Journal of Molecular Graphics and Modelling 80 (2018) 95–103

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

# FilTer BaSe: A web accessible chemical database for small compound libraries

Baban S. Kolte[a], Sanjay R. Londhe[a,1], Bhushan R. Solanki[b], Rajesh N. Gacche[c], Rohan J. Meshram[a,*]

[a] *Bioinformatics Centre, Savitribai Phule Pune University, Pune, 411007, India*
[b] *Tech Mahindra Pvt Ltd, Pune, 411006, India*
[c] *Department of Biotechnology, Savitribai Phule Pune University, Pune, 411007, India*

ARTICLE INFO

*Article history:*
Received 27 August 2017
Received in revised form
22 December 2017
Accepted 28 December 2017
Available online 6 January 2018

*Keywords:*
Compound libraries
FilTer BaSe
Virtual screening
Substructure searching
Chemoinformatics
Bioinformatics
Drug discovery

ABSTRACT

Finding novel chemical agents for targeting disease associated drug targets often requires screening of large number of new chemical libraries. *In silico* methods are generally implemented at initial stages for virtual screening. Filtering of such compound libraries on physicochemical and substructure ground is done to ensure elimination of compounds with undesired chemical properties. Filtering procedure, is redundant, time consuming and requires efficient bioinformatics/computer manpower along with high end software involving huge capital investment that forms a major obstacle in drug discovery projects in academic setup. We present an open source resource, FilTer BaSe- a chemoinformatics platform (http://bioinfo.net.in/filterbase/) that host fully filtered, ready to use compound libraries with workable size. The resource also hosts a database that enables efficient searching the chemical space of around 348,000 compounds on the basis of physicochemical and substructure properties. Ready to use compound libraries and database presented here is expected to aid a helping hand for new drug developers and medicinal chemists.

## 1. Introduction

For many years in the past, the drug discovery process prioritized chemical synthesis and in *vitro/in vivo* testing at earlier stages and amendment of compound's Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties at later stages in the drug discovery pipeline. The failure reason of new molecules in that time was attributed to indigent pharmacokinetic properties, lack of efficacy and undesired toxicity effects [1,2]. Reports are published demonstrating that over 90% of compound dropping out of drug discovery pipeline at various stages are attributed to acute toxicity, while two out of three drug candidates are withdrawn from market due to either hepatotoxicity and/or cardiovascular associated complications [3]. However, due to drastic paradigm shift in pharmaceutical research, ADMET testing is now being given central importance at the earlier stages in the drug discovery pipeline. Several *in vitro* and *in silico* approaches have been devised for prediction of some key ADMET properties [4]. In terms of predicting compound properties, the pioneering "rule-of five" can be considered as a cornerstone that handles the issue of oral bioavailability [5]. However, many anti infectious agents, some anti-cancer drugs and natural compounds tend to escape these rules [6]. Therefore, more extensive set of rules are now identified and added to the list of descriptors that are regularly used to precisely predict not only ADMET properties, but also identify drug like properties [7].

There are numerous resources that provide compound collections online [8,9] 4SC (http://www.4sc.com/), Ambinter (http://www.ambinter.com/libraries), ChemDiv (http://www.chemdiv.com/resources/downloads/). Either being commercial and/or provided in unfiltered form limits their utility in academic labs (or small biotech companies) since subsequent filtering procedure require significant amount of time, expert bioinformatics and computational manpower that constitute a major bottle neck. The escalating cost of commercial software that handle large compound datasets and their subsequent virtual screening largely restrains effective drug discovery programs in academic setup where usage of open source software is preferred. To address these issues, we present a chemical resource "FilTer BaSe" that hosts fully filtered compound libraries with manageable size. Short compound

*Abbreviations:* ADMET, absorption distribution metabolism excretion and toxicity; MW, molecular weight; tPSA, topological polar surface area; PAINS, pan assay interference compounds; QASR, quantitative structure–activity relationship.

[*] Corresponding author.
*E-mail address:* rohan@bioinfo.net.in (R.J. Meshram).

[1] Author contributed equally with first author.

https://doi.org/10.1016/j.jmgm.2017.12.020
1093-3263/© 2017 Elsevier Inc. All rights reserved.

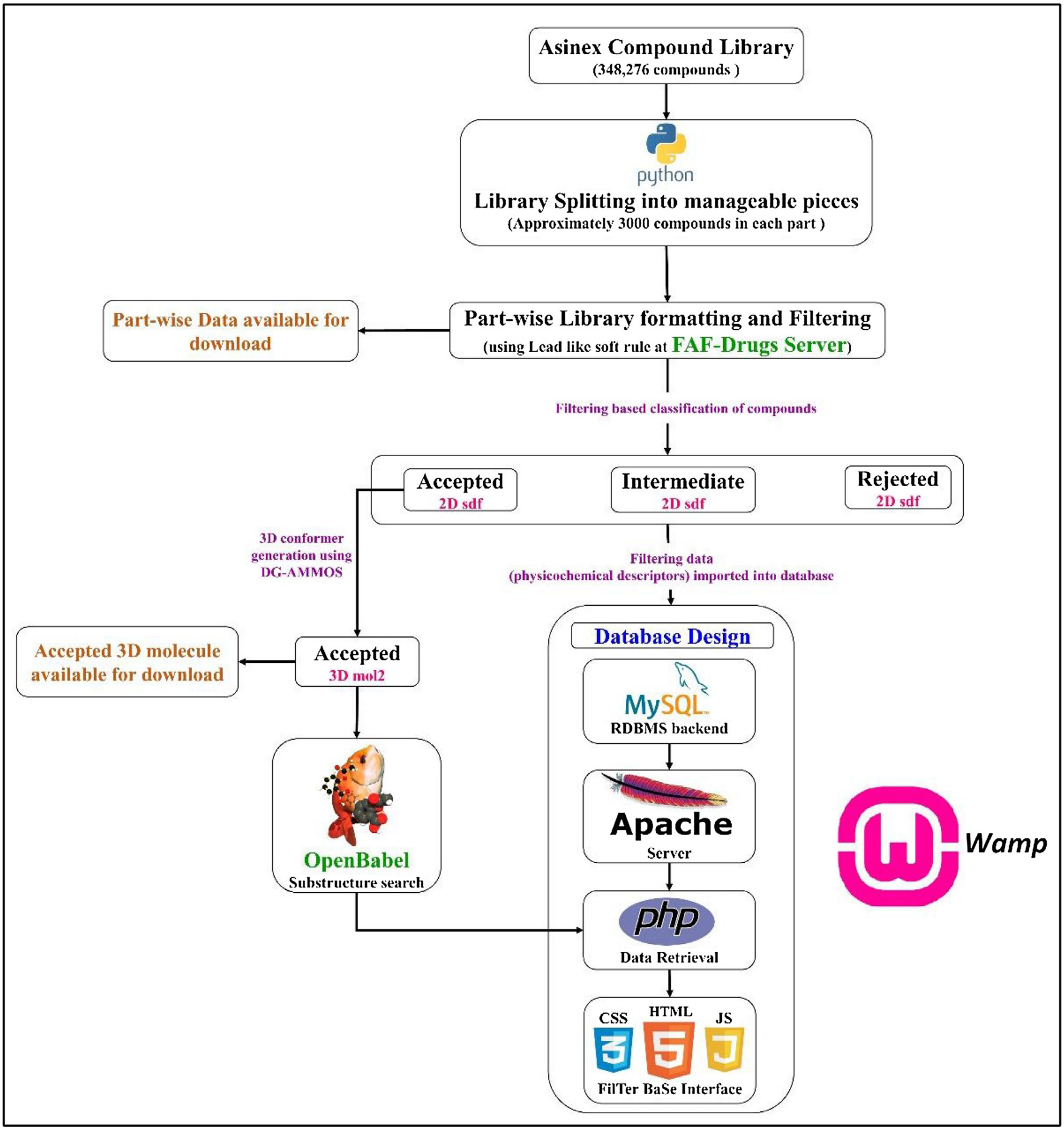

Asinex Compound Library
(348,276 compounds )
python
Library Splitting into manageable pieces
(Approximately 3000 compounds in each part )
Part-wise Library formatting and Filtering
(using Lead like soft rule at FAF-Drugs Server)
Part-wise Data available for download
Filtering based classification of compounds
Accepted
2D sdf
Intermediate
2D sdf
Rejected
2D sdf
3D conformer generation using DG-AMMOS
Filtering data
(physicochemical descriptors) imported into database
Accepted
3D mol2
Accepted 3D molecule available for download
Database Design
MySQL
RDBMS backend
Apache
Server
php
Data Retrieval
CSS
HTML
JS
FilTer BaSe Interface
OpenBabel
Substructure search
Wamp

