Journal of Microbiology (2020) Vol. 58, No. 3, pp. 235–244
DOI 10.1007/s12275-020-9563-z
eISSN 1976-3794
pISSN 1225-8873

PROTOCOL

User guide for the discovery of potential drugs *via* protein structure prediction and ligand docking simulation[§]

Bilal Shaker, Myung-Sang Yu, Jingyu Lee, Yongmin Lee, Chanjin Jung, and Dokyun Na*

School of Integrative Engineering, Chung-Ang University, Seoul 06974, Republic of Korea

(Received Nov 25, 2019 / Revised Jan 17, 2020 / Accepted Jan 22, 2020)

Due to accumulating protein structure information and advances in computational methodologies, it has now become possible to predict protein-compound interactions. In biology, the classic strategy for drug discovery has been to manually screen multiple compounds (small scale) to identify potential drug compounds. Recent strategies have utilized computational drug discovery methods that involve predicting target protein structures, identifying active sites, and finding potential inhibitor compounds at large scale. In this protocol article, we introduce an *in silico* drug discovery protocol. Since multi-drug resistance of pathogenic bacteria remains a challenging problem to address, *UDP-N-acetylmuramate-L-alanine ligase* (*murC*) of *Acinetobacter baumannii* was used as an example, which causes nosocomial infection in hospital setups and is responsible for high mortality worldwide. This protocol should help microbiologists to expand their knowledge and research scope.

Keywords: drug discovery, docking, ADMET, protein structure prediction

Overview

The multi-drug resistance of pathogenic bacteria has been an alarming issue for decades because of its fatal effects. Given that drugs are essential for the treatment and prevention of diseases, many attempts have been made to identify drugs to treat multi-drug resistance. Traditional approaches, which involve screening chemical compounds to detect hits, require a great investment in time, cost, and labor (Myers and Baker, 2001; Zizalova *et al.*, 2015). However, recent advances in computational technologies and accumulating protein structure information have allowed for the prediction of potential drugs from the structure of a target protein (Sliwoski *et al.*, 2014). In this protocol article, we introduce steps to obtain protein structures or to predict protein structures and thus predict the structures of potential drug compounds using molecular docking.

Classically, high-throughput screening (HTS) experiments that evaluate a large number of compounds with automated tools have been commonly used to discover potent drug compounds, but HTS requires a great investment in resources, time, and cost (Lavecchia and Di Giovanni, 2013). Thus, pharmaceutical companies have been looking for plausible solutions to avoid such screening, and consequently computational technologies have received a lot of attention (Shoichet, 2004).

Computer-aided drug design (CADD) offers methods to discover and optimize potent drugs *in silico* (Reddy, 2012). Specifically, the purpose of CADD is to virtually screen millions of compounds to identify chemical compounds that can bind both geometrically and chemically to a specific cavity (active site or regulatory site) on a target protein (Reddy, 2012). Thus, CADD tools, which are mostly based on structures, have acquired popularity and have become an essential part of lead discovery, lead optimization, and preclinical trials (Tang *et al.*, 2006; Bharath *et al.*, 2011).

In silico drug discovery involves multiple steps to identify potent drug candidates for a selected target (Fig. 1). Briefly, if the 3D structure of the target protein has not yet been experimentally determined, *in silico* structure prediction tools can be employed to predict the valid 3D structure of the protein. Likewise, the active site of the target protein can be predicted if it has not yet been experimentally identified. Then, the molecular docking of a compound with the target allows for the identification of potential inhibitors (i.e., drug candidates). Thus, compound-target molecular docking is one of the most common methods used to predict interactions (Meng *et al.*, 2011).

The *in silico* approach is able to increase the hit rate because it screens far more compounds than traditional experimental screening approaches and is thereby able to find more druggable compounds (Tang *et al.*, 2006). In addition, the structure-based approach is also able to explain the molecular basis of drug activity and predict possible derivatives that may improve that activity. *In silico* technologies require three major steps: (1) filtering a large number of compounds to identify active compounds (lead compounds), (2) optimizing the lead compounds to enhance their binding affinity, and (3) calculating ADMET (Absorption, Distribution, Meta-

*For correspondence. E-mail: blisszen@cau.ac.kr
[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956

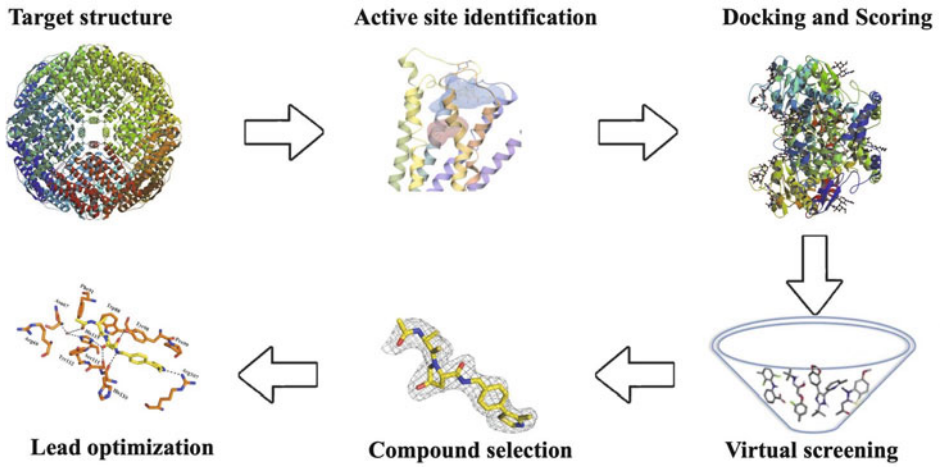

Fig. 1. **Workflow for the structure-based prediction of compound affinity towards a target receptor.**

bolism, Excretion, and Toxicity) properties that describe the physicochemical properties of a given compound that greatly reduce the risk of failure at the preclinical stage (Topliss, 1995).

This protocol introduces how to use tools, online servers, and databases for *in silico* drug discovery. Since this protocol is a simple guide for experimentalists, we introduce only one tool or one webserver for each task. However, there are many available tools performing similar tasks. It is recommended to run several similar tools and see if their results are reproduced. Such comparison of results from several different tools will improve the accuracy of the predictions.

