Virtual Screening of Potential Anticancer Drugs based on Microbial Products

Gaspar P. Pinto[1,2], Natalie M. Hendrikse[1], Jan Stourac[1,2], Jiri Damborsky[1,2], David Bednar[1,2]*

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5/A13, 625 00 Brno, Czech Republic; [2]International Clinical Research Center, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic.

Corresponding author: davidbednar1208@gmail.com

Abstract

The development of microbial products for cancer treatment has been in the spotlight in recent years. In order to accelerate the lengthy and expensive drug development process, *in silico* screening tools are systematically employed, especially during the initial discovery phase. Moreover, considering the steadily increasing number of molecules approved by authorities for commercial use, there is a demand for faster methods to repurpose such drugs. Here we present a review on virtual screening web tools, publicly available databases of molecular targets and libraries of ligands, with the aim to facilitate the discovery of potential anticancer drugs based on microbial products. We provide an entry-level step-by-step description of the workflow for virtual screening of microbial metabolites with known protein targets, as well as two practical examples using freely available web tools. The first case presents a virtual screening study of drugs developed from microbial products using Caver Web, a web tool that performs docking along a tunnel. The second case comprises a comparative analysis between a healthy isocitrate dehydrogenase 1, a mutant that results in cancer, using the recently developed web tool PredictSNP[Onco]. In summary, this review provides the basic and essential background information necessary for virtual screening experiments, which may accelerate the discovery of novel anticancer drugs.

Keywords:

Databases, Microbial products, Virtual screening, Web tools, PredictSNP[Onco]

Introduction

Cancer is the umbrella name given to a group of diseases that manifest themselves in a very similar fashion, namely by abnormal, uncontrolled cell multiplication and growth.[1–4] Certain cancers have the potential to spread from their place of origin to other parts in the body,[5,6] at which point treatment by surgically removing the diseased cell agglomeration (tumour) becomes very difficult. This spread of diseased cells, or metastasis,[7] is what differentiates malignant tumours from benign tumours.[8–11] The latter type may also multiplicate in an uncontrolled fashion but does not metastasise and can therefore often be removed through surgery relatively easily.[12] According to the World Health Organization, the yearly number of deaths caused by cancer worldwide is approximately 10 million.[13,14] Currently available treatment options for cancer include surgery, chemo and radiotherapy[15], RNA-binding proteins[16], personalised treatment in the form of targeted therapy, and immunotherapy. The latter includes the use of microbial products: anything from secondary metabolites of microorganisms, or even their cellular components, to viral particles. The virtual screening of such microbial products against known molecular targets will be the focus of this review. One of the main advantages[17] of microbial-based therapies is their relatively low cost compared to other cancer therapies, which could allow for improved access to effective

treatments in low- and medium-income countries. Although this approach had already been suggested and used as early as the 1880s[18–21], it was not until recently that the usage of microbial-based cancer drugs gained increasing attention.[22] Approval of the drug Talimogene laherparepvec[23], a genetically engineered herpes virus, by the Food and Drug Administration and the European Medicines Agency in 2015 marked the beginning of a series of publications and special issues dedicated to microbial products in cancer treatment.

Microbial secondary metabolites represent a wide range of complex molecules, targeting and inhibiting proteins involved in carcinogenesis. Molecular recognition and binding of these metabolites to the clefts and cavities of target proteins is essential for their biological activity. Given the number of secondary metabolites of microbial origin, virtual screening is an essential tool to increase drug design efficiency or drug repurposing studies. This computational technique enables the docking[24–27] of thousands or even millions of biologically active compounds for binding to various cancer-related target proteins at a very low cost. Such virtual screening efforts can identify potential inhibitors of individual proteins or entire signalling pathways involved in the relevant cancer pathology. The most promising hits can then be studied *in vitro* and later *in vivo* using cellular assays, which not only improves cost efficiency, but also shortens the time of the study and allows for a much more extensive library of compounds to be tested. There are different approaches to virtual screening, and the choice of the most suitable one depends on the kind of information available (**Figure 1A**). This review will focus mostly on the scenario in which the user has access to a 3D structure of the molecular target of interest, and virtual screening of the protein and ligand can be performed (**Figure 1A**, left).[28] However, when no such target structure is available, there are two ligand-based approaches that can be used: (i) the similarity search method and (ii) machine learning (**Figure 1A**, right). The former is based on the assumption that similar ligands have similar characteristics and bind to the same target. In the latter, knowledge of ligands that are active, as well as ligands that are inactive against a given target, is used to train a dataset in order to predict the activity of newly tested ligands. Machine learning is also the basis for Quantitative-Structure Activity Relationships studies and has been widely used for some time, but has recently gained broader applicability with increasing computational power and research advances.[29–31]

A typical workflow for the scenario where a 3D structure of the target is available is depicted in Figure 1B. The first step includes target specification, followed by the selection of a corresponding structure as discussed in **Section 2**, "Selection of molecular target structures". The second step of the workflow consists of choosing a ligand library and preparing the ligands for the screening, which will be discussed in **Section 3**, "Selection and preparation of the ligand library". The subsequent virtual screening calculation step can be performed using different web tools, which will be the topic of **Section 4**, "Web-based software tools for virtual screening". This review is meant to guide beginners or non-experts in conducting virtual screening studies, critically interpreting the results, and analysing limitations of the studies. It focuses on databases, web tools and free stand-alone software with graphical user interfaces that support the user in different steps of the virtual screening process. Finally, **Section 5**, "Applications of Virtual Screening", includes two case studies with web tools and workflows that were developed *in-house* that can be used as tutorials.
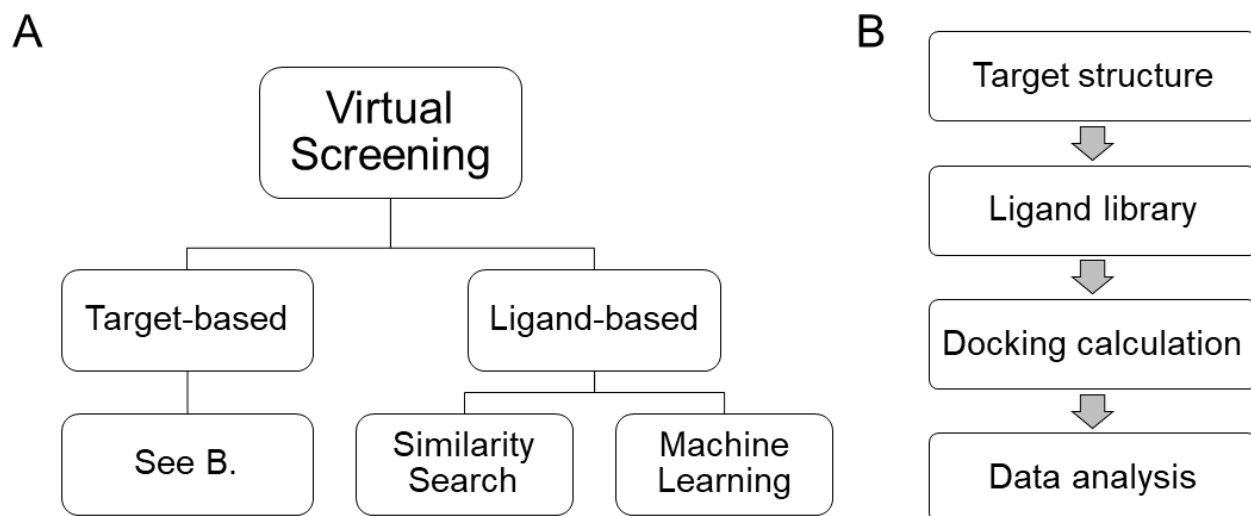
A

Virtual Screening

Target-based

See B.

