J. Chem. Sci. Vol. 129, No. 5, May 2017, pp. 515–531. © Indian Academy of Sciences.
DOI 10.1007/s12039-017-1268-4

REGULAR ARTICLE

Assessing therapeutic potential of molecules: molecular property diagnostic suite for tuberculosis (MPDS^{TB})

ANAMIKA SINGH GAUR[a], ANSHU BHARDWAJ[b], ARUN SHARMA[b], LIJO JOHN[a], M RAM VIVEK[a], NEHA TRIPATHI[c], PRASAD V BHARATAM[c], RAKESH KUMAR[b], SRIDHARA JANARDHAN[a], ABHAYSINH MORI[c], ANIRBAN BANERJI[a,†], ANDREW M LYNN[e], ANMOL J HEMROM[e], ANURAG PASSI[b], APARNA SINGH[a], ASHEESH KUMAR[g], CHARUVAKA MUVVA[d], CHINMAI MADHURI[f], CHINMAYEE CHOUDHURY[a], D ARUN KUMAR[a], DEEPAK PANDIT[f], DEEPAK R. BHARTI[e], DEVESH KUMAR[g], ER AZHAGIYA SINGAM[d], GAJENDRA PS RAGHAVA[b], HARI SAILAJA[h], HARISH JANGRA[c], KAAMINI RAITHATHA[h], KARUNAKAR TANNEERU[a], KUMARDEEP CHAUDHARY[b], M KARTHIKEYAN[f], M PRASANTHI[a], NANDAN KUMAR[a], N YEDUKONDALU[a], NEERAJ K RAJPUT[b], P SRI SARANYA[a], PANKAJ NARANG[e,†], PRASUN DUTTA[h], R VENKATA KRISHNAN[c], REETU SHARMA[a], R SRINITHI[a], RUCHI MISHRA[g], S HEMASRI[a], SANDEEP SINGH[b], SUBRAMANIAN VENKATESAN[d], SURESH KUMAR[g], UCA JALEEL[h], VIJAY KHEDKAR[f], YOGESH JOSHI[f] and G NARAHARI SASTRY[a,*]

[a]Centre for Molecular Modeling, CSIR-Indian Institute of Chemical Technology, Tarnaka, Hyderabad 500 007, India
[b]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh 160 036, India
[c]Department of Medicinal Chemistry, National Institute of Pharmaceutical Education and Research (NIPER), Mohali 160 062, India
[d]Chemical Laboratory, CSIR-Central Leather Research Institute, Chennai 600 020, India
[e]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110 067, India
[f]Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, Pune 411 008, India
[g]Department of Applied Physics, Babasaheb Bhimrao Ambedkar University, Lucknow 226 025, India
[h]Open Source Drug Discovery Consortium, New Delhi, India
E-mail: gnsastry@gmail.com

MS received 5 March 2017; revised 20 March 2017; accepted 22 March 2017

Abstract. Molecular Property Diagnostic Suite (MPDS^{TB}) is a web tool (http://mpds.osdd.net) designed to assist the in silico drug discovery attempts towards Mycobacterium tuberculosis (Mtb). MPDS^{TB} tool has nine modules which are classified into data library (1–3), data processing (4–5) and data analysis (6–9). Module 1 is a repository of literature and related information available on the Mtb. Module 2 deals with the protein target analysis of the chosen disease area. Module 3 is the compound library consisting of 110.31 million unique molecules generated from public domain databases and custom designed search tools. Module 4 contains tools for chemical file format conversions and 2D to 3D coordinate conversions. Module 5 helps in calculating the molecular descriptors. Module 6 specifically handles QSAR model development tools using descriptors generated in the Module 5. Module 7 integrates the AutoDock Vina algorithm for docking, while module 8 provides screening filters. Module 9 provides the necessary visualization tools for both small and large molecules. The workflow-based open source web portal, MPDS^{TB} 1.0.1 can be a potential enabler for scientists engaged in drug discovery in general and in anti-TB research in particular.

Keywords. Tuberculosis; chemoinformatics; open science; neglected diseases; drug discovery portal; web-based technology.

*For correspondence
† Deceased: ANIRBAN BANERJI and PANKAJ NARANG.

1. Introduction

Data and knowledge generated in drug discovery have been escalating exponentially in recent years owing to the demanding nature of pharmaceutical industry to deliver affordable and safer drugs for existing and emerging diseases.[1–10] How to make the knowledge thus generated available to the practicing scientists is an issue of great significance as it reduces the redundancy, enables research activity and focuses on the grand challenges in the healthcare sector.[11–17] Open science and open innovation are extremely important in the drug discovery approaches in general and those directed towards neglected and orphan diseases in particular.[18–27] How the existing knowledge helps medicinal chemists can be addressed by answering the following two questions: a) What is the value or relevance of molecules that were synthesized? and (b) which of those molecules are the most promising? Because currently chemist's ability to synthesize complex molecules has increased tremendously and more often than not, the question is which molecule to synthesize rather than how to synthesize.

Tuberculosis (TB) has become a global threat killing nearly 1.4 million people with 10.4 million new cases in 2015.[28] The disease-causing bacteria Mtb is a rather challenging microorganism that takes over six months of treatment with multiple drugs to curb its infection.[29] The therapeutic interventions further get confounded due to the fact that most of the time Mtb remains in latent phenotype, which is not well understood.[30] The emergence of multi-drug resistant, extensively drug-resistant and totally drug-resistant forms have resulted in long duration therapies with various side effects and toxicity issues with a risk of non-compliance.[31–38] Therefore, the need of new therapy to combat this dreadful disease is inevitable and demands exploration of new chemical space. Computational approaches to obtain and optimize anti-tubercular leads have been extensively employed in this area.[39–44] There are several public databases having diverse chemical classes of the compounds including PubChem, ZINC, KEGG, DrugBank, ASINEX, ChEMBL and NCI.[45–62] The computational tools and databases are required to facilitate rational prioritization and analysis of compounds from the available large chemical space. Further, the development of new algorithms/scripts is required to classify the chemical space for searching or extracting the useful information for each molecule. The tools are needed to predict the physical, chemical and biological properties of small molecules in order to improve search capabilities.

Before synthesizing any molecule, the medicinal chemist should have prior knowledge to optimize various physicochemical properties, structural alerts, in addition to the understanding of protein-ligand interactions that help in improving drug-likeness and avoiding toxicity issues. There are a number of open source scripts and algorithms developed by various developers for solving drug discovery issues.[63,64] Ideally, developing a disease-specific web portal which integrates the publicly available tools could provide a right platform to conduct drug discovery research in TB.

A variety of chemoinformatics analysis tools have been previously implemented in workflow systems. Steinbeck *et al.,* have implemented the chemoinformatics library of Chemistry Development Kit (CDK)[65,66] in the Taverna workflow suite.[67] Steinbeck *et al.*[68] have also implemented CDK in Konstanz information miner (KNIME), which is an open source workflow platform. It contains functions like format conversion, signatures, fingerprints and molecular properties generation. The Galaxy platform[69–71] is another workflow management system that provides easy to use interfaces of tools to the users and allows easy connection of the tools as well. Various instances of Galaxy have already been established.[72] For example, Ballaxy[73] is a Galaxy instance for structural bioinformatics wherein functions like protein preparation, ligand and protein checker, docking and many other tools have been implemented.

The current manuscript presents a web-based MPDS^{TB} Galaxy tool, which provides an open source platform for the chemoinformaticians, bioinformaticians, medicinal chemists, computational biologists, pharmacologists and others scientists to work on the design of anti-tuberculosis (anti-TB) drugs. The Galaxy based web tool is conveniently designed to integrate with any other software or script and can be used by designing user-defined workflows, a feature conveniently exploited by the users of Galaxy in many cases. The MPDS^{TB} tool provides three class of modules: a) Data Library (modules: 1. Literature, 2. Target library, 3. Compound library); b) Data Processing (4. File format conversion, 5. Descriptor calculation); and c) Data Analysis

Principal investigator: G NARAHARI SASTRY
Co-principal investigators: P ANSHU BHARDWAJ, PRASAD V BHARATAM, ANDREW M LYNN, DEVESH KUMAR, GAJENDRA P S RAGHAVA, M KARTHIKEYAN, SUBRAMANIAN VENKATESAN
Core developers: ANAMIKA SINGH GAUR, ANSHU BHARDWAJ, ARUN SHARMA, LIJO JOHN, M RAM VIVEK, NEHA TRIPATHI, PRASAD V BHARATAM, RAKESH KUMAR, SRIDHARA JANARDHAN, G NARAHARI SASTRY
Co-developers: ABHAYSINH MORI, ANIRBAN BANERJI, ANMOL J HEMROM, ANURAG PASSI, APARNA SINGH, ASHEESH KUMAR, CHARUVAKA MUVVA, CHINMAI MADHURI, CHINMAYEE CHOUDHURY, D ARUN KUMAR, DEEPAK PANDIT, DEEPAK R BHARTI, ER AZHAGIYA SINGAM, HARI SAILAJA, HARISH JANGRA, KAAMINI RAITHATHA, KARUNAKAR TANNEERU, KUMARDEEP CHAUDHARY, M PRASANTHI, NANDAN KUMAR, N YEDUKONDALU, NEERAJ K RAJPUT, P SRI SARANYA, PANKAJ NARANG, PRASUN DUTTA, R VENKATA KRISHNAN, REETU SHARMA, R SRINITHI, RUCHI MISHRA, S HEMASRI, SANDEEP SINGH, SURESH KUMAR, UCA JALEEL, VIJAY KHEDKAR, YOGESH JOSHI.