The *UDP-N-acetylmuramate-L-alanine ligase* (*murC*) of a multi-drug resistant pathogen, *Acinetobacter baumannii*, was used as an exemplary model, because *A. baumannii* causes nosocomial infection in hospital setups and is responsible for high mortality worldwide (Roy *et al.*, 2018). All available tools and databases are briefly introduced, and their usage is explained step by step, so experimentalists are able to follow without difficulty.

Applications

CADD plays a key role in the discovery of small molecules of therapeutic significance (Song *et al.*, 2009). CADD tools are employed in almost every stage of the drug discovery process, from the generation of small molecule libraries and hit finding to the optimization of the affinity and sensitivity of the hit molecule. Several early successes in structural design include the carbonic anhydrase inhibitor Dorzolamide and the HIV protease inhibitors Indinavir, Nelfinavir, Ritonavir, and Saquinavir (Kubinyi, 1999). The *in silico* tools outlined in this protocol constitute a simple and easy method to identify potent drug compounds from a large compound library.

PDB database

Information regarding protein structures is important for drug discovery. If the location of each atom and amino acid, particularly amino acids around an active site, is known, chemical compounds may be designed that fit the pocket of a particular active site. Protein databank (PDB) has compiled the structures of proteins and protein-chemical complexes for decades (Burley *et al.*, 2017). In particular, the number of deposited protein structures increased rapidly in the late 1990s. As of November 2019, the database contained about 50,000 distinct protein structures, bringing the total number of deposited structures to 150,000.

3D Structure prediction

Homology modeling : Homology modeling, also termed as comparative modeling, uses experimentally determined 3D structures as a template to predict the conformation of a query sequence (Johnson *et al.*, 1994). Homology modeling assumes that proteins that share high sequence identity also share similar structures with similar functions. Therefore, the accuracy of this prediction method highly depends on the sequence identity between the query and template sequences. If the query sequence shares more than 80% of identity with the template sequence, the resulting predicted structure is usually of high quality (Kopp and Schwede, 2004). If there are no similar proteins with known structures, an *ab initio* modeling approach can be used.

Homology modeling requires four steps to predict structures (Sali and Blundell, 1993): (1) the identification of known 3D structures that show high sequence identity; (2) the alignment of these sequences with a target sequence, the selection of the most identical sequence, and the use of the structure of the identical protein as a template; (3) the prediction of the structure of the query sequence; and (4) the evaluation of the structure using a variety of criteria. If necessary, structure prediction is iterated with different templates until a satisfactory structure is obtained.

***Ab initio* modeling :** Template-free modeling, called *ab initio* modeling, employs a conformational search under the guidance of a designed energy function. Therefore, successful ab initio modeling depends on three factors (Edwards and Cottage, 2001): (1) an accurate energy function with which the protein folds into the most thermodynamically stable state, (2) an efficient search method that can quickly identify the low energy states through a conformational search, and (3) a strategy that can select near-native structures from many

thermodynamically similar structures.

Homology modeling and *ab initio* modeling have their own pros and cons. Thus, a fused approach has been developed, which uses local structure information. Locally conserved sequences can be found through multiple sequence alignment (MSA), and the local sequences share similar structures and functions. MSA can be used to find local regions and to predict the backbone structure of a query sequence like homology modeling. Furthermore, the spatial contacts of amino acid residues can be predicted by correlated mutations (Göbel *et al*., 1994). All local structure information and residue contact information guide the 3D structure prediction with various techniques, such as gradient-based optimization, distance geometry, or fragment assembly. Initial 3D structures are then built with a minimum representation and coarse-grained energy function, and the structures are further refined with energy functions.

Active site prediction

If the active site or regulatory site is not already known, binding site information may be obtained by comparing the target protein with other proteins that share similar functions. Various *in silico* tools for binding site prediction are introduced in Table 4.

The interaction of a protein with other chemicals (e.g., ligand and inhibitors) is critical to its function. Ideally, the target binding site is a small region similar to a pocket that is usually concave with a variety of possible hydrogen bond donors and acceptors and hydrophobic properties (Kalyaanamoorthy and Chen, 2011). Thus, binding site prediction is crucial to understand molecular interactions, and physiochemical properties are the main characteristics required to predict binding sites (Nisius *et al*., 2012).

Compound databases

Numerous compound libraries are freely available online (Villoutreix *et al*., 2007; Jahn *et al*., 2009). Compound databases are used to screen potent drug candidates for the target protein. Compound databases for virtual screening include drug-like small molecules available for purchase or synthesis. Computational methods screen the database for molecules with desirable characteristics, such as stability and solubility in aqueous media, the presence of appropriate functional groups to interact with the biological targets, and the absence of toxic or undesirable moieties. The compound databases are listed in Table 5.

