

FAF-Drugs3: a web server for compound property calculation and chemical library design

David Lagorce^{1,2}, Olivier Sperandio^{1,2}, Jonathan B. Baell³, Maria A. Miteva^{1,2} and Bruno O. Villoutreix^{1,2,*}

¹Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques In Silico, Paris 75013, France, ²Inserm U973, Molécules Thérapeutiques In Silico, Paris 75013, France and ³Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria 3052, Australia

Received January 28, 2015; Revised March 20, 2015; Accepted April 02, 2015

ABSTRACT

Drug attrition late in preclinical or clinical development is a serious economic problem in the field of drug discovery. These problems can be linked, in part, to the quality of the compound collections used during the hit generation stage and to the selection of compounds undergoing optimization. Here, we present FAF-Drugs3, a web server that can be used for drug discovery and chemical biology projects to help in preparing compound libraries and to assist decision-making during the hit selection/lead optimization phase. Since it was first described in 2006, FAF-Drugs has been significantly modified. The tool now applies an enhanced structure curation procedure, can filter or analyze molecules with user-defined or eight predefined physicochemical filters as well as with several simple ADMET (absorption, distribution, metabolism, excretion and toxicity) rules. In addition, compounds can be filtered using an updated list of 154 hand-curated structural alerts while Pan Assay Interference compounds (PAINS) and other, generally unwanted groups are also investigated. FAF-Drugs3 offers access to user-friendly html result pages and the possibility to download all computed data. The server requires as input an SDF file of the compounds; it is open to all users and can be accessed without registration at <http://fafdrugs3.mti.univ-paris-diderot.fr>.

INTRODUCTION

Chemical biology and even more so drug discovery are challenging endeavors that usually involve high-throughput screening computations and/or experiments, prioritization of the hit compounds and different levels of compound optimization. As such, the nature/composition of the compound collection used in the early phases has a significant

impact in determining both, the quantity and quality of identified hits/leads and ultimately to the overall success of the project (1). There are obviously different ways to prepare a compound collection depending on the disease type, the stage of the project, whether the screening is target-based or phenotypic-based and the goals (e.g. drug discovery or chemical biology) (2). Numerous rules have been developed over the years to guide the preparation of a compound collection or to select molecules for optimization (3–5), yet, all these rules, warnings, etc., have to be used with caution as blindly applying such recipes can discard from development many interesting molecules (6–8).

The quality of a compound collection can be defined in many different ways but very often, physicochemical properties and the presence of some unwanted chemical groups (e.g. toxic groups or chemicals that interfere with experimental readouts) are used in the field at the beginning of the project. For examples, some rules correlate physicochemical properties with oral administration (like the rule-of-five (RO5): molecular mass ≤ 500 ; calculated log P (cLogP) ≤ 5 ; number of hydrogen bond donors (HBD) ≤ 5 ; number of hydrogen bond acceptors (HBA) ≤ 10 ; a molecule whose properties fell outside these boundaries would be less likely orally absorbed and it was stated that a compound with two parameters out of these ranges would be subject to a flag (3)). Other rules suggest possible toxicity, anticipate difficulties with compound development as well as off-target interactions, for instance, the GSK 4/400 rule (higher risks of toxicity, interactions with off-targets or difficulties during development if $\log P > 4$ and $MW > 400$) (9); the Pfizer 3/75 rule (the rule states that a compound has a 6-fold reduction in preclinical toxicity when $\text{ClogP} < 3$ and a topological polar surface area (tPSA) $> 75 \text{ \AA}^2$ (and 24-fold reduction for basic compounds), the rule is agnostic to the toxicity mechanisms as it is expected that off-target issues are often responsible for the observed toxicity (10)) and the Fsp3 rule (molecular complexity, defined as number of sp³ hybridized carbons/total carbon count) that correlates molecular complexity with success in drug develop-

*To whom correspondence should be addressed. Tel: +33 1 57 27 83 88; Fax: +33 1 57 27 83 72; Email: bruno.villoutreix@inserm.fr

ment (11). Other *in silico* strategies to help in preparing a compound collection or selecting compounds for optimization involve the detection of potentially toxic chemical groups (or larger substructures) often referred to as toxicophores (8,12–16). In addition, several observations led to the definition of rules that flag compounds or substructures likely to interfere with biological assays, for instance Pan Assay Interference Compounds or PAINS and aggregators (17–19). Quantitative structure-activity relationship (QSAR) models, that are trained to predict specific properties or toxicological endpoints, can also assist compound selection and optimization (2).