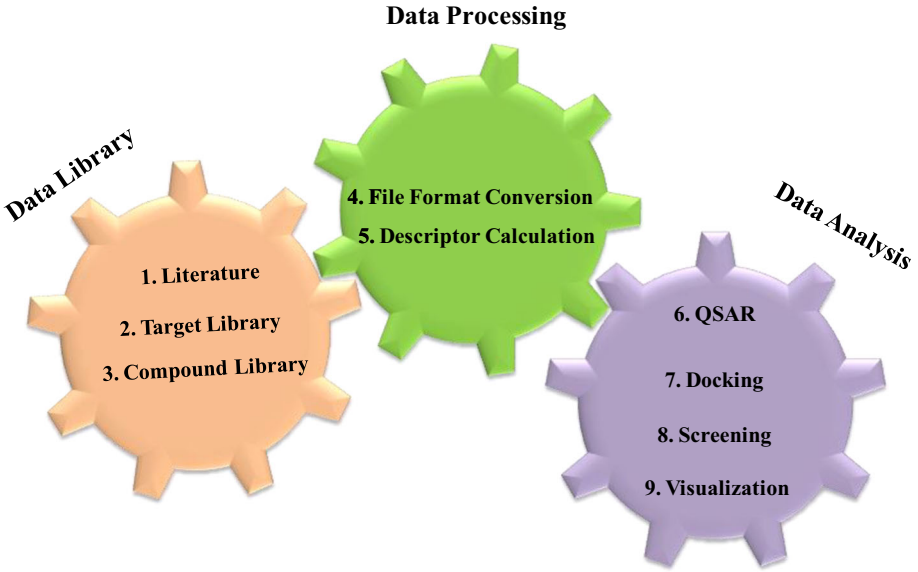


Figure 1. MPDS is structured into data library (literature, target library, compound library), data processing (file format conversion, descriptor calculation) and data analysis (QSAR, docking, screening and visualization).

(6. QSAR, 7. Docking, 8. Screening, 9. Visualization) (Figure 1). While modules 1, 2, 7 and 8 are specific to a particular disease (in this case TB), it is quite possible to make at least three out of the four modules, 2, 7, and 8 as generic. Efforts are underway to achieve this in the near future. However, in MPDS^{TB}1.0.1, only five modules Compound library, QSAR, Docking, Screening, Visualization are generic in nature and can be used 'as is' in drug discovery platforms directed towards other diseases. As the main focus of the current endeavor is to store all the data that is generated for each of the molecules, we decided to generate a unique MPDS ID, which is akin to Aadhar number in India, or social security number in the USA where all the information pertaining to a molecule is stored. A structure-based classification tool has been employed which facilitates the navigation through the chemical space.

The current work has the following major objectives: a) Quantitatively evaluating the multifarious aspects of drug-likeness of a given molecule, in order to diagnose its potential application as a drug; b) Calculate various drug-like physico-chemical properties for prioritization of compounds; c) Provide necessary framework to employ virtual screening of large dataset of compounds; d) Help synthetic medicinal chemists for the design of novel compounds; e) Coordinate the strategic development and integration of chemoinformatics efforts; f) Develop new multi-disciplinary collaborative projects. The current work focuses on the anti-tubercular lead discovery, design and optimization. Expectedly, a suitably altered protocol can be generated for other disease specific Galaxy web MPDS portals by customizing some of the modules. Thus, MPDS in the long run, can emerge as a general purpose open source drug discovery web portal.

2. Methods and modules

Galaxy (http://galaxyproject.org/) is a web-based workflow management system implemented in Python programming language, which is widely used for making data libraries, data integration, data processing and data analysis. In the current work, MPDS^{TB} is developed using Linux (CentOS 6.4) operating system having the python version 2.7. It provides a graphical user interface (GUI) to many computational tools that helps in computational chemistry, drug design, image analysis, climate modeling, linguistics, and biomedical research. Basically, it was developed to analyze the genomic data including gene expression, proteomics, transcriptomics, and gene assembly. Galaxy can be used directly on the web or can be installed in local machines which gives freedom to the users to integrate their own tools. It has flexibility in using diverse biological, chemical data formats and it allows the integration of tools that is written in any programming language or script for which a command line invocation can be constructed. Once the piece of code is written, a tool definition file should be written in XML that describes the working of the tool and its input/output parameters (Table 1).

For each module, the XML code and its complete path should be incorporated into the main configuration file of the Galaxy. The new tool implemented gets displayed in the

Table 1. Description of XML file as implemented in Galaxy MPDS[TB].

Tool ID	Description
<tool id>	Gives a unique name to the tool whose description is mentioned in the XML file.
<name>	It has the name of the tool that will be displayed as hyperlink in Galaxy.
<description>	This is displayed just after the hyperlinked name.
<command>	It describes how the tool (which compiler) will be executed and its input and output parameters.
<inputs>	Defines the input parameters.
<outputs>	Defines the output parameters.
<help>	Describes what the tool does.

tool panel of the Galaxy home page. The Galaxy workspace mainly consists of four areas, the first one is the navigation bar which provide links to Galaxy's major components, analysis workspace, workflows, data libraries, and user repositories (histories, shared data, workflows, pages), the second one is the tool panelist containing the analysis tools and data sources available to the user, the third one is detailed panel display interfaces for tools selected by the user and the fourth one is history panel that shows data and the results of analyses performed by the user, as well as automatically tracked metadata and user-generated annotations.

3. Structure of MPDS[TB]

MPDS[TB] is structured into data library (literature, target library, compound library), data processing (file format conversion, descriptor calculation), and data analysis (QSAR, docking, screening, and visualization) (Table 2). Each of these modules is customized for TB drug discovery and will be described in the following sections.

3.1 *Data library*

3.1a *Module 1: Literature*: Module 1 provides information of druggable protein targets/gene information, FDA-approved drugs, and polypharmacology for Mtb. The genetic information provided includes RvID, gene name, gene product, class of protein, structural details from PDB, active site, function, metabolic pathway, localization, method of validation, drug/inhibitor information, druggability index, and mechanism of action. The literature module provides the list of available FDA approved drugs along with their identification, pharmacology, potential targets and corresponding references. The polypharmacology covers the structure of Mtb cell wall, biosynthetic, metabolic pathways (cell wall, chorismate, amino acids, lipids, carbohydrate, cofactor, nitrogen/sulphur, DNA, protein), bibliography and hyperlinks were given to various servers related to TB.

Table 2. Description of various modules in MPDS[TB] 1.0.1.

Category	Modules	Description
Data Library	Module 1: Literature	Contains Mtb proteins and its genetic information; FDA approved drug information and polypharmacological information.
	Module 2: Target Library	Contains crystal structures and homology models for Mtb proteins.
	Module 3: Compound Library	Contains a single window interface for searching a compound in MPDS compound database.
Data Processing	Module 4: File Format Conversion	Conversions of files from one chemical format to another chemical format, 2D to 3D file conversion using Open Babel.
	Module 5: Descriptor Calculation	Calculation of descriptors and fingerprints using PaDEL and CDK tools.
Data Analysis	Module 6: QSAR	Generation of QSAR models using the data mining tools, McQSAR and SVMlight.
	Module 7: Docking	Ligand Optimization; Conformer Generation and Protein-Ligand docking.
	Module 8: Screening	Prioritization of compounds for drug-like features using DruLiTo tool; Biopharmaceutical Classification System (BCS); Identification of toxicophoric groups in a compound.
	Module 9: Visualization	Visualizing protein-ligand interactions using Jmol and Ligplot.

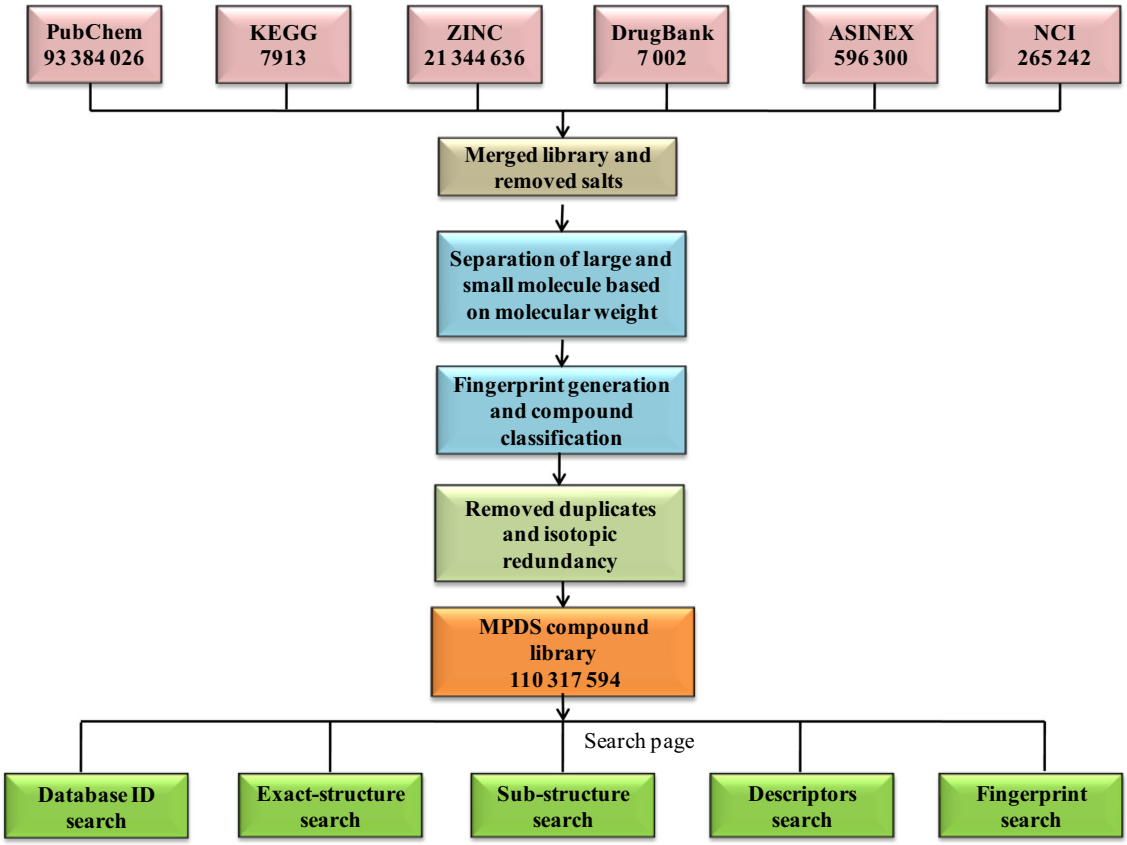


Figure 2. Schema for the generation of MPDS compound library from various data sources and representation of compound search engine.