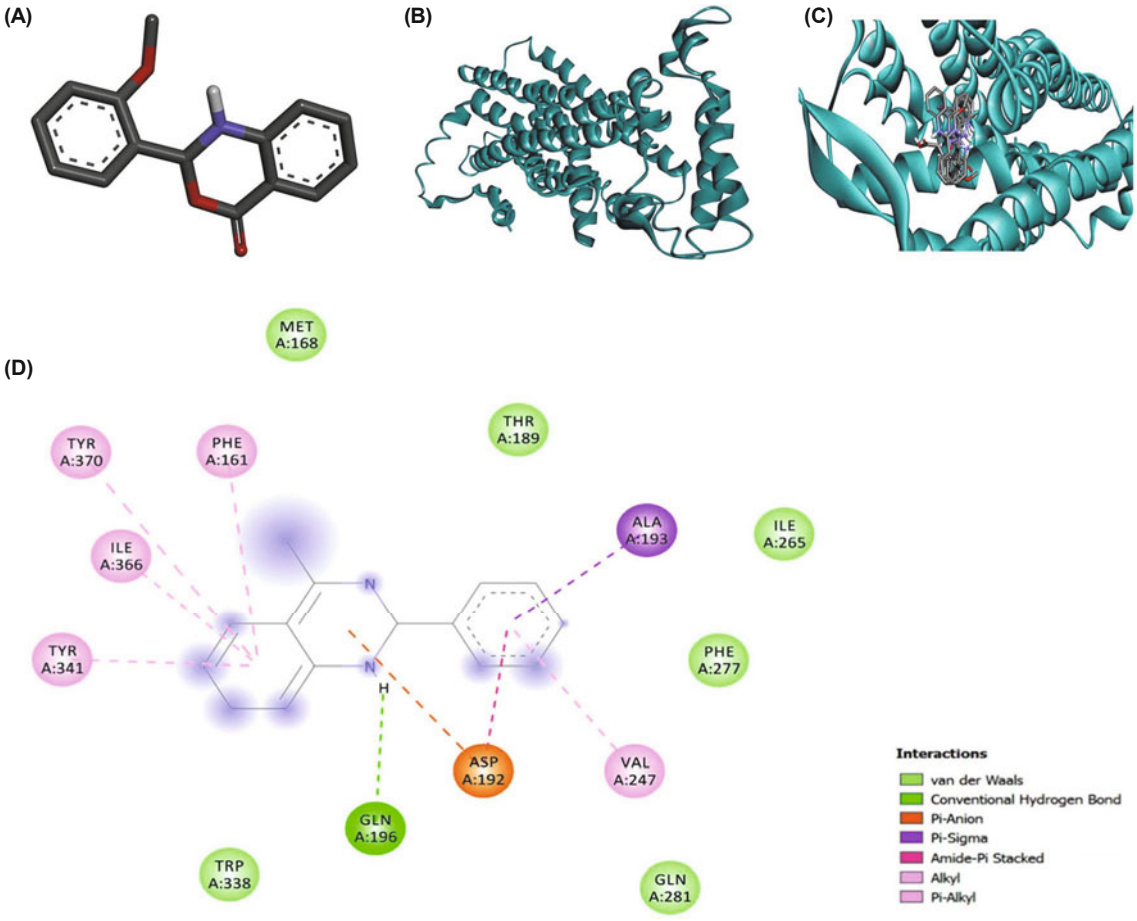

Fig. 2. Molecular docking process. (A) 3D structure of a compound, (B) 3D structure of a target protein, (C) multiple compounds docked on the binding site of a protein, (D) the best-docked compound and its interaction with protein residues.

Molecular docking

Molecular docking is one of the most widely used approaches in structure-based drug design because it can efficiently predict interactions between a compound and the binding site of the target protein (Fig. 2 and Meng *et al.* [2011]). These interactions allow for the elucidation of the behavior of the compounds and the fundamental biochemical processes of the docked complex (Mcconkey *et al.*, 2002). Molecular docking involves two steps: (1) the prediction of the binding affinity of a compound and its target protein and (2) the prediction of the binding pose of the compound within the binding site.

Docking analysis with a large number of compounds increases the probability of obtaining highly potent chemical compounds that interact with the target. The available docking tools are listed in Table 6.

Protocols

If you have no prior knowledge on the target protein, start from section 1 (*Search for your enzyme in PDB*). If you already possess a 3D structure of the target protein and its active site has also been identified, proceed to section 6 (*Molecular Docking*).

1. Search for your enzyme in PDB

Step 1. Visit the PDB database at www.rcsb.org. PDB is a database that compiles protein structure information (Berman *et al.*, 2002).
Step 2. *MurC ligase* is used in this example (Ahmad *et al.*, 2019). Enter its protein name *UDP-N-acetylmuramate-L-alanine ligase* and click the *Go* button (Supplementary data Fig. S1). A few structures will be returned. At this moment, the 3D structure of MurC for *A. baumannii* has not yet been experimentally determined, but the 3D structures of the *mur* family are available.
Step 3. For more information, click on *4HV4* (Supplementary data Fig. S2). Its structure information will thus be shown and comprises the following: a structure image, a download menu, literature, macromolecule information, experimental data, and validation information. Under *Macromolecules* (Supplementary data Fig. S3), the whole protein structure is identified except for the gray regions, which are disordered regions for which the specific structure was not determined. In this case, since the structure has not yet been discovered, the structure of the enzyme should be predicted before for further analysis.

2. 3D Structure prediction

2.1. Homology modeling

If the protein structure has not yet been determined, as in the case of MurC, the structures of similar proteins (in terms of sequence and function) are utilized to predict the structure by a process called homology modeling. There are several tools for homology modeling (Table 1).

This protocol uses SWISS-MODEL, a web-based tool of high reliability (Bertoni *et al.*, 2017). SWISS-MODEL takes an amino acid sequence as a query and predicts its 3D structure from the sequence. Follow the steps below for structure prediction.
Step 1. Go to the UniProt website (www.uniprot.org) to obtain the MurC sequence (Apweiler *et al.*, 2004). UniProt is a database of protein sequences and other protein-related information. The database has collected 561,000 manually curated proteins.
Step 2. Enter the name of the target enzyme with a specific organism name for sequence retrieval (Supplementary data Fig. S4). For example, *UDP-N-acetylmuramate-L-alanine ligase A. baumannii*, and then press *Search*. Many enzymes from the same family will be returned (Supplementary data Fig. S5).
Step 3. The result page shows the entry (UniProt accession number), protein name, and gene name, organism, and sequence length. Use the information to find *UDP-N-acetylmuramate-L-alanine ligase*. In the first column, click on the entry *A3M9Y0*. This is MurC of *A. baumannii*. More specific information of the MurC ligase will be shown on the web page (Supplementary data Fig. S6).
Step 4. To obtain the MurC sequence, click on *Sequence* in the left menu, and the sequence of the MurC ligase will be returned (Supplementary data Fig. S7). Now click on *FASTA* and save the sequence.
Step 5. Go to the SWISS-MODEL web page (swissmodel.expasy.org). Click on *Start Modeling* (Supplementary data Fig. S8) and paste the sequence into the *Target sequence*(*s*). The user can also enter the UniProt accession number (A3-M9Y0) instead of the sequence. Enter the information for *Project Title* and *Email* (Supplementary data Fig. S9).
Step 6. Click on *Search for Templates* and select the sequence with highest GMQE score (an estimated accuracy score of a predicted structure) and resolution < 2Å (Supplementary data Fig. S10). In this protocol, the sequence with the maxi-