**Fig. 1.** Workflow implemented in generation of resource FilTer BaSe.

libraries ready for direct virtual screening can be downloaded from here in popular format (sdf and mol2) to reduce the redundant and time consuming work of library filtering.

The resource also host a database that enables user to search for specific compounds. Queries can also be made on physicochemical parameters. Additionally, we also implement a substructure search utility for efficient structure based querying for identification of compounds with specific functional group.

## 2. Methodology

A compound library was obtained, formatted and filtered on basis of standard physicochemical and substructure filtering rules. Filtering classified the compounds into accepted, rejected and intermediate categories. The 3D coordinates for the accepted compounds were generated and provided online in mol2 format. Filtering data is included in a separate database for the efficient physicochemical property based searching. Substructure searching utility was implemented for conducting chemical fragment based searching in accepted class of compounds. The complete workflow implemented in current study is represented in Fig. 1.

### 2.1. Compound library acquisition and splitting

Asinex compound library was obtained from ligand.info Meta-Database (http://ligand.info/) containing 348,276 compounds [10]. Managing large sized compound libraries becomes a difficult task in academic setup with limited computational/bioinformatics human

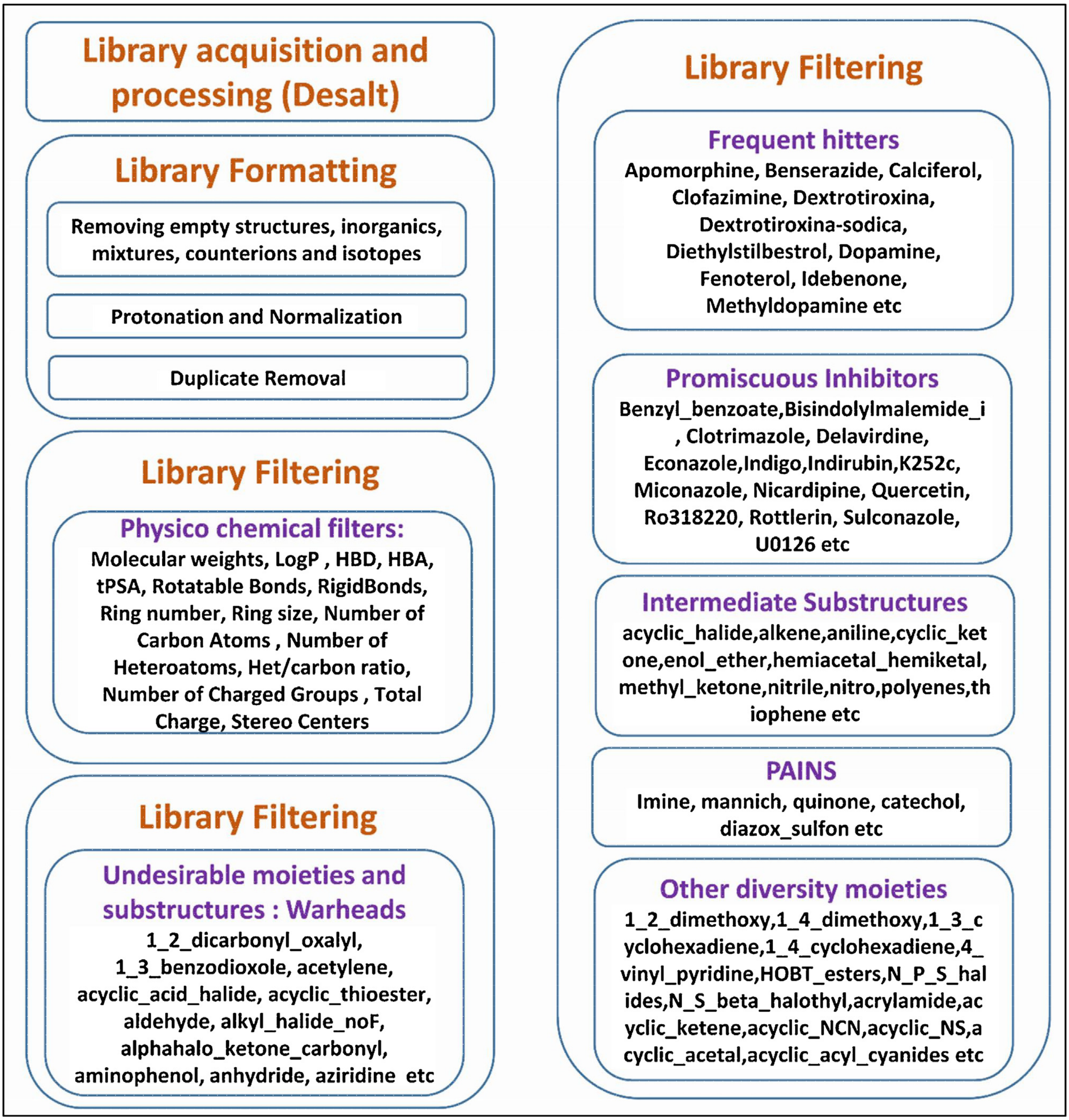
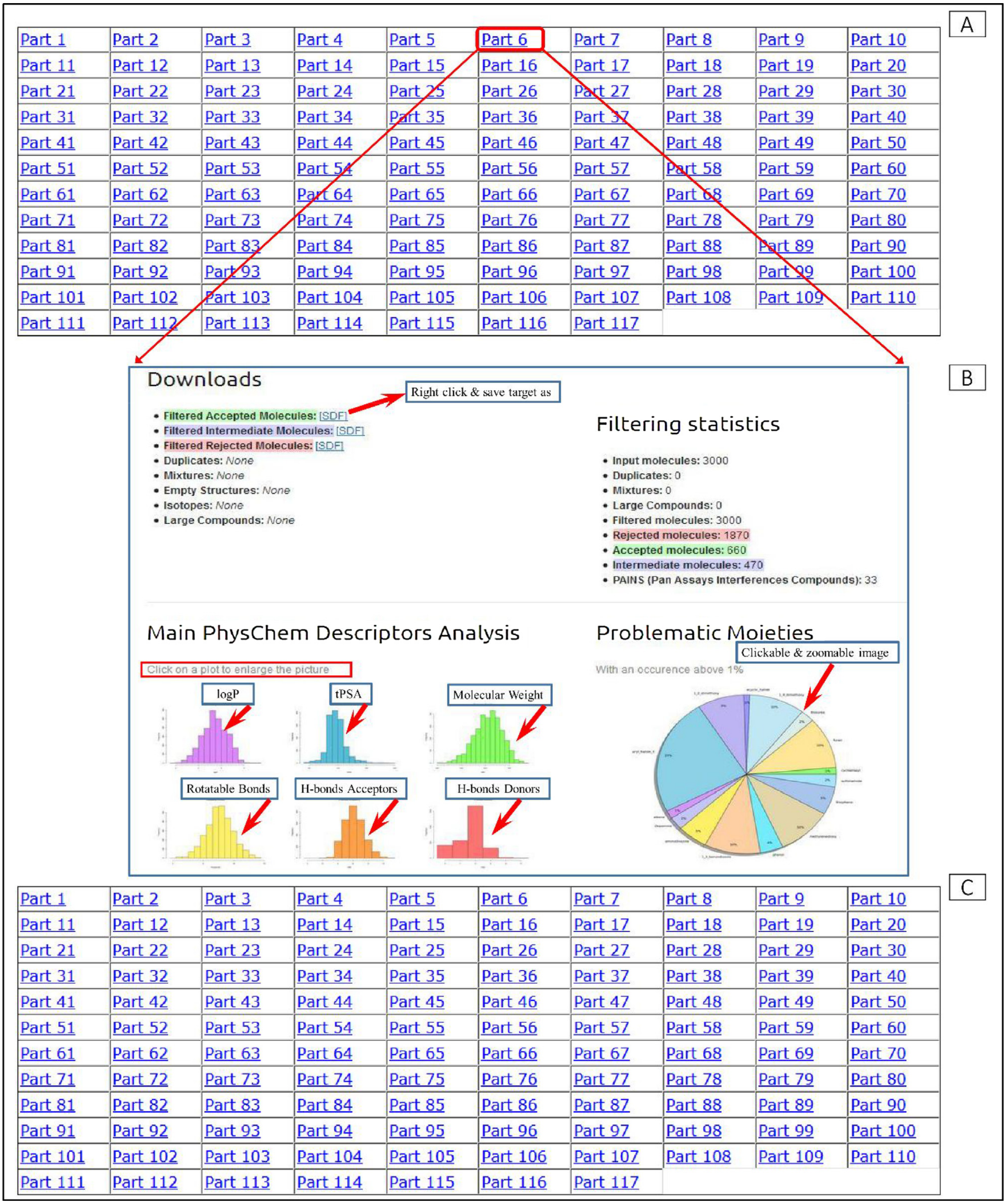

