Managing, Profiling and Analyzing a Library of 2.6 Million Compounds Gathered from 32 Chemical Providers

Aurélien Monge, Alban Arrault, Christophe Marot, Luc Morin-Allory

Managing, Profiling and Analyzing a Library of 2.6 Million Compounds Gathered from 32 Chemical Providers

Aurélien Monge, Alban Arrault, Christophe Marot and Luc Morin-Allory*

Institut de Chimie Organique et Analytique, UMR CNRS 6005, Université d'Orléans

BP 6759, 45067 Orléans Cedex 2.

* Corresponding author:

phone: +33 (0) 2 38 41 70 42

fax: +33 (0) 2 38 41 72 81

e-mail: aurelien.monge@univ-orleans.fr

ABSTRACT

3.8 million compounds from structural databases of 32 providers were gathered and stored in a single chemical database. Duplicates are removed using the IUPAC International Chemical Identifier. After this, 2.6 million compounds remain. Each database and the final one were studied in term of uniqueness, diversity, frameworks, ‘drug-like’ and ‘lead-like’ properties. This study also shows that there are more than 87 000 frameworks in the database. It contains 2.1 million ‘drug-like’ molecules among which, more than one million are ‘lead-like’. This study has been carried out using ‘ScreeningAssistant’, a software dedicated to chemical databases management and screening sets generation. Compounds are stored in a MySQL database and all the operations on this database are carried out by Java code. The druglikeness and leadlikeness are estimated with ‘in-house’ scores using functions to estimate convenience to properties; unicity using the InChI code and diversity using molecular frameworks and fingerprints. The software has been conceived in order to facilitate the update of the database. ‘ScreeningAssistant’ is freely available under the GPL license.

Keywords: chemical databases, chemoinformatics, diversity, drug-like, lead-like, screening.

ABBREVIATIONS

HBA: H Bond Acceptor.

HBD: H Bond Donor.

HTS: High-Throughput Screening.

InChI: IUPAC International Chemical Identifier.

JNI: Java Native Interface.

MW: Molecular Weight.

RO5: rule-of-five.

SCA: Stochastic Clustering Analysis.

SSSR: Smallest Set of Smallest Rings.

INTRODUCTION

For a project of virtual or real screening, choosing the set of molecules that are to be tested is a really important and difficult step. It must be as representative as possible of the potential ligands of the studied biological target. To perform these operations in an efficient way, the medicinal chemist needs:

- a suitable software. However, as far as we know, there is no affordable structure data management system which allows the end-user to easily manage a database of millions of screening compounds coming from various providers. This tool must store only unique structures (and in consequence it has to use an efficient unique code for structures) but also keep information of all the providers of a given structure. Furthermore the software must be able to assist the user to select compounds. Its function is to analyze the compounds in the database in terms of physicochemical properties, druglikeness, leadlikeness and diversity.
- a good knowledge of the features of the commercially available chemical libraries. Except 'Big Pharmas', companies usually prefer to buy compounds to a very limited number of providers for convenience and also to get a discount on those compounds. In consequence, a choice must be done among numerous providers. Druglikeness and diversity of a database are the two main features generally considered (even if a database with a low diversity and few 'drug-like' compounds can be a good database when it is focalized on a particular target). Despite the relevance of such information, there is a lack of exhaustive commercially available chemical libraries comparison in the literature [1]. Even if some analysis have already been published, they are either limited to a small number of providers or do not compare both 'drug-like' properties and diversity [1, 2, 3, 4, 5, 6].

The paper will discuss how to manage and profile a large size of compound library with reasonable cost, and how to distinguish ‘drug-like’ and ‘lead-like’ compounds from a huge chemical space. A program has been developed for these applications, and is publicly available [7].

MATERIALS AND METHODS

The database server MySQL is used on a Linux PC. All the code is programmed in Java except one method that is coded in C using Java Native Interface (JNI) which allows launching batch jobs using Windows. Operations on chemical structures and descriptors calculation were carried out with the JOELib [8] Java library and Java code. The structural data (SDF, SMILES, InChI...) are separated from descriptors in order to speed up SQL queries on descriptors. The 3D structures are automatically generated in MOL2 format using Corina [9]. All structures are normalized by Corina software.

PREPROCESS

Each imported compound has, if necessary, its counter-ion removed (only the largest contiguous fragment is kept) and is protonated at physiological pH using JOELib. This new structure is added if not already present in the database. To characterize the chemical structures we chose to use the IUPAC International Chemical Identifier (InChI) [10]. This new unique code has been used recently in several projects [11, 12, 13]. InChI is free and conceived in the perspective to become a standard unique code for molecules. Structures are represented by a code made of one text line taking into account sp2 and sp3 stereochemistries, isotopes and tautomerism. It has very good functionalities compared to other softwares which use ‘unique’ SMILES approaches. A comparison of the InChI, MOE [14], OEChem [15] and

Marvin [16] functionalities is available in Table 1. It is the only algorithm in our tests that manages simple tautomerism and moveable positive charge detection.