3.1b *Module 2: Target library*: The target library consists of crystal structures of 140 Mtb proteins which are reported targets and also the ones prioritized through systems level analysis of Mtb interactome. Each protein is also annotated for key residues around the active site. Most of these target structures were collected from protein data bank (PDB) and some of them were homology models. If the structural information is not available, then mutagenesis study results were collected from literature. Multiple sequence alignment of the same family of protein was employed for identification of active site residues for the protein. One of the principal objectives of the target library is to provide a list of prepared proteins in Mtb suitable for molecular docking, and give hyperlink of the data source, if available. The collected PDB structures were prepared by using standard protocols such as assigning bond orders, adding hydrogens, and minimizing protein complexes.

The targets were selected from Mtb H37Rv genome family and they majorly belong to various enzyme classes, such as oxidoreductase, transferase, hydrolase, lyase, isomerase, and ligase. These targets are involved in various biological functions that include signal transduction, peptidoglycan and cell wall synthesis, amino acid synthesis, drug metabolism, DNA precursor synthesis, post-translational modifications and nitrogen metabolism, *etc.* The protein structures in MPDS^{TB} target library can be potentially exploited in structure-based drug design approaches.

3.1c *Module 3: Compound library*: The compound library is generated with the objective of establishing a single window interface to search compounds available across different public domain databases. An efficient small molecule search, implemented using multiple search strategies, facilitates the comprehensive analysis of the available chemical space that may be utilized for identification of novel anti-TB compounds.

Preparation: For the preparation of MPDS^{TB} compound library, existing small molecule databases such as PubChem, Drugbank, KEGG, NCI, ZINC and ASINEX were downloaded. To ensure that globally acceptable standards are followed to store and search this data, each compound in the library is converted into SMILES, InchI and InchIKey using OpenBabel 2.3.2.[74] Indexing of the compound library is done using InchIKey (Figure 2).

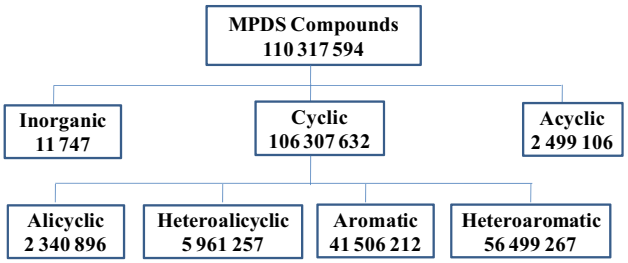


Figure 3. A structure-based chemical classification system of MPDS compound library.

Storage and search: As mentioned above, each compound in the library is stored using InchI standards along with SMILES string and the IDs from where the compound is sourced. Multiple search options are provided through MPDS^{TB} web interface which includes structure-based search, property-based search, and fingerprint-based search (Figure 3). Compound property data is mostly calculated using PaDEL. The pKa values and IUPAC names are calculated using ChemAxon command line tools. The substructure search is implemented using RDkit.[75] A novel fingerprinting algorithm is developed to classify compounds at various levels of structural features. This binary fingerprinting algorithm is developed and employed for the classification and clustering of all the molecules included in the MPDS^{TB} compound library database. Currently, a 30-bit qualitative and partially quantitative fingerprinting scheme is adopted to cluster similar compounds together and reduce the search space. Furthermore, two open source tools, 'molecule cloud' and 'open molecule generator' are also implemented in compound library module to create a molecular cloud of scaffolds and to generate molecule library respectively.[76,77]

3.2 *Data processing*

3.2a *Module 4: File format conversion*: A core requirement of any workflow management system is the connecting links between different analysis tools. Most often these links are based on the file formats that are read by these tools. As there are various molecular file formats to represent chemical structures, an open source file format converter is implemented to facilitate the creation of workflows over the web. Different tools require specific input file formats and will produce output in another specific format. In order to maintain a smooth flow of data between different tools in a workflow, the file format converter module utilizes one of the tools from the Galaxy toolshed to convert one file format to another based on user requirements. As of now, a few input formats like mol2, mol, sdf, SMILES, *etc.,* and output formats, mol2, sdf, mol, pdb are incorporated. This module also contains a tool that generates 3D coordinate from 2D structural file or SMILES which utilizes OpenBabel 2.3.2. The 3D structure generated follows geometrical rules based on hybridization of atoms. Once the structure is generated, the stereochemistry of the structure is taken care of and the lowest free energy conformer is generated by MMFF94 force field using weighted rotor search.

3.2b *Module 5: Descriptor calculation*: In order to computationally assess the properties of the compounds present in MPDS^{TB} compound library or those provided by end users, two descriptor calculation tools, namely, PaDEL and CDK, are incorporated in MPDS^{TB}. These tools may be used to calculate different compound properties. The input format for both of these descriptor tools is sdf and provides the output in CSV format. The descriptor module may read the output from compound library search or user uploaded sdf. The output may be used to build machine-learning models for target specific filters, predicting anti-TB properties, drug-like properties or toxicity of the compounds.

3.3 *Data analysis*

3.3a *Module 6: QSAR*: Two methods of data mining, quantitative structure activity relationships (QSAR) and support vector machine (SVM), are incorporated in MPDS^{TB}. A Multi-conformational Quantitative Structure-Activity Relationship (McQSAR) using Genetic algorithms is implemented for developing QSAR models.[78] The 'Build_QSAR_Model' tool of the QSAR module takes the descriptor file of the compounds with known activity and prompts the user to enter the name of the column whose value needs to be predicted (activity, in this case). In order to remove the redundancy and unwarranted features, the user has been given six options. The feature selection options are as follows: exclude correlated descriptors, exclude identical conformers, exclude inactive compounds, exclude sparse conformers, exclude sparse descriptors and exclude descriptors with zeros. The user also has the flexibility to set the number of times user wants cross-validation step to be repeated. The user can perform four kinds of cross-validation using the percentage of bins 3, 5, 7, 10 folds to divide the compounds. McQSAR can be used to predict the activity of new compounds using the model created by the previously mentioned tool. This accepts two input files: one that contains the descriptors of the compounds whose activity needs to be predicted and the second is a model file created by the 'Build_QSAR_Model' tool (Figure 4).

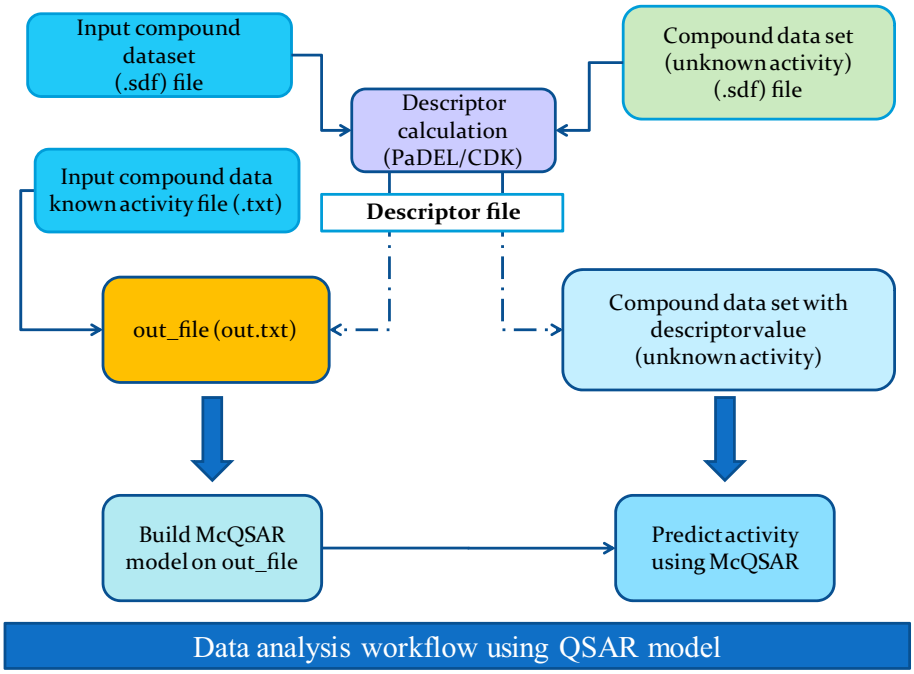


Figure 4. Workflow for QSAR model generation.

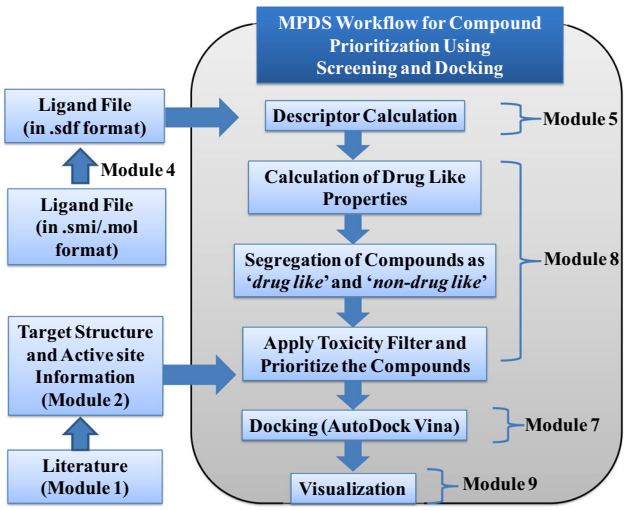


Figure 5. Workflow for compound prioritization using screening and docking modules.

Along with McQSAR, SVMlight[79] has also been incorporated in MPDS[TB] which helps in classifying the compound data into actives and inactives. The 'Build QSAR Model: SVMlight' builds a QSAR model based on descriptor files of active and inactive molecules. Then 'Classify Data: SVMlight' tool takes the model file and the descriptor file of compounds with unknown activity, and classifies them into actives and inactives.