Library acquisition and processing (Desalt)
Library Formatting
Removing empty structures, inorganics, mixtures, counterions and isotopes
Protonation and Normalization
Duplicate Removal
Library Filtering
Physico chemical filters:
Molecular weights, LogP , HBD, HBA, tPSA, Rotatable Bonds, RigidBonds, Ring number, Ring size, Number of Carbon Atoms , Number of Heteroatoms, Het/carbon ratio, Number of Charged Groups , Total Charge, Stereo Centers
Library Filtering
Undesirable moieties and substructures : Warheads
1_2_dicarbonyl_oxalyl, 1_3_benzodioxole, acetylene, acyclic_acid_halide, acyclic_thioester, aldehyde, alkyl_halide_noF, alphahalo_ketone_carbonyl, aminophenol, anhydride, aziridine etc
Library Filtering
Frequent hitters
Apomorphine, Benserazide, Calciferol, Clofazimine, Dextrotiroxina, Dextrotiroxina-sodica, Diethylstilbestrol, Dopamine, Fenoterol, Idebenone, Methyldopamine etc
Promiscuous Inhibitors
Benzyl_benzoate,Bisindolylmalemide_i , Clotrimazole, Delavirdine, Econazole,Indigo,Indirubin,K252c, Miconazole, Nicardipine, Quercetin, Ro318220, Rottlerin, Sulconazole, U0126 etc
Intermediate Substructures
acyclic_halide,alkene,aniline,cyclic_ketone,enol_ether,hemiacetal_hemiketal, methyl_ketone,nitrile,nitro,polyenes,thiophene etc
PAINS
Imine, mannich, quinone, catechol, diazox_sulfon etc
Other diversity moieties
1_2_dimethoxy,1_4_dimethoxy,1_3_cyclohexadiene,1_4_cyclohexadiene,4_vinyl_pyridine,HOBT_esters,N_P_S_halides,N_S_beta_halothyl,acrylamide,acyclic_ketene,acyclic_NCN,acyclic_NS,acyclic_acetal,acyclic_acyl_cyanides etc

**Fig. 2.** Summary of protocol utilized for filtering compound library.

resource and usage of open source software for virtual screening are common scenario. Huge compound libraries can become manageable in terms of subsequent filtering as well as virtual screening if fragmented in workable pieces. Therefore, the initial compound library, obtained in sdf format, was fragmented in manageable pieces by parsing the compound library using a Python script. Each fragmented part consist of 3000 compounds, thus generating 117 sub libraries. Each sub library was named as Part-1 (containing compounds 1–3000), Part-2 (containing compounds 3001–6000) and so on. Thus, we have named first compound from first part as Compound_1-P1.The last compound library (i.e. Part-117) contains 276 compounds.

## *2.2. Compound library formatting and filtering*

Each sub-library was subject to extensive filtering process using web based filtering tools at the FAF-Drugs server (http://fafdrugs4.mti.univ-paris-diderot.fr/) [11]. This server contains tools that accepts inputs in web based forms and provides output in compressed archives. For a process such as ADMET filtering, the presence of salt associated to a compounds in its description can bias the filter (molecular weight (MW), logP calculation, etc.). After previous, fragmentation, the individual pieces were also in sdf format, hence they were first converted to SMILES, since Desalt service requires input in SMILES format. This task is carried out using Open Babel software [12]. Desalt service apply a series of rules to remove



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Part 1 | Part 2 | Part 3 | Part 4 | Part 5 | Part 6 | Part 7 | Part 8 | Part 9 | Part 10 |
| Part 11 | Part 12 | Part 13 | Part 14 | Part 15 | Part 16 | Part 17 | Part 18 | Part 19 | Part 20 |
| Part 21 | Part 22 | Part 23 | Part 24 | Part 25 | Part 26 | Part 27 | Part 28 | Part 29 | Part 30 |
| Part 31 | Part 32 | Part 33 | Part 34 | Part 35 | Part 36 | Part 37 | Part 38 | Part 39 | Part 40 |
| Part 41 | Part 42 | Part 43 | Part 44 | Part 45 | Part 46 | Part 47 | Part 48 | Part 49 | Part 50 |
| Part 51 | Part 52 | Part 53 | Part 54 | Part 55 | Part 56 | Part 57 | Part 58 | Part 59 | Part 60 |
| Part 61 | Part 62 | Part 63 | Part 64 | Part 65 | Part 66 | Part 67 | Part 68 | Part 69 | Part 70 |
| Part 71 | Part 72 | Part 73 | Part 74 | Part 75 | Part 76 | Part 77 | Part 78 | Part 79 | Part 80 |
| Part 81 | Part 82 | Part 83 | Part 84 | Part 85 | Part 86 | Part 87 | Part 88 | Part 89 | Part 90 |
| Part 91 | Part 92 | Part 93 | Part 94 | Part 95 | Part 96 | Part 97 | Part 98 | Part 99 | Part 100 |
| Part 101 | Part 102 | Part 103 | Part 104 | Part 105 | Part 106 | Part 107 | Part 108 | Part 109 | Part 110 |
| Part 111 | Part 112 | Part 113 | Part 114 | Part 115 | Part 116 | Part 117 | | | |

### Downloads

- Filtered Accepted Molecules: [SDF]
- Filtered Intermediate Molecules: [SDF]
- Filtered Rejected Molecules: [SDF]
- Duplicates: *None*
- Mixtures: *None*
- Empty Structures: *None*
- Isotopes: *None*
- Large Compounds: *None*

### Filtering statistics

- Input molecules: 3000
- Duplicates: 0
- Mixtures: 0
- Large Compounds: 0
- Filtered molecules: 3000
- Rejected molecules: 1870
- Accepted molecules: 660
- Intermediate molecules: 470
- PAINS (Pan Assays Interferences Compounds): 33