Some of the above-mentioned properties can be computed using commercial packages provided by most major software companies working in the field of drug design while some freely available web services, provided by academic groups or by the private sector, are also available, for instance Molinspiration (Molinspiration Cheminformatics, for instance for the RO5 computations), PROTOX (20) (Prediction of Rodent Oral TOXicity) or the Aggregator Advisor (to search for molecules that aggregate, <http://advisor.bkslab.org/>). The URLs for most online tools in the field are listed at our website www.vls3d.com (21). We have developed an online tool in 2006 (22) named Free ADMET Filtering for Drugs (FAF-Drugs) that was optimized in 2011 (23) (FAF-Drugs2) and that is now much further enhanced (FAF-Drugs3). It should be noted here that freely available tools most often compute data for only one compound and/or do not compute several of the rules mentioned above or, else, compute properties that are not implemented in FAF-Drugs and as such are complementary to our server.

FAF-Drugs3: SERVICE OVERVIEW AND ENHANCEMENTS

Since its first release in 2006 (22), FAF-Drugs has been used by many groups worldwide (more than 30 000 connections) to prepare compound collections or to analyze a small list of chemical compounds. In 2006, FAF-Drugs was designed to perform only physicochemical filtering while the version 2 reported in 2011 (23) was the first free web-based package capable of preparing compound libraries by combining physicochemical rules, undesirable functional group searches and detection of PAINS (18). The major changes introduced in FAF-Drugs3 are: a new input data curation procedure including new ways to search for salts, new ways to predict solubility, optimized computations of properties to for instance predict blood brain barrier penetration or administration by inhalation among others, development of new pre-defined drug-like and lead-like filters, computations of the 3/75 and of the GSK 4/400 rules, search for toxicophores using a hand-curated list of structural alerts, identification of likely protein–protein interaction inhibitors, the prediction of drug-induced phospholipidosis (24), the implementation of the Eli-Lilly open drug discovery medicinal chemistry filter (12) and many new graphical windows such as a chart representing compound complexity. Several selected changes are discussed below.

WEB SERVER

FAF-Drugs3 is user-centered as it has a new user-friendly interface with new graphical windows that facilitate the analysis of the compounds online. The FAF-Drugs3 web server is an easy-to-use service consisting of a set of seven object-oriented Python modules embedded in the RPBS' Mobylye framework (25). Mobylye is a centralized workspace for the end-user and an on-the-fly program results pipelining. This portal also allows users to open a personal session upon registration (not mandatory) where data are stored. Each compound processed by FAF-Drugs3 is represented as a molecular object importing methods from the OpenBabel toolkit through its Python wrapper Pybel which allows to access to the OpenBabel C++ library (26). Furthermore, FAF-Drugs3 Python modules act as generators of methods belonging to this chemical object and needed for the filtering steps. The Mobylye frontend submits all the FAF-Drugs3 processes to an 800 core cluster running on Debian Operating System and managed by SUN Grid Engine.

In the front-page, the user is invited to upload a file and to select the type of filtering computations that should be carried out. As illustrated in Figure 1, FAF-Drugs3 performs the following main procedures: (i) an input data curation stage that removes large molecules and compounds containing some types of inorganic atoms and a desalting procedure, a structure normalization step and the removal of duplicates, (ii) the computation of several physicochemical properties and rules and the filtering of the small molecules, (iii) the detection of substructures such as potential toxic groups, aggregators and PAINS and (iv) the output section where the results are reported and from where files can be downloaded. Molecules treated by FAF-Drugs3 can then be piped into other software packages like our 3D generator software Frog2 (27) (<http://bioserv.rpbs.univ-paris-diderot.fr/services/Frog2>).

Input

The FAF-Drugs3 web server only accepts SDF files (we limit the number of compounds to 50 000) where each molecule has a unique identification number (ID). In addition, we have implemented the service Bank-Formatter on Mobylye (see <http://fafdrugs3.mti.univ-paris-diderot.fr/links>) to facilitate the preparation of the input file. The Bank-Formatter service to convert SMILES input file (using Openbabel libraries) to a suitable SDF file for FAF-Drugs3. Similarly if the input SDF file does not have the right format with an ID field, then the service can help in preparing the appropriate input SDF file. The Bank-Formatter SDF output file can be piped into the FAF-Drugs3 service through the Mobylye interface. Note that RPBS' Mobylye also allows the user to draw one compound with the ChemAxon's Marvin Sketch applet (www.chemaxon.com) and authorizes the output SDF file to be sent to FAF-Drugs3.

Output

Once the FAF-Drugs3 process is finished, the user is redirected to the result pages. All data can be downloaded and