Ligand-based

Similarity Search

Machine Learning

B

Target structure

Ligand library

Docking calculation

Data analysis

Figure 1 Virtual screening strategies depend on the knowledge and available data. (**A**) Virtual screening with a known molecular target structure (left), where the goal is to inhibit the target with one or several ligands. Virtual screening against an unknown target structure (right) can be performed using similarity search and pharmacophore mapping or by means of machine learning methods employing both positive and negative hits. (**B**) Step-by-step workflow for virtual screening with a known molecular target structure, which is the focus of this review.

2. Selection of molecular target structures

The first step of a virtual screening project consists of identifying molecular targets (**Figure 1B**) - typically proteins such as receptors or enzymes. However, in principle, any macromolecule that changes its behaviour upon ligand binding can be considered a biological target. Structures of macromolecules that have been experimentally solved by means of X-ray crystallography, Nuclear Magnetic Resonance spectroscopy or cryogenic electron microscopy are publicly available in databases such as RSCB – PDB or PDBe (**Table 1A-B**). In order to ensure the maximum possible validity of the virtual screening study, several properties of the chosen target structure should be carefully considered. The resolution at which the structure was solved and constructed should be as high as possible. Although there is no exact cut-off value for obtaining reliable results, a general consensus exists that says that a structure solved with a resolution above 2.5Å can be used, but with care.

In case there is no solved structure available for the target of interest, it is often possible to obtain one by means of homology modelling[32] and recently also to use machine learning.[33–35] Homology modelling makes use of the fact that protein structure is more conserved than sequence[36,37] and predicts a structure of the target protein based on its amino acid sequence alone. Homology modelling tools will either ask the user to provide a template structure of a protein that is homologous to the target of interest or search for possible templates in available protein databases. A distinction can thus be made between

molecular target databases that contain solved 3D structures (**Table 1A-B**) and databases that store structures built by homology models (**Table 1C**) using the structures from the databases in **Table 1A-B** as templates.

In addition to performing virtual screening studies on one single structural conformation (the one that has been experimentally solved), it is also possible to run virtual screening studies using several structures of one molecular target.[29] There are also web tools solely dedicated to building homology models **(Table 1D)**, but that usually do not store the models built by the users for more than two months. As a consequence, if a user has a similar request to a previous user, the model will be rebuilt from scratch and might differ slightly from the previous model.[38–41] To this end, a couple of databases contain data from molecular dynamics (MD) simulations that have been performed with various families of proteins **(Table 1E)** and are available for non-experienced users or users without access to a supercomputer.

In summary, a number of aspects need to be considered when using databases and web tools for the selection of the molecular target structure. Firstly, if there is more than one experimentally solved structure available, it is wise to choose the one with the highest resolution. It is also important to confirm that the structure contains all (relevant) amino acids from the sequence, or at least the amino acids lining the active site. In case there is no complete 3D structure available in the databases, certain concessions can be made, but one needs to keep in mind that the quality of the results will be highly dependent on the quality of the input.

Table 1 List of sources with available molecular targets. All databases present on the list are free to access, but some request the user to register with an academic verifiable email account.

Database	Link	Type of Database
A. General database of experimental protein structures		
wwPDB[42–44]	wwpdb.org	General Database
PDBe[45,46]	ebi.ac.uk/pdbe	General Database
RCSB PDB[47]	rcsb.org	General Database
PDBJapan[48]	pdbj.org	General Database
Biological Magnetic Resonance DB[49]	bmrb.io	General Database
SCOP[50]	scop.mrc-lmb.cam.ac.uk	General Database
B. Specialized databases of experimental protein structures		
AmyPDB[51]	amypdb.genouest.org	Amyloid proteins
CAZy[52]	cazy.org	Carbohydrate-Active enzymes Database
CyBase[53]	cybase.org.au	Cyclic proteins
CYPED[54]	cyped.biocatnet.de	Cytochrome P450 Engineering Database
GLYCO3D[55]	glyco3d.cermav.cnrs.fr	Carbohydrates
GlyGen[56]	glygen.org	Glycosylated Proteins
Homeodomain Resource[57]	research.nhgri.nih.gov/homeodomain	Homeodomain family
KNOTIN[58]	dsimb.inserm.fr/KNOTTIN	Disulphide Rich Proteins

MEROPS[59]	ebi.ac.uk/merops	Peptidases
RNA FRABASE[60]	rnafrabase.ibch.poznan.pl	RNA fragments database
SAAPdb[61]	bioinf.org.uk/mutations/saapdb/	Mapping SNPs and pathogenic deviations
SDAP[62]	fermi.utmb.edu/SDAP	Allergenic proteins
C. Database of protein models from homology modelling		
Swissmodel[63]	swissmodel.expasy.org/repository	General Database
PMP[64]	proteinmodelportal.org	General Database
MODBASE[65]	modbase.compbio.ucsf.edu	General Database
D. Web tools for protein structure prediction		
iTasser[38]	zhanglab.ccmb.med.umich.edu/I-TASSER	Homology modelling
Modeller[39]	salilab.org/modeller	Homology modelling
Phyre[2] [40]	sbg.bio.ic.ac.uk/~phyre	Homology modelling
E. Database of trajectories from molecular dynamics		
Dynameomics[66]	dynameomics.org	Dynamics database

3. Selection and preparation of the ligand library

Once a suitable target structure has been obtained, a collection of potential ligands can be selected for virtual screening. Similarly to the case of molecular targets described in the previous section, several databases exist that contain libraries of ligands.[67] These libraries include information about the ligand molecules, such as their molecular formula, 3D structure, physicochemical properties, as well as other features that will be further discussed in this section.[68] A library of ligands can be designed with a specific aim or a general-purpose. Whereas the former is typically done with a specific protein target in mind and with some previous knowledge, the latter is based on the idea that the library might be used for a variety of targets. As the aim of this review is to assist non-experienced users in their virtual screening projects, the main focus will be on the second type of library.

An essential feature of a library is that it contains molecular representations of the ligands that can easily be translated into 3D structures by means of software such as open babel.[69] The *de facto* standard, simplified molecular-input line-entry system (SMILES), was first proposed in 1980. It reads a specific chemical notation and uses a software tool to translate them into 3D structures.[70] In addition to SMILES, there are other notations that can be used with similar outcomes, and many libraries contain more than one notation. For instance, the International Chemical Identifier (InChI) is a notation developed by the International Union of Pure and Applied Chemistry to create a standard for molecular information encoding.[71] Notations such as SMILES and InChI also have some drawbacks, most notably the representation of only one tautomeric or ionisation state per molecule. On the other hand, they allow for non-experienced users to make use of these notations, and experienced users will have the know-how to use additional tools[69,72,73]. Other ligand-specific features that should be included in the library are characteristics related to absorption, distribution, metabolism, and excretion characteristics, e.g., polarity, pH range or even toxicity features such as LD_{50}. One very important piece of information is whether molecules have been approved for use in humans, are currently included in clinical trials, or are possible to purchase. This information is crucial if the virtual screening is done with the aim of immediate use in a patient, in contrast to being part of an exploratory protocol to develop new drugs for a certain

target. Although there are many ligand databases available (**Table 2**), only some of them will have libraries that contain the type of molecules that is relevant for the selected target; anything from biomimetic peptides to natural product-like molecules to lead-like molecules and drug-like molecules. Within these libraries, sub-libraries may exist for particular targets such as kinase-specific libraries, which are the most important for cancer treatment. Providing details on all available databases is beyond the scope of this review, but we refer the reader to the more comprehensive studies on this subject.[74–78]

Once a ligand library has been selected, it needs to be prepared for the docking calculations. This preparation step entails converting the ligand structures from a common 3D file format (.mol2, .xyz, .pdb) to a format that is accepted by the docking algorithm, e.g. the .pdbqt format. Some scripting knowledge may be recommended when performing a conversion for thousands of ligands.