Table 1. Homology-based structure prediction tools

Name	Description	Link
SWISS-MODEL (Bertoni *et al.*, 2017)	Consists of three modeling modes: 1) automated mode, 2) alignment mode, and 3) project mode.	https://swissmodel.expasy.org
PSIPRED (Buchan and Jones, 2019)	Comparative modeling and fold recognition incorporates two feed-forward neural networks and uses stringent cross validation to evaluate performance.	http://bioinf.cs.ucl.ac.uk/psipred/
3D-Jigsaw (Moal *et al.*, 2018)	Automatic and interactive mode, distant homology detection, and profile-profile comparison algorithm for sequence alignment.	https://bmm.crick.ac.uk/~svc-bmm-3djigsaw/
Phyre-2 (Kelley *et al.*, 2015)	Predicts and analyzes protein structure, function, and mutation.	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
ModWeb (Pieper *et al.*, 2014)	Uses modpipe for template collection from the PDB and UniProt databases to perform homology modeling.	https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

Table 2. Tools for *ab initio* structure prediction

Name	Description	Link
I-Tasser (Yang and Zhang, 2015)	Uses three steps for modeling: 1) identification of similar folds from PDB, 2) the reassembly of folds into a full-length model by replica-exchange MC simulation, and 3) fragment assembly from the SPICKER cluster centroid	https://zhanglab.ccmb.med.umich.edu/I-TASSER/
Raptor-x (Ma *et al.*, 2013)	Predicts 2D structure, 3D structure, solvent accessibility, and disordered regions.	http://raptorx.uchicago.edu/
Robetta (Conchúir *et al.*, 2015)	Performs homology modeling or *de novo* structure prediction. Protein domain prediction.	http://robetta.bakerlab.org/
PSIPRED (Buchan and Jones, 2019)	Performs fold recognition and homology modeling and incorporates two feed-forward neural networks that perform an analysis of the output obtained from PSI-BLAST.	http://bioinf.cs.ucl.ac.uk/psipred/

mum GMQE and resolution of 2.2Å was selected as a template. Click on *Build Model* to generate the 3D structure. Prediction results will be returned on the screen (Supplementary data Fig. S11).

Template selection is a critical step in homology modeling, because the prediction is made based on the structure of a selected template. Thus, it is recommended selecting the structure as a template with highest sequence identity and best resolution to perform more accurate homology modeling.

Step 7. Click on *Structure Assessment* to validate the predicted structures. More than 85% of the residues should appear in the favored region (*Ramachandran Favored*), few outliers (*Ramachandran Outliers*) should be present, a positive *QMEAN* in the *Quality Estimate* should also be present, and a low *MolProbity Score* should be visible (Supplementary data Fig. S12).

Structure assessment tools are used to assure the reliability of 3D structure prediction. Structure validation with multiple assessment tools can improve the structure assessment analysis.

Step 8. Click on the *Download* icon to download the prediction details along with the *PDB* file of the predicted model.

2.2. *Ab initio* modeling

If the structures of similar proteins have also not yet been determined, the protein structure should be predicted from scratch. This is called *ab initio* modeling, and molecular dynamics is a main feature of this type of modeling. There are several tools for *ab initio* modeling (Table 2).

This protocol uses I-Tasser, an online tool to perform *ab initio* modeling. I-Tasser is easy to use and produces more reliable results (Yang and Zhang, 2015). To perform *ab initio* modeling with I-Tasser, follow the steps below.

Step 1. To use I-Tasser services, an academic email is required and the user should register before first use. Visit zhanglab.ccmb.med.umich.edu/I-TASSER/ and look for *Registration* (Supplementary data Fig. S13). After registering, the user will receive a password that will be used to submit queries in I-Tasser.

Step 2. Paste the sequence into the textbox and enter the *email* address and *password* obtained after registration. Then click *Run I-TASSER* (Supplementary data Fig. S14).

Step 3. The user will receive an email from I-Tasser with a web link containing the predicted structure results. Structure prediction generally requires one day and depends on the length of the query sequence. Open the web link to view the prediction details and structures (Supplementary data Fig. S15). Download the PDB files of the predicted structures by clicking on *Download Model*.

I-Tasser server uses C-Score (a combined measure for evaluating global and local similarity between query and template protein) values to rank predicted structures. The score ranges from 0 to 1, and a higher value represents more confident prediction result. Download the model with the highest C-Score for further analysis.

3. Structure validation

Once the structure has been predicted, it should be validated prior to further analysis. Several web servers are available for structure validation (Table 3).

This protocol uses SWISS-MODEL by which predicted structures can be evaluated using Ramachandran plot, Quality features, and MolProbity. To perform the validation, follow the steps below.

Step 1. Go to the SWISS-MODEL main page (swissmodel.expasy.org), click on *Tools*, and select *Structure Assessment*. Upload the predicted PDB file by clicking on *Upload Coordinate File* and then press *Start assessment* (Supplementary data Fig. S16).

Step 2. The validation results are presented graphically to facilitate their understanding. More than 85% of the residues should be in the favored region (*Ramachandran Favoured*), there should be few outliers (*Ramachandran Outliers*), a positive *QMEAN* in the *Quality Estimate*, and a low *MolProbity* score (Supplementary data Fig. S12).

Table 3. Structure assessment tools

Name	Description	Link
SWISS-MODEL (Bertoni *et al.*, 2017)	Calculates the Ramachandran plot with model quality at global and local levels.	https://swissmodel.expasy.org/
Verify-3D (Eisenberg *et al.*, 1997)	Operates on the "3D-1D profile" of the protein structure, which includes the statistical preferences for the following criterial: 1) the buried residue area, 2) fraction of the side-chain area that is covered by polar atoms, and 3) the local secondary structure.	http://servicesn.mbi.ucla.edu/Verify3D/
Ramachandran Plot (Arlaud *et al.*, 2002)	Describes amino acid torsional angles (phi and psi). Visualizes the favored backbone regions.	http://mordred.bioc.cam.ac.uk/~rapper/rampage.php

Table 4. Online webservers for binding site prediction

Name	Description	Link
3Dligandsite (Wass *et al.*, 2010)	Finds structural similarity with known ligand-binding sites by using a MAMMOTH search of the structure library.	http://www.sbg.bio.ic.ac.uk/~3dligandsite/
Metapocket (Zhang *et al.*, 2011)	Performs three steps: 1) based-method calling, 2) meta-pocket site generation, and 3) binding-residue mapping.	https://projects.biotec.tu-dresden.de/metapocket/
Fpocket (Schmidtke *et al.*, 2011)	Detects cavities based on Voronoi tessellation. Designed to search for conserved pockets in protein families	http://fpocket.sourceforge.net/
CASTp (Tian *et al.*, 2018)	Includes annotated functional information of the active residues in the protein structure for active pocket prediction. Annotations are derived from PDB, Swiss-Prot, and OMIM, etc.	http://sts.bioe.uic.edu/castp/index.html?2pk9

4. Protein active site prediction

If the active site of a target protein has not yet been identified experimentally, there are several approaches to predict its active site. The classical identification of the cavity in the protein structure is the roughest approach. For a more accurate prediction, the key amino acids involved in the reaction and the nearby region should be predicted to find an active site. Several computational tools and online servers can be used for this purpose (Table 4).