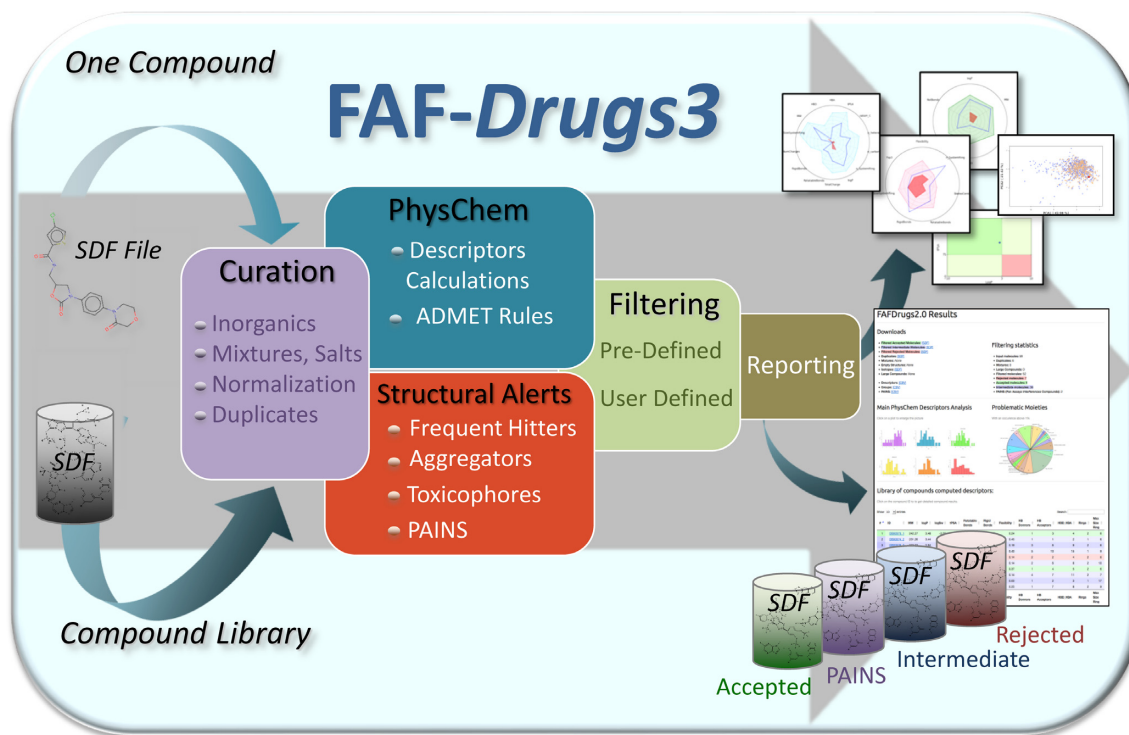


Figure 1. FAF-Drugs3 flow chart. Users can upload a single molecule or a compound collection. The molecules are treated following the selected filter parameters and structural alerts. Several output files can be visualized and/or downloaded such as the Accepted, Rejected and Intermediate SDF files and the file containing the PAINS compounds.

two key web-based interfaces are offered (see Figure 1). The first one can be defined as a summary result page where users can find links for downloading all filtered compound files Accepted.sdf (compounds with no structural alerts and satisfying the physicochemical filter), Intermediate.sdf (low-risk structural alerts), Rejected.sdf (the molecules that do not pass the selected or user-defined physicochemical filter which include a high-risk structural alert and/or exceed the threshold of occurrence of low-risk structural alerts) and PAINS.sdf (molecules flagged due to the presence of some chemical groups that belong to the PAINS category). Several tabulated files are also available: results.csv (containing the computed physico-chemical descriptors), groups.csv (containing the results of the structural alert searches) and pains.csv (containing the results of the search for PAINS). Then, a brief statistical summary of the filtering process is shown, graphical representations of the distributions of several properties, the ranges of the physicochemical filters applied and a sortable table containing all compounds analyzed with their ADMET filtering state, color-coded according to the compound status (green, blue and red for, respectively, Accepted, Intermediate and Rejected). A click on the compound ID opens the second web-based interface that can be defined as a detailed result page. On this new page, one can retrieve all computed values, a list of all detected problems justifying the classification of the compound in the Rejected or Intermediate basket, a 2D depiction picture of the selected compound (based on ChemAxon molconvert (www.chemaxon.com)), and a principal component analysis mapping the selected compound

into the oral chemical space. This oral space contains 916 oral drugs obtained from Dr Douguet (28). Furthermore, three radar plots are presented to illustrate how the compound fits into the used physicochemical filters, to highlight the complexity of the molecule and to graphically give a crude estimation of possible oral administration. In addition, several rules of thumb are computed (e.g. 3/75, 4/400, etc.) and the results are shown for each selected compound.

COMPUTATIONS OF MOLECULAR PROPERTIES AND DETECTION OF CHEMICAL GROUPS

Curation procedure

A data curation procedure is applied with several chained steps:

- (i) Removal of large compounds (i.e. molecules with more than 120 heavy atoms), of compounds that contain an atom other than H, C, N, O, F, P, S, Cl, Br, I, B and isotopes.
- (ii) A desalting procedure, which was optimized in order to correctly detect the salt part of a mixture. We collected 211 most used salts in medicinal chemistry and designed the corresponding SMARTS strings. These strings are then compared with the different parts of the mixtures identified in the input file and only the part recognized as a salt is removed. Otherwise, if no known salts are recognized, the process categorizes the molecules as mixtures and the molecules are not evalu-

ated further. Mixtures are written in a specific SDF file named *Mixtures.sdf* that the user can analyze.