Table 2 List of freely accessible databases of ligands with classification. Some of the databases also contain microbial metabolic products.

Database	Link	Type of Database
ZINC[68]	zinc.docking.org/	Generalist
Drugbank[79]	go.drugbank.com/	Generalist
3DMET[80]	integbio.jp/dbcatalog/en/record/nbdc00351	Generalist
BiGG[81]	bigg.ucsd.edu/	Metabolites
Binding DB[82]	bindingdb.org/bind/index.jsp	Drug-Like
BRENDA[83]	brenda-enzymes.org/	Metabolites
ChEBI[84]	ebi.ac.uk/chebi/	Chemicals
ChEMBL[85]	ebi.ac.uk/chembl/	Chemicals
ChemSpider[86]	chemspider.com/	Chemicals
DNP	dnp.chemnetbase.com/	Generalist
BIOFAQUIM[87]	biofacquim.herokuapp.com/	Plant, Fungi
ChemDB[88]	cdb.ics.uci.edu/	Plants
NPASS[89]	bidd.group/NPASS/	Plants, Bacteria, Metazoan, Fungi
NPAtlas[90]	npatlas.org/joomla/	Bacteria, Fungi
StreptomeDB[91]	132.230.56.4/streptomedb/	Bacteria

4. Performing virtual screening with web-based software tools

Once the selected target structure and ligand library are in place, the virtual screening can be set up. This section explains the workflow of starting, running, and analysing results of a virtual screening calculation, focusing specifically on inexperienced users. For more experienced users, we refer to the following articles and reviews on molecular docking,[92,93] virtual screening,[94] molecular docking algorithms,[95] molecular docking stand-alone tools,[96–98] and the bridging between molecular docking and MD simulations.[99]

There are two different algorithms that run sequentially in a docking calculation (**Figure 1B**). First, a search algorithm[92,100–103] looks for possible conformations in which the ligand fits the active pocket. Several possible conformations and positions are saved, which are subsequently used by the scoring

algorithm[104–108] to calculate the associated binding energy.[109] In the search algorithm, flexibility of ligand and target can be implemented in different ways. The most common approach is the combination of a rigid target and a flexible ligand,[104,110,111], which is often used because of its speed. In this scenario, ligand flexibility can be incorporated by using libraries of rotamers to create different conformations or by MD simulations. Another approach is one that accounts for the flexibility of both the molecular target and the ligand,[112–114] using the aforementioned strategies to simulate flexibility for both components. However, using this approach significantly increases the degrees of freedom, resulting in lengthy calculations and corresponding demands of computational power. Alternatively, target flexibility can be incorporated by using several snapshots of the protein from an MD simulation or several different structures available on databases, as discussed in **Section 2**. For some targets, snapshots from MD simulations are readily available, and the structures can be retrieved from the database of choice. The structures act as different molecular targets, and the binding energies per ligand can be averaged. Since enzymes' active sites are sometimes buried,[115,116] the need to describe tunnels and how ligands travelled from the surface of the protein to the active site. Since classical docking tries to find the best binding energy of a ligand to a target within the active site, the trajectory travelled by the ligand is disregarded. This has been addressed recently by the advent of trajectory docking, meaning that a ligand is docked in several positions along the tunnel from the surface into the active site, or vice versa.[117–119] It is similar to molecular docking technically, but it searches for a pathway of a ligand through the tunnel of a molecular target into the active pocket while calculating the binding energies. Most web tools that perform virtual screening calculations use the flexible ligand-rigid receptor method, while the number of web tools doing trajectory docking is very limited.[117,120,121]

After running the search and scoring algorithms, the user is provided with calculated binding energies for each combination of ligand and target. Whereas virtual screening calculations using classical docking are likely to produce more than one binding energy per ligand, studies using trajectory docking provide one binding energy value for each step along the calculated pathway (**Case study 5.1**). It is important to note that a virtual screening study should not be regarded as a one-hit result study, and data should be carefully analysed and interpreted. The strength of virtual screening is most apparent as an exploratory approach, in which the top hits can be subjected to further analysis. In addition, the screening can be applied to several different targets to highlight the effect of changes in protein structure on ligand binding. Analysis of the effect of point mutations on the binding of all FDA-approved drugs to a well-known oncology target will be illustrated in the next section (**Case study 5.2**).

Table 3 Web tools for molecular docking and virtual screening that are free to use for academic purposes. Some of these tools have other bioinformatics tools integrated to provide a more comprehensive analysis.

WebTool	Link	Method
iStar iDock[122]	istar.cse.cuhk.edu.hk/idock	Classical docking
DockNmine[123]	ufip.univ-nantes.fr/tools/docknmine/	Database for docking publicly available data
Prodigy-Lig[124]	bianca.science.uu.nl/prodigy/	Classical docking
Caver Web[120]	loschmidt.chemi.muni.cz/caverweb/	Trajectory docking

SystemsDock[125]	systemsdock.unit.oist.jp/	Classical docking
CHARMMing[126]	charmming.org/	Fragment based docking
DINC[127]	dinc.kavrakilab.org/	Meta docking
MoMA-LigPath[117]	hal.inria.fr/LAAS-IDEA/hal-00843321v1	Unbinding trajectory
Covalent[128]	docking.sce.ntu.edu.sg/	Classical docking
Pose and Rank[31]	modbase.compbio.ucsf.edu/poseandrank/	Calculates the pose or the rank
GalaxyPepDock[26]	seoklab.github.io/GalaxyPepDock/	Docking based on interaction similarity
CABS-dock[27]	biocomp.chem.uw.edu.pl/CABSdock	Flexible docking with no active site knowledge
BSP-SLIM[129]	zhanglab.ccmb.med.umich.edu/BSP-SLIM/	Docking using low resolution molecular targets
SwissDock[130]	swissdock.ch/	Classical docking
FlexPepDock[131]	flexpepdock.furmanlab.cs.huji.ac.il/	Models protein-ligand complexes
DockBlaster[132]	covalent.docking.org/	Searches ZINC database to find a ligand for the target

5. Practical applications of virtual screening

5.1 Virtual screening of potential inhibitors of Bruton's tyrosine kinase using Caver Web

In this case study, we used Caver Web[120] to perform virtual screening on molecular target Bruton's tyrosine kinase (BTK). BTK is a non-receptor tyrosine kinase involved in B-cell activation and development[133] that is typically overexpressed in various haematological malignancies.[134] BTK is commonly overexpressed in … and is a common target of inhibition by drugs, which has lead to increased drug resistance.[135] The web tool Caver Web[120] (https://loschmidt.chemi.muni.cz/caverweb/) analyses tunnels in proteins in the first step, followed by the analysis of ligand transportation via these tunnels in the second step. The input for the calculation is a structure of the molecular target and the library of ligands representing a potential inhibitor. Inhibitors can be uploaded but can also be searched in the specialised databases, written in SMILES code or drawn in 2D directly within the web interface. The outputs of the calculations are visualisations of tunnels and channels, trajectories of ligand passage and energy profiles allowing prioritisation of potential drugs. Calculations are fast and take 2-20 min per job, making them appropriate for the virtual screening purposes.

