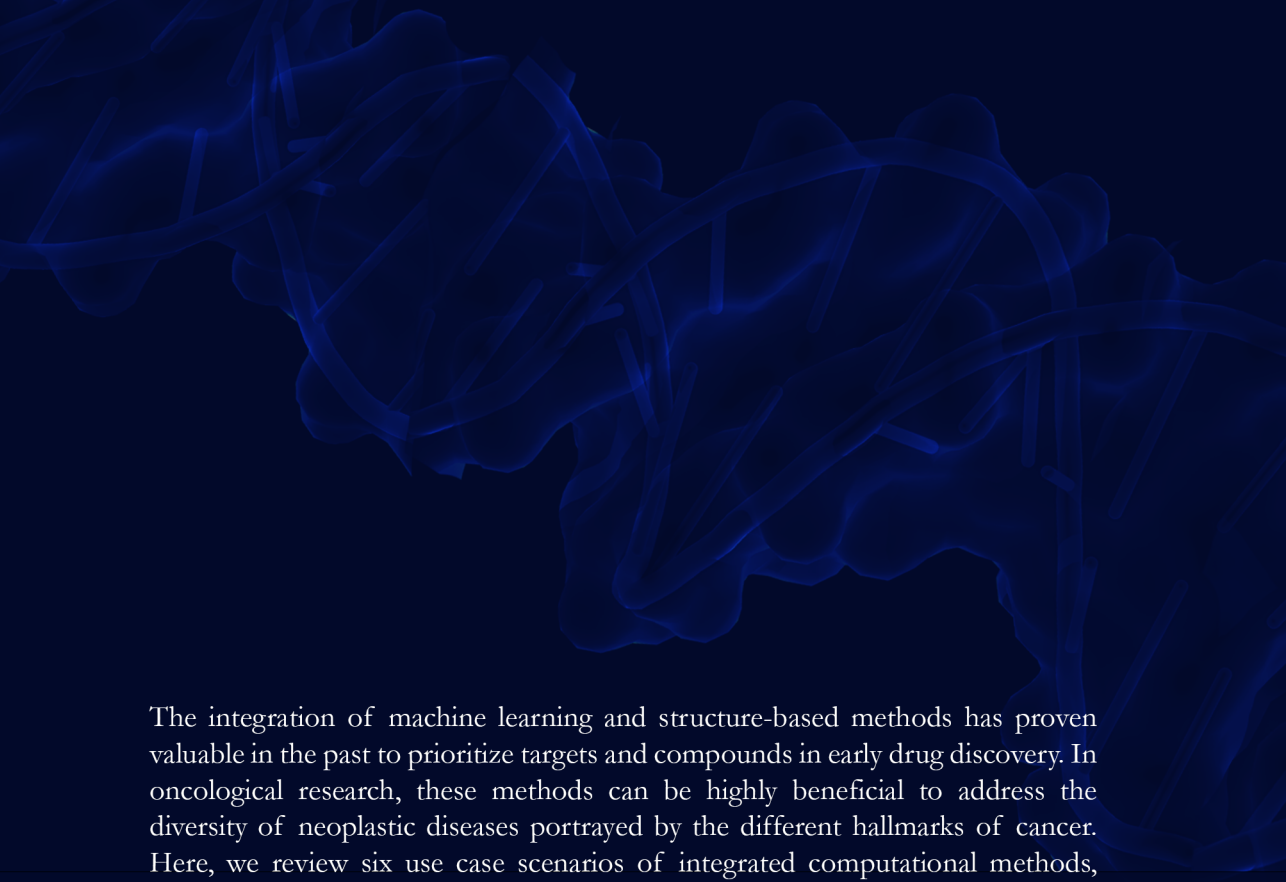
# Chapter 2

## 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 results[4] 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.

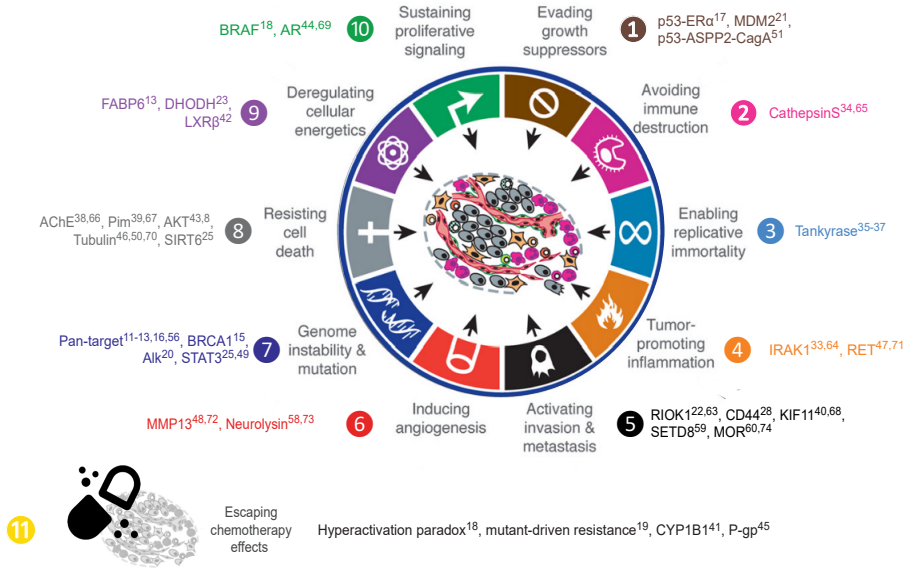
## Driver prediction

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 variations[12], epigenetic[12], or RNAseq[13] 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 characteristics[8]. 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.