- (iii) After a deprotonation step, the normalization of eight key chemical functions (amine, nitro, carboxylic acid, phosphonamide, amide, sulfonamide, phosphate and sulfate) is performed and a ring aromatization procedure is applied. This step is performed with the ChemAxon's Standardizer Academic package (www.chemaxon.com) and in-house SMARTS depicting patterns.
- (iv) A duplicates removal step is performed by comparing each normalized compound against all the other molecules contained in the file. The CANSMILES strings internally generated by OpenBabel representing molecules in a unique manner are compared and duplicates are removed. As stereoisomers can have different bioactivities, the different forms of the same compound are kept only when the stereochemistry is noted in the input file, and as such, these compounds are not considered as duplicates.

It is important to note that if users only need to curate a small compound library (without computing other properties), then the Bank-Cleaner service (<http://fafdrugs3.mti.univ-paris-diderot.fr/links>) can be used.

To illustrate this initial step, we processed the third release of the NCI Open Database (September 2003) that contains 260 071 structures. We counted around 1% of internal duplicates, 0.3% of large compounds (more than 300 atoms), 6.6% of molecules containing inorganic atoms, 10% of compounds associated with a salt and 1% of mixtures.

Physicochemical descriptors

In order to generate a compound collection with acceptable physicochemical properties, FAF-Drugs3 computes 17 descriptors: MW, number of rigid and rotatable bonds, log P (with XLOGP3 (29)), log D (using the ChemAxon package (www.chemaxon.com)), tPSA (30), number of RO5 HBA and HBD (3) (HBA and HBD are detected according to the publication with an in-house SMARTS definition depicting, respectively, the number of N and O atoms and the number of OH and NH), number of heavy atoms, heteroatoms/carbon atoms ratio, number of rings, maximum ring size, number of stereocenters and Fsp3 (11), number of charges and the formal charge of the molecule. In order to correctly compute descriptors depending on the charged state of the groups, a protonation procedure, as implemented in ChemAxon, is used. It is managed by the *cxcalc* command of the ChemAxon's Calculator Plugins Academic package (www.chemaxon.com) and our protocol selects the major microspecies at pH 7.4.

Physicochemical filtering parameters

According to the type of screening assays and/or the nature of the target, the desired physicochemical properties of the compounds can vary. Thus, FAF-Drugs3 allows user to design a custom filter by using the Filter-Editor form where one can choose project-dependent physicochemical ranges. Alternatively, FAF-Drugs3 proposes a large list of

pre-defined physicochemical filters (see Supplementary Table S2 for ranges): an RO5-like filter (3), an RO3-filter useful when constructing fragment libraries (31), a Probe-like filter (5), the REOS filter (32), the ZINC drug-like filter (33), a CNS filter (34) based on the knowledge of the physicochemical properties of drugs and molecules that are known to penetrate the blood-brain barrier and a respiratory filter (35) developed after analysis of inhaled or intranasal administered drugs. In addition, we defined a drug-like and a lead-like filters based on the analysis of the physicochemical properties of a set of 916 oral FDA approved drugs (28) and on a consensus list of already published filters (3–5,33,36).

Physicochemical, ADMET rules and related to help in decision-making

Several rules estimating oral bioavailability are implemented in FAF-Drugs3, including the well-known RO5 mentioned above (3). In the same manner, the Veber rule (37) (≤ 10 rotatable bonds and $tPSA \leq 140 \text{ \AA}^2$ (or $\leq 12 \text{ HBA} + \text{HBD}$)), the Egan rule (38) ($-1 \leq \log P \leq 5.8$ and $tPSA \leq 130 \text{ \AA}^2$) and the Bayer TrafficLights (39) (involving tPSA, log P, MW, rotatable bonds and solubility (ESOL model)) are calculated to estimate oral bioavailability. Regarding the estimation of aqueous solubility, a property that is very difficult to predict correctly (40,41), two approaches have been implemented in FAF-Drugs3: the ESOL model giving a rate of log S (42) and the Solubility Forecast index (43) indicating if a compound has a reasonable chance to be soluble in water (i.e. when its log D (pH7.4) + number of Aromatic Rings < 5 (the log D is computed with the ChemAxon's Calculator Plugins Academic package (www.chemaxon.com))). Our web server, along with descriptor calculations, computes also several other rules: the GSK 4/400 (9) and Pfizer 3/75 (10). In addition, FAF-Drugs3 embeds the freely available Eli-Lilly open drug discovery medicinal chemistry software (12). This package was developed to flag, accept or reject compounds according to the presence of difficult chemical groups and of some other properties including their resemblance to known drugs. A compound that passes this filter can enter the open drug discovery program offered by the company.