This protocol uses CASTp to identify active sites and is easy to use (Tian *et al.*, 2018). Follow the steps below for the analysis.

Step 1. Open the CASTp main page (sts.bioe.uic.edu/castp/index.html) and click on *Calculation* from the menu (Supplementary data Fig. S17).

Step 2. Upload the PDB file of the target protein and press the *SUBMIT* button (Supplementary data Fig. S17).

Step 3. The predicted results will be displayed in a 3D representation, and the active residues for the binding site will be highlighted in grey under *Sequence* (Supplementary data Fig. S18).

Step 4. To show all the other predicted active sites, click on *SHOW POCKETS* and check all the boxes in the *Show* column. Then click *UPDATE* (Supplementary data Fig. S19). All the active pockets will be displayed in the view screen (Supplementary data Fig. S20).

CastP uses α shape and the pocket algorithm to delineate and measure the active pockets in proteins. CastP ranks predicted active pockets on the basis of atomic and quantitative characterization (interacting residues inside the pocket, accessibility of ligand, volume and area of pocket) including active pocket formation. Hence, the top-ranked predicted active pocket is regarded as the best active site and can be used for further analysis. Moreover, the user can use prior knowledge on the active pocket if available.

5. Compound databases

Compound databases can be used for the high-throughput screening of new drug compounds. The available databases are listed in Table 5. Download the *SDF* or *Mol2* file from the compound libraries of the databases.

This protocol uses the Asinex database (Lipinski, 2004) as the target database for screening.

Step 1. Visit the website www.asinex.com and select the compound category from the *Research Area* according to the nature of the target (Supplementary data Fig. S21).

Step 2. As this protocol uses a bacterial protein, select *Antibacterial*. Click *SDF* in the *Download* menu and save the *SDF* file.

6. Molecular docking

Computational approaches and machine learning techniques can play a notable role in the prediction of potential interactions between chemical compounds and target proteins. Several web-based and commercial tools for docking are listed in Table 6.

In order to investigate the reproducibility of compound-bound structure, blind docking is generally used. It puts a compound in any position randomly and do the calculation to find the binding region. If the binding pose is the same regardless of starting positions, it is considered reproducible. Otherwise, the parameters of docking software should be tuned.

The PyRx tool is used for docking analysis in this protocol. Follow the steps below for molecular docking.

Step 1. Download PyRx (sourceforge.net/projects/pyrx/), run the *setup*, and open the program (Supplementary data Fig. S22).

Step 2. Add the target *library* by clicking *File – Import*, select *Chemical Table File – SDF*, and click *NEXT* (Supplementary data Fig. S23).

Table 5. Compound databases

Name	Description	Link
Asinex (Lipinski, 2004)	Lead-like, natural, and commercially available compound database.	http://www.asinex.com/
Zinc database (Irwin and Shoichet, 2005)	Commercially available chemical database.	https://zinc.docking.org/
PubChem (Kim *et al.*, 2018)	Contains small and large molecules, such as nucleotides, carbohydrates, lipids, and peptides. Data are collected from chemical vendors, journal publishers, and government agencies.	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL (Davies *et al.*, 2015)	Manually curated chemical database of bioactive molecules with drug-like properties maintained by the European Bioinformatics Institute (EBI).	https://www.ebi.ac.uk/chembl/
ChemSpider (Ayers, 2012)	Chemical structure database providing fast access to over 77 million structures, properties and associated information collected from over 270 data sources.	http://www.chemspider.com/

Table 6. Tools for Molecular docking

Name	Description	Link
PyRx (Dallakyan and Olson, 2015)	Virtual screening software. Uses AutoDock, AutoDock vina, and open Babel.	https://sourceforge.net/projects/pyrx/
GEMDOCK (Yang and Chen, 2004)	Integrated environment of heavily modified tools: protein-ligand profiles, pharmacological interactions, compound clusters, scoring functions, and evolutionary algorithms.	http://gemdock.life.nctu.edu.tw/dock/
Glide (Friesner *et al.*, 2006)	Uses two different scoring functions to rank compounds: 1) SP scoring and 2) XP GlidScore.	https://www.schrodinger.com/glide
GOLD (Jones *et al.*, 1997)	Uses genetic algorithms for protein-ligand docking and allows for full ligand and partial protein flexibility.	https://www.ch.cam.ac.uk/computing/software/gold-suite
ICM-Dock (Fernandez-Recio *et al.*, 2005)	Provides a unique set of tools for accurate ligand-protein docking, peptide-protein docking, and protein-protein docking.	http://www.molsoft.com/docking.html
Molecular Operating Environment MOE (Vilar *et al.*, 2008)	An integrated drug discovery software that is able to design ligands and ligand modifications, providing correlation plots, activity relationships, and the visualization of hydrophobic and charged protein surfaces.	https://www.chemcomp.com/Products.htm
Patch Dock (Schneidman-Duhovny *et al.*, 2005)	Implemented geometry-based molecular docking algorithm with a server for rigid and symmetric docking that performs protein-protein docking and predicts protein-small molecule complexes.	https://bioinfo3d.cs.tau.ac.il/PatchDock/php.php
BIOVIA Discovery Studio (Schuster *et al.*, 2010)	Software suite for small molecules and macromolecule simulation. Performs MD simulations, QM simulations, ligand design, pharmacophore modeling, QSAR, and ADME.	https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/