Briefly, the workflow starts with the user selecting a target structure, either by uploading one directly or by providing a PDB ID code, which is then automatically prepared for the following steps. These steps entail the selection of a starting point for tunnel calculation, which can be done manually by the user or automatically by the tool, followed by identification and analysis of the tunnels in the target protein structure. Finally, the identified tunnels provide the basis for docking the selected ligands and for studying their possible trajectories. The output of Caver Web consists of a comprehensive report that contains the binding energy of the bound state E_{Bound}, the maximum binding energy E_{Max}, the binding energy at the surface $E_{Surface}$, the binding energy of activation E_a (E_{Max} - $E_{Surface}$), and the binding energy of

the bound state subtracted from the binding energy at the surface ΔE_{BS}. Here, E_{Bound} means the binding energy of the ligand in the active site, and E_{Max} represents the highest binding energy anywhere along the tunnel.

In order to reduce complexity and calculation time (14 min) of the case study, BTK was chosen as a single molecular target, and a mix of known BTK ligands and microbial products was selected as corresponding ligand library. The comprehensive output report can be found on Caver Web's website following the link: https://loschmidt.chemi.muni.cz/caverweb/?action=results&job=3rquhs. **Table 4** shows a sample of nine calculated binding energies, illustrating the high variability between different drugs. Energetically favourable binding to the active site cavity was observed with Mafenide, Triadimefon, Gantanol and Tyrosol. Most drugs bind better in the first part of the tunnel, represented by lower binding energies, highlighting the benefits of trajectory docking versus classical docking when screening for inhibitors. Whereas classical docking only considers binding in the targets active site cavity, the trajectory docking approach enables identification of inhibitors that bind strongly to the tunnel, thereby blocking the way to the active site. It should be noted that, since this virtual screening was performed using a single snapshot of the protein crystal structure, the calculated tunnel might be unreasonably narrow, and relaxation of the structure by the MD simulations could be appropriate.

Table 4: BTK binding energies calculated for nine drugs of microbial origin using Caver Web. The top three compounds originate from archaea, the middle three originate from bacteria, and the bottom three originate from viruses.

Ligand[a]	E_{Bound} (kcal/mol)	$E_{Surface}$ (kcal/mol)
Mafenide	0.8	-3.6
Cox	13.9	-5.3
Daranide	2.2	-3.5
Piritrexim	3.2	-5.4
Triadimefon	0.6	-4.1
Norfloxacin	11.5	-5.0
Gantanol	-1.8	-4.5
Tyrosol	-3.5	-3.5
Homidium	21.1	-5.6

[a] Full table and report obtained from Caver Web are stored on the website with the link: https://loschmidt.chemi.muni.cz/caverweb/?action=results&job=3rquhs.

5.2 Analysis of mutations and virtual screening of isocitrate dehydrogenase 1 using PredictSNP[Onco]

The second case study demonstrates the use of web tool PredicSNP[Onco] (currently under development) for prediction of the effect of mutations on protein stability, function and, binding to FDA-approved drugs. Given the high heterogeneity of cancer, treatment focus is gradually shifting from a "one size fits all" approach   towards   personalised   medicine.   The   PredictSNP[onco]

(http://loschmidt.chemi.muni.cz/predictsnp-onco/) is the tool developed specifically for precision oncology. The only inputs for the calculation are the target protein and mutations found previously by sequencing the patient's samples. The outputs of the calculation are estimates of the effects of studied mutation on the stability and function of the target protein. Virtual screening of both wild type enzyme and the mutant against the whole FDA-approved drugs navigates the users towards possible "non-conventional" treatment, reflecting the potential impact of the mutation on the binding. The workflow (**Figure 2**) was designed in a way to provide all the results within two weeks; the calculation for one target currently takes about 8-10 days.

The results are visualised directly in the web interface, in the form of a summary table (**Table 5**) and a report where the interpretation of relevant values is highlighted. This template facilitates its interpretation by Medical Doctors and readily provides them with the essential information, while those that are interested can access all raw data that have been calculated.

In this case study, mutations in isocitrate dehydrogenase 1 (IDH1) that were identified in cancer patients by transcriptome sequencing were analysed using PredictSNP[Onco]. IDH1 is an enzyme encoded by the *idh1* gene located on chromosome 2 and, along with IDH2 and IDH3,[136,137] is a key enzyme involved in cellular metabolism.[138,139] Its structure has been studied extensively [42,47,140] and structures of the wild type enzyme, as well as two single-point mutants (R132H and R132C), are available in the RCSB PDB database. Several mutations have previously been studied, including G70D, G123R, R132H, R132C, and A134D. Especially position 132 is of interest, as five deleterious mutations are known to date. The most commonly observed mutation in this position, R132H,[38–41] accounts for 90% of the mutations in IDH1 in lower-grade glioma.[142,143] The residue is located in the enzyme's active site[144], and mutation results in loss of enzymatic function, which in turn leads to a concomitant gain of neomorphic activity and production of the oncometabolite 2-HG. In 2018, the drug Tibsovo (generic name Ivosidenib) was approved[145] by the FDA to treat mutations in IDH1 that led to acute myeloid leukaemia and was granted orphan drug designation.

Using PredictSNP[Onco], we analysed the effects of mutation R132H on the enzyme's stability, biological activity, and interaction with all FDA-approved drugs. Firstly, the tool correctly identified the mutation in question as being deleterious with 87% of confidence, as reported in the literature.[146–149] As it would be expected with a mutation in the active site, the function of the protein was predicted to be altered in the variant. This was based on relatively large changes in pocket size and pKa of the catalytic residues and was further supported by the reports outputted by the protein interactions calculator[150] (PIC) and HOPE, a protein structure analysis of mutations causing inheritable diseases. These two tools analyse intraprotein interactions and predict the effect that a given mutation has on the protein and its function. Finally, we performed virtual screening with the dataset of all FDA approved drugs that are present in the ZINC[68] database. Since the mutation occurs in a catalytic residue, we expected the virtual screening to show at least small differences in the binding energies. This was not the case, however, suggesting that the interactions of the drugs with the active site are mostly conserved in the variant. Yet, the trajectory docking energies obtained for the wild type and the variant differed up to 0.9 kcal/mol (~10%). Given that the treatment for this mutation is known, we highlight its binding energy yielded by both AutoDock Vina[25] and CaverDock[28,118] in the table.