Furthermore, if the users are interested in preparing a compound collection potentially enriched in inhibitors of protein-protein interactions, we have previously reported a decision tree model named PPI-HitProfiler (44) that analyzes the 3D shape of a compound and search for a critical number of multiple bonds, this tool is now implemented online. In this case however, the users should upload molecules in 3D. As it is known that many protein-protein interactions inhibitors often have a high log P and high MW (45), we anticipate that users wishing to prepare such specialized collections would benefit coupling the PPI-HitProfiler with the other filters that we have implemented in our server.

Toxicophores and molecules interfering with biological assays

Structural alerts here include several known toxicophores, which are chemical moieties directly and/or indirectly linked to toxicity, and also substructures or molecules interfering with biological assays. Searches for such molecules

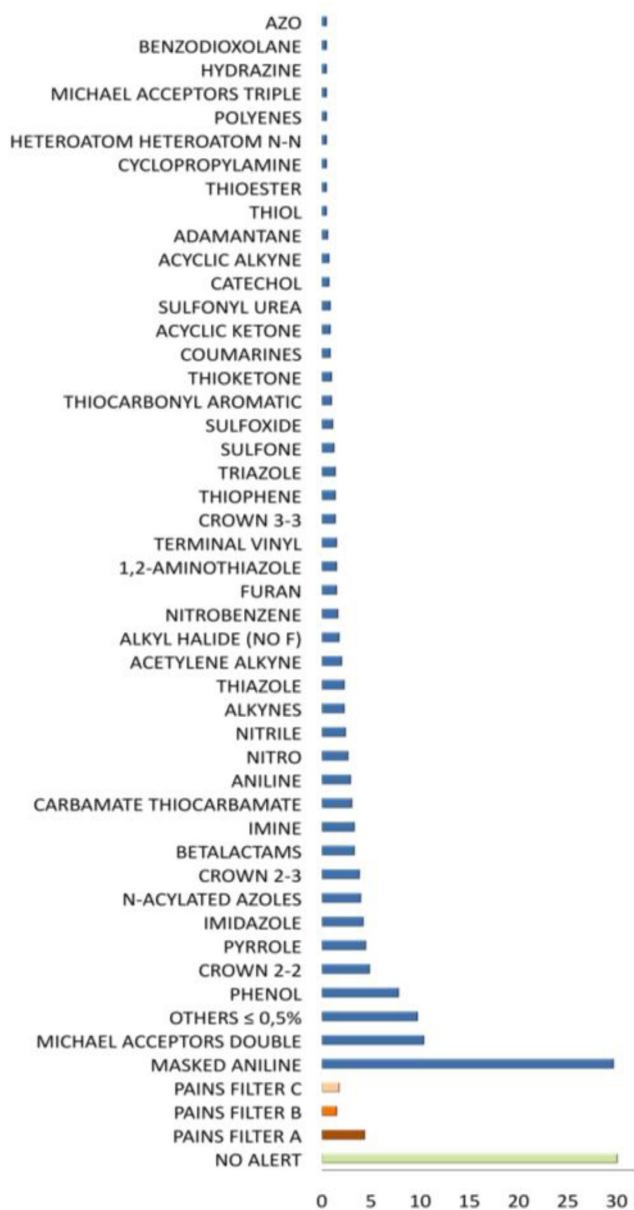


Figure 2. FAF-Drugs3 alerts identified on oral drugs. Search for the presence of structural alerts and PAINS in 778 oral drugs (i.e. the oral drugs that had clear annotation in terms of therapeutic areas). Thirty percent of drugs (in green) do not show any structural alert; 7.8% of the alerts revealed a phenol group (in blue) and 7% of the drugs are flagged as PAINS (in orange).

or toxicophores were already performed in FAF-Drugs2 but we have introduced several changes and optimized the process.

Regarding toxicophores, a survey of the literature was made in order to optimize the definition of the SMARTS strings used for the search as compared to the FAF-Drugs2 version. We gathered a consensus list of 154 documented toxicophores (see Supplementary Table S1 for the name of the chemical groups and corresponding SMARTS codes), i.e. if there are no or very little published data on the groups, it is not implemented in FAF (2,4,8,12–16,18,46–47). This

final list includes chemical functions, for instance warheads where the toxicity is linked to an inherent direct chemical reactivity (such as epoxides) but also to structural groups that require metabolism with subsequent generation of a reactive function (e.g. anilines) or yet alerts that usually imply tight binding to CYP450 enzymes. This list also contains mutagenics like quinones or coumarins that can act as DNA intercalators. For some of them, according to reports in the literature and a structural analysis of approved drugs, we decided to apply a threshold cutoff on the number of occurrences of a given substructure in the same compound (e.g. we accept three occurrences of the nitro structural alert because such group could easily be substituted while we authorize only one furan five-membered aromatic ring). Depending on such analysis, the molecule is then flagged (intermediate basket) or removed (rejected basket). Each structural alert SMARTS definition was depicted with the online web-server SMARTSViewer (48) in order to graphically validate the pattern. Then, the chemical groups of interest were benchmarked against test compounds that contained the desired substructures; these test structures were downloaded from PubChem (49). Another validation step (see below) was performed to further evaluate our SMARTS on a list of annotated drugs (13).