3.3b *Module 7: Docking*: The docking protocol involves multiple steps including energy minimization (chemical structure optimization), conformer generation, docking and its analysis, and visualization. In MPDS[TB], the docking module contains four tools to carry out these steps. The ligand optimization tool incorporates the Phenix electronic Ligand Builder and Optimisation Workbench (eLBOW).[80] The tool uses semi-empirical quantum chemistry based calculations using AM1 (Austin Model 1).[81] Hydrogen atoms are automatically added to the compound by eLBOW. The conformer generation tool utilizes the OpenBabel 2.3.2. genetic algorithm approach to generate diverse conformers based on RMSD. AutoDock Vina[82] has been implemented as the docking software in module 7. Thus, the use of docking module allows the user to carry out efficient and robust docking calculations (Figure 5).

The proteins can be uploaded as a PDB file, or can directly be accessed from target library available in the MPDS[TB] shared library. The ligands can be obtained from the compound library of MPDS[TB] package or a user can upload their choice of ligand in sdf or pdb format with 3D coordinates. The docking calculation can be started using the default parameters and with some user defined parameters. Users can also download the docking results as a zip file containing the complex files in PDB format and the Vina log file containing the ranked binding free energy scores. The docked complexes can be visualized through the visualization module for examining the protein-ligand interactions.

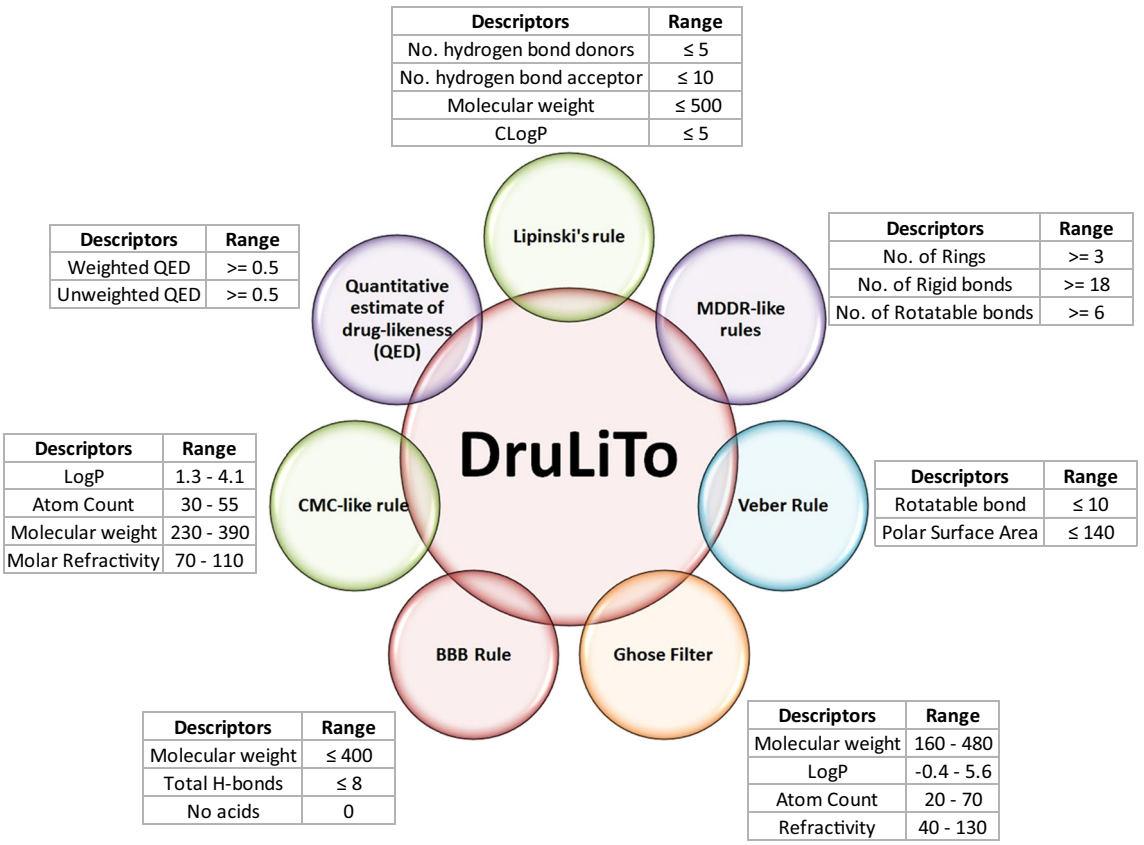


Figure 6. Drug-Likeness prediction tool (DruLiTo) for screening chemical compounds, databases or libraries.

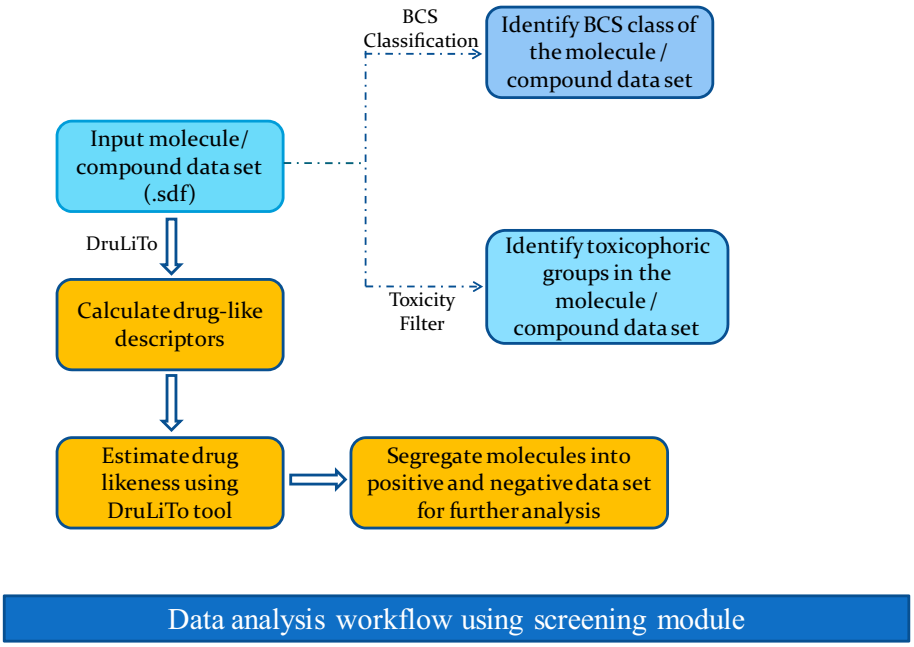


Figure 7. Workflow for compound screening using DruLiTo and filters.

3.3c *Module 8: Screening*: To prioritize compounds for their drug-like features, DruLiTo[83] is integrated into MPDS[TB]. This tool provides the assessment of drug-like features based on eight different published algorithms.[84–88] To calculate drug-likeness of a compound, DruLiTo needs descriptor data, which may be calculated using CDK. DruLiTo uses these descriptors to determine the drug-likeness of the molecule,

based on the different rules as defined by the end user. The user can then segregate the input dataset into passed and failed groups based on one or more rules selected (Figure 6, 7). In addition to this, the biopharmaceutics classification system (BCS) is implemented using in-house scripts to evaluate the solubility and the permeability of compounds. These predictions can subsequently be used to process the positive ligand set or negative ligand set or the original input dataset, all of which are in sdf format, to determine the BCS class (I, II, III, IV) into which the molecules fall based on their predicted permeability and intrinsic solubility values.

Another feature provided in MPDS^{TB} is to perform toxicity analysis. A comprehensive list of structural alerts including data from FAF-Drugs2 (204 substructures), OChem and literature are generated. The toxicity filter performs a substructure search on the query molecule and highlights the matched toxicophoric groups from this data and displays it in an image format. It also produces the mol format of the molecules with toxicophoric groups that can be saved. Along with this, a text file listing the molecule ID, the total number of toxicophoric matches found and the SMARTS pattern of the matched groups can also be generated with this module.