Supporting references are cited for the target of each hallmark; references cited in italic text provide additional support on the connection of a certain target to a hallmark. Figure adapted from Hanahan and Weinberg[8].

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 Commons[14]. 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 studies[17], or MD[18]. 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 Chitrala *et al.*[17] for the p53-ERα 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 interactions[17], ligand binding affinity and dynamics[18,19], and clinical significance[15]. 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 interactions[23]. 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 drugs[23] or repurposing[22]. 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 Hoksza[24]. 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 Hoksza[24] 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 tools[25]. Some of these methods build on top of general binding site predictors with e.g. an added layer of ML classification[26]. 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 challenging[27]. Protein-protein interactions have been shown to be crucial in certain aspects of cancer pathogenicity[8]. 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 AlphaFold[31] 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 AlphaFill[32], 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 so[38]. One of the simplest approaches is to include docking scores as features for an ML classifier[33]. 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 function[35] but also MD-derived features[36]. 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-Baboli[39] 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 mapping[43], ΔG calculation with MM-GBSA[44]), 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 activity[50], or mode of action[45]. 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 *et al.* [11] | Pan-target / TCGA-MC3 set | 7 [11] | Various | Various | Ⓓ |
| Knijnenburg *et al.* [12] | Pan-target / TCGA-MC3 set | 7 [12] | Logistic regression classifier | FoldX, MD | Ⓓ |
| Liñares-Blanco *et al.* [13] | Pan-target (FABP6) / TCGA | 7 9 [13] | RF and generalized linear classifiers | Docking | Ⓓ |
| **Computational mutagenesis** | | | | | |
| Masso *et al.* [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 | Ⓐ |
| Chitrala *et al.* [17] | P53-ERα / NA | 1 [17] | One-layer NN | Protein-protein docking | Ⓐ |
| Babbitt *et al.* [18] | BRAF / FDA | 10 11 [8,18] | Seven stacked classifiers | MD | Ⓒ |
| Aldeghi *et al.* [19] | Abl / Platinum database, in-house set | 11 [19] | Extremely randomized regression trees | FEP | Ⓓ |
| Patil *et al.* [20] | Kinome (Alk) / UniProt, literature | 7 | SVM, RF, NeuralNet, LR | MD (metadynamics) | Ⓓ |
| **(Off)-target prediction** | | | | | |
| Pande *et al.* [21] | Pan-target (MDM2) / Literature | 1 [21] | CoMFA/CoMSIA PLS regressor, DT, RF, KNN, MLP, SVM classifiers | Docking, MD | Ⓐ |