With regard to molecules interfering with assays, the FAF-Drugs2 web server (23) had a total of 326 SMARTS definitions for frequent hitters (50,51) (15 patterns) plus aggregators (52) and promiscuous inhibitors (19) (we have implemented a total of 311 patterns for these two groups). Also, 511 patterns to search for PAINS compounds (18) were implemented. All these SMARTS are present in FAF-Drugs3 (the SMARTS codes are available in Supplementary Table S1) and were tested again. For example, a new benchmark for PAINS recognition was carried out against the original control data set of 10 000 structures from the WEHI 93K HTS library (18) in order to verify that the compound preparation and curation steps had no impact on the accuracy of the PAINS detections.

VALIDATION AND CASE STUDIES

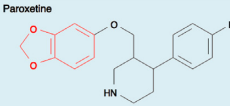
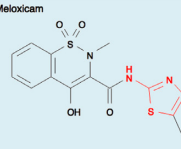
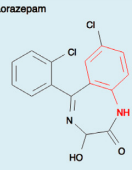
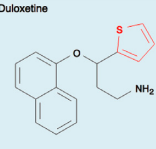
Detection of toxicophores

As noted above, we have tested extensively FAF-Drugs3. In order to assess further the SMARTS codes that search for toxicophores, we compared the output of FAF-Drugs3 with the results of a previous report (13) that looked for the presence of structural alerts in marketed and withdrawn drugs. In the present study, we decided to study a subset of 40 low molecular weight compounds selected at random in the study of Stepan *et al.* FAF-Drugs3 detected a problematic moiety in 36 out of the 40 compounds, with an efficacy of 100% with regard to the nature of the alerts (see Table 1 and Supplementary Table S3).

Analysis of drug compounds

In order to test further the FAF-Drugs3 service, we decided to analyze about 1500 annotated FDA-approved drugs. We selected only oral drugs and compounds that were annotated in terms of therapeutic areas. We made this selection following the established Tufts Center for

Table 1. Detection of toxicophores on some selected molecules

	Paroxetine	Meloxicam	Lorazepam	Duloxetine
				
Stepan <i>et al.</i>	1,3-Benzodioxole	2-Aminothiazole	Aniline	Thiophene
FAF-Drugs3	Benzodioxolane	(a) 1,2-aminothiazole (b) Michael Acceptor	Masked Aniline (low risk aniline)	Thiophene

Four drugs taken from the top 200 drugs marketed in the USA and their structural alerts are shown. These molecules were annotated by Stepan *et al.* (13) and the results obtained with FAF-Drugs3 are shown (see also Supplementary Table S3 for the analysis of 40 drugs). A recent study suggests that 2-Aminothiazoles could be considered as frequent-hitting fragments (51).

the Study of Drug Development and FDA's definitions. Thus, we collected 778 compounds categorized in 19 therapeutic areas: Allergology, Analgesics, Anesthetics, Anti-Inflammatory, Cardiology, Dermatology, Endocrinology, Family, Gastroenterology, Hematology, Immunology, Infectious, Musculoskeletal, Neurology, Nutrition, Obstetrics, Oncology, Ophthalmology, Otolaryngology, Pharmacology, Pulmonary, Rheumatology and Urology. We analyzed the physico-chemical properties of these compounds and we also checked if FAF-Drugs3 could identify the structural alerts present in these molecules. The analysis showed that only 30% of these oral drugs do not contain structural alerts (Figure 2) while 29% of the alerts are low-risk anilines (i.e. the anilinic nitrogen atom is not terminal, but can be incorporated in an aromatic ring). We also detailed which structural alerts are the most frequently detected in four therapeutic areas (see Supplementary Figure S4). One notes that the infectious and oncologic areas are the worst with regard to structural alerts with only 14% of the drugs without detected toxicophores. Overall, the most frequent alert is the aniline, especially the masked aniline. Michael Acceptors, phenol or pyrrole can be problematic moieties that are frequently retrieved in the different therapeutic areas. Regarding the physico-chemical properties, these oral drugs globally satisfy the RO5. Indeed, 630 drugs out of the 778 (81%) have no RO5 violation. Again, the infectious area is among the worst where 52 drugs out of 159 (33%) show at least one RO5 violation. It should be underlined that in the seminal Lipinski *et al.* analysis (3), antibiotics (and natural products) were not considered. When investigating the dermatologic area, it is interesting to note that 50% of the molecules do not satisfy the RO5. Analysis of the four descriptors involved in the RO5 and their values computed for the approved compounds that violate this rule underline

(see Supplementary Figure S5) that log P and MW are most often responsible for the rejection, 80 and 78%, respectively.