Step 3. Select the downloaded *SDF* file of the antibacterial library and click *Open* (Supplementary data Fig. S24).
Step 4. All compounds will show up in the *Open Babel* menu below the 3D view with the column headers: *Title, Formula, Weight,* and *Number of atoms* (Supplementary data Fig. S25).
Step 5. Place the mouse curser on any compound in the *Controls* view and right-click. Then select *Minimize All* in the pop-up menu (Supplementary data Fig. S26). Minimization time depends on the number of compounds, and the estimated time will be displayed on the screen.
Step 6. After minimization, right-click on any compound in the *Controls* view once again and select *Convert All to AutoDock Ligand (pdbqt*; Supplementary data Fig. S27). The estimated time will be displayed on the screen.
Step 7. Once the compounds are converted, all compounds will then appear in the *Ligands* box. Click *Vina Wizard* in the *Controls* view and then click *Start* (Supplementary data Fig. S28)
Step 8. To add the target protein, click *Add Macromolecule,* select the target protein, and click *Open*. The protein will be displayed in 3D view and in the *Macromolecules* box (Supplementary data Fig. S29). The user can use the *pdb, pdbqt*, or *mol2* file of the protein in the PyRx tool as an input.
Step 9. To select all ligands, single click on the first ligand then press and hold the *SHIFT* button on the keyboard, drag the bar down to the last ligand, click on the last ligand, and release the *SHIFT* button (Supplementary data Fig. S30). All ligands will be selected, and the total number of selected ligands will be displayed in the *Controls* menu box.
Step 10. Click *Forward* in the *Controls* menu box. The protein with a *Coordinate Selection box* will appear in the *3D View* menu. This coordinate box is used to define the binding site of the protein. The user can move and change the size of the box with the help of the mouse cursor (Supplementary data Fig. S31).
Step 11. Place the box at a predicted binding site of the protein or try to place the box where it can surround the predicted binding site by increasing or decreasing the size of the coordinate box and then click *Forward* (Supplementary data Fig. S32).

The PyRx tool will predict the binding sites inside the box and generate nine models for each ligand with different poses and positions inside the coordinate box. It may take hours to complete docking with each ligand.
Step 12. The results will be displayed in tabular form in the *Controls* view box with different column headers to ensure easy interpretation for the user (Supplementary data Fig. S33). Lower negative values indicate higher binding affinity, and the *Mode* column contains the generated model number for each ligand.
Step 13. Click in the *Binding Affinity (kcal/mol)* column to sort the results in ascending order (Supplementary data Fig. S34). Note down the ligand names that show high binding affinity and mode number.
Step 14. Click the *File – Export* menu to retrieve the docked models and save them in the desire destination folder with the default file name *PyRx_AutoDocl4.tar.gz* by clicking on *Browse* and then click *Finish* (Supplementary data Fig. S35).
Step 15. To save the docking score in Excel, click on the *blue disk* button in the upper-right corner, set the destination folder, and click *Save* after designating the desired file name (Supplementary data Fig. S36).
Step 16. Open the destination folder and extract *PyRx_AutoDocl4.tar.gz.*
Step 17. Open the *Macromolecules* folder and navigate to the folder *murC.* Find the *pdbqt* file of the protein. In this protocol, *murC.pdbqt* is the target protein file and the rest of the files are compound files (Supplementary data Fig. S37). The Pyrx docking tool converts all input files into the *pdbqt* format for use and provides the results in the same format.

To remove the ligands and macromolecules for the next use, select all the desired ligands, right-click on the selection, and click *Delete* (Supplementary data Fig. S38).
Step 18. To visualize the results in 3D, download BIOVIA Discovery Studio Visualizer (www.3dsbiovia.com/products/

collaborative-science/biovia-discovery-studio/) and click on *Download: Discovery Studio Visualizer* (Supplementary data Fig. S39).

Step 19. Provide the required information and click *Submit*. Select the *Client* based on the computer operating system (e.g., Windows or Linux) and click on the available link below the desired operating system to obtain the setup file (Supplementary data Fig. S40).

Step 20. Run the setup file, install the program, and open DS visualizer (Supplementary data Fig. S41).

Step 21. Drag and drop the files to open the *protein* and *compound* files in the same tab of DS studio. First drag the protein file *murC.pdbqt* from the result folder and drop into the welcome screen of *DS visualizer*. The protein will be displayed in 3D view (Supplementary data Fig. S42).

Step 22. Select the compound based on the *binding affinity score*, and drag and drop the *pdbqt* file of the compound into the same *screen* of discovery studio to visualize the docked pose with the protein. In this protocol, the compound *murC_BDA_25045353_uff_3=757.63* was selected for visualization based on its binding affinity (Supplementary data Fig. S43). Use one compound at a time to visualize the docking pose with the protein.

Step 23. Right-click and use the mouse wheel to rotate, zoom in, and zoom out of the 3D complex. Clicking on the different models of the compound will show the different binding locations and poses. Compare the compound binding position with the predicted binding site and choose the compound model that shows binding with the predicted binding position and high binding affinity. In this protocol *murC_BDA_25045353_uff_3=757_model_0* was selected based on its high binding affinity and binding position.

Step 24. Click on *Check box* of the compound model and click *File* and then *Save As*. Set the destination folder, enter the desired file name, and select the Protein Data Bank file extension (*.pdb:*.pdb1:*.ent; Supplementary data Fig. S44). This will generate a single complex file (protein-ligand) with the selected model of the compound. Repeat these steps for each compound.

Step 25. Open the saved complex file in DS visualizer again and click *Show 2D Diagram*. This will show all interaction sites and the binding patterns of the compound and protein (Supplementary data Fig. S45).

7. ADMET properties predictions

During drug development ADMET properties analysis of a lead compound is important given that most compounds are withdrawn because of ADMET deficiencies during preclinical stages. Computational approaches help predict these properties of a compound, and online servers are available to predict ADMET properties (Table 7). PreADMET is used in this protocol for ADMET analysis. Most of the tools accept MOL files, which are generally classified as data files that contain molecular data, atom, bond, coordinate, and connectivity information, or SMILES files, which are simple, concise, and contain rather readable molecule structure specification formats, as inputs. To create the *MOL* or *SMILES* file of a compound, follow these steps.

Step 1. Visit chemdrawdirect.perkinelmer.cloud/js/sample/index.html#, which is an online platform to draw 2D structures and to convert a structure into multiple formats. Users can easily draw the 2D structure of compounds using the drawing menu or can obtain MOL files with the compound name (Supplementary data Fig. S46).