With the outputted report from this tool, it is our goal to give more information about a patient's case to Medical Doctors and, with their previous knowledge of the case, help in the decision-making process.
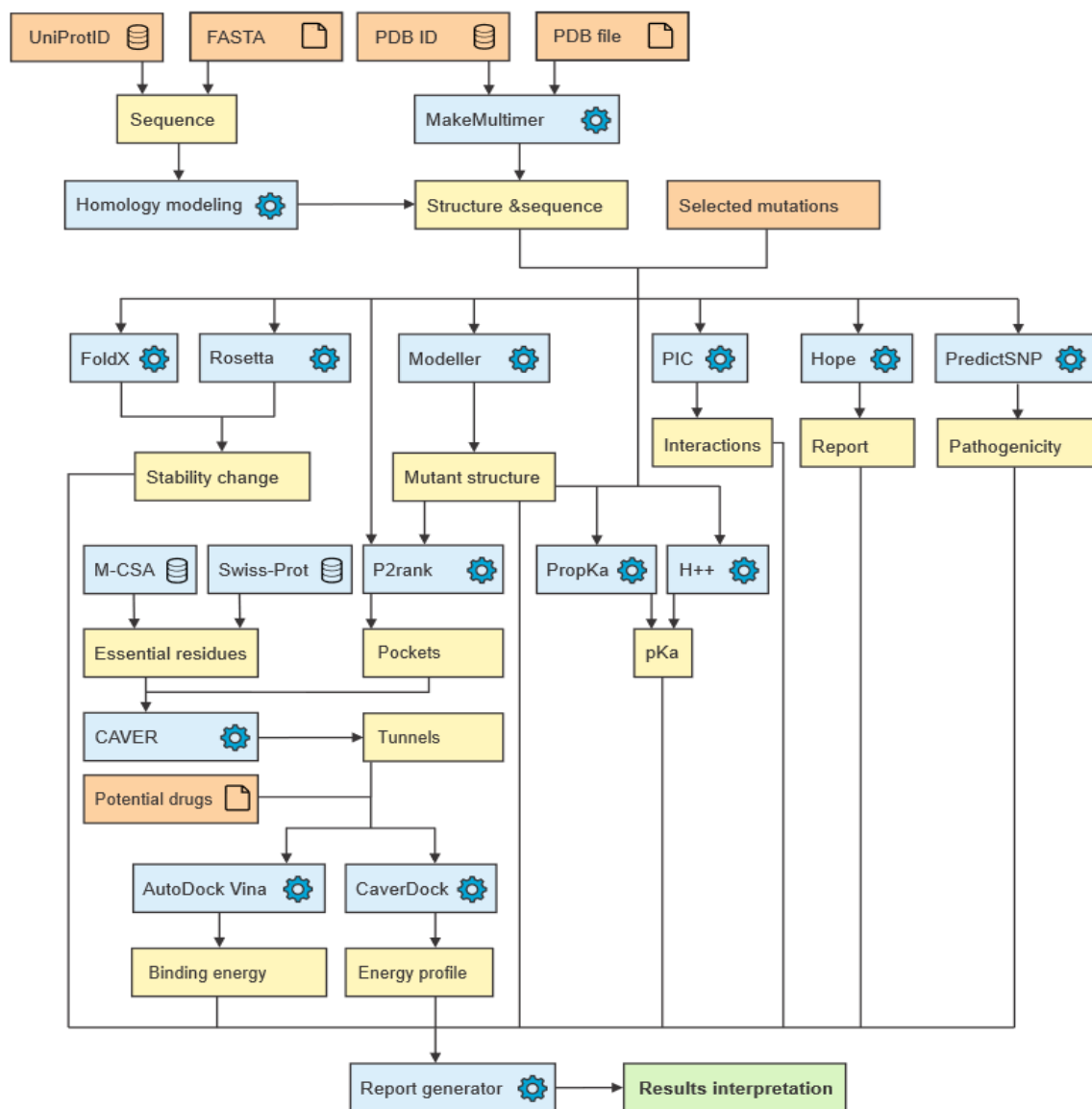
Figure 2 Workflow used by PredictSNP[Onco] to study any mutation leading to cancer in any molecular target. The workflow starts with the input of the system of choice in one of four ways: UniProtID, FASTA sequence, PDB ID or PDB file created by the user. The first two options will be used when there is no structure available, homology modelling will ensure that a structure will be predicted. When PDB ID or PDB file are provided, a tool will identify the biologically relevant monomer. A second user input will be needed to identify the mutation suspect for the development of cancer in a patient. From left to right, FoldX and Rosetta will be used to identify if the mutation is stabilising or destabilising. The modeller tool is used to creating a model of the mutant protein, PIC and HOPE output reports on general function and interactions comparing the wild-type and the mutant protein. PredictSNP predicts if the mutation is neutral or deleterious. A series of tools are then used to assess the difference in function inserted by the mutation, such as tunnels (Caver), pockets (P2Rank), essential residues and their pKa's (PropKa, H++). Finally, a virtual screening based on AutoDock Vina is run with the whole FDA-approved dataset and CaverDock with the top 100 binding ligands.

Table 5 Table with summarised results for R132H of IDH1 from tool PredictSNP[Onco] to facilitate decision making by Medical Doctors. The table is colour coded with stability in blue at the top of the table, function in bright green in the middle of the table and inhibition in bordeaux. The results are also colour coded with deleterious modifications and variations being represented with bright red colour, and beneficial modifications and variations bring represented in bright green. Ivosidenib is known drug targeting IDH1 and is highlighted. RMSD – root means square deviation. NA – not applicable.

	Wild Type		Mutant R132H		Variations	
PredictSNP					87% deleterious	
	$\Delta\Delta$G (kcal/mol)		$\Delta\Delta$G (kcal/mol)		$\Delta\Delta$G (kcal/mol)	
STABILITY - Rosetta	NA		1.1		1.1	
STABILITY - FoldX	NA		0.8		0.8	
	Bottleneck Å	Throughput	Bottleneck Å	Throughput	Bottleneck ΔÅ	Throughput
	2.1	0.84	2.1	0.89	0.0	0.05
FUNCTION	1.8	0.70	1.9	0.8	0.1	0.10
Tunnels	1.6	0.60	1.8	0.79	0.2	0.19
	1.7	0.57	1.7	0.57	0.0	0.00
	Relevance %	Volume Å^3	Relevance %	Volume Å^3	Relevance Δ %	ΔVolume Å^3
	100	3623	100	4283	0	660
FUNCTION	51	2131	49	2152	-2	21
Cavities	32	1375	31	1321	-1	-54
	Residue name and number		Residue name and number		Residue name and number	
	Arginine 132		Histidine 132		R132H	
FUNCTION	Tyrosine 139					
Catalytic Residues	Lysine 212					
	Aspartate 275					
	Residue	Number	pKa	Residue	Number	pka	ΔpKa
FUNCTION	Arginine 132 - 14.23			Histidine 132 - 6.47			-7.76
Catalytic Residues pKa	Tyrosine 139 - 12.58			Tyrosine 139 - 12.58			0.00

		Lysine 212 - 11.56			Lysine 212 - 10.73			-0.83		
		Aspartate 275 - 4.60			Aspartate 275 - 3.32			-1.28		
		Binding Energy (kcal/mol)			Binding Energy (kcal/mol)			Δ Binding Energy (kcal/mol)		
INHIBITORS Autodock Vina	Alfentanil	-10.3			-10.3			0.0		
	Prostaglandin E2	-10.2			-10.2			0.0		
	Propoxyphene	-10.0			-10.0			0.0		
	Sulfamethazine	-9.3			-9.3			0.0		
	Rivastigmine	-9.3			-9.3			0.0		
	Droperidol	-9.2			-9.2			0.0		
	isoMisoprostol	-9.2			-9.2			0.0		
	Fluvastatin	-9.2			-9.2			0.0		
	Dehydrocholic	-9.0			-9.0			0.0		
	Ivosidenib	-8.1			-8.1			0.0		
		Min Energy (kcal/mol)	Max Energy (kcal/mol)	RMSD Å	Min Energy kcal/mol	Max Energy kcal/mol	RMSD Å	ΔMin (kcal/mol)	ΔMax (kcal/mol)	ΔRMSD (Å)
INHIBITORS CaverDock	Alfentanil	-8.6	-4.2	0.2	-8.9	-4.4	0.1	-0.3	-0.2	-0.1
	Prostaglandin E2	-10.9	-7.9	0.9	-10.7	-7.4	0.4	0.2	0.5	-0.5
	Propoxyphene	-10.0	-5.8	1.5	-10.6	-5.4	1.4	-0.6	0.4	-0.1
	Sulfamethazine	-10.1	-8.8	0.4	-10.9	-9.1	0.7	-0.8	-0.3	0.3
	Rivastigmine	-10.1	-7	0.6	-9.4	-7.4	0.9	0.7	-0.4	0.3
	Droperidol	-9.3	-6.8	1.1	-9.1	-6.4	0.7	0.2	0.4	-0.4
	isoMisoprostol	-9.4	-5.2	1.4	-9.8	-5.8	1.1	-0.4	-0.6	-0.3
	Fluvastatin	-9.6	-5.3	0.7	-10.4	-5.4	1.4	-0.8	-0.1	0.7
	Dehydrocholic	-8.8	-6	0.2	-9.7	-6.3	0.7	-0.9	-0.3	0.5
	Ivosidenib	-7.4	5.7	0.5	-6.5	2.1	0.9	-0.9	3.6	0.4

6. Conclusions

This review highlights the potential of virtual screening, which can be a useful aid in the prioritisation of promising drug compounds for experimental testing. As illustrated by the case studies presented herein, this technique can also be employed in drug repurposing efforts, enabling the efficient discovery of microbial products that can be used for cancer treatment. Therefore, our goal is to increase awareness of available virtual screening tools within the medical and scientific community, in particular among those working in the area of experimental and clinical oncology. Virtual screening against known targets involved in a specific cancer pathology may help to identify small molecules, including products of microbial origin, that are likely to bind to it.