Most of the databases contain some duplicated products. A reason for detecting duplicates in databases is the presence of undefined or badly-defined stereochemistries (e.g. two diastereoisomers without the indicators of the chirality). We also want to note that counter-ions are not taken into account to check duplicates (then, two different salts of the same acid are considered as duplicates). The result is a slight overestimation of the number of duplicates.

We used the 1.12 Beta version of InChI which was the only available version at the time of this work. It had a basic support for aromatics bonds because this bond type is specified to be only “for query” in the MDL mol format description. As JOELib codes some structures using the aromatic bond type, these structures cannot have an InChI code computed and are considered as unique. The structures without unique codes represent less than 2 % of the database and then, the percentage of duplicates due to this problem is very low. The support of aromatic bond type was improved in the now available final version of InChI.

In order to check for presence of a structure in the database, the MD5 hashcode of the InChI of this molecule is compared to the indexed MD5 hashcodes of the InChI of all the compounds in the database. If two structures possess the same MD5 hashcode, their InChI codes are compared to check whether they are actual duplicates, even if the probability that two structures have the same MD5 hashcode but not the same InChI code is very low (there are 16^{32} possibilities for MD5 hashcodes, so the probability of collision is very low). The use of the MD5 hashcode allows research of duplicates in a smaller table with fixed size rows, which is much more effective in term of computational time.

DATABASES

The different databases were gathered on the web or obtained by collaboration. We obtained the databases of 32 providers or institutions (Table 2).

The database files of these 32 providers contain 3.8 million molecules. These compounds are made using classical organic synthesis, combinatorial chemistry or natural compounds extraction. The ICOA database is our corporate database, and is included in the french 'Chimiothèque Nationale' (Chem. Nat.) which gathers the databases of 17 french public laboratories [17].

DIVERSITY

We used the dissimilarity step of the Stochastic Clustering Analysis (SCA) [18] to identify the number of clusters as it had previously been done in a study of the NCI database [6]. Since the clusters are created by diversity, the number of clusters gives us information about the diversity of the database. The descriptors used and stored in the database are the SSKey-3DS [19] fingerprints. The SSKey-3DS are constituted of 32 bits coding for the presence or absence of 32 fragments, and 22 bits for the number of H bond acceptors, number of aromatic bonds, and fraction of rotatable bonds. These keys are computed during the insertion step of new compounds in the database. We used Tanimoto coefficient as a metric for pairwise comparison of molecules. The similarity cut-off has been set to 0.8. We programmed the SCA in Java. The number of clusters of the whole database (2.6 million unique compounds) is investigated within one hour on a standard PC.

FRAMEWORKS

In a study of shape of 'drug-like' compounds, Bemis used the notion of graph frameworks corresponding to ring systems connected to each other by linkers [20]. We chose to use frameworks because we think that the number of frameworks of a database can give

information about its diversity. To obtain the graph framework of a hydrogen depleted structure of a compound, all the atoms of the molecule are to be replaced by ‘non-typed’ atoms and all the bonds are replaced by ‘non-typed’ bonds. Then, all the atoms connected to only one other atom are removed. This step is repeated until no atom is deleted (Figure 1).

Our implementation of this algorithm replaces ‘non-typed’ atoms by C atoms and ‘non-typed’ bonds by single bonds. However, unlike the Bemis method, we differentiate aromatic from non aromatic bonds. In our opinion, it is important to distinguish between aromatic and non aromatic compounds because they belong to two very different chemical families. For instance, compounds based on a cyclohexan framework or on a benzene one are very different in terms of shape and flexibility, and it is important to keep this information. The advantage of this representation is that it can be stored as a simplified structure, able to be visualized with a molecular viewer. Furthermore, InChI can be computed for the frameworks which will be stored in a unique frameworks list.

‘DRUG-LIKE’ AND ‘LEAD-LIKE’ PROPERTIES

Several reviews concerning druglikeness were published in the last years [21, 22, 23, 24]. A major contributor in the area of the characterization of ‘drug-like’ compounds was Lipinski with the rule-of-five [25]. This rule is the most widely used to identify ‘drug-like’ compounds [26]. It deals with orally active compounds that have achieved phase II clinical status. So it is not a method to distinguish between drugs and non-drugs, but rather a method to predict compounds with poor absorption or permeability. The results published by Frimurer [27] illustrate this idea: the rule-of-five accepts 74 % of the ACD compounds but only 66 % of the MDDR compounds. The inability of the rule-of-five to distinguish between drugs and non-drugs was also demonstrated by Oprea [28].

Since the first publication of Lipinski, many methods have been published to identify 'drug-like' compounds. Some of them are also based on limits of physical properties [3, 28, 29, 30]. In a more recent publication, the authors use more complex descriptors to recognize 'drug-like' products [31]. Counting pharmacophore points was also used to predict druglikeness [32]. Machine learning methods such as Support Vector Machines [33] and Neural Networks [34, 35] were also successfully used in this area. But even if machine learning based methods give good results, these methods are "black-boxes". Simple and comprehensive rules are generally preferred to identify 'drug-like' compounds and so we chose to use 'drug-like' rules based on limits on physicochemical properties and on structural filtering. However, to avoid the drawbacks of strict cut-off, we implemented a progressive 'drug-like' score.