**Table 2.1 (continues)**

| | | | | | |
|---|---|---|---|---|---|
| Lim *et al.* [22] | Pan-target (RIOK1, PDE3) / ChEMBL, DrugBank, literature datasets, TCGA-CCLE | 5 [63] | ElasticNet, SVR regressors | Ligand binding space search in genome, docking | Ⓐ |
| Zhi *et al.* [23] | DHODH / STRING, KEGG, ChEMBL, ZINC | 9 [23] | Multi-GNN | Docking, MD | Ⓓ |
| **Binding site prediction** | | | | | |
| Kawaguchi *et al.* [28] | CD44 / NA (pre-trained) | 5 [28] | Bayesian active learning | Protein-protein docking | Ⓑ |
| Taherzadeh *et al* [29] | Pan-target / BioLip | 1 (protein-protein binding) | RF classifier, DBSCAN | Structure-derived features | Ⓐ |
| **Virtual screening** | | | | | |
| Che *et al.* [33] | IRAK1 / ChEMBL, DUD-E | 4 [64] | SVM classifier | Docking | Ⓑ |
| Yang *et al.* [34] | Cathepsin S / PDBbind, CSAR, GC3/4, ChEMBL | 2 [65] | XGBoost regressor | Similarity-based docking | Ⓑ |
| Berishvili *et al.* [35–37] | Pan-target, Tankyrase / ZINC | 3 [37] | DNN | Docking, MD, FEP | Ⓑ |
| Adeshina *et al.* [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 *et al.* [40] | KIF11 / KEGG BRITE, DrugBank, STITCH | 5 [68] | Bayesian Additive Regression Trees | Bow-pharmacological space (protein-ligand interactions) | Ⓐ |
| Raju *et al.* [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 *et al.* [42] | LXRβ / ChEMBL, Binding DB, in-house library, GSMTL | 9 [42] | SVM, Naïve Bayes classifiers | Docking, MD | Ⓓ |
| Halder and Cordeiro [43] | AKT / ChEMBL, Asinex library | 8 [8] | LDA, XGBoost and other classifiers | MD | Ⓓ |
| Azhagiya Singam *et al.* [44] | AR / Tox21, CompTox | 10 [69] | SVM classifiers | Docking | Ⓓ |
| Kadioglu and Efferth [45] | P-gp / ChEMBL | 11 [45] | RF classifier | Docking | Ⓓ |
| Guo *et al.* [46] | Tubulin / ChEMBL | 8 [70] | Naïve Bayes classifiers | Docking, MD | Ⓓ |
| Burggraaff *et al.* [47] | RET / ChEMBL, ZINC | 4 [71] | RF classifiers | (Induced-fit) docking, metadynamics | Ⓓ |
| Chen *et al.* [48] | MMP13 / Traditional Chinese medicine database | 6 [72] | RF, gradient boosting, AdaBoost, deep learning | MD | Ⓓ |
| Chen *et al.* [49] | STAT3 / Literature set, ZINC | 7 [49] | Nine regressors, 3D QSAR | Docking, MD | Ⓓ |
| Guo *et al.* [50] | Tubulin / ChemDiv | 8 [70] | Discovery studio prediction models | Docking, MD | Ⓓ |
| Junaid *et al.* [51] | p53-ASPP2-CagA / Rationally designed | 1 [51] | ML module in MOE | MD | Ⓓ |
| **Allosteric modulation analysis** | | | | | |
| Lu *et al.* [25] | SIRT6, STAT3 / PDB, commercial | 8, 7 [25] | SVM | Geometric binding site predictor | Ⓐ |
| Song *et al.*[56] | Pan-target / PDB | 7 [56] | RF, neural networks | Structure-derived features | Ⓐ |
| Uyar *et al.* [58] | Neurolysin / PDB | 6 [73] | ElasticNet, PCA, LDA | MD | Ⓒ |
| Chen *et al.* [59] | SETD8 / cBioPortal | 5 [59] | Markov state model, tICA, clustering | MD | Ⓒ |
| Hu *et al.* [60] | MOR / Rationally designed | 5 [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 overfitting[52]. 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 compounds[53]. In that sense, the D-COID dataset[38] 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 counterparts[54], 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 prediction[26], allosteric interaction scoring[55], and allosteric effect analysis of mutations[56]. 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, respectively[25]. 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 AlloDriver[56], 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 decades[25]. Of the 19 currently FDA-approved allosteric modulators, three are indicated as anticancer drugs[61]. 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.

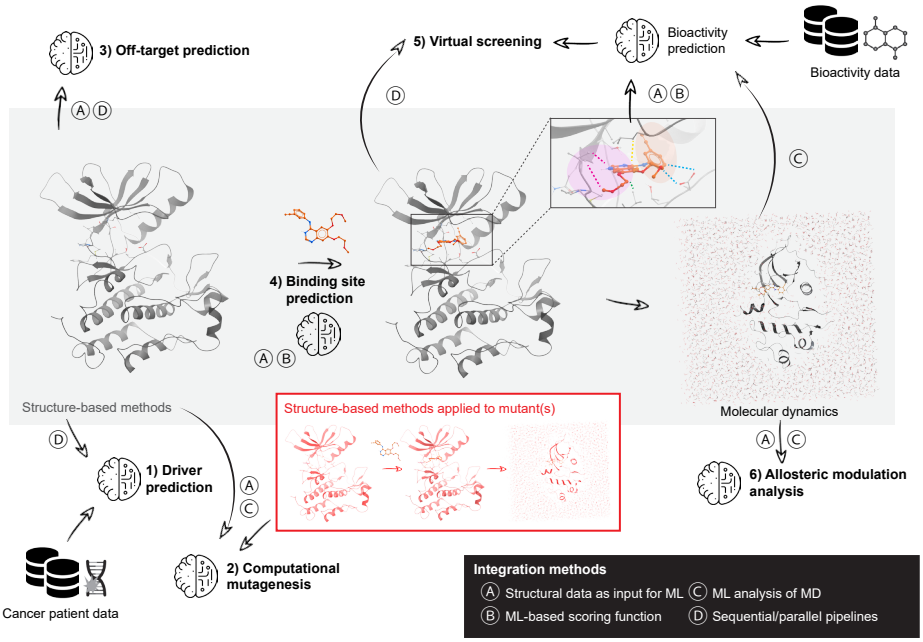



**Figure 2.1.** Use case scenarios of integrated structure-based (SB) and machine learning (ML) methods in oncological drug discovery and the integration methods employed. In this review we address six use case scenarios, namely 1) driver prediction, 2) computational mutagenesis, 3) (off)-target prediction, 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 published[62]. 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 AlphaFold[31] 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. 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)
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. & Karanicolas, 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 LXRβ 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 ASPP2 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 RIOK1 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)