The aim of this review was to provide an introduction and overview of virtual screening, in particular for those with little to no experience. It includes tables that list databases of molecular target structures, libraries of ligand molecules, and various software tools for docking calculations, which may assist users in their future virtual screening studies. Although using stand-alone tools for virtual screening has advantages, such as calculation speed and the ability to control every step of the workflow, using available web tools can be preferred for various reasons. First and foremost, they are often more user friendly and do not require any previous experience, as tutorials are available that provide inexperienced users with the necessary knowledge to perform the virtual screening. The fact that they typically do not require installations or access to a cluster or advanced computer further contributes to the user-friendliness of web tools. In the case studies, we introduced *in house* software tools Caver Web and PredictSNP^{onco}, which perform highly automated virtual screening and may serve as a starting point for exploring this technique. However, it is important to keep in mind that using automated web tools brings challenges of its own, and, as for most computational methods, the results will only be as good as the input data.

References

1. Shah, R. B. *et al.* Androgen-Independent Prostate Cancer Is a Heterogeneous Group of Diseases:

    Lessons from a Rapid Autopsy Program. *Cancer Res* **64**, 9209–9216 (2004).

2. Weinberg, R. A. How cancer arises. *Sci Am* **275**, 62–70 (1996).

3. Health (US), N. I. of & Study, B. S. C. *Understanding Cancer*. (National Institutes of Health (US), 2007).

4. Bertram, J. S. The molecular biology of cancer. *Mol Aspects Med* **21**, 167–223 (2000).

5. Cooper, G. M. The Development and Causes of Cancer. *The Cell: A Molecular Approach. 2nd edition* (2000).

6. Sarkar, S. *et al.* Cancer Development, Progression, and Therapy: An Epigenetic Overview. *Int J Mol Sci* **14**, 21087–21113 (2013).

7. Pantel, K., Alix-Panabières, C. & Riethdorf, S. Cancer micrometastases. *Nature Reviews Clinical Oncology* **6**, 339–351 (2009).

8. Sugiyama, H. *et al.* Characteristics of Benign and Malignant Bone Tumors Registered in the Hiroshima Tumor Tissue Registry, 1973-2012. *JB JS Open Access* **3**, (2018).

9. Kassi, M. *et al.* Differentiating benign from malignant cardiac tumors with cardiac magnetic resonance imaging. *The Journal of Thoracic and Cardiovascular Surgery* **157**, 1912-1922.e2 (2019).

10. Aaltonen, L. A. *et al.* Replication Errors in Benign and Malignant Tumors from Hereditary Nonpolyposis Colorectal Cancer Patients. *Cancer Res* **54**, 1645–1648 (1994).

11. England, D. M., Hochholzer, L. & McCarthy, M. J. Localized benign and malignant fibrous tumors of the pleura. A clinicopathologic review of 223 cases. *Am J Surg Pathol* **13**, 640–658 (1989).

12. Sciacovelli, M., Schmidt, C., Maher, E. R. & Frezza, C. Metabolic Drivers in Hereditary Cancer Syndromes. *Annu. Rev. Cancer Biol.* **4**, 77–97 (2020).

13. Ritchie, H. & Roser, M. Causes of Death. *Our World in Data* (2018).

14. Global Health Estimates: Life expectancy and leading causes of death and disability. https://www.who.int/data/maternal-newborn-child-adolescent/monitor.

15. Laham-Karam, N., Pinto, G. P., Poso, A. & Kokkonen, P. Transcription and Translation Inhibitors in Cancer Treatment. *Front. Chem.* **8**, (2020).

16. Pereira, B., Billaud, M. & Almeida, R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer* **3**, 506–528 (2017).

17. Forbes, N. S. *et al.* White paper on microbial anti-cancer therapy and prevention. *Journal for ImmunoTherapy of Cancer* **6**, 78 (2018).

18. Dougan, M. & Dougan, S. K. Programmable bacteria as cancer therapy. *Nature Medicine* **25**, 1030–1031 (2019).

19. Starnes, C. O. Coley's toxins in perspective. *Nature* **357**, 11–12 (1992).

20. McCarthy, E. F. The Toxins of William B. Coley and the Treatment of Bone and Soft-Tissue Sarcomas. *Iowa Orthop J* **26**, 154–158 (2006).

21. Richardson, M. A., Ramirez, T., Russell, N. C. & Moye, L. A. Coley toxins immunotherapy: a retrospective review. *Altern Ther Health Med* **5**, 42–47 (1999).

22. Hoffman, R. *Bacterial Therapy of Cancer*. (Humana Press, 2016).

23. Fukuhara, H., Ino, Y. & Todo, T. Oncolytic virus therapy: A new era of cancer treatment at dawn. *Cancer Sci* **107**, 1373–1379 (2016).

24. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **30**, 2785–2791 (2009).

25. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem* **31**, 455–461 (2010).

26. Lee, H., Heo, L., Lee, M. S. & Seok, C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res* **43**, W431-435 (2015).

27. Kurcinski, M. *et al.* CABS-dock standalone: a toolbox for flexible protein-peptide docking. *Bioinformatics* **35**, 4170–4172 (2019).

28. Pinto, G. P. *et al.* Fast Screening of Inhibitor Binding/Unbinding Using Novel Software Tool CaverDock. *Front. Chem.* **7**, (2019).

29. Pinto, G. *et al.* Screening of World Approved Drugs against Highly Dynamical Spike Glycoprotein SARS-CoV-2 using CaverDock and Machine Learning. (2020) doi:10.26434/chemrxiv.13168883.v2.

30. Holderbach, S., Adam, L., Jayaram, B., Wade, R. C. & Mukherjee, G. RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physicochemical Features. *Front. Mol. Biosci.* **7**, (2020).

31. Morrone, J. A., Weber, J. K., Huynh, T., Luo, H. & Cornell, W. D. Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach. *J. Chem. Inf. Model.* **60**, 4170–4179 (2020).

32. Kleywegt, G. J. & Jones, T. A. Model building and refinement practice. *Methods Enzymol* **277**, 208–230 (1997).

33. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**, 823–826 (1986).

34. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).

35. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

36. Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* **13**, 190–202 (2004).

37. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **77**, 499–508 (2009).

38. Yang, J. & Zhang, Y. Protein structure and function prediction using I-TASSER. *Current Protocols in Bioinformatics* **52**, 5.8.1-5.8.15 (2015).

39. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).

40. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845–858 (2015).

41. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).

42. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology* **10**, 980–980 (2003).

43. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research* **35**, D301–D303 (2007).

44. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47**, D520–D528 (2019).

45. PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Research* **48**, D344–D353 (2020).

46. Gutmanas, A. *et al.* PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* **42**, D285-291 (2014).

47. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

48. Kinjo, A. R. *et al.* Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Research* **45**, D282–D288 (2017).

49. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res* **36**, D402-408 (2008).

50. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research* **48**, D376–D382 (2020).

51. Pawlicki, S., Le Béchec, A. & Delamarche, C. AMYPdb: A database dedicated to amyloid precursor proteins. *BMC Bioinformatics* **9**, 273 (2008).

52. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **42**, D490–D495 (2014).

53. Wang, C. K. L., Kaas, Q., Chiche, L. & Craik, D. J. CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research* **36**, D206–D210 (2008).

54. Sirim, D., Wagner, F., Lisitsa, A. & Pleiss, J. The Cytochrome P450 Engineering Database: integration of biochemical properties. *BMC Biochem* **10**, 27 (2009).

55. Pérez, S. *et al.* Glyco3D: A Suite of Interlinked Databases of 3D Structures of Complex Carbohydrates, Lectins, Antibodies, and Glycosyltransferases. *A Practical Guide to Using Glycomics Databases* 133–161 (2016) doi:10.1007/978-4-431-56454-6_7.

56. York, W. S. *et al.* GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* **30**, 72–73 (2019).

57. Moreland, R. T., Ryan, J. F., Pan, C. & Baxevanis, A. D. The Homeodomain Resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family. *Database (Oxford)* **2009**, (2009).

58. Postic, G., Gracy, J., Périn, C., Chiche, L. & Gelly, J.-C. KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Research* **46**, D454–D458 (2018).

59. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res* **38**, D227–D233 (2010).

60. Popenda, M. *et al.* RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* **11**, 231 (2010).

61. Hurst, J. M. *et al.* The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Hum Mutat* **30**, 616–624 (2009).

62. Ivanciuc, O., Schein, C. H. & Braun, W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* **31**, 359–362 (2003).

63. Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res* **45**, D313–D319 (2017).

64. Haas, J. *et al.* The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database (Oxford)* **2013**, bat031 (2013).

65. Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **42**, D336-346 (2014).

66. Jonsson, A. L., Schaeffer, R. D., Kamp, M. W. van der & Daggett, V. Dynameomics: protein dynamics and unfolding across fold space. *Biomolecular Concepts* **1**, 335–344 (2010).

67. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).

68. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery | Journal of Chemical Information and Modeling. https://pubs.acs.org/doi/10.1021/acs.jcim.0c00675.

69. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33 (2011).

70. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

71. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5**, 7 (2013).

72. Miteva, M. A., Guyon, F. & Tuffï¿½ry, P. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Research* **38**, W622–W627 (2010).

73. Vainio, M. J. & Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **47**, 2462–2474 (2007).

74. Lionta, E., Spyrou, G., Vassilatis, D. K. & Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr Top Med Chem* **14**, 1923–1938 (2014).

75. Banegas-Luna, A.-J., Cerón-Carrasco, J. P. & Pérez-Sánchez, H. A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. *Future Medicinal Chemistry* **10**, 2641–2658 (2018).

76. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).

77. Renom-Carrasco, M. & Lefort, L. Ligand libraries for high throughput screening of homogeneous catalysts. *Chem. Soc. Rev.* **47**, 5038–5060 (2018).

78. Gennari, C. & Piarulli, U. Combinatorial Libraries of Chiral Ligands for Enantioselective Catalysis. *Chem. Rev.* **103**, 3071–3100 (2003).

79. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**, D668-672 (2006).

80. Maeda, M. H. & Kondo, K. Three-dimensional structure database of natural metabolites (3DMET): a novel database of curated 3D structures. *J Chem Inf Model* **53**, 527–533 (2013).

81. Norsigian, C. J. *et al.* BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Research* **48**, D402–D406 (2020).

82. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **35**, D198-201 (2007).

83. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* **30**, 47–49 (2002).

84. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36**, D344–D350 (2008).

85. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2019).

86. ChemSpider: An Online Chemical Information Resource | Journal of Chemical Education.

    https://pubs.acs.org/doi/10.1021/ed100697w.

87. Pilón-Jiménez, B. A., Saldívar-González, F. I., Díaz-Eufracio, B. I. & Medina-Franco, J. L. BIOFACQUIM:

    A Mexican Compound Database of Natural Products. *Biomolecules* **9**, 31 (2019).

88. Chen, J., Swamidass, S. J., Dou, Y., Bruand, J. & Baldi, P. ChemDB: a public database of small

    molecules and related chemoinformatics resources. *Bioinformatics* **21**, 4133–4139 (2005).

89. Zeng, X. *et al.* NPASS: natural product activity and species source database for natural product

    research, discovery and tool development. *Nucleic Acids Res* **46**, D1217–D1222 (2018).

90. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial

    Natural Products Discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).

91. Moumbock, A. F. A. *et al.* StreptomeDB 3.0: an updated compendium of streptomycetes natural

    products. *Nucleic Acids Research* **49**, D600–D604 (2021).

92. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search

    algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics* **47**,

    409–443 (2002).

93. Pinzi, L. & Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *Int J Mol Sci* **20**,

    (2019).

94. Seifert, M. H. J., Kraus, J. & Kramer, B. Virtual high-throughput screening of molecular databases.

    Curr Opin Drug Discov Devel **10**, 298–307 (2007).

95. Brooijmans, N. & Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annual Review of

    Biophysics and Biomolecular Structure* **32**, 335–373 (2003).

96. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys Rev* **9**, 91–

    102 (2017).

97. Wang, Z. *et al.* Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* **18**, 12964–12975 (2016).

98. Oferkin, I. V. *et al.* Evaluation of Docking Target Functions by the Comprehensive Investigation of Protein-Ligand Energy Minima. *Advances in Bioinformatics* https://www.hindawi.com/journals/abi/2015/126858/ (2015) doi:https://doi.org/10.1155/2015/126858.

99. Salmaso, V. & Moro, S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. *Front. Pharmacol.* **9**, (2018).

100. Shoichet, B. K., Stroud, R. M., Santi, D. V., Kuntz, I. D. & Perry, K. M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **259**, 1445–1450 (1993).

101. Mustard, D. & Ritchie, D. W. Docking essential dynamics eigenstructures. *Proteins* **60**, 269–274 (2005).

102. Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **46**, 499–511 (2003).

103. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727–748 (1997).

104. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* **3**, 935–949 (2004).

105. Jain, A. N. Scoring functions for protein-ligand docking. *Curr Protein Pept Sci* **7**, 407–420 (2006).

106. Rajamani, R. & Good, A. C. Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. *Curr Opin Drug Discov Devel* **10**, 308–315 (2007).

107. Kinnings, S. L. *et al.* A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J Chem Inf Model* **51**, 408–419 (2011).

108.   Li, H., Sze, K.-H., Lu, G. & Ballester, P. J. Machine-learning scoring functions for structure-based

drug lead optimization. *WIREs Computational Molecular Science* **10**, e1465 (2020).

109.   Pantsar, T. & Poso, A. Binding Affinity via Docking: Fact and Fiction. *Molecules* **23**, (2018).

110.   Wang, Q. & Pang, Y.-P. Preference of Small Molecules for Local Minimum Conformations when

Binding to Proteins. *PLOS ONE* **2**, e820 (2007).

111.   Bruns, R. F. Conformational induction versus conformational selection: evidence from allosteric

enhancers. *Trends Pharmacol Sci* **17**, 189; discussion 190-191 (1996).

112.   Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between

flexible peptides and globular proteins. *Proteins* **78**, 2029–2040 (2010).

113.   Klebe, G. & Mietzner, T. A fast and efficient method to generate biologically relevant

conformations. *J Computer-Aided Mol Des* **8**, 583–606 (1994).