In the first step of this study, compounds with atoms other than C, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, or Li are flagged. Then, we have used filters described in another study from our laboratory to evaluate the druglikeness [2] established from publications [22, 36]:

- $100 \leq$ molecular weight ≤ 800 g.mol^{-1}
- logP ≤ 7
- HBA ≤ 10
- HBD ≤ 5
- rotatable bonds ≤ 15
- no reactive functions (eliminate false positives in biochemical tests)
- halogen atoms ≤ 7
- alkyl chains $\leq$ -$(CH_2)_6CH_3$
- no perfluorinated chains: -$CF_2CF_2CF_3$
- SSSR ≤ 6
- no big size ring with more than 7 members

- at least one N or O atom

The following definitions are used:

- reactive functions: modified version by Oprea [28] of the list published by Rishton [37]. In addition to the list, we add vinyl sulfones as reactive function.
- HBA: nitrogen, oxygen, phosphor and sulfur, except the following cases: aromatic oxygen and sulfur, aromatic nitrogen connected to three other atoms, nitrogen with valence 5, sulfur with valence of 6 or 7.
- HBD: heteroatom with a minimum of one hydrogen and without negative charge.
- rotatable bonds: The definition of JOELib [8] is used "Number of rotatable bonds, where the atoms are heavy atoms with bond orders one and a hybridization which is not one (no sp). Additionally the bond is a non-ring-bond."
- logP: SlogP is used for all this study [38].

The 'lead-like' concept is similar to the 'drug-like' one, but is more restrictive for some terms. This idea is the result of the fact that optimization of a lead compound often results in an increase of molecular weight, logP and complexity [39, 40]. In consequence, a 'lead-like' filter needs to select polar compounds with simple chemical structures [41]. Hann and Oprea recently gave rules to select 'lead-like' molecules, established from properties based analyses for preclinical drug discovery [42]: molecular weight (MW) $\leq$ 460, -4 $\leq$ logP $\leq$ 4.2, LogSw $\geq$ -5, rotatable bonds $\leq$ 10, number of rings $\leq$ 4, H bond donors $\leq$ 5 and H bond acceptors $\leq$ 9. We use a 'lead-like' filter based on these rules. In our database a compound is regarded as 'lead-like' if it is 'drug-like', with H bond acceptors (HBA) $\leq$ 9, molecular weight (MW) $\leq$ 460, -4 $\leq$ logP $\leq$ 4.2, rotatable bonds $\leq$ 10, and smallest set of smallest rings (SSSR) $\leq$ 4.

On the basis of these rules, we have evaluated the 'drug-like' and 'lead-like' properties using two methods. The first one simply counts the number of non-fitted criteria. The second method computes a progressive score based on these criteria. For each criterion, a penalty is calculated. For eight criteria this penalty varies from 0 to 1 and is computed from empiric functions based on the former cut-off values. These functions are described in Table 3 and an example is given in Figure 2. For HBD, HBA, rotatable bonds, SSSR, maximum ring size and halogens, we defined an intermediate zone centered on the limits of our previous 'drug-like' filters. The intermediate zone covers 60 % of the limit value. For example, the 'drug-like' limit for HBA is 10. So an intermediate function for HBA will stretch from 7 (10 - 30 %) to 13 (10 + 30 %). If a molecule has less than 7 HBA, the penalty for this property will be 0, and if it has more than 13, the maximum penalty of 1 will be applied for this penalty. For MW and logP, published distributions of these properties were used. For MW, the lower intermediate penalty zone extends from 100 to 150 (based on the marketed drug weight distribution [43]). The upper intermediate penalty zone stretches from 350 to 800 for the druglikeness (500 – 30 % and the former limit 800 is kept because it was already very permissive) and from 322 to 588 for the leadlikeness (former cut-off: 460 [42]).

For logP, the lower intermediate penalty zone extends from -5 to -1.5, the upper intermediate penalty zone stretches from 4.5 to 7.5 for the druglikeness (based on the marketed drug logP distribution [43]) and from 2.9 to 5.5 for the leadlikeness (former cut-off : 4.2 [42]).

All these functions result either from the distribution of properties of known drugs or from previously proposed limits. We will see later that they are able to efficiently identify compounds which present a bad absorption or a low solubility. But they are empiric and future studies should allow refining them.

This method has two advantages against the sum of the number of unsatisfied 'drug-like' criteria. Firstly, threshold effects are avoided. For example, these effects may be important in

logP calculations by different methods. Secondly, this progressive score permits to sort compounds by druglikeness and leadlikeness.

Four criteria are still used in a 'cut-off way':

- the presence of a reactive function
- the presence of a single chain > $-(CH_2)_6CH_3$
- the presence of a perfluorinated chain
- no O or N atom

The 'lead-like' score is designed exactly like the 'drug-like' score but using the 'lead-like' properties when they differ from the 'drug-like' ones.