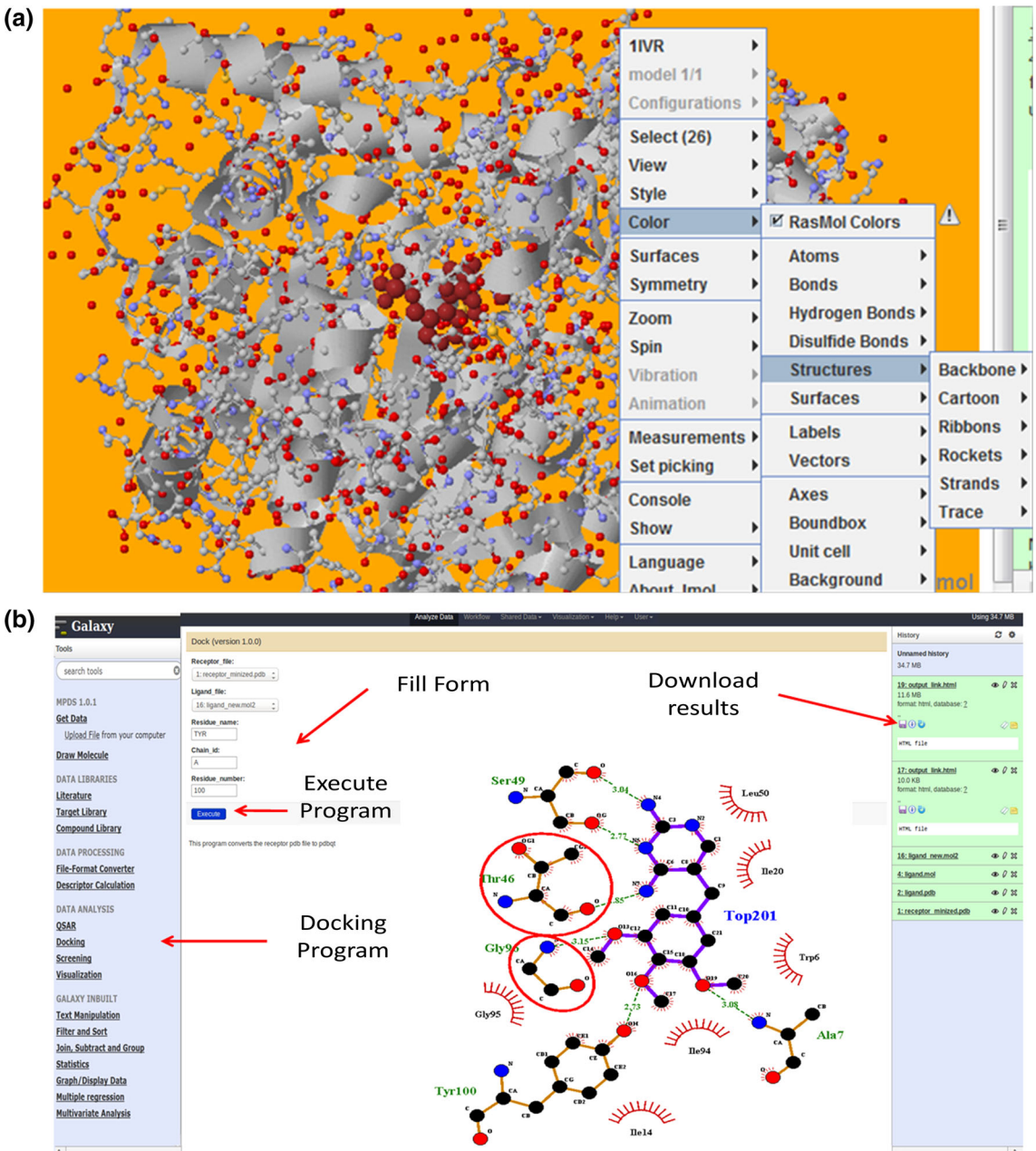


Figure 8. (**a**) Visualization of protein-ligand complex by various structural representations, (**b**) LigPlot for visualizing protein-ligand interactions (hydrogen bonding and hydrophobic contacts).

3.3d *Module 9: Visualization*: Protein-ligand complex can be viewed in LigPlot^{+} tool available in this module in order to examine the protein-ligand interactions. The key interactions of the ligand with active site amino acids are displayed by hydrogen bonding (dashed lines between the atoms) and hydrophobic contacts (an arc with spokes radiating towards the ligand atoms). The protein-ligand complex can be visualized in Jmol by various structural representations (Figure 8).

4. Results and Discussion

4.1 *The compound library and search page*

MPDS^{TB} compound library is a compiled and curated database containing 110.31 million unique compounds (as on 31st December 2016) from various publicly available databases including PubChem, KEGG, ZINC, DrugBank, ASINEX and NCI. Before making unique compounds, salt containing compounds were removed and further, the compounds were separated on the basis of molecular weight taking 750 Da as cut-off. Fingerprints were generated and compounds were classified into classes and clusters. Chemically well-defined classes obtained from fingerprinting algorithm were subjected to redundancy as well as removal of isotopes based on the InchIKey and truncated InChI respectively (Table 3). A class is a chemically well-defined group, whereas cluster is a set of compounds with a size limit of 0.25 million. After classification, the compound library majorly consists of 106.30 million cyclic compounds (2.34 million alicyclic, 5.96 million heteroalicyclic, 41.51 million aromatic and 56.49 million heteroaromatic). For aromatic compounds, class 12 (aromatic compounds contain two rings) has a maximum number of compounds (8.97 million) having 42 clusters, whereas class 20 (aromatic compounds contains more than or equal to four rings) has 0.91 million compounds in clusters. Highest number of compounds is present in the heteroaromatic class 22 (10 million), which are stored in 46 clusters (Figure 9). MPDS^{TB} fingerprinting algorithm and search engine offer a number of advantages over the available fingerprinting algorithms. The various available algorithms match a set of SMARTS patterns against each molecule to calculate each bit of the fingerprint that leads to reduced search time.[89] MPDS-Database uses SMILES notation of the molecules for storage and fingerprinting which makes it faster than the other approaches utilizing SMARTS notation for pattern search and matching. The available fingerprinting algorithms are not suitable for the large data size (110.31 million) as the speed is compromised due to matching a large set of SMARTS patterns against each molecule. The limitation of the current version of MPDS-fingerprinting is that it is not at the stage of being molecule specific. The purpose of designing this fingerprinting algorithm was to reduce the search space to a level where the 2D structural comparison can be performed without compromising the computational power and time utilization. Therefore, more efforts are required to make this fingerprinting robust and molecule specific. Fingerprinting of inorganic molecules is an issue with most of the available fingerprinting algorithms. However, MPDS^{TB} fingerprinting algorithm adopts the structure-based fingerprint classifying the inorganic compounds efficiently. The program for the generation of 30-bit fingerprinting is written in Java (using NetBeans IDE 7.2.1).[90] Further,

Table 3. Distribution of number of compounds in structural classes of MPDS^{TB} compound library.

Chemical Classes	Class‡	No. of clusters$^{£}$	No. of compounds
Acyclic	1	8	1,790,302
	2	4	708,804
Alicyclic	3	7	1,436,757
	4	3	483,877
	5	1	127,301
	6	2	292,961
Heteroalicyclic	7	12	2562,119
	8	11	2,289,075
	9	4	772,643
	10	2	337,420
Aromatic	11	31	6,842,221
	12	42	8,978,649
	13	32	6,898,540
	14	13	2,649,637
	15	31	6,276,919
	16	10	2,071,127
	17	5	1,024,251
	18	15	3,089,744
	19	14	2,771,697
	20	5	903,427
Heteroaromatic	21	16	3,461,762
	22	46	10,086,573
	23	25	5,484,547
	24	32	6,580,855
	25	45	9,268,612
	26	13	2,588,133
	27	12	2,228,509
	28	39	7,829,460
	29	45	8,970,816
Large Molecules	30	7	1,499,109
Inorganic	31	1	11747

‡Class is a chemically well-defined group, $^{£}$Cluster is a set of compounds with a size limit of 0.25 million.

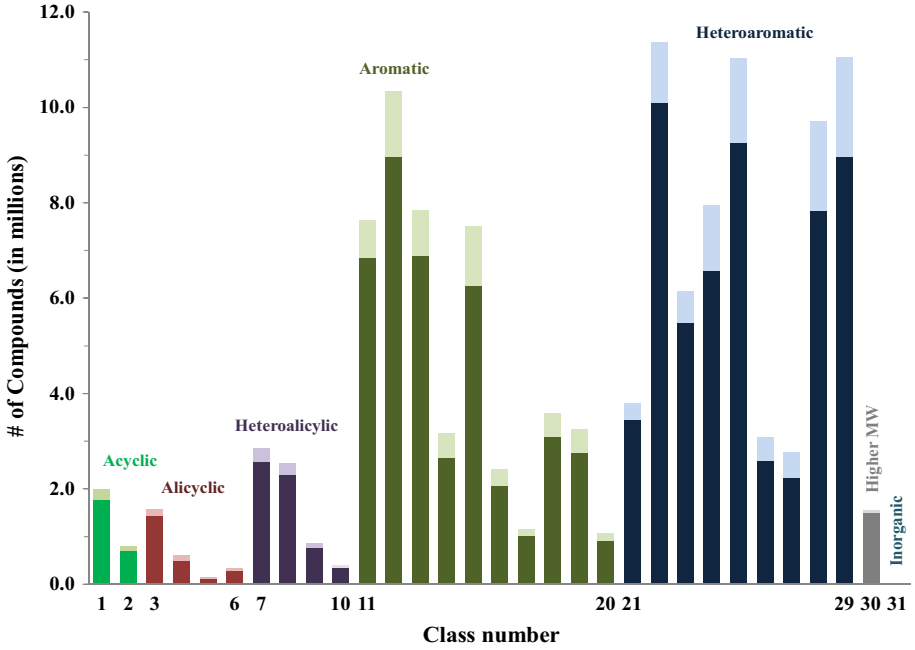


Figure 9. Distribution of unique (dark color) and duplicate (light color) compounds in different chemically well-defined classes.

the choice of SMILES notation for the assignment of fingerprinting and structure search was attributed to the easy and fast handling of the SMILES patterns rather than SMARTS patterns (Table 4).

4.2 *The search page*

MPDS^{TB} provides an interface for searching molecule based on their diverse properties through compound library search tool designed on the Galaxy platform. The user can search the compound library by database ID, exact structure, substructure, descriptor and fingerprint. The result of search page shows a list of molecules fulfilling the criteria, which are further linked to MPDS ID card, displaying the basic information of molecule. The in-house program will calculate molecular fingerprints and searches the query molecules in the database for generating MPDS ID card. For assisting the molecular drawing in MPDS-Galaxy portal, JS draw has been incorporated.

4.3 *Workflow system*

This is one of the most important features of the current platform, which paves the way to generate custom design workflows. The Galaxy workflow system can be used by a graphical drag-and-drop interface in which tools are configured and interconnected between the appropriate input and output points. The user can also extract/download the workflow from the history pane. MPDS^{TB} presents a platform where a data library, data processing and data analysis are configured and interconnected to assist *in silico* drug discovery. The user can utilize the inbuilt workflows or design a new one to carry out chemoinformatics analysis for any given molecule. The MPDS^{TB} workflow system can be interconnected between different modules for a specific task.

4.4 *MPDS ID card*

The output from MPDS^{TB} portal is MPDS ID card (which is akin to Aadhar card in India, or social security number in the USA) that represents a molecular profile report specific to a given molecule. This card reports the vital physico-chemical properties of a molecule essential to estimate the drug-likeness and activity of the molecule in the early stages of the drug discovery and development pipeline. MPDS ID which provides all pertinent information can have several pages. First page is standard and essentially aids in registering the molecule in the database, besides providing vital physico-chemical parameters. However, if more pertinent data are available on the molecule, such as their biological activity, spectroscopic data, toxicity, PK/PD data, *etc*., subsequent pages will be created. The first page of the molecule can be viewed publicly from the portal; second and subsequent pages are stored in the external hard disks at the development site, due to apparent disk storage limitations (Figures 10 and 11).

Table 4. 30-bit fingerprinting scheme and its various levels employed for the classification of molecules in MPDS^{TB} 1.0.1.