114.   Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B. & Johnson, A. P. eHiTS: A new fast, exhaustive

flexible ligand docking system. *Journal of Molecular Graphics and Modelling* **26**, 198–212 (2007).

115.   Kingsley, L. J. & Lill, M. A. Substrate Tunnels in Enzymes: Structure-Function Relationships and

Computational Methodology. *Proteins* **83**, 599–611 (2015).

116.   Pavlova, M. *et al.* Redesigning dehalogenase access tunnels as a strategy for degrading an

anthropogenic substrate. *Nat Chem Biol* **5**, 727–733 (2009).

117.   Devaurs, D. *et al.* MoMA-LigPath: a web server to simulate protein–ligand unbinding. *Nucleic

Acids Research* **41**, W297–W302 (2013).

118.   Filipovic, J. *et al.* CaverDock: A Novel Method for the Fast Analysis of Ligand Transport.

IEEE/ACM Transactions on Computational Biology and Bioinformatics 1–1 (2019)

doi:10.1109/TCBB.2019.2907492.

119.   Lucas, M. F. & Guallar, V. An Atomistic View on Human Hemoglobin Carbon Monoxide Migration

Processes. *Biophysical Journal* **102**, 887–896 (2012).

120.   Stourac, J. *et al.* Caver Web 1.0: identification of tunnels and channels in proteins and analysis of

ligand transport. *Nucleic Acids Res.* **47**, W414–W422 (2019).

121.   Guallar, V., Lu, C., Borrelli, K., Egawa, T. & Yeh, S.-R. Ligand Migration in the Truncated

Hemoglobin-II from Mycobacterium tuberculosis THE ROLE OF G8 TRYPTOPHAN. *J. Biol. Chem.* **284**,

3106–3116 (2009).

122.   Li, H., Leung, K.-S., Ballester, P. J. & Wong, M.-H. istar: A Web Platform for Large-Scale Protein-

Ligand Docking. *PLOS ONE* **9**, e85678 (2014).

123.   Gheyouche, E. *et al.* DockNmine, a Web Portal to Assemble and Analyse Virtual and

Experimental Interaction Data. *International Journal of Molecular Sciences* **20**, 5062 (2019).

124.   Vangone, A. *et al.* Large-scale prediction of binding affinity in protein–small ligand complexes:

the PRODIGY-LIG web server. *Bioinformatics* **35**, 1585–1587 (2019).

125.   Hsin, K.-Y. *et al.* systemsDock: a web server for network pharmacology-based prediction and

analysis. *Nucleic Acids Res* **44**, W507–W513 (2016).

126.   Pevzner, Y., Frugier, E., Schalk, V., Caflisch, A. & Woodcock, H. L. Fragment-based docking:

development of the CHARMMing Web user interface as a platform for computer-aided drug design.

J Chem Inf Model **54**, 2612–2620 (2014).

127.   Dhanik, A., McMurray, J. S. & Kavraki, L. E. DINC: A new AutoDock-based protocol for docking

large ligands. *BMC Structural Biology* **13**, S11 (2013).

128.   Ouyang, X., Zhou, S., Ge, Z., Li, R. & Kwoh, C. K. CovalentDock Cloud: a web server for automated

covalent docking. *Nucleic Acids Research* **41**, W329–W332 (2013).

129.   Lee, H. S. & Zhang, Y. BSP-SLIM: a blind low-resolution ligand-protein docking approach using

predicted protein structures. *Proteins* **80**, 93–110 (2012).

130.   Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web

service based on EADock DSS. *Nucleic Acids Res* **39**, W270–W277 (2011).

131.   London, N., Raveh, B., Cohen, E., Fathi, G. & Schueler-Furman, O. Rosetta FlexPepDock web

server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Res* **39**, W249-253

(2011).

132.   Irwin, J. J. *et al.* Automated Docking Screens: A Feasibility Study. *J Med Chem* **52**, 5712–5720

(2009).

133.   Turner, H. & Kinet, J. P. Signalling through the high-affinity IgE receptor Fc epsilonRI. *Nature* **402**,

B24-30 (1999).

134.   Hendriks, R. W., Yuvaraj, S. & Kil, L. P. Targeting Bruton's tyrosine kinase in B cell malignancies.

Nat Rev Cancer **14**, 219–232 (2014).

135.   Xu, L. *et al.* Acquired mutations associated with ibrutinib resistance in Waldenström

macroglobulinemia. *Blood* **129**, 2519–2525 (2017).

136.   Kim, S., Kim, D. H., Jung, W.-H. & Koo, J. S. Succinate dehydrogenase expression in breast cancer.

SpringerPlus **2**, 299 (2013).

137.   Dimitrov, L., Hong, C. S., Yang, C., Zhuang, Z. & Heiss, J. D. New developments in the

pathogenesis and therapeutic targeting of the IDH1 mutation in glioma. *Int J Med Sci* **12**, 201–213

(2015).

138.   Geisbrecht, B. V., Liang, X., Morrell, J. C., Schulz, H. & Gould, S. J. The mouse gene PDCR encodes

a peroxisomal delta(2), delta(4)-dienoyl-CoA reductase. *J Biol Chem* **274**, 25814–25820 (1999).

139.   Sun, P. *et al.* Molecular basis for the function of the αβ heterodimer of human NAD-dependent

isocitrate dehydrogenase. *Journal of Biological Chemistry* **294**, 16214–16227 (2019).

140.   Bernstein, F. C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular

structures. *J Mol Biol* **112**, 535–542 (1977).

141.   Hemerly, J. P., Bastos, A. U. & Cerutti, J. M. Identification of several novel non-p.R132 IDH1

variants in thyroid carcinomas. *European Journal of Endocrinology* **163**, 747–755 (2010).

142.   Huang, J. *et al.* Isocitrate Dehydrogenase Mutations in Glioma: From Basic Discovery to

Therapeutics Development. *Front. Oncol.* **9**, (2019).

143.   Javadi, S. A., Hartmann, C., Walter, G. F., Banan, R. & Samii, A. IDH1 Mutation in Brain Stem

Glioma: Case Report and Review of Literature. *Asian J Neurosurg* **13**, 414–417 (2018).

144.   Xu, X. *et al.* Structures of Human Cytosolic NADP-dependent Isocitrate Dehydrogenase Reveal a

Novel Self-regulatory Mechanism of Activity *. *Journal of Biological Chemistry* **279**, 33946–33957

(2004).

145.   Commissioner, O. of the. FDA approves first targeted treatment for patients with relapsed or

refractory acute myeloid leukemia who have a certain genetic mutation. *FDA*

https://www.fda.gov/news-events/press-announcements/fda-approves-first-targeted-treatment-

patients-relapsed-or-refractory-acute-myeloid-leukemia-who (2020).

146.   Huang, L. E. Friend or foe-IDH1 mutations in glioma 10 years on. *Carcinogenesis* **40**, 1299–1307

(2019).

147.   Tejera, D. *et al.* Ivosidenib, an IDH1 inhibitor, in a patient with recurrent, IDH1-mutant

glioblastoma: a case report from a Phase I study. *CNS Oncol* **9**, CNS62 (2020).

148.   Mellinghoff, I. K. *et al.* Ivosidenib in Isocitrate Dehydrogenase 1-Mutated Advanced Glioma. *J

Clin Oncol* **38**, 3398–3406 (2020).

149.   Lu, V. M. & McDonald, K. L. Isocitrate dehydrogenase 1 mutation subtypes at site 132 and their

translational potential in glioma. *CNS Oncol* **7**, 41–50 (2018).

150.   Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res* **35**,

W473-476 (2007).