The score of compounds is obtained by the sum of these penalties. A low score (≤ 1) indicates a molecule which can be considered as 'drug-like' or 'lead-like'. A score ≥ 2 means that the compound is not 'drug-like' or 'lead-like'.

The selection of compounds with nitro group is not recommended for the risk of causing 'false positives' in HTS tests [36]. Although nitro group is not in our default frequent hitters list, all nitro compounds are flagged and can easily be removed by the software if wanted.

We want to highlight that it there is no absolute 'lead-like' and 'drug-like' rules. It is strongly dependant of the nature of the project. Although we have chosen parameters for each of these rules, our system allows us to change them very easily in order to extract a new dataset of compounds with different properties. In addition to the classical parameters, we can eliminate molecules with unwanted substructures from the dataset.

The two approaches, the rule-of-five (RO5) and our score have been applied to the Prestwick database in order to compare their results. We chose the Prestwick database for this analyze because 85 % of the compounds in this database are marketed drugs. Among the 876

compounds of the database, 92 % (804) are accepted by the RO5 and 8 % (72) are rejected. This result is quite normal for products which are mainly administered orally.

Our ‘drug like’ score is more restrictive; it accepts 80 % (700) of the products and rejects 20 % (176). Among the 176 rejected compounds, 44 are rejected by rules not present in the RO5, mainly by the notion of reactive functions (43). These compounds can be good drug candidates, but they are unwanted during the HTS process. As reactive functions have nothing to do with water solubility or absorption, we will compare only the progressive part of the ‘drug-like’ score to the RO5. Using this new score, 85 % (744) of the products are accepted and 15 % rejected (132). All the compounds accepted by the progressive part of our ‘drug-like’ score are also accepted by the rule-of-five, but 7 % of the compounds of the database are rejected by our score and accepted by the rule-of-five. Then, from these results, it is possible to create three sets of products:

- Set A (744 products) which contains products accepted by both approaches,
- Set B (60 products) with products accepted by the RO5 but rejected by our score,
- Set C (72 products) with products refused by both methods.

The issue is to determine if our ‘drug-like’ score has identified compounds, the ones of set B, with potential solubility or absorption issues not identified by RO5, or if it is just too restrictive. As we didn’t have the experimental values of these properties for these compounds, we chose to use predicted water solubility [44] implemented in MOE 2005.06, and Topological Polar Surface Area (TPSA) introduced by Ertl et al. [45] for the prediction of absorption. We considered that compounds with water solubility < 1µM [4] have a low water solubility, and that compounds with TPSA > 140 A^2 have a poor absorption. The TPSA limit is based on the work of Palm et al. who established a good sigmoidal correlation between dynamic PSA and passive drug transport ($r^2 = 0.94$) [46].

Using these criteria, 7% of set A are rejected; which proves that this set is mainly made of compounds with good properties. For set C, 96% are rejected, a proof of bad properties of the products of this set.

The set B contains 68% of products rejected. It shows that the products of this set are mainly products with bad properties and then, our score can be a better way to filter products than the simple application of the RO5. Thanks to the progressive limits, it is able to detect products which have many properties just under the limits of the RO5. These products will probably have problems of solubility or absorption.

The importance of the progressive limits can be illustrated with an example. In our current set B, the compound with the highest value (then the worst) of our progressive 'drug-like' score (2.6) is neamine (CAS number 3947-65-7). This compound is very polar (logP = -5.1), but it only has one criterion of the rule-of-five that is not validated (HBD = 8), and one criterion that is at the limit value (HBA = 10). In consequence the rule-of-five is validated for this compound but the probability that it will have absorption issues is high (TPSA = 210). Actually neamine is a component of neomycin, a topical and gastrointestinal antibiotic. Neomycin is known to have a bad intestinal permeability. This is a good example of how a compound, that just passes the binary filters of a rule, can be identified with progressive limits.

RESULTS AND DISCUSSION

DISTRIBUTION OF THE SUM OF THE PROVIDERS' DATABASES

The origin of the 3.8 million compounds is given in Table 2. The four providers with the greatest number of available compounds are ChemDiv, InterBioScreen, ChemBridge and Enamine. However, the originality of the structures of each provider must also be assessed in order to compare them.

INTERNAL DUPLICATES

Before the creation of the fused database, a first step is the removal of internal duplicates. The databases Sigma-Aldrich (9.3 %), NCI (6.0 %), MDPI (5.3 %) and Arkive (3.6 %) have the highest percentages of internal duplicates (Table 4). No duplicates were found in the databases of ACB Blocks, AnalytiCon Discovery, BioFocus and Tripos. The size of the library is definitively not linked to the number of internal duplicates. The best example being the ChemDiv library which is the biggest library. It has only 0.02 % of internal duplicates.

PROPORTION OF UNIQUE COMPOUNDS

The proportion of unique structures in a provider's database, *i.e.* the products only present in this database and not in the other providers' one, may vary in a great extend. These values are given in Table 4.