Fingerprint Level	Fingerprint bits	Fingerprint bits and Molecular features
Level 1 (Skeleton)	1 2 3 4 5 6	1st bit: Acyclic/Cyclic **Acyclic:** 2nd and 3rd bits: Saturation & Conjugation 4th bit: Straight/Branched 5th bit: Homo/Hetero atomic **Cyclic:** 2nd and 3rd bits: Number of rings 4th and 5th bits: Alicyclic/Aromatic 6th bit: Large/Small
Level 2 (Atom type)	7 8 9 10 11 12	**Acyclic and Cyclic:** 7th bit: Geometrical isomerism 8th bit: Hydrogen Bond Donor 9th bit: Hydrogen Bond Acceptor 10th bit: Halogens 11th bit: Heteroatom in side chain/backbone **Acyclic:** 12th bit: Presence of metal ion **Cyclic:** 12th bit: Presence of fused/unfused rings
Level 3 (Functional group)	13 14 15 16 17 18 19 20	**Acyclic and Cyclic:** 13th bit: Carbonyl group 14th bit: Phosphate containing group 15th bit: Cyanide group 16th bit: Nitro group 17th bit: Sulphur group 18th bit: Amino group 19th bit: Alcohol/Ether group 20th bit: Boron
Level 4 (Size of molecule)	21 22 23 24 25 26	**Acyclic and Cyclic:** 21st, 22nd, 23rd and 24th bits: Number of heavy atoms 25th bit: Chirality 26th bit: Connected/Disconnected structure
Level 5 (Heteroatom type in ring)	27 28 29 30	**Acyclic:** 27th, 28th, 29th and 30th bits: Even/Odd number of carbons **Cyclic:** 27th bit: Nitrogen in ring 28th bit: Oxygen in ring 29th bit: Sulphur in ring 30th bit: Phosphorus in ring
Level 6	–	Number of heteroatom containing rings

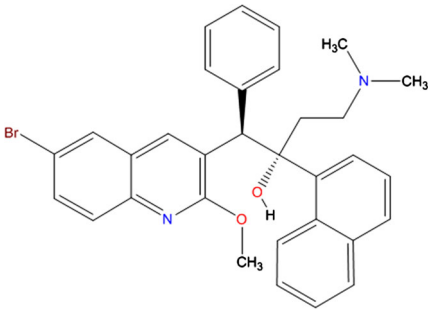
5. Conclusions

MPDS^{TB} 1.0.1 is a comprehensive open source Galaxy-based web tool, which provides a platform to integrate the data collection, processing, and analysis customized towards anti-TB drug discovery and design. One of the main features of this program is its ability to assess and estimate the activity of a given molecule using chemoinformatics, bioinformatics tools, and the existing knowledge in the area. The major attainment of MPDS^{TB} is the creation of a unique compound library (110.31 million) by removing duplicates, isotopic redundancy and salts. A structure-based classification program was developed, which classifies the compound library into 31 classes and 533 clusters.

We believe that the web-based chemoinformatics and modeling tools are great enablers to tackle grand challenges in healthcare in general and drug discovery in particular. One of the major bottlenecks for carrying out research in drug discovery in the academia can be

Molecular Property Diagnostic Suite

MPDS ID:27-09-137939



Molecular Formula:

$C_{32}H_{31}BrN_2O_2$

IUPAC Name:

(1R,2S)-1-(6-Bromo-2-methoxy-3-quinolinyl)-4-(dimethylamino)-2-(1-naphthyl)-1-phenyl-2-butanol

Remarks:

Name/Synonyms: Bedaquilina, Bedaquilinum, TMC207, Sirturo, TMC-207, R207910, TMC207, R207910

Molecular Properties:			
Mol. Wt	555.5	LogP	6.37
HBD	1	LogS	-6.5
HBA	4	pKa	pKa1: 11.64; pKa2: 13.47; pKa3: 7.67; pKa4: 2.67
Molar refractivity	154.02	Polar surface area	45.59
Heavy atom count	37	Aromatic rings count	5
Rotatable bonds	8	Polarizability	57.29

Figure 10. Molecular Property Diagnostic Suite ID card (MPDS ID Card).

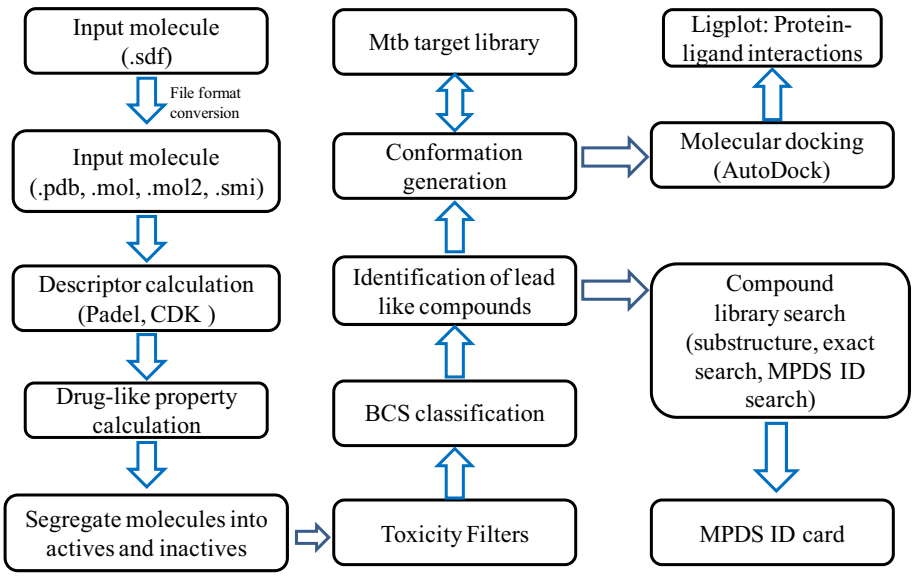


Figure 11. A workflow system for the generation of MPDS ID card by integration of data library, data processing and data analysis modules.

traced to the lack of proper software, access to the comprehensive information on the disease, and also what the contemporary challenges are in developing therapeutics in a given area. Developed countries are not significantly affected by TB and thus the onus of coordinating and conducting frontline research on neglected diseases will be on the developing world. It is very important that India, possessing a tremendous pool of talent and human resource, takes up the leadership role in undertaking research on TB and other neglected diseases. The

current web-based tool, MPDS^{TB} is expected to provide exactly such a platform to drive the research in this area.

Abbreviations

AM1: Austin Model 1; **BCS**: Biopharmaceutics Classification System; **CDK**: Chemistry Development Kit; **CSV**: Comma-Separated Values; **DruLiTo**: Drug Likeness Tool; **eLBOW**: electronic Ligand Builder and Optimisation Workbench; **FAF-Drugs3**: Free ADME-Tox Filtering Tool; **FDA**: Food and Drug Administration; **GUI**: Graphical User Interface; **ID**: Identification number; **InChI**: International Chemical Identifier; **InchIKey**: International Chemical Identifier key; **IUPAC**: International Union of Pure and Applied Chemistry; **KEGG**: Kyoto Enclyclopedia of Genes and Genomes; **KNIME**: Konstanz Information Miner; **McQSAR**: Multi-conformational Quantitative Structure-Activity Relationship; **MMFF94**: Merck Molecular Force Field 94; **MPDS**: Molecular Property Diagnostic Suite; **Mtb**: Mycobacterium tuberculosis; **NCI** National Cancer Institute; **PaDEL**: Pharmaceutical Data Exploration Laboratory; **PDB**: Protein Data Bank; **PK/PD**: PharmacoKinetic/Pharmaco Dynamic; **QSAR**: Quantitative Structure Activity Relationship; **RMSD**: Root Mean Square Deviation; **SDF**: Structure Data File; **SMARTS**: SMiles ARbitrary Target Specification; **SMILES**: Simplified Molecular-Input Line-Entry System; **SVM**: Support Vector Machine; **TB**: Tuberculosis; **XML**: Extensible Markup Language

Acknowledgements

We are thankful to OSDD, CSIR and Sir Dorabji TATA trust for providing TCOF fellowships to some of the authors in the study. CSIR 12th five year program GENESIS (BSC 0121), Department of Science and Technology (New Delhi) and Department of Biotechnology (New Delhi) are also thanked for funding. Code development has taken about 5 years of time starting from 2012 and has witnessed 5 Workshops in IICT, IMTECH, OSDD centre, Bangalore, and NCL. Besides there were several exchange of students between various institutes. We thank CSIR OSDD consortium, NIPER, JNU, and BBAU for providing support. GNS thank J C Bose fellowship of DST. This manuscript is dedicated to the memory of Dr. Anirban Banerji and Dr. Pankaj Narang who have provided a lot of energy and enthusiasm during the kick-start stages of the MPDS teamwork.