### Main PhysChem Descriptors Analysis

Click on a plot to enlarge the picture

### Problematic Moieties

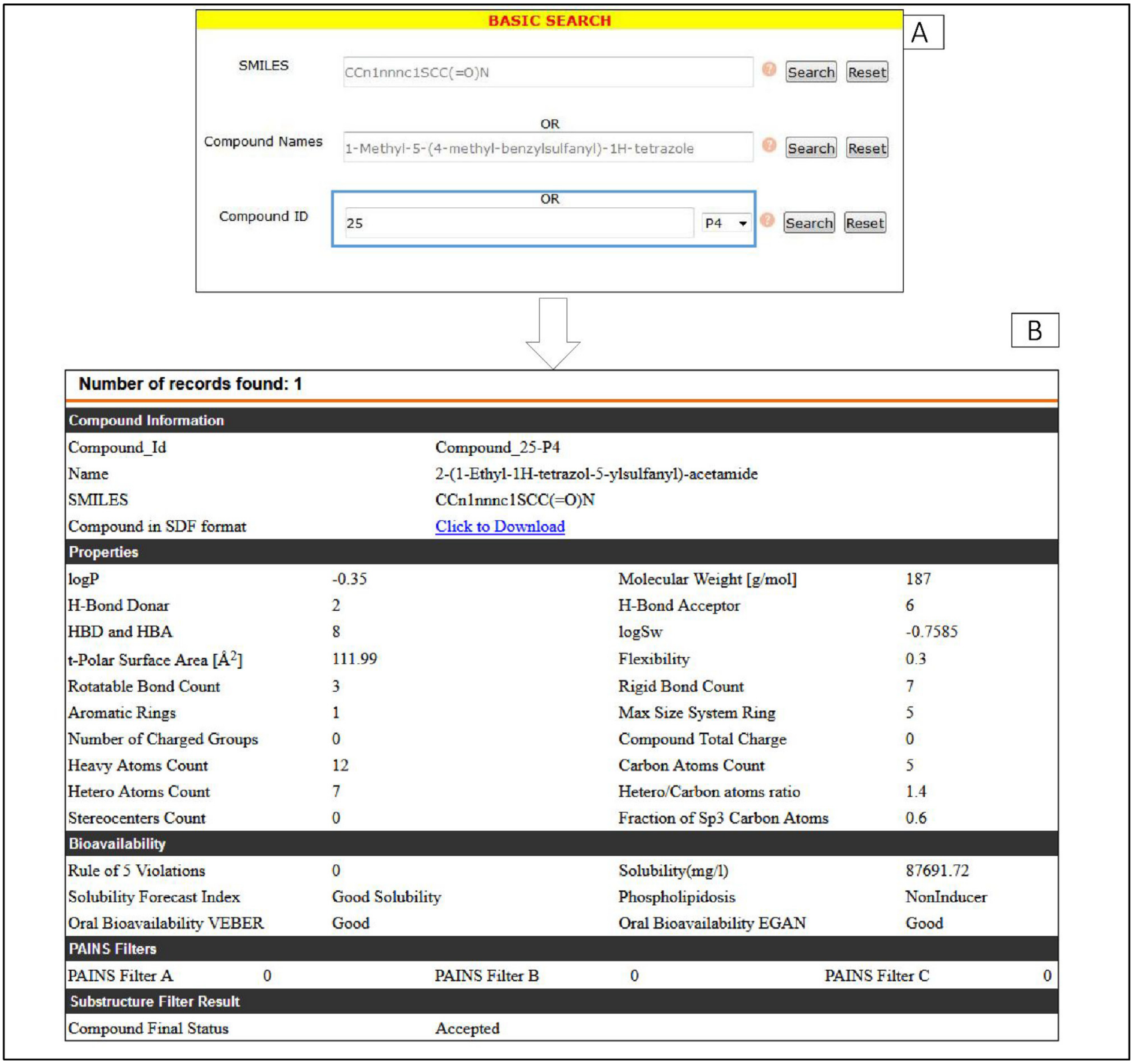
With an occurence above 1%

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Part 1 | Part 2 | Part 3 | Part 4 | Part 5 | Part 6 | Part 7 | Part 8 | Part 9 | Part 10 |
| Part 11 | Part 12 | Part 13 | Part 14 | Part 15 | Part 16 | Part 17 | Part 18 | Part 19 | Part 20 |
| Part 21 | Part 22 | Part 23 | Part 24 | Part 25 | Part 26 | Part 27 | Part 28 | Part 29 | Part 30 |
| Part 31 | Part 32 | Part 33 | Part 34 | Part 35 | Part 36 | Part 37 | Part 38 | Part 39 | Part 40 |
| Part 41 | Part 42 | Part 43 | Part 44 | Part 45 | Part 46 | Part 47 | Part 48 | Part 49 | Part 50 |
| Part 51 | Part 52 | Part 53 | Part 54 | Part 55 | Part 56 | Part 57 | Part 58 | Part 59 | Part 60 |
| Part 61 | Part 62 | Part 63 | Part 64 | Part 65 | Part 66 | Part 67 | Part 68 | Part 69 | Part 70 |
| Part 71 | Part 72 | Part 73 | Part 74 | Part 75 | Part 76 | Part 77 | Part 78 | Part 79 | Part 80 |
| Part 81 | Part 82 | Part 83 | Part 84 | Part 85 | Part 86 | Part 87 | Part 88 | Part 89 | Part 90 |
| Part 91 | Part 92 | Part 93 | Part 94 | Part 95 | Part 96 | Part 97 | Part 98 | Part 99 | Part 100 |
| Part 101 | Part 102 | Part 103 | Part 104 | Part 105 | Part 106 | Part 107 | Part 108 | Part 109 | Part 110 |
| Part 111 | Part 112 | Part 113 | Part 114 | Part 115 | Part 116 | Part 117 | | | |

**Fig. 3.** (A) Illustration for arrangement of part-wise filtering data. (B) Details of part-wise filtering and statistics. (C) Tabulated link of part-wise accepted compounds in 3D mol2 format (Results obtained from FAFDrugs server [11].).

salt from associated with the compounds. An internal SMILES definition of Pybel [13] of each compound is analysed in order to detect separating points (structure smaller than 6 atoms is considered as a salt), the biggest structure is retained. After elimination of salts from the compound library, it was further subjected to filtering process.

Before the application of filtering protocol, the library is required to be appropriately formatted to remove empty structures, inorganics, mixtures, counter ions, isotopes etc. In the next stage of formatting, issues like protonation and normalization of compound libraries are generally dealt. Duplicate compounds were detected and eliminated using internal definitions stated in Pybel module of each normalized compounds.