We must notice that our corporate database "ICOA" is not represented here because it is included in the french "Chimiothèque Nationale" (Chem. Nat.). Biofocus (100 %), Analyticon Discovery (100 %), ACB Blocks (98 %) and Tripos (97 %) have the highest percentages of original compounds. Except for the huge databases, there is no direct relationship between the databases sizes and the percentages of original compounds. Indeed, Analyticon Discovery is relatively small with 5438 compounds, but ACB Blocks and Tripos are larger with 61237 and 82370 compounds. The four biggest databases have between 36 and 85 % of unique compounds.

DIVERSITY

The chemical space covered by a database is an essential information. We used the dissimilarity step of the SCA algorithm with SSKey-3DS fingerprints to compute the number

of clusters for the whole database and for each provider (Figure 3 and Table 4). The NCI database is clearly the most representative of the chemical space and covers 59 % of the chemical space of the whole database. However this database can't be considered as a commercial database. After the NCI, Enamine (37 %), ChemDiv (36 %), InterBioScreen (35 %), Sigma-Aldrich (35 %) and ChemBridge (34 %) are the databases which are the most representative of the global diversity. The less representative database is ArrayBioPharma (0.5 %), but it is also the smallest database (517 compounds).

We have also studied the relationship between the number of compounds in a database and the diversity (number of clusters) of this database. We can see, in Figure 4, a rapid linear increase for the databases with less than 100 000 compounds but, for the databases of more than 150 000 compounds, the increase of the diversity is slower. The NCI with 10 623 clusters for 250 000 compounds is once more specific, with a very high diversity.

FRAMEWORKS

The whole database contains 87 000 frameworks (the structures are available as supplementary materials). Figure 3 shows the percentage of the frameworks of the whole database for each provider. Unlike the results obtained by the diversity study, the NCI is not the most representative of the whole database and comes in fifth position (19 %). Enamine is the first (33 %) followed by ChemDiv (26 %) and InterBioScreen (23 %). Among the commercial databases, the three with the most important number of frameworks are also the most diverse in the cluster approach.

The less representative databases are obviously the small databases: ArrayBioPharma (0.02 %), Chemical Block (0.29 %) and Prestwick (0.39 %). We can see in Figure 5 that the number of frameworks is highly correlated to the size of the databases ($r^2 = 0.89$) ; this correlation explains the previous results.

‘DRUG-LIKE’

We have studied the ‘drug-like’ properties of the bases using two approaches. A classic method computes the number of violations of the limits of the rules for each product. The second one uses the score presented in a previous section.

For each provider, the numbers of molecules with 0, 1, 2 and more than 2 ‘drug-like’ failures are available as supplementary materials (Figure 1 and Table 1).

All the libraries have a high ratio of molecules with 0 or 1 ‘drug-like’ failure. The library with the lowest percentage of molecules with none ‘drug-like’ failure are AnalytiCon Discovery, NCI and MDPI with 71 %. Biofocus (97 %) and ChemBridge (95 %) are the libraries with the highest percentages of compounds without ‘drug-like’ failure.

Among the libraries, only Array BioPharma has no molecules with 2 or more ‘drug-like’ failures.

It is interesting to note that Prestwick has not a specific low number of ‘drug-like’ failures (79 %), although it is a library containing mainly marketed drugs. It shows that ‘drug-like’ notion is not an absolute rule to filter potential drugs.

The other method to estimate ‘drug-like’ properties is the ‘drug-like’ score (Figure 6 and supplementary materials Table 2). If we consider as ‘drug-like’ the compounds with a ‘drug-like’ score ≤ 1, the providers with the largest percentage of ‘drug-like’ compounds are ChemBridge (93 %), Aurora (92 %) and Chemical Blocks (90 %). There are 2.1 million (83 %) ‘drug-like’ compounds in the whole database according to this score. The distribution of the ‘drug-like’ scores in the whole database is shown in Figure 7.

This score is the result of the sum of various criteria. The relative importance of the criteria is shown in Table 5. Much of the compounds (6 %) are removed because they contain reactive

functions. For virtual screening purposes, these reactive functions have no meaning and can be easily ignored by the software.

‘LEAD-LIKE’

At the beginning of a drug discovery project it is more convenient to have ‘lead’ compounds. A lead compound has a molecular weight and a logP smaller than a final drug compound which allows it to be optimized by adding chemical groups. As the criteria to select ‘lead-like’ compounds are more stringent than for ‘drug-like’ ones, the ‘lead-like’ space is smaller than the ‘drug-like’ space.

If we consider the molecules without ‘lead-like’ failures as ‘lead-like’, ChemicalBlock (80 %) and Array BioPharma (73 %) are the most ‘lead-like’ libraries. Analyticon Discovery (21 %) is the database with the lowest proportion of ‘lead-like’ compounds (supplementary materials Figure 2 and Table 1). These results are coherent with the ‘lead-like’ score presented in Figure 8.