References

1. Searls D B 2005 Data integration: Challenges for drug discovery *Nat. Rev. Drug Discovery* **4** 45
2. Nwaka S, Ramirez B, Brun R, Maes L, Douglas F and Ridley R 2009 Advancing drug innovation for neglected diseases-criteria for lead progression *PLoS Negl. Trop. Dis.* **3** e440
3. Sachs J D 2001 A new global commitment to disease control in Africa *Nat. Med.* **7** 521
4. Jagarlapudi S A and Kishan K V 2009 Database systems for knowledge-based discovery *Methods Mol. Biol.* **575** 159
5. Winter M J, Owen S F, Murray-Smith R, Panter G H, Hetheridge M J and Kinter L B 2010 Using data from drug discovery and development to aid the aquatic environmental risk assessment of human pharmaceuticals: Concepts, considerations, and challenges *Integr. Environ. Assess Manage.* **6** 38
6. Lushington G H, Dong Y and Theertham B 2013 Chemical informatics and the drug discovery knowledge pyramid *Comb. Chem. High Throughput Screening* **16** 764
7. Bajorath J 2017 Compound Data Mining for Drug Discovery *Methods Mol. Biol.* **1526** 247
8. Boran A D and Iyengar R 2010 Systems approaches to polypharmacology and drug discovery *Curr. Opin. Drug Discovery Dev.* **13** 297
9. Badrinarayan P and Sastry G N 2011 Virtual high throughput screening in new lead identification *Comb. Chem. High Throughput Screening* **14** 840
10. Reddy A S, Pati S P, Kumar P P, Pradeep H N and Sastry G N 2007 Virtual screening in drug discovery – a computational perspective *Curr. Protein Pept. Sci.* **8** 329
11. Collins P Y, Patel V, Joestl S S, March D, Insel T R, Daar A S; Scientific Advisory Board and the Executive Committee of the Grand Challenges on Global Mental Health, Anderson W, Dhansay M A, Phillips A, Shurin S, Walport M, Ewart W, Savill S J, Bordin I A, Costello E J, Durkin M, Fairburn C, Glass R I, Hall W, Huang Y, Hyman S E, Jamison K, Kaaya S, Kapur S, Kleinman A, Ogunniyi A, Otero-Ojeda A, Poo M M, Ravindranath V, Sahakian B J, Saxena S, Singer P A and Stein D J 2011 Grand challenges in global mental health *Nature* **475** 27
12. Varmus H, Klausner R, Zerhouni E, Acharya T, Daar A S and Singer P A 2003 Public health. Grand Challenges in Global Health *Science* **302** 398
13. Paul S M, Mytelka D S, Dunwiddie C T, Persinger C C, Munos B H, Lindborg S R and Schacht A L 2010 How to improve R&D productivity: The pharmaceutical industry's grand challenge *Nat. Rev. Drug Discov.* **9** 203
14. Dubois D J 2010 Grand Challenges in Pharmacoeconomics and Health Outcomes *Front. Pharmacol.* **1** 7
15. Yildirim O, Gottwald M, Schüler P and Michel MC M 2016 Opportunities and Challenges for Drug Development: Public–Private Partnerships, Adaptive Designs and Big Data *Front. Pharmacol.* **7** 461
16. Gostin L O and Mok E A 2009 Grand challenges in global health governance *Br. Med. Bull.* **90** 78
17. Pai M, Daftary A and Satyanarayana S 2016 TB control: Challenges and opportunities for India *Trans. R. Soc. Trop. Med. Hyg.* **110** 158
18. Wells T N, Willis P, Burrows J N and Hooft V H R 2016 Open data in drug discovery and development: Lessons from malaria *Nat. Rev. Drug Discov.* **15** 661

19. Van Voorhis W C, Adams J H, Adelfio R, Ahyong V, Akabas M H, Alano P, Alday A, Alemán Resto Y, Alsibaee A, Alzualde A, Andrews K T, Avery S V, Avery V M, Ayong L, Baker M, Baker S, Ben Mamoun C, Bhatia S, Bickle Q, Bounaadja L, Bowling T, Bosch J, Boucher L E, Boyom F F, Brea J, Brennan M, Burton A, Caffrey C R, Camarda G, Carrasquilla M, Carter D, Belen Cassera M, Chih-Chien Cheng K, Chindaudomsate W, Chubb A, Colon B L, Colón-López D D, Corbett Y, Crowther G J, Cowan N, D'Alessandro S, Le Dang N, Delves M, DeRisi J L, Du A Y, Duffy S, Abd El-Salam El-Sayed S, Ferdig M T, Fernández Robledo J A, Fidock D A, Florent I, Fokou P V, Galstian A, Gamo F J, Gokool S, Gold B, Golub T, Goldgof G M, Guha R, Guiguemde W A, Gural N, Guy R K, Hansen M A, Hanson K K, Hemphill A, Hooft van Huijsduijnen R, Horii T, Horrocks P, Hughes T B, Huston C, Igarashi I, Ingram-Sieber K, Itoe M A, Jadhav A, Naranuntarat Jensen A, Jensen L T, Jiang R H, Kaiser A, Keiser J, Ketas T, Kicka S, Kim S, Kirk K, Kumar V P, Kyle D E, Lafuente M J, Landfear S, Lee N, Lee S, Lehane A M, Li F, Little D, Liu L, Llinás M, Loza M I, Lubar A, Lucantoni L, Lucet I, Maes L, Mancama D, Mansour N R, March S, McGowan S, Medina Vera I, Meister S, Mercer L, Mestres J, Mfopa A N, Misra R N, Moon S, Moore J P, Morais Rodrigues da Costa F, Müller J, Muriana A, Nakazawa Hewitt S, Nare B, Nathan C, Narraidoo N, Nawaratna S, Ojo K K, Ortiz D, Panic G, Papadatos G, Parapini S, Patra K, Pham N, Prats S, Plouffe D M, Poulsen S A, Pradhan A, Quevedo C, Quinn R J, Rice C A, Abdo Rizk M, Ruecker A, St Onge R, Salgado Ferreira R, Samra J, Robinett N G, Schlecht U, Schmitt M, Silva Villela F, Silvestrini F, Sinden R, Smith D A, Soldati T, Spitzmüller A, Stamm S M, Sullivan D J, Sullivan W, Suresh S, Suzuki B M, Suzuki Y, Swamidass S J, Taramelli D, Tchokouaha L R, Theron A, Thomas D, Tonissen K F, Townson S, Tripathi A K, Trofimov V, Udenze K O, Ullah I, Vallieres C, Vigil E, Vinetz J M, Voong Vinh P, Vu H, Watanabe N A, Weatherby K, White P M, Wilks A F, Winzeler E A, Wojcik E, Wree M, Wu W, Yokoyama N, Zollo P H, Abla N, Blasco B, Burrows J, Laleu B, Leroy D, Spangenberg T, Wells T and Willis P A 2016 Open Source Drug Discovery with the Malaria Box Compound Collection for Neglected Diseases and Beyond *PLoS Pathog.* **28** e1005763
20. Williamson A E, Ylioja P M, Robertson M N, Antonova-Koch Y, Avery V, Baell J B, Batchu H, Batra S, Burrows J N, Bhattacharyya S, Calderon F, Charman S A, Clark J, Crespo B, Dean M, Debbert S L, Delves M, Dennis A S, Deroose F, Duffy S, Fletcher S, Giaever G, Hallyburton I, Gamo F J, Gebbia M, Guy R K, Hungerford Z, Kirk K, Lafuente-Monasterio M J, Lee A, Meister S, Nislow C, Overington J P, Papadatos G, Patiny L, Pham J, Ralph S A, Ruecker A, Ryan E, Southan C, Srivastava K, Swain C, Tarnowski M J, Thomson P, Turner P, Wallace I M, Wells T N, White K, White L, Willis P, Winzeler E A, Wittlin S and Todd M H 2016 Open Source Drug Discovery: Highly Potent Antimalarial Compounds Derived from the Tres Cantos Arylpyrroles *ACS Cent. Sci.* **2** 687
21. Rottmann M, McNamara C, Yeung B K, Lee MC, Zou B, Russell B, Seitz P, Plouffe D M, Dharia N V, Tan J, Cohen S B, Spencer K R, González-Páez GE, Lakshminarayana S B, Goh A, Suwanarusk R, Jegla T, Schmitt E K, Beck H P, Brun R, Nosten F, Renia L, Dartois V, Keller T H, Fidock D A, Winzeler E A and Diagana T T 2010 Spiroindolones, a potent compound class for the treatment of malaria *Science* **329** 1175
22. Meister S, Plouffe D M, Kuhen K L, Bonamy G M, Wu T, Barnes S W, Bopp S E, Borboa R, Bright A T, Che J, Cohen S, Dharia N V, Gagaring K, Gettayacamin M, Gordon P, Groessl T, Kato N, Lee M C, McNamara C W, Fidock D A, Nagle A, Nam T G, Richmond W, Roland J, Rottmann M, Zhou B, Froissard P, Glynne R J, Mazier D, Sattabongkot J, Schultz P G, Tuntland T, Walker J R, Zhou Y, Chatterjee A, Diagana T T and Winzeler E A 2011 Imaging of Plasmodium liver stages to drive next-generation antimalarial drug discovery *Science* **334** 1372
23. Gamo F J, Sanz L M, Vidal J, de Cozar C, Alvarez E, Lavandera J L, Vanderwall D E, Green D V, Kumar V, Hasan S, Brown J R, Peishoff C E, Cardon L R and Garcia-Bustos J F 2010 Thousands of chemical starting points for antimalarial lead identification *Nature* **465** 305
24. Guiguemde W A, Shelat A A, Bouck D, Duffy S, Crowther G J, Davis P H, Smithson D C, Connelly M, Clark J, Zhu F, Jiménez-Díaz M B, Martinez M S, Wilson E B, Tripathi A K, Gut J, Sharlow E R, Bathurst I, El Mazouni F, Fowble J W, Forquer I, McGinley P L, Castro S, Angulo-Barturen I, Ferrer S, Rosenthal P J, Derisi J L, Sullivan D J, Lazo J S, Roos D S, Riscoe M K, Phillips M A, Rathod P K, Van Voorhis W C, Avery V M and Guy R K 2010 Chemical genetics of Plasmodium falciparum *Nature* **465** 311
25. Wells T N 2010 Microbiology. Is the tide turning for new malaria medicines? *Science* **329** 1153
26. Rees S 2015 The promise of open innovation in drug discovery: An industry perspective *Future Med. Chem.* **7** 1835
27. Allarakhia M 2014 The successes and challenges of open-source biopharmaceutical innovation *Expert Opin. Drug Discovery* **9** 459
28. *Global Tuberculosis report*http://apps.who.int/iris/bitstream/10665/250441/1/9789241565394-eng.pdf?ua=1 (accessed on 31st January 2017)
29. *Guidelines for treatment of tuberculosis, fourth edition*http://apps.who.int/iris/bitstream/10665/44165/1/9789241547833_eng.pdf?ua=1&ua=1 (accessed on 31st December 2016)
30. Esmail H, Barry C E, Young D B and Wilkinson R J 2014 The ongoing challenge of latent tuberculosis*Philos. Trans. R. Soc. London, Ser. B* **369** 20130437
31. Davis C E, Carpenter J L, McAllister C K, Matthews J, Bush B A and Ognibene A J 1985 Tuberculosis. Cause of death in antibiotic era *Chest* **88** 726
32. Frieden T R, Sterling T R, Munsiff S S, Watt C J and Dye C 2003 Tuberculosis *Lancet* **362** 887
33. Dye C, Scheele S, Dolin P, Pathania V and Raviglione M C 1999 Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project *JAMA* **282** 677
34. Norton B L and Holland D P 2012 Current management options for latent tuberculosis: a review *Infect. Drug Resist.* **5** 163