BASIC SEARCH

SMILES: CCn1nnnc1SCC(=O)N

OR

Compound Names: 1-Methyl-5-(4-methyl-benzylsulfanyl)-1H-tetrazole

OR

Compound ID: 25 P4

Number of records found: 1

| Compound Information | | | |
|---|---|---|---|
| Compound_Id | Compound_25-P4 | | |
| Name | 2-(1-Ethyl-1H-tetrazol-5-ylsulfanyl)-acetamide | | |
| SMILES | CCn1nnnc1SCC(=O)N | | |
| Compound in SDF format | Click to Download | | |
| **Properties** | | | |
| logP | -0.35 | Molecular Weight [g/mol] | 187 |
| H-Bond Donar | 2 | H-Bond Acceptor | 6 |
| HBD and HBA | 8 | logSw | -0.7585 |
| t-Polar Surface Area [Å$^2$] | 111.99 | Flexibility | 0.3 |
| Rotatable Bond Count | 3 | Rigid Bond Count | 7 |
| Aromatic Rings | 1 | Max Size System Ring | 5 |
| Number of Charged Groups | 0 | Compound Total Charge | 0 |
| Heavy Atoms Count | 12 | Carbon Atoms Count | 5 |
| Hetero Atoms Count | 7 | Hetero/Carbon atoms ratio | 1.4 |
| Stereocenters Count | 0 | Fraction of Sp3 Carbon Atoms | 0.6 |
| **Bioavailability** | | | |
| Rule of 5 Violations | 0 | Solubility(mg/l) | 87691.72 |
| Solubility Forecast Index | Good Solubility | Phospholipidosis | NonInducer |
| Oral Bioavailability VEBER | Good | Oral Bioavailability EGAN | Good |

| PAINS Filters | | | | | |
|---|---|---|---|---|---|
| PAINS Filter A | 0 | PAINS Filter B | 0 | PAINS Filter C | 0 |

| Substructure Filter Result | |
|---|---|
| Compound Final Status | Accepted |

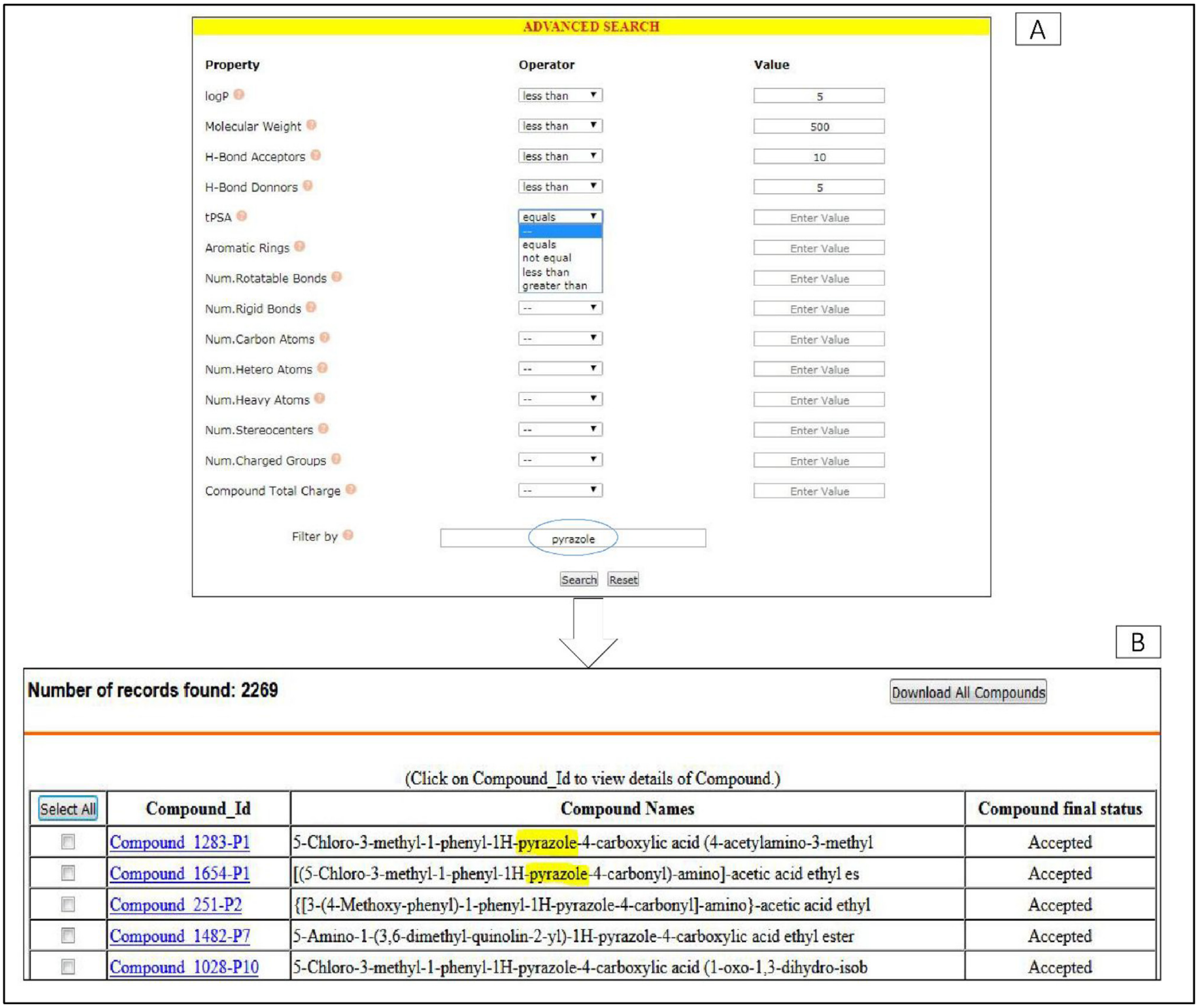
**Fig. 4.** Performing Basic Search. (A) Illustrating selection of compound id and its corresponding part. (B) The result of the query is divided into five sections containing Compound information like compound id, name and SMILES notation. The compound can also be downloaded in sdf format. Properties section highlights physicochemical properties of compounds like logP, molecular weight flexibility etc [11]. Bioavailability section provides information on compound's performance on three general rules like, Lipinski's Rule of five [5], Veber Rules [14] and Egan Rule along with solubility [28] and phospholipidosis status [29]. PAINS Filter section furnish PAINS status of the compound [24]. The last section substructure filter results presents final compound status, whether it is accepted, rejected or it is an intermediate compound. If the compound falls in either of rejected or intermediate category, the reasons for the status are also shown. Similarly, compounds can also be searched using their SMILES notation or exact compound name from the library.

Library filtering typically entails two major aspects. The first aspect deals with the inspection and filtering on basis of physicochemical properties. These parameters are often used to deduce compounds oral bioavailability and drug likeness properties. The second aspect of filtering is scrutinizing the library for presence of undesirable functional groups and moieties. Presences of such undesirable groups adversely affects the experimental high throughput screening assays and are thus essential to be eliminated.

### 2.2.1. Physicochemical filters

In addition to four descriptors stated by Lipinski (i.e. MW, logP and number of hydrogen bonds donor and acceptors), numerous other physicochemical filtering rules are routinely applied for efficient screening of compounds. Following section deals with brief explanations of various physicochemical descriptors popularly used in screening huge compound libraries.

Studies have demonstrated that compounds having topological polar surface area (tPSA) $\leq$ 160 prove to be good lead candidates [5]. Analysis involving molecular properties of drug candidates suggests that compounds with $\leq$ 9 rotatable bonds positively influences the oral bioavailability of drug candidates [14]. Compounds with rigid bonds $\leq$ 30 are characterized to possess good drug-like properties [15]. Compounds containing $\leq$ 4 number of rings are found to be a better lead candidates while compounds with ring size of $\leq$ 18 atoms are observed to possess optimal lead like prop-



ADVANCED SEARCH

| Property | Operator | Value |
|---|---|---|
| logP | less than | 5 |
| Molecular Weight | less than | 500 |
| H-Bond Acceptors | less than | 10 |
| H-Bond Donnors | less than | 5 |
| tPSA | equals<br>--<br>equals<br>not equal<br>less than<br>greater than | Enter Value |
| Aromatic Rings | | Enter Value |
| Num.Rotatable Bonds | | Enter Value |
| Num.Rigid Bonds | -- | Enter Value |
| Num.Carbon Atoms | -- | Enter Value |
| Num.Hetero Atoms | -- | Enter Value |
| Num.Heavy Atoms | -- | Enter Value |
| Num.Stereocenters | -- | Enter Value |
| Num.Charged Groups | -- | Enter Value |
| Compound Total Charge | -- | Enter Value |

Filter by: pyrazole

Number of records found: 2269

(Click on Compound_Id to view details of Compound.)