If we consider the compounds with scores ≤ 1 as ‘lead-like’, the conclusions are similar. ChemicalBlock (82 %) and Array BioPharma (76 %) have the largest percentage of ‘lead-like’ compounds and Analyticon Discovery (10 %) has the fewer ‘lead-like’ compounds. The high percentages of ‘lead-like’ compounds in ChemicalBlock and Array BioPharma can be explained by the fact that they are building blocks databases. In the whole database, one million compounds (38 %) have a ‘lead-like’ score ≤ 1.

We can see in Figure 9 that the distribution of the ‘lead-like’ scores is linearly progressive on the whole database. As a consequence this function can be very useful to sort compounds by leadlikeness.

As for the ‘drug-like’ score, the criteria have not the same influence. The Table 6 shows that the logP filter is the most selective of the ‘lead-like’ filters and removes 42 % of the compounds.

DIVERSITY IN THE ‘LEAD-LIKE’ SPACE

We have compared in a previous section the chemical space covered by each database, using the number of clusters created by diversity in this database. However, the diversity in a database can be increased by compounds which are not good drug candidates. We present here a second analysis of the diversity space coverage, but this time we have limited our study to the ‘lead-like’ space (Figure 10). Then each result is the percentage of the ‘lead-like’ space of the whole database covered by the ‘lead-like’ compounds of a provider.

The NCI (58 %) is first of this ranking, then Chembridge (39 %), InterBioScreen (38 %), Enamine (38 %) and ChemDiv (37 %). In Figure 3 the order was NCI, Enamine, ChemDiv, InterBioScreen, Sigma-Aldrich and Chembridge. So we can see that the sorting of the databases by diversity is dependent of the chemical space studied. It appears that NCI is clearly the most diverse database, even in the lead-like space. The last database is Analyticon Discovery with 0.5 %, which is simply due to the nature of this database (we have used the NatDiverse database of Analyticon Discovery, in which one natural product scaffold can be used to synthesize 500-1500 compounds).

CONCLUSION

We have developed a platform to manage and analyze chemical databases. This system allows to combine easily databases and to analyze them. Furthermore we used progressive ‘drug-like’ and ‘lead-like’ scores to limit threshold effects of the classical rules based on criteria counts. We have gathered chemical libraries from 32 providers to obtain a database of 2.6

millions compounds. Among these unique compounds, 2.1 million compounds are found to be ‘drug-like’ and one million ‘lead-like’ according to our scores. The chemical libraries of each provider were analyzed in terms of druglikeness, leadlikeness, fingerprints based diversity and frameworks. The results can be very useful for the choice of a database at the beginning of a drug-discovery project.

Several improvements of our chemical database management system are in progress, such as the introduction of ‘warheads’ and ‘promiscuous aggregating inhibitors’ filters and the identification of privileged structures. ‘ScreeningAssistant’, the software used for this study, is freely available online [7].

ACKNOWLEDGMENT

We gratefully acknowledge Philippe Guedat and François Bellamy for fruitful discussions about ‘drug-like’ filters.

REFERENCES

1. Bradley, M.P., *An overview of the diversity represented in commercially-available databases*, J. Comput. Aided Mol. Des., 16 (2002) 299-300.

2. Mozziconacci, J.C., Arnoult, E., Baurin, N., Marot, C., Morin-Allory, L., *Preparation of a molecular database from a set of 2 million compounds for virtual screening applications : gathering, structural analysis and filtering*, 9th Electronic Computational Chemistry Conference, World Wide Web, march 2003.

3. Sirois, S., Hatzakis, G., Wei, D., Du, Q., Chou, K.C., *Assessment of chemical libraries for their druggability*, Comput. Biol. Chem. 29 (2005) 55-67.

4. Baurin, N., Baker, R., Richardson, C., Chen, I., Foloppe, N., Potter, A., Jordan, A., Roughley, S., Parratt, M., Greaney, P., Morley, D., Hubbard, R.E., Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds, J. Chem. Inf. Comput. Sci., 44 (2004) 643-657.

5. Cummins, D.J., Andrews, C.W., Bentley, J.A., Cory, M., *Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds,* J. Chem. Inf. Comput. Sci., 36 (1996) 750-763.

6. Voigt, J.H., Bienfait, B., Wang, S., Nicklaus, M.C., *Comparison of the NCI Open Database with Seven Large Chemical Structural Databases*, J. Chem. Inf. Comput. Sci., 41 (2001) 702-712.

7. Monge, A., *Screening Assistant*, http://screenassistant.sourceforge.net/

8. Wegner, J.K., *JOELib*, http://joelib.sourceforge.net/

9 . *Corina*. Molecular Networks GmbH. http://www.mol-net.com

10. *The IUPAC International Chemical Identifier Project*, http://www.iupac.org/inchi/

11. Murray-Rust, P., Rzepa, H.S., Stewart, J.J., Zhang, Y. *A global resource for computational chemistry*, J. Mol. Model., 11 (2005) 532-41.

12. Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S., Zhang, Y. *Enhancement of the chemical semantic web through the use of InChI identifiers*, Org. Biomol. Chem., 3 (2005) 1832-4.

13. Prasanna, M.D., Vondrasek, J., Wlodawer, A., Bhat, T.N. *Application of InChI to curate, index, and query 3-D structures*, Proteins, 60 (2005) 1-4.

14. *Molecular Operating Environment (MOE)*, Chemical Computing, http://www.chemcomp.com

15. *OEChem*, OpenEye Scientific Software, http://www.eyesopen.com

16. *Marvin*, ChemAxon. http://www.chemaxon.com

17. *Groupement De Service Chimiothèque Nationale*, http://chimiotheque-nationale.enscm.fr/

18. Reynolds, C.H., Druker, R., Pfahle, L.B., *Lead discovery using Stochastic Cluster Analysis (SCA): a new method for clustering structurally similar compounds*, J. Chem. Inf. Comput. Sci., 38 (1998) 305-312.

19. Xue, L., Godden, J.W., Bajorath, J., *Database searching for compounds with similar biological activity using short binary bit string representations of molecules*, J. Chem. Inf. Comput. Sci., 39 (1999) 881-886.

20. Bemis, G.W., Murcko, M.A., *The properties of known drugs. 1. Molecular frameworks*, J. Med. Chem., 39 (1996) 2887-2893.

21. Lajiness, M.S., Vieth, M., Erickson, J. *Molecular properties that influence oral drug-like behavior*, Curr. Opin. Drug Discov. Devel., 7 (2004) 470-477.

22. Walters,W.P., Murcko, M.A., *Prediction of 'drug-likeness'*, Adv. Drug Delivery Rev., 54 (2002) 255–271.

23. Clark, D.E., Pickett, S.D., *Computational methods for the prediction of 'druglikeness'*. Drug Discov. Today, 5 (2000), 49–58.

24. Muegge, I., *Selection criteria for drug-like compounds*, Med. Res. Rev., 23 (2003), 302–321.

25. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Deliv. Rev., 23 (1997) 3-25.

26. Lipinski, C.A., *Lead- and drug-like compounds: the rule-of-five revolution*, Drug Discov. Today, 1 (2004) 337-341.

27. Frimurer, T.M., Bywater, R., Nærum, L., Lauritsen, L.N., Brunak, S. *Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds*, J. Chem. Inf. Comput. Sci., 40 (2000), 1315-1324.

28. Oprea, T.I., *Property distribution of drug-related chemical databases*, J. Comput. Aided Mol. Des., 14 (2000) 251-264.

29. Xu, J., Stevenson, J. *Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity*, J. Chem. Inf. Comput. Sci., 40 (2000) 1177-1187.

30. Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W., Kopple, K.D., *Molecular Properties That Influence the Oral Bioavailability of Drug Candidates*, J. Med. Chem., 45 (2002), 2615-2623.

31. Zheng, S., Luo, X., Chen, G., Zhu, W., Shen, J., Chen, K., Jiang, H., *A New Rapid and Effective Chemistry Space Filter in Recognizing a Druglike Database*, J. Chem. Inf. Comput. Sci., 45 (2005), 856-862.

32. Muegge, I., Heald, S.L., Brittelli, D., *Simple Selection Criteria for Drug-like Chemical Matter*, J. Med. Chem., 44 (2001), 1841-1846.

33. Zernov, V.V., Balakin, K.V., Ivaschenko, A.A., Savchuk, N.P., Pletnev, I.V., *Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions*, J. Chem. Inf. Comput. Sci., 43 (2003), 2048-2056.

34. Ajay, A., Walters, W.P., Murcko, M.A., *Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?*, J. Med. Chem., 41 (1998), 3314-3324.

35. Sadowski, J., Kubinyi, H., *A scoring scheme for discriminating between drugs and nondrugs*, J. Med. Chem., 41 (1998), 3325-3329.

36. Charifson, P.S., Walters, W.P., *Filtering databases and chemical libraries*, J. Comput. Aided Mol. Des., 16 (2002), 311-323.

37. Rishton, G.M., *Reactive compounds and in vitro false positives in HTS*, Drug Discov. Today, 2 (1997) 382-384.

38. Wildman, S.A., Crippen, G.M., *Prediction of physicochemical parameters by atomic contributions*, J. Chem. Inf. Comput. Sci., 39 (1999) 868-873.

39. Hann, M.M., Leach, A.R., Harper, G, *Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery*, *J. Chem. Inf. Comput. Sci.*, 41 (2001), 856-864.

40. Oprea, T.I., *Current trends in lead discovery: Are we looking for the appropriate properties?*, J. Comput. Aided Mol. Des., 16 (2002), 325-334.

41. Davis, A.M., Teague, S.J., Kleywegt, G.J., *Application and limitations of X-ray crystallographic data in structure-based ligand and drug design*, J. Chem. Inf. Comput. Sci., 42 (2003) 2718-2736.