Step 2. To draw a structure, select the desire chemical component (e.g., ring or bond) from the *tool bar* with a single left-click and then click on the sheet to draw. To join two components together, select the component from the *tool bar*, place the mouse cursor on the joining point, and *left-click* to draw when the blue color appears (Supplementry data Fig. S47). Click on the *A* button to add a single letter representation of an atom in the compound. Users can use the *2D interaction diagram* of DS studio to draw the compound structure.

Step 3. The user can use the compound name to obtain a CDXML file by clicking on *Utilities* and selecting *Convert Name to CDXML*. Enter the name in the search box and click *OK* or draw the structure of the compound (Supplementary data Fig. S48).

Step 4. A CDXML file will show in the text box. Click *Copy* and then *Close* (Supplementary data Fig. S49).

Step 5. Click *Structure* and select *Load CDXML* and *Paste* the CDXML file in the box and click *OK* (Supplementary data Fig. S50). A 2D structure will appear on the screen.

Step 6. Once the structure is displayed on the screen, click *Structure* again and select *Get MOL*. Then, click *Copy* and *Close* (Supplementary data Fig. S51). Users can convert the structure into multiple formats, such as *Smiles* or *InChlKey*, from the same menu.

Use this text of the MOL file as a query tool for ADMET prediction. In this protocol, preADMET was used to evaluate the compound. Follow these steps to predict ADMET properties.

Table 7. ADMET property prediction tools

Name	Description	Link
SwissADME (Daina *et al.*, 2017)	Computes physicochemical descriptors, ADMET parameters, pharmacokinetic properties, the druglike nature, and medicinal chemistry friendliness of small molecules.	http://www.swissadme.ch/index.php
PROTOX (Banerjee *et al.*, 2018)	Incorporates molecular similarity, pharmacophores, fragment propensities, and machine-learning models to predict the toxicity, hepatotoxicity, cytotoxicity, carcinogenicity, mutagenicity, and immunotoxicity of a compound.	http://tox.charite.de/protox_II/
ADMET Modeler (Gola *et al.*, 2006)	A QSAR/QSPR model building module in the ADMET predictor. Model building algorithms are artificial neural networks that employ regression, kernel partial least square, and multiple linear regression techniques.	https://www.simulations-plus.com/software/admetpredictor/admet-modeler/
PreADMET (Lee *et al.*, 2004)	Builds drug-like library and predicts ADMET data. The four principle components are: 1) a molecular descriptor calculation, 2) a drug-likeness prediction, 3) the ADMET prediction, 4) and a toxicity prediction.	https://preadmet.bmdrc.kr/

Step 1. Visit preadmet.bmdrc.kr/_and select the prediction target from the menu. As ADMET properties were selected to be predicted, click *ADME Prediction* (Supplementary data Fig. S52).
Step 2. Click on the *Open* button (yellow folder), paste the text of the MOL file of the compound into the input box, and click *Load* (Supplementary data Fig. S53).
Step 3. This tool translates the MOL file into a 2D structure that will be displayed in the input box. Click *submit* to predict properties (Supplementary data Fig. S54)
Step 4. In the results window, each property has a specific value. Based on these values, a compound can be evaluated (e.g., whether the compound is absorbed in the intestine). Click on the *Green Plus* button for a more detailed description of each property (Supplementary data Fig. S55).

The user can predict toxicity and drug-likeness using the menus *Toxicity Prediction* and *Drug-Likeness Prediction*.

Conclusion

For decades, *in silico* drug discovery has paved a new path for the discovery new drug candidates. The available computational tools have become easy to use, even for experimental biologists. These tools will expand the scope of research areas by allowing microbiologists to perform *in silico* drug discovery.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (No. NRF-2018R1A5A1025077 and No. NRF-2019M3-E5D4065682). This work was also supported by the National Research Foundation of Korea (NRF) and the Center for Women in Science, Engineering and Technology (WISET) Grant funded by the Ministry of Science and ICT (MSIT) under the Program for Returners into R&D.

Conflicts of Interest

There's no conflict of interest.

References

Ahmad, S., Murtaza, U.A., Raza, S., and Azam, S.S. 2019. Blocking the catalytic mechanism of MurC ligase enzyme from *Acinetobacter baumannii*: An *in silico* guided study towards the discovery of natural antibiotics. *J. Mol. Liq.* **281**, 117–133.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* 2004. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119.

Arlaud, G.J., Gaboriaud, C., Garnier, G., Circolo, A., Thielens, N.M., Budayova-Spano, M., Fontecilla-Camps, J.C., and Volanakis, J.E. 2002. Structure, function and molecular genetics of human and murine C1r. *Immunobiology* **205**, 365–382.

Ayers, M. 2012. ChemSpider: The free chemical database. *Ref. Rev.* **26**, 45–46.

Banerjee, P., Eckert, A.O., Schrey, A.K., and Preissner, R. 2018. ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res.* **46**, W257–W263.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., *et al.* 2002. The protein data bank. *Acta Crystallogr. Sect. D Biol.* **58**, 899–907.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. 2017. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7**, 10480.

Bharath, E.N., Manjula, S.N., and Vijaychand, A. 2011. *In silico* drug design tool for overcoming the innovation deficit in the drug discovery process. *Int. J. Pharm. Pharm. Sci.* **3**, 8–12.

Buchan, D.W.A. and Jones, D.T. 2019. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407.

Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., and Velankar, S. 2017. Protein data bank (PDB): The single global macromolecular structure archive. *Methods Mol. Biol.* **1607**, 627–641.

Conchúir, S.Ó., Barlow, K.A., Pache, R.A., Ollikainen, N., Kundert, K., O'Meara, M.J., Smith, C.A., and Kortemme, T. 2015. A web resource for standardized benchmark datasets, metrics, and rosetta protocols for macromolecular modeling and design. *PLoS One* **10**, e134033.

Daina, A., Michielin, O., and Zoete, V. 2017. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **7**, 42717.

Dallakyan, S. and Olson, A.J. 2015. Small-molecule library screening by docking with PyRx. *Methods Mol. Biol.* **1263**, 243–250.

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., and Overington, J.P. 2015. ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **43**, W612–W620.