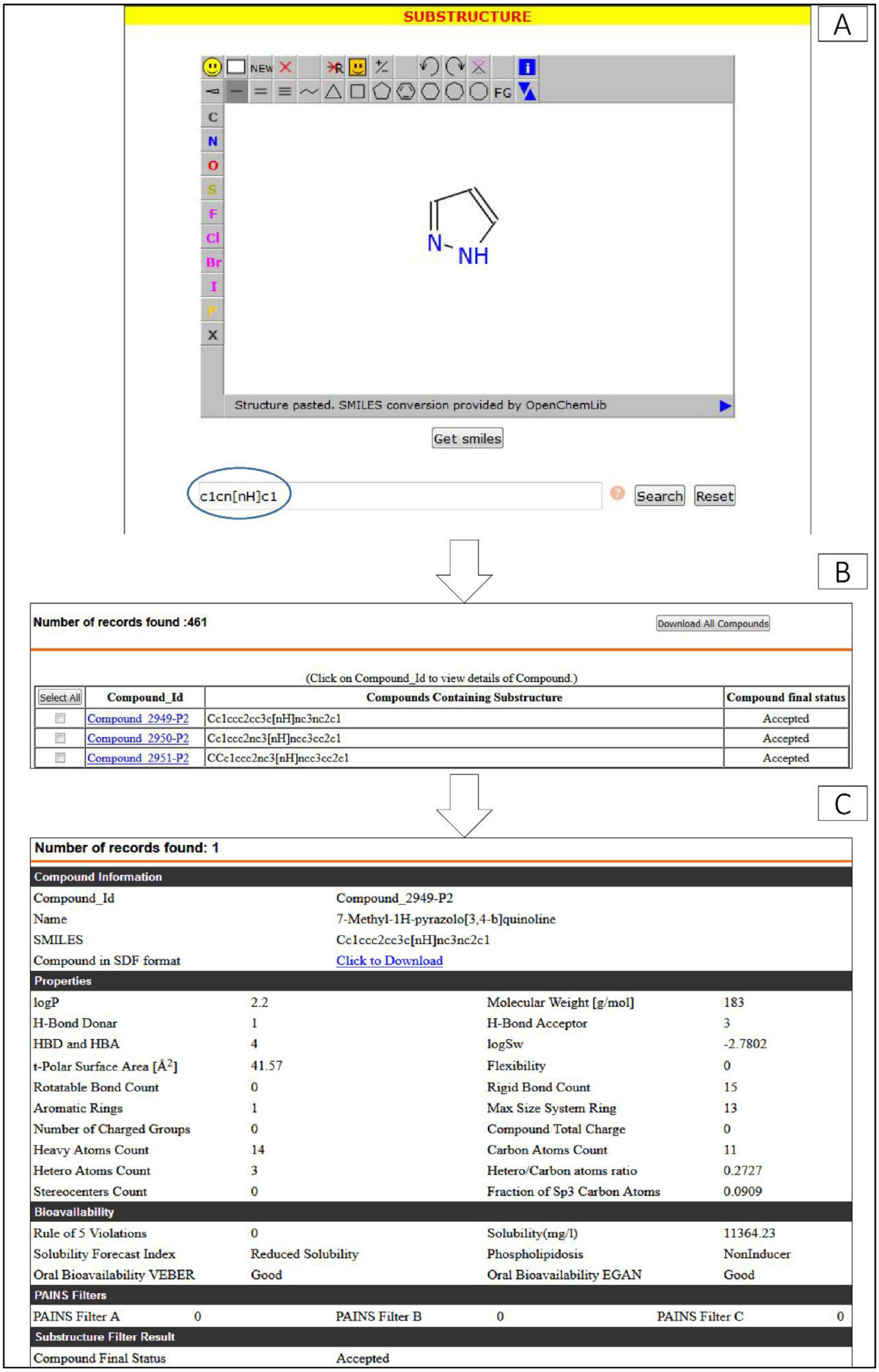
| Select All | Compound_Id | Compound Names | Compound final status |
|---|---|---|---|
| □ | Compound_1283-P1 | 5-Chloro-3-methyl-1-phenyl-1H-pyrazole-4-carboxylic acid (4-acetylamino-3-methyl | Accepted |
| □ | Compound_1654-P1 | [(5-Chloro-3-methyl-1-phenyl-1H-pyrazole-4-carbonyl)-amino]-acetic acid ethyl es | Accepted |
| □ | Compound_251-P2 | {[3-(4-Methoxy-phenyl)-1-phenyl-1H-pyrazole-4-carbonyl]-amino}-acetic acid ethyl | Accepted |
| □ | Compound_1482-P7 | 5-Amino-1-(3,6-dimethyl-quinolin-2-yl)-1H-pyrazole-4-carboxylic acid ethyl ester | Accepted |
| □ | Compound_1028-P10 | 5-Chloro-3-methyl-1-phenyl-1H-pyrazole-4-carboxylic acid (1-oxo-1,3-dihydro-isob | Accepted |

**Fig. 5.** Performing advanced search. (A) Illustration demonstrating construction of complex queries (B) The intermediary result page.

erties [15]. Organic compounds with number of carbons ranging between 3 and 35 and number of hetero atoms in range of 1–15 are characterized to have drug like propensity [8,16]. Experimental findings suggest that compounds with Hetero/Carbon atoms ratio in range of 0.1–1.1 are found to act as drug like molecules [16–18]. Experimental findings also suggest that most of the drug compounds contains $\leq 3$ charges with total formal charge ranging between $-2$ to +2 [8,17,18]. Moreover, more than 90 percent of oral drugs are known to possess $\leq 2$ stereo centers [8,19,20]. These values are optimized by various experimental and comparative studies derived after carrying out extensive statistical analysis of existing drug molecules. Therefore, all the above mentioned parameters were used to filter compound library in present resource.

### 2.2.2. *Undesirable moieties and substructures*

Rishton et al. introduced classification of certain moieties as “Warheads” [21]. Their analysis concluded that compounds containing such moieties must be removed. Roche et al. identified special class of compounds as frequent hitters. Frequent hitters are set of compounds which are identified as hits in many different biological assays covering a wide range of targets, since activity of these compound is not specific for the target and hence acquire a tendency to bind every target used in biological assay; thus perturbing the assay or detection method. Compounds with such moieties are poor starting points for drug discovery programs and in general must be removed [22]. McGovern et al. identified group of compounds as ‘Promiscuous Inhibitors’ that act noncompetitively, show little relationship between structure and activity, and have poor selectivity. Such molecules are difficult to further characterize thus failing to synthesize a viable leads; resulting in futile efforts in lead development process and wastage of time, capital and manpower. Compounds which contains that kind of moieties, warrants complete removal from the dataset [23]. There are certain structural moieties and compounds that are identified as ‘Intermediate Substructures’. This class of substructures/compounds is found to be interfering in some assays but remains benign in most. Removal of such compounds thus remains circumstantial and thus could be potentially problematic. Therefore such compounds are needed to be identified separately to warn the presence of a potentially undesirable group/compound. The choice of elimination or inclusion of such compounds thus is kept up to users. Baell and Holloway et al. on basis of their single assay detection technology introduced group of compounds called Pan Assay Interference compounds (PAINS). This series of compounds appear as frequent hitters (or promiscuous compounds) in many biochemical high throughput screens. These compounds are also advised to be removed during compound library processing [24]. Besides above classified moieties, there are other set of compounds which interferes with an assay if present in



Structure pasted. SMILES conversion provided by OpenChemLib

c1cn[nH]c1

Number of records found :461

(Click on Compound_Id to view details of Compound.)

| Select All | Compound_Id | Compounds Containing Substructure | Compound final status |
|---|---|---|---|
| □ | Compound_2949-P2 | Cc1ccc2cc3c[nH]nc3nc2c1 | Accepted |
| □ | Compound_2950-P2 | Cc1ccc2nc3[nH]ncc3cc2c1 | Accepted |
| □ | Compound_2951-P2 | CCc1ccc2nc3[nH]ncc3cc2c1 | Accepted |