CONCLUSIONS AND FUTURE DIRECTIONS

We believe that FAF-Drugs3 can contribute to the design of high-quality compound collections. Further, FAF-Drugs3 outputs should assist decision-making and help users in selecting the molecules that could have a higher probability of success and that are best suited for optimization. Our tool should help in saving time and money to the scientific community working in the field of drug discovery, chemical biology and environmental sciences. In order to keep the service at a high standard, updates are planned twice a year. This will involve code optimizations to deal with larger chemical collections as well as adding new functionalities.

AVAILABILITY

Help page: fafdrugs3.mti.univ-paris-diderot.fr.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank INSERM and Paris Diderot University for supports. They are also grateful to Drs P. Tufféry, A. Saladin and J. Rey for helpful discussions on the Mobylye Portal.

FUNDING

Funding for open access charge: INSERM.

Conflict of interest statement. None declared.

REFERENCES

- Keseru, G.M. and Makara, G.M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.*, **8**, 203–212.
- Cumming, J.G., Davis, A.M., Muresan, S., Haeberlein, M. and Chen, H. (2013) Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.*, **12**, 948–962.
- 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 Deliv. Rev.*, **46**, 3–26.
- Brenk, R., Schipani, A., James, D., Krasowski, A., Gilbert, I.H., Frearson, J. and Wyatt, P.G. (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem*, **3**, 435–444.
- Workman, P. and Collins, I. (2010) Probing the probes: fitness factors for small molecule tools. *Chem. Biol.*, **17**, 561–577.
- Doak, B.C., Over, B., Giordanetto, F. and Kihlberg, J. (2014) Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chem. Biol.*, **21**, 1115–1142.
- Mok, N.Y., Maxe, S. and Brenk, R. (2013) Locating sweet spots for screening hits and evaluating pan-assay interference filters from the performance analysis of two lead-like libraries. *J. Chem. Inf. Model.*, **53**, 534–544.
- Blagg, J. (2010) Structural Alerts for Toxicity. In: Abraham, D.J. and Rotella, D.P. (eds). *Burger's Medicinal Chemistry, Drug Discovery, and Development*. 7th edn. John Wiley & Sons, Inc, NJ, 301–334.
- Gleeson, M.P. (2008) Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.*, **51**, 817–834.
- Hughes, J.D., Blagg, J., Price, D.A., Bailey, S., Decrescenzo, G.A., Devraj, R.V., Ellsworth, E., Fobian, Y.M., Gibbs, M.E., Gilles, R.W. et al. (2008) Physicochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.*, **18**, 4872–4875.
- Lovering, F., Bikker, J. and Humblet, C. (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.*, **52**, 6752–6756.
- Bruns, R.F. and Watson, I.A. (2012) Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.*, **55**, 9763–9772.
- Stepan, A.F., Walker, D.P., Bauman, J., Price, D.A., Baillie, T.A., Kalgutkar, A.S. and Aleo, M.D. (2011) Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.*, **24**, 1345–1410.
- Benigni, R. and Bossa, C. (2011) Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. *Chem. Rev.*, **111**, 2507–2536.
- Enoch, S.J., Ellison, C.M., Schultz, T.W. and Cronin, M.T. (2011) A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. *Crit. Rev. Toxicol.*, **41**, 783–802.
- Kalgutkar, A.S., Gardner, I., Obach, R.S., Shaffer, C.L., Callegari, E., Henne, K.R., Mutlib, A.E., Dalvie, D.K., Lee, J.S., Nakai, Y. et al. (2005) A comprehensive listing of bioactivation pathways of organic functional groups. *Curr. Drug Metab.*, **6**, 161–225.
- Baell, J. and Walters, M.A. (2014) Chemistry: chemical con artists foil drug discovery. *Nature*, **513**, 481–483.
- Baell, J.B. and Holloway, G.A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, **53**, 2719–2740.
- McGovern, S.L., Caselli, E., Grigorieff, N. and Shoichet, B.K. (2002) A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.*, **45**, 1712–1722.
- Drwal, M.N., Banerjee, P., Dunkel, M., Wettig, M.R. and Preissner, R. (2014) ProTox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic Acids Res.*, **42**, W53–W58.
- Villoutreix, B.O., Lagorce, D., Labbe, C.M., Sperandio, O. and Miteva, M.A. (2013) One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. *Drug Discov. Today*, **18**, 1081–1089.
- Miteva, M.A., Violas, S., Montes, M., Gomez, D., Tuffery, P. and Villoutreix, B.O. (2006) FAF-Drugs: free ADME/tox filtering of compound collections. *Nucleic Acids Res.*, **34**, W738–W744.
- Lagorce, D., Maupetit, J., Baell, J., Sperandio, O., Tuffery, P., Miteva, M.A., Galons, H. and Villoutreix, B.O. (2011) The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound collections. *Bioinformatics*, **27**, 2018–2020.
- Przybylak, K.R., Alzahrani, A.R. and Cronin, M.T. (2014) How does the quality of phospholipidosis data influence the predictivity of structural alerts? *J. Chem. Inf. Model.*, **54**, 2224–2232.
- Neron, B., Menager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P. and Letondal, C. (2009) Moby: a new full web bioinformatics framework. *Bioinformatics*, **25**, 3005–3011.
- O'Boyle, N.M., Morley, C. and Hutchison, G.R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5–12.
- Miteva, M.A., Guyon, F. and Tuffery, P. (2010) Frog2: efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res.*, **38**, W622–W627.
- Pihan, E., Colliandre, L., Guichou, J.F. and Douguet, D. (2012) e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design. *Bioinformatics*, **28**, 1540–1541.
- Cheng, T., Zhao, Y., Li, X., Lin, F., Xu, Y., Zhang, X., Li, Y., Wang, R. and Lai, L. (2007) Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.*, **47**, 2140–2148.
- Ertl, P., Rohde, B. and Selzer, P. (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, **43**, 3714–3717.
- Congreve, M., Carr, R., Murray, C. and Jhoti, H. (2003) A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today*, **8**, 876–877.
- Charifson, P.S. and Walters, W.P. (2002) Filtering databases and chemical libraries. *Mol. Divers.*, **5**, 185–197.
- Irwin, J.J. and Shoichet, B.K. (2005) ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, **45**, 177–182.
- Jeffrey, P. and Summerfield, S. (2010) Assessment of the blood-brain barrier in CNS drug discovery. *Neurobiol. Dis.*, **37**, 33–37.
- Ritchie, T.J., Luscombe, C.N. and Macdonald, S.J. (2009) Analysis of the calculated physicochemical properties of respiratory drugs: can we design for inhaled drugs yet? *J. Chem. Inf. Model.*, **49**, 1025–1032.
- Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. Aided Mol. Des.*, **14**, 251–264.
- 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**, 2615–2623.
- Egan, W.J., Merz, K.M. Jr and Baldwin, J.J. (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.*, **43**, 3867–3877.
- Lobell, M., Hendrix, M., Hinzen, B., Keldenich, J., Meier, H., Schmeck, C., Schohe-Loop, R., Wunberg, T. and Hillisch, A. (2006) In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem*, **1**, 1229–1236.
- Chevillard, F., Lagorce, D., Reynes, C., Villoutreix, B.O., Vayer, P. and Miteva, M.A. (2012) In silico prediction of aqueous solubility: a multimodel protocol based on chemical similarity. *Mol. Pharm.*, **9**, 3127–3135.
- Wang, J. and Hou, T. (2011) Recent advances on aqueous solubility prediction. *Comb. Chem. High Throughput Screen.*, **14**, 328–338.
- Delaney, J.S. (2004) ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.*, **44**, 1000–1005.
- Hill, A.P. and Young, R.J. (2010) Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discov. Today*, **15**, 648–655.
- Reynes, C., Host, H., Camproux, A.C., Laconde, G., Leroux, F., Mazars, A., Deprez, B., Fahraeus, R., Villoutreix, B.O. and Sperandio, O. (2010) Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. *PLoS Comput. Biol.*, **6**, e1000695.
- Labbe, C.M., Laconde, G., Kuenemann, M.A., Villoutreix, B.O. and Sperandio, O. (2013) iPPI-DB: a manually curated and interactive