42. Hann, M.M., Oprea, T.I., *Pursuing the leadlikeness concept in pharmaceutical research*, Curr. Opin. Chem. Biol., 8 (2004**)** 255-263.

43. Wenlock, M.C., Austin, R.P., Barton, P., Davis, A.M., Leeson P.D., *A comparison of physiochemical property profiles of development and marketed oral drugs*, J. Med. Chem., 46 (2003) 1250-1256.

44. Hou, T.J., Xia, K., Zhang, W., Xu, X.J., *ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach*, J. Chem. Inf. Comput. Sci. 44 (2004) 266-275.

45. Ertl, P., Rohde, B., Selzer, P., *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 (2000) 3714-3717.

46. Palm, K., Stenberg, P., Luthman, K., Artursson, P., *Polar molecular surface properties predict the intestinal absorption of drugs in humans*, Pharm. Res., 14 (1997) 568-571.

FIGURE LEGENDS

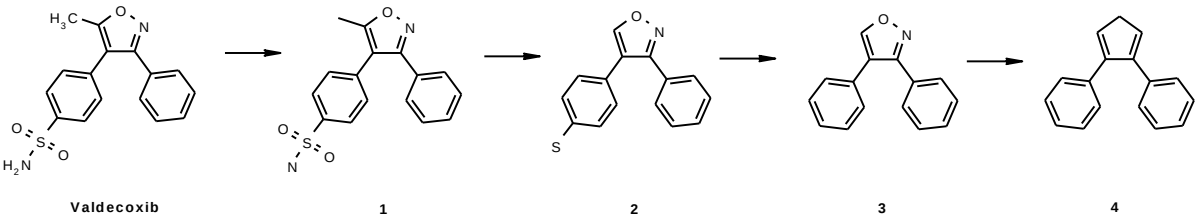

Figure 1. Visualization of our framework algorithm: 1) hydrogens are removed, 2) atoms with only one bond are removed, 3) step 2 is repeated until it only remains atoms with two bonds or more, 4) all atoms' types are set to C, and all bonds' types except aromatic are set to single.

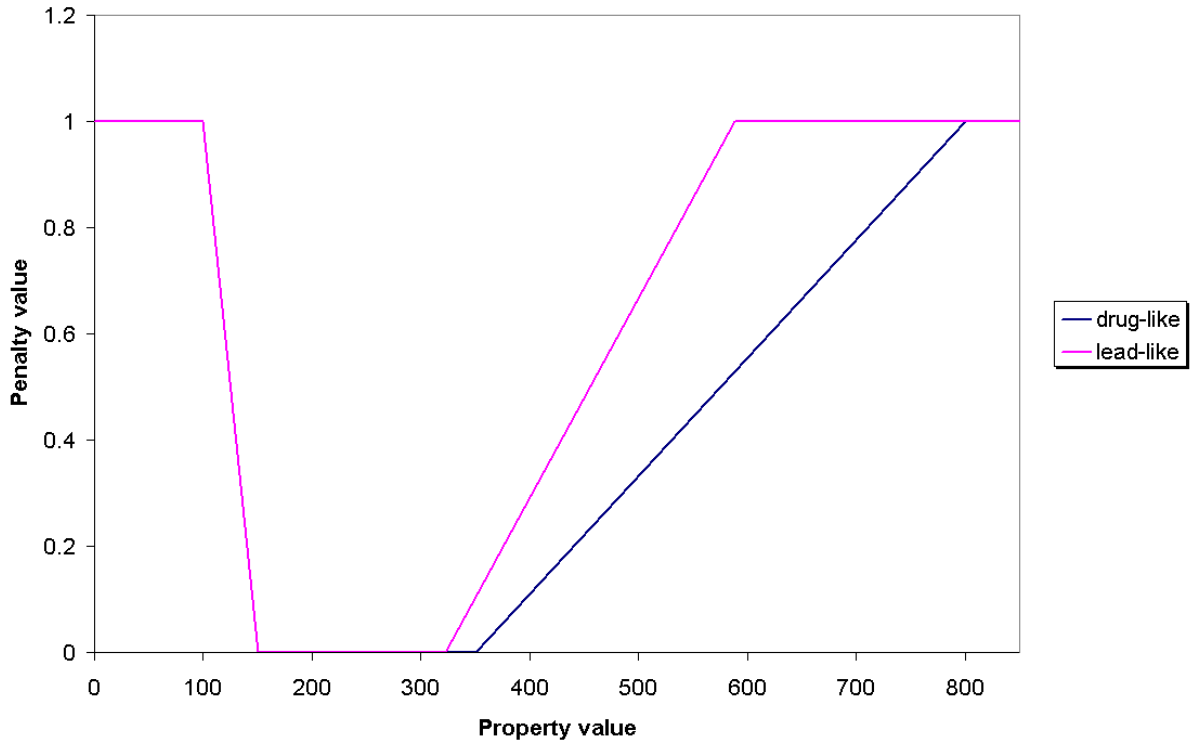

Figure 2. Graphic representation of the 'drug-like' and 'lead-like' penalties functions for MW.