Number of records found: 1

| Compound Information | | | |
|---|---|---|---|
| Compound_Id | Compound_2949-P2 | | |
| Name | 7-Methyl-1H-pyrazolo[3,4-b]quinoline | | |
| SMILES | Cc1ccc2cc3c[nH]nc3nc2c1 | | |
| Compound in SDF format | Click to Download | | |
| **Properties** | | | |
| logP | 2.2 | Molecular Weight [g/mol] | 183 |
| H-Bond Donar | 1 | H-Bond Acceptor | 3 |
| HBD and HBA | 4 | logSw | -2.7802 |
| t-Polar Surface Area [Å$^2$] | 41.57 | Flexibility | 0 |
| Rotatable Bond Count | 0 | Rigid Bond Count | 15 |
| Aromatic Rings | 1 | Max Size System Ring | 13 |
| Number of Charged Groups | 0 | Compound Total Charge | 0 |
| Heavy Atoms Count | 14 | Carbon Atoms Count | 11 |
| Hetero Atoms Count | 3 | Hetero/Carbon atoms ratio | 0.2727 |
| Stereocenters Count | 0 | Fraction of Sp3 Carbon Atoms | 0.0909 |
| **Bioavailability** | | | |
| Rule of 5 Violations | 0 | Solubility(mg/l) | 11364.23 |
| Solubility Forecast Index | Reduced Solubility | Phospholipidosis | NonInducer |
| Oral Bioavailability VEBER | Good | Oral Bioavailability EGAN | Good |

| PAINS Filters | | | | | |
|---|---|---|---|---|---|
| PAINS Filter A | 0 | PAINS Filter B | 0 | PAINS Filter C | 0 |

| Substructure Filter Result | |
|---|---|
| Compound Final Status | Accepted |

**Fig. 6.** Performing substructure search. (A) Illustration demonstrating formulation of substructure query using JSME molecule editor and obtaining canonical SMILES for pyrazole group using 'Get smiles' button. (B): The intermediate page listing number of compounds with queried substructure along with link for detailed information. (C) Detailed information of the compound with desired substructure.

more than a threshold number (called as 'Other diversity moieties'). Such threshold values of detection are fine-tuned using in-house knowledge and medicinal chemistry literature. Compounds with such moieties are thus required to be flagged separately for their presence. Fig. 2 summarizes the overall filtering process.

The physicochemical and substructure filtering rules described above are coded on FAF drugs service as "Lead like soft" filtering rules that were implemented for filtering purpose. The "Lead like soft" rules are constructed on the seminal literature [16–19]. The lead like soft rules were developed with an intention to screen starting compounds with possibility for further property optimization. Thus, compound libraries obtained after filtering from such a generalized rules are thus applicable for any conventional enzyme assay protocol. This fact thus increases the scope of filtered compound libraries to be implemented widely rather than focusing on specific set of targets. In summary, compound processed here can be effectively used on any general enzyme/receptor target.

### 2.3. Database design and implementation

The database was developed on traditional three layer architecture [25] that consist of data layer, an intermediate or middle layer and presentation layer. The data layer is constituted on MySQL (a relational database management system, RDBMS) that functions on a server enabling multiple users to access the database simultaneously. The intermediate or middle layer was designed using PHP. The user interface (presentation layer) was coded using HTML, Java Scripts, CSS and AJAX. The resource is implemented on the windows based server and uses APACHE as main web-server engine.

### 2.4. Substructure searching utility

The substructure searching tools on this resource expects user inputs in the form of canonical SMILES. In order to make this tool more user friendly, we implemented JAVA based molecular editor JSME [26] that enables on-the-fly 2D structure drawing and subsequent conversion to canonical SMILES. All the compounds passed in filtering process were initially compiled in a single sdf file. For substructure similarity search, FilTer BaSe executes babel program from the Open Babel suit (http://openbabel.org). The substructure searching algorithm in babel program is based on molecular fingerprint method [12,27]. Finger prints can be defined as sequentially arranged binary bits in form of zeros and ones. In molecular fingerprinting, the molecular structures are encoded into binary bits, and position of individual bits correlates to molecular information like existence or absence of a particular atom or a functional group or even a substructure. For the sake of improving speed during substructure searching, a fast search index of the accepted sdf file was created. When a substructure query is executed on the server, a binary finger print from the query compound is generated and matched with precompiled fingerprint data from fast search index file. The output is generated in a temporary file and finally tabulated on the HTML page.

## 3. Results and discussions

The initial compound library obtained from Asinex Ltd contained 348276 compounds (three lakh, forty eight thousand, two hundred and seventy six). During the compound library formatting process, number of cleansing procedures was applied on library, like removal of empty structures, inorganic mixtures and duplicate compounds. This process removed 1078 compounds and remaining 347198 compounds actually entered the filtering protocol. Total 187481 compounds were rejected owing to either incompatibility due to undesirable physiochemical properties or because of presence of undesirable functional groups. The number of compound that were classified as intermediate were 86244. Near about 28715 compounds were detected as PAINS. Remaining, 44758 compounds were found to be accepted. Therefore, upon successful filtering process, about 10 percent of the compounds were found to be accepted and can be carried out for further virtual screening.

Part wise data of the compound collection (Part-1 to Part-117) can be accessed from the 'Compound library' tab from the FilTer BaSe resource. On the compound library tab page, part wise data is tabulated in upper table Fig. 3(A); while, lower table represent part wise 3D coordinates of accepted compounds Fig. 3(C).

Part wise details of filtering statistics, information on main physiochemical descriptors (e.g. logP, tPSA, MW, number of rotatable bonds, hydrogen bond acceptors and donors) statistics along with report on problematic and undesirable substructure moieties can be retrieved using links from upper table. The organization of data has been retained in original form as per the FAF-Drugs server [11] Fig. 3 (B). For example, filtered accepted, rejected, intermediate etc. compounds can be downloaded in 2D sdf format from the 'Download' section. 'Filtering statistics' section provides the details of the number of compounds that were filtered, number of duplicates, mixtures and large compounds if any; along with the number of compounds that were rejected or categorized as intermediate, accepted or PAINS etc. The 'Main PhysChem Descriptor Analysis' section highlights the graphical representation of main physiochemical descriptors like logP, tPSA, MW, number of rotatable bonds, hydrogen bond acceptors and donors in a clickable and zoom able images. Similarly, 'Problematic moieties' section details the number of problematic moieties that were observed during substructure filtering in the form of a pie-chart.

The part-wise data obtained from filtering process can be efficiently fetched from the Compound library Table However, filtering data of individual compounds cannot be obtained from this section thus limiting its utility. Therefore, a separate database was designed to cater compound wise data. The applications and features of the database can be best demonstrated using following model examples.

### 3.1. Example for basic search

Basic search utility is aimed to search the FilTer BaSe resource for a specific compound using its compound_id and its respective 'part' as input. For example, detailed compound information, physicochemical properties, bioavailability status, sub-structure filter data along with final filtering status (accepted or rejected or intermediate) can be obtained using this option. We have implemented 'gen3d' option from obabel program that generates 3D coordinates for the queried compound that can be downloaded in sdf format using 'Click to download' link from result page. The result for query made for compound 25 from part 4 is illustrated in Fig. 4.