35. Johnson R, Streicher E M, Louw G E, Warren R M, van Helden P D and Victor T C 2006 Drug resistance in Mycobacterium tuberculosis *Curr. Issues Mol. Biol.* **8** 97
36. Kremer L S and Besra G S 2002 Current status and future development of antitubercular chemotherapy *Expert Opin. Invest. Drugs* **11** 1033
37. Chan E D and Iseman M D 2008 Multidrug-resistant and extensively drug-resistant tuberculosis: a review *Curr. Opin. Infect. Diseases* **21** 587
38. Daley C L and Caminero J A 2013 Management of multidrug resistant tuberculosis *Semin. Respir. Crit. Care Med.* **34** 44
39. Choudhury C, Priyakumar U D and Sastry G N 2014 Molecular dynamics investigation of the active site dynamics of mycobacterial cyclopropane synthase during various stages of the cyclopropanation process *J. Struct. Biol.* **187** 38
40. Choudhury C, Priyakumar U D and Sastry G N 2015 Dynamics based pharmacophore models for screening potential inhibitors of mycobacterial cyclopropane synthase *J. Chem. Inf. Model.* **55** 848
41. Choudhury C, Priyakumar U D and Sastry G N 2016 Dynamic ligand-based pharmacophore modeling and virtual screening to identify mycobacterial cyclopropane synthase inhibitors *J. Chem. Sci.* **128** 719
42. Janardhan S, Ram Vivek M and Sastry G N 2016 Modeling the permeability of drug-like molecules through the cell wall of Mycobacterium tuberculosis: an analogue based approach *Mol. Biosyst.* **12** 3377
43. Reddy A S, Amarnath H S, Bapi R S, Sastry G M and Sastry G N 2008 Protein ligand interaction database (PLID) *Comput. Biol. Chem.* **32** 387
44. Srivastava H K, Choudhury C and Sastry G N 2012 The efficacy of conceptual DFT descriptors and docking scores on the QSAR models of HIV protease inhibitors *Med. Chem.* **8** 811
45. Dobson C M 2004 Chemical space and biology *Nature* **432** 824
46. Lipinski C and Hopkins A 2004 Navigating chemical space for biology and medicine *Nature* **432** 855
47. Barker A, Kettle J G, Nowak T and Pease J E 2013 Expanding medicinal chemistry space *Drug Discovery Today* **18** 298
48. Reymond J L and Awale M 2012 Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database *ACS Chem. Neurosci.* **3** 649
49. Oprea T I and Gottfries J 2001 Chemography: The art of navigating in chemical space *J. Com. Chem.* **3** 157
50. Xu J and Stevenson J 2000 Drug-like index: A new approach to measure drug-like compounds and their diversity *J. Chem. Inf. Comput. Sci.* **40** 1177
51. 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
52. Bolton E E, Wang Y, Thiessen P A and Bryant S H 2008 PubChem: Integrated platform of small molecules and biological activities *Annu. Rep. Comput. Chem.* **4** 217
53. Wang Y, Xiao J, Suzek T O, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S and Shoemaker B A 2012 PubChem's BioAssay database *Nucleic Acids Res.* **40** D400
54. Vasilevich N I, Kombarov R V, Genis D V and Kirpichenok M A 2012 Lessons from natural products chemistry can offer novel approaches for synthetic chemistry in drug discovery *J. Med. Chem.* **55** 7003
55. Milne G W and Miller J 1986 The NCI drug information system. 1. System overview *J. Chem. Inf. Comput. Sci.* **26** 154
56. Wishart D S, Knox C, Guo A C, Shrivastava S, Hassanali M, Stothard P, Chang Z and Woolsey J 2006 DrugBank: A comprehensive resource for in silico drug discovery and exploration *Nucleic Acids Res.* **34** D668
57. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M and Tanabe M 2014 Data, information, knowledge and principle: Back to metabolism in KEGG *Nucleic Acids Res.* **42** D199
58. Pence H E and Williams A 2010 ChemSpider: An online chemical information resource *J. Chem. Educ.* **87** 1123
59. Chen C Y 2011 TCM Database@ Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico *PLoS One* **6** e15939
60. Kiss R, Sandor M and Szalai F A 2012 http://Mcule.com: A public web service for drug discovery *J. Cheminf.* **4** P17
61. Olah M, Rad R, Ostopovici L, Bora A, Hadaruga N, Hadaruga D, Moldovan R, Fulias A, Mractc M and Oprea T I 2008 In *Small Molecules to Systems Biology and Drug Design -WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery Chemical Biology* S L Schreiber, T M Kapoor and G Wess (Eds.) (Weinheim: Wiley-VCH Verlag GmbH) Vol. **1–3** p. 760
62. Anna G, Louisa J B, Bento A P and Jon C 2012 ChEMBL: A large-scale bioactivity database for drug discovery *Nucleic Acids Res.* **40** D1100
63. Jiang C, Jin X, Dong Y and Chen M 2016 Kekule.js: An Open Source JavaScript Chemoinformatics Toolkit *J. Chem. Inf. Model.* **56** 1132
64. Wojcikowski M, Zielenkiewicz P and Siedlecki P 2015 Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field *J. Cheminf.* **7** 26
65. Kuhn T, Willighagen E L, Zielesny A and Steinbeck C 2010 CDK-Taverna: An open workflow environment for chemoinformatics *BMCBioinformatics* **11** 159
66. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E and Willighagen E 2003 The Chemistry Development Kit (CDK): An open-source Java library for Chemo- and Bioinformatics *J. Chem. Inf. Comput. Sci.* **43** 493
67. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F, Hardisty A, Nieva H A, Balcazar V M P, Sufi S and Goble C 2013 The Taverna workflow suite: Designing and executing workflows of Web Services on the desktop, web or in the cloud *Nucleic Acids Res.* **41** W557
68. Beisken S, Meinl T, Wiswedel B, de Figueiredo L F, Berthold M and Steinbeck C 2013 KNIME-CDK: Workflow-driven cheminformatics *BMC Bioinf.* **14** 257
69. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A and Taylor J 2010 Galaxy: a web-based genome analysis tool for

experimentalists *Curr. Protoc. Mol. Biol.* Chapter 19 Unit 19.10.1-21
70. Afgan E, Baker D, Beek M V D, Blankenberg D, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Eberhard C, Gruning B, Guerler A, Jackson J H, Kuster G V, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A and Goecks J 2016 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update *Nucleic Acids. Res.* **44** W3
71. Blankenberg D, Kuster G V, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy Team, Taylor J and Nekrutenko A 2014 Dissemination of scientific software with Galaxy ToolShed *Genome Biol.* **15** 403
72. *Publicly Accessible Galaxy Servers* https://wiki.galaxyproject.org/PublicGalaxyServers (accessed on 31st December 2016)
73. Hildebrandt A K, Stockel D, Fischer N M, de la Garza L, Kruger J, Nickels S, Rottig M, Scharfe C, Schumann M, Thiel P, Lenhof H P, Kohlbacher O and Hildebrandt A 2015 ballaxy: web services for structural bioinformatics *Bioinformatics* **31** 121
74. O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T and Hutchison G R 2011 Open Babel: An open chemical toolbox *J. Cheminf.* **3** 33
75. *Landrum G RDKit: Open-Source Cheminformatics* http://www.rdkit.org (accessed on 31st December 2016)
76. Ertl P and Rohde B 2012 The Molecule Cloud - compact visualization of large collections of molecules *J. Cheminf.* **4** 12
77. Peironcely J E, Cherto M R, Fichera D, Reijmers T, Coulier L, Faulon J L and Hankemeier T 2012 OMG: Open Molecule Generator *J. Cheminf.* **4** 21
78. Vainio M J and Johnson M S 2005 McQSAR: a multiconformational quantitative structure-activity relationship engine driven by genetic algorithms *J. Chem. Inf. Model.* **45** 1953
79. Joachims T 1999 *Advances in Kernel Methods- Making Large-Scale SVM Learning Practical* B Scholkopf, C Burges and A Smola (Eds.) (Cambridge: MIT-Press) p. 169
80. Moriarty N W, Grosse-Kunstleve R W and Adams P D 2009 electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation *Acta Crystallogr., D: Biol. Crystallogr.* **65** 1074
81. Dewar M J S, Zoebisch E G, Healy E F and Stewart J J P 1985 Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model *J. Am. Chem. Soc.* **107** 3902
82. Trott O and Olson A J 2010 AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading *J. Comput. Chem.* **31** 455
83. *Drug Likeness Tool (DruLiTo)* http://www.niper.ac.in/pi_dev_tools/DruLiToWeb/DruLiTo_index.html (accessed on 31st December 2016)
84. Lipinski C A, Lombardo F, Dominy B W and Feeney P J 2001 Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings *Adv. Drug Delivery Rev.* **46** 3
85. Oprea T I 2000 Property distribution of drug-related chemical databases *J. Comput. -Aided. Mol. Des.* **14** 251
86. Ghose A K, Viswanadhan V N and Wendoloski J J 1999 A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases *J. Comb. Chem.* **1** 55
87. Bickerton G R, Paolini G V, Besnard J, Muresan S and Hopkins A L 2012 Quantifying the chemical beauty of drugs *Nat. Chem.* **4** 90
88. Veber D F, Johnson S R, Cheng H Y, Smith B R, Ward K W and Kopple K D 2002 Molecular properties that influence the oral bioavailability of drug candidates *J. Med. Chem.* **45** 2615
89. Yap C W 2011 PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints *J. Comput. Chem.* **32** 1466
90. Jensen C and Scacchi W 2005 *Collaboration, leadership, control, and conflict negotiation and the netbeans.org open source software development community IEEE* 196b