- database of small non-peptide inhibitors of protein-protein interactions. *Drug Discov. Today*, **18**, 958–968.
46. Pearce, B.C., Sofia, M.J., Good, A.C., Drexler, D.M. and Stock, D.A. (2006) An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.*, **46**, 1060–1068.
47. Ashby, J., Basketter, D.A., Paton, D. and Kimber, I. (1995) Structure activity relationships in skin sensitization using the murine local lymph node assay. *Toxicology*, **103**, 177–194.
48. Schomburg, K., Ehrlich, H.C., Stierand, K. and Rarey, M. (2010) From structure diagrams to visual chemical patterns. *J. Chem. Inf. Model.*, **50**, 1529–1535.
49. Bolton, E. and W.Y.W.Y., Thiessen, P.A. and Bryant, S.H. (2008) *Annual Reports in Computational Chemistry*, American Chemical Society, Washington, DC, Vol. 4.
50. Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht, E.M., Rogers-Evans, M. *et al.* (2002) Development of a virtual screening method for identification of 'frequent hitters' in compound libraries. *J. Med. Chem.*, **45**, 137–142.
51. Devine, S.M., Mulcair, M.D., Debono, C.O., Leung, E.W., Nissink, J.W., Lim, S.S., Chandrashekar, I.R., Vazirani, M., Mohanty, B., Simpson, J.S. *et al.* (2015) Promiscuous 2-aminothiazoles (PrATs): a frequent hitting scaffold. *J. Med. Chem.*, **58**, 1205–1214.
52. Rishton, G.M. (2008) Aggregator compounds confound amyloid fibrillization assay. *Nat. Chem. Biol.*, **4**, 159–160.