### 3.2. Example for advanced search

Advanced search utility enables the user to construct and execute complex queries. Various logical operators (e.g. equals, not equals, less than and greater than) are implemented for constructing the query with complex logic to obtain user defined results. Queries can either be made individually or clubbed together. This feature enables user to search the database for any user defined rules or existing popular rules like Lipinski's Rule of five [5], Veber [14], Ghose [30], Egan [28], Opera Drug-like [15], Opera Lead-like [16],Walters [31], Martin [32], REOS [33] etc. For example, all pyrazole compounds passing Lipinski's rule of five can be obtained (Fig. 5). The result of such queries yields an intermediate page that reports the number of compound hits and the compounds are tabulated with their compound ids and names. The detailed compound information can be viewed as per Fig. 5(B) by clicking 'compound_id' on the page.

### 3.3. Example for substructure search

Searching and subsequent design/synthesis of the compounds with a particular functional group is a routine task in medicinal chemistry. The substructure search utility on the FilTer BaSe resource is developed to make such queries user friendly. Continuing the previous example, user can search the pyrazole class of compounds that are accepted during filtering process. The outputs is generated with an intermediate page listing the number of compounds with desired functional group or substructure. Finally, compounds are tabulated showing compound id, SMILES and compound's final status (Fig. 6).

## 4. Conclusions

Here we report an efficient filtering of 348,276 chemical compounds and this chemical space is made freely available for drug screening purpose. We have developed ready to use, easily manageable, small sized compound libraries that are expected to add a service for new drug developers even in academic settings. However, it is to be noted that no precise set of rules exist, that can be universally implemented for filtering compound libraries, only few rule of thumb are generally followed. Therefore, using the standard filtering rules we project set of compound libraries that are easily amenable for further optimization. 3D structures of compounds are made available here that can be directly implemented for virtual screening purposes. The resource is aimed to provide start-up compounds libraries that have scope for further user defined modifications if required. The database presented here enables efficient property based as well as sub-structure based searching of compounds that can be used to pinpoint novel chemical scaffolds as starting point for new drug discovery campaigns. The descriptors data from current resource can be used in defining and optimizing QSAR studies.

## Acknowledgements

We sincerely thank, Dr. Sangeeta Sawant, Director, Bioinformatics Centre, Savitribai Phule Pune University, Pune for providing infrastructure and support throughout the project. We also acknowledge Mrs. Smita Saxena, for critical suggestions in database design and development. We are grateful to Mr. Rajiv Karbhal and Mr. Rahul Khandare for kind help and support. We are thankful to anonymous reviewers for their critical comments and constructive suggestions that resulted in improvement of the presented resource.

## References

[1] R.A. Prentis, Y. Lis, S.R. Walker, Pharmaceutical innovation by the seven UK owned pharmaceutical companies (1964–1985), J. Clin. Pharmacol. 25 (3) (1988) 387–396.
[2] T. Kennedy, Managing the drug discovery/development interface, Drug Discov. Today 2 (10) (1997) 436–444.
[3] D. Schuster, C. Laggner, T. Langer, Why drugs fail: a study on side effects in new chemical entities, Curr. Pharm. Des. 11 (27) (2005) 3545–3559.
[4] E.H. Kerns, L. Di, Drug-Like Properties: Concepts, Structure Design and Methods From ADME to Toxicity Optimization, 1st ed., Academic Press, 2008.
[5] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (1–3) (2001) 3–26.
[6] J.F. Blake, Identification and evaluation of molecular properties related to preclinical optimization and clinical fate, Med. Chem. 1 (6) (2005) 649–655.
[7] D.J. Huggins, A.R. Venkitaraman, D.R. Spring, Rational methods for the selection of diverse screening compounds, ACS Chem. Biol. 6 (3) (2011) 208–217.
[8] J.J. Irwin, B.K. Shoichet, ZINC–a free database of commercially available compounds for virtual screening, J. Chem. Inf. Model. 45 (1) (2005) 177–182.
[9] Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B.A. Shoemaker, T.O. Suzek, J. Wang, J. Xiao, J. Zhang, S.H. Bryant, An overview of the PubChem BioAssay resource, Nucleic Acids Res. 38 (2010) D255–D266.
[10] M. von Grotthuss, G. Koczyk, J. Pas, L.S. Wyrwicz, L. Rychlewski, Ligand.Info small-molecule meta-database, Comb. Chem. High Throughput Screen. 7 (8) (2004) 757–761.
[11] D. Lagorce, O. Sperandio, H. Galons, M.A. Miteva, B.O. Villoutreix, FAF-Drugs2: a free ADME/tox filtering tool to assist drug discovery and chemical biology projects, BMC Bioinf. 9 (2008) 396.
[12] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminform. 3 (1) (2011) 33.
[13] N.M. O'Boyle, C. Morley, G.R. Hutchison, Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit, Chem. Cent. J. 2 (2008) 5.
[14] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, J. Med. Chem. 45 (12) (2002) 2615–2623.
[15] T.I. Oprea, Property distribution of drug-related chemical databases, J. Comput. Aided Mol. Des. 14 (3) (2000) 251–264.
[16] T.I. Oprea, A.M. Davis, S.J. Teague, P.D. Leeson, Is there a difference between leads and drugs? A historical perspective, J. Chem. Inf. Comput. Sci. 41 (5) (2001) 1308–1315.
[17] P. Workman, I. Collins, Probing the probes: fitness factors for small molecule tools, Chem. Biol. 17 (6) (2006) 561–577.
[18] J.B. Baell, Broad coverage of commercially available lead-like screening space with fewer than 350, 000 compounds, J. Chem. Inf. Model. 53 (1) (2013) 39–55.
[19] R. Brenk, A. Schipani, D. James, A. Krasowski, I.H. Gilbert, J. Frearson, P.G. Wyatt, Lessons learnt from assembling screening libraries for drug discovery for neglected diseases, ChemMedChem 3 (3) (2008) 435–444.
[20] E. Pihan, L. Colliandre, J.F. Guichou, D. Douguet, e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design, Bioinformatics 28 (11) (2012) 1540–1541.
[21] G.M. Rishton, Nonleadlikeness and leadlikeness in biochemical screening, Drug Discov. Today 8 (2) (2003) 86–96.
[22] O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E.M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjogren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, Development of a virtual screening method for identification of frequent hitters in compound libraries, J. Med. Chem. 45 (1) (2002) 137–142.
[23] S.L. McGovern, E. Caselli, N. Grigorieff, B.K. Shoichet, A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening, J. Med. Chem. 45 (8) (2002) 1712–1722.
[24] J.B. Baell, G.A. Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, J. Med. Chem. 53 (7) (2010) 2719–2740.
[25] A. Jadhav, V. Ezhilarasan, O. Prakash Sharma, A. Pan, Clostridium-DT(DB): a comprehensive database for potential drug targets of Clostridium difficile, Comput. Biol. Med. 43 (4) (2013) 362–367.
[26] B. Bienfait, P. Ertl, JSME: a free molecule editor in JavaScript, J. Cheminform. 21 (5) (2013) 24.
[27] S. Loharch, I. Bhutani, K. Jain, P. Gupta, D.K. Sahoo, R. Parkesh, EpiDBase: a manually curated database for small molecule modulators of epigenetic landscape, Database 2015 (2015) bav013.
[28] W.J. Egan, K.M. Merz Jr., J.J. Baldwin, Prediction of drug absorption using multivariate statistics, J. Med. Chem. 43 (21) (2000) 3867–3877.
[29] K.R. Przybylak, A.R. Alzahrani, M.T. Cronin, How does the quality of phospholipidosis data influence the predictivity of structural alerts? J. Chem. Inf. Model. 54 (8) (2014) 2224–2232.
[30] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, 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 (1999) 55–68.
[31] W.P. Walters, M.A. Murcko, Library filtering systems and prediction of drug-like properties, Methods Principles Med. Chem. 10 (2000) 15–30.
[32] Y.C. Martin, A bioavailability score, J. Med. Chem. 48 (2005) 3164–3170.
[33] W.P. Walters, M.T. Stahl, M.A. Murcko, Virtual screening – an overview, Drug Discov. Today 3 (1998) 160–178.