Edwards, Y.J. and Cottage, A. 2001. Prediction of protein structure and function by using bioinformatics. *Methods Mol. Biol.* **175**, 341–375.

Eisenberg, D., Luthy, R., and Bowie, J.U. 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* **277**, 396–404.

Fernandez-Recio, J., Totrov, M., Skorodumov, C., and Abagyan, R. 2005. Optimal docking area: A new method for predicting protein-protein interaction sites. *Proteins* **58**, 134–143.

Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., and Mainz, D.T. 2006. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **49**, 6177–6196.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317.

Gola, J., Obrezanova, O., Champness, E., and Segall, M. 2006. ADMET property prediction: the state of the art and current challenges. *Qsar Comb. Sci.* **25**, 1172–1180.

Irwin, J.J. and Shoichet, B.K. 2005. Zinc - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182.

Jahn, A., Hinselmann, G., Fechner, N., and Zell, A. 2009. Optimal assignment methods for ligand-based virtual screening. *J. Cheminform.* **1**, 14.

Johnson, M.S., Srinivasan, N., Sowdhamini, R., and Blundell, T.L. 1994. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.* **29**, 1–68.

Jones, G., Willett, P., Glen, R.C., Leach, A.R., and Taylor, R. 1997. Development and validation of a genetic algorithm for flexible

docking. *J. Mol. Biol.* **267**, 727–748.
Kalyaanamoorthy, S. and Chen, Y.P. 2011. Structure-based drug design to augment hit discovery. *Drug Discov. Today* **16**, 831–839.
Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858.
Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., *et al.* 2018. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109.
Kopp, J. and Schwede, T. 2004. The SWISS-MODEL repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.* **32**, D230–D234.
Kubinyi, H. 1999. Chance favors the prepared mind--from serendipity to rational drug design. *J. Recept. Signal Transduct. Res.* **19**, 15–39.
Lavecchia, A. and Di Giovanni, C. 2013. Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* **20**, 2839–2860.
Lee, S.K., Chang, G.S., Lee, I.H., Chung, J.E., Sung, K.Y., and No, K.T. 2004. The PreADME: PC-based program for batch prediction of ADME properties. *EuroQSAR 2004.* **9**, 5–10.
Lipinski, C.A. 2004. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discov. Today Technol.* **1**, 337–341.
Ma, J., Wang, S., Zhao, F., and Xu, J. 2013. Protein threading using context-specific alignment potential. *Bioinformatics* **29**, i257–i265.
Mcconkey, B.J., Sobolev, V., and Edelman, M. 2002. The performance of current methods in ligand-protein docking. *Curr. Sci.* **83**, 845–856.
Meng, X.Y., Zhang, H.X., Mezei, M., and Cui, M. 2011. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.* **7**, 146–157.
Moal, I.H., Chaleil, R.A.G., and Bates, P.A. 2018. Flexible protein-protein docking with swarmdock. *Methods Mol. Biol.* **1764**, 413–428.
Myers, S. and Baker, A. 2001. Drug discovery - an operating model for a new era. *Nat. Biotechnol.* **19**, 727–730.
Nisius, B., Sha, F., and Gohlke, H. 2012. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **159**, 123–134.
Pieper, U., Webb, B.M., Dong, G.Q., Schneidman-Duhovny, D., Fan, H., Kim, S.J., Khuri, N., Spill, Y.G., Weinkam, P., Hammel, M., *et al.* 2014. Modbase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **42**, D336–D346.
Reddy, M.R. 2012. Use of computer aided drug design methods in the discovery of a new class of clinical candidates for diabetes. *Abstr. Pap. Am. Chem. S.* **243**.
Roy, R., Tiwari, M., Donelli, G., and Tiwari, V. 2018. Strategies for combating bacterial biofilms: a focus on anti-biofilm agents and their mechanisms of action. *Virulence* **9**, 522–554.
Sali, A. and Blundell, T.L. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
Schmidtke, P., Bidon-Chanal, A., Luque, F.J., and Barril, X. 2011. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **27**, 3276–3285.
Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* **33**, W363–W367.
Schuster, D., Waltenberger, B., Kirchmair, J., Distinto, S., Markt, P., Stuppner, H., Rollinger, J.M., and Wolber, G. 2010. Predicting cyclooxygenase inhibition by three-dimensional pharmacophoric profiling. Part I: Model generation, validation and applicability in ethnopharmacology. *Mol. Inform.* **29**, 75–86.
Shoichet, B.K. 2004. Virtual screening of chemical libraries. *Nature* **432**, 862–865.
Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W.Jr. 2014. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395.
Song, C.M., Lim, S.J., and Tong, J.C. 2009. Recent advances in computer-aided drug design. *Brief Bioinform.* **10**, 579–591.
Tang, Y., Zhu, W., Chen, K., and Jiang, H. 2006. New technologies in computer-aided drug design: Toward target identification and new chemical entity discovery. *Drug Discov. Today Technol.* **3**, 307–313.
Tian, W., Chen, C., Lei, X., Zhao, J.L., and Liang, J. 2018. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* **46**, W363–W367.
Topliss, J.G. 1995. Computer-aided drug design in industrial research - a management perspective. *Ernst Schering Res. Found. Workshop* **15**, 11–38.
Vilar, S., Cozza, G., and Moro, S. 2008. Medicinal chemistry and the molecular operating environment (MOE): Application of QSAR and molecular docking to drug discovery. *Curr. Top. Med. Chem.* **8**, 1555–1572.
Villoutreix, B.O., Renault, N., Lagorce, D., Sperandio, O., Montes, M., and Miteva, M.A. 2007. Free resources to assist structure-based virtual ligand screening experiments. *Curr. Protein Pep. Sci.* **8**, 381–411.
Wass, M.N., Kelley, L.A., and Sternberg, M.J.E. 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **38**, W469–W473.
Yang, J.M. and Chen, C.C. 2004. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **55**, 288–304.
Yang, J. and Zhang, Y. 2015. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181.
Zhang, Z., Li, Y., Lin, B., Schroeder, M., and Huang, B. 2011. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27**, 2083–2088.
Zizalova, J., Rrahmaniova, D., Svorcikova, J., and Vrubel, F. 2015. The relation between real costs of drugs temporarily reimbursed in mode of coverage with evidence development and budget impact analysis submitted as a mandatory requirement of the application. *Value Health* **18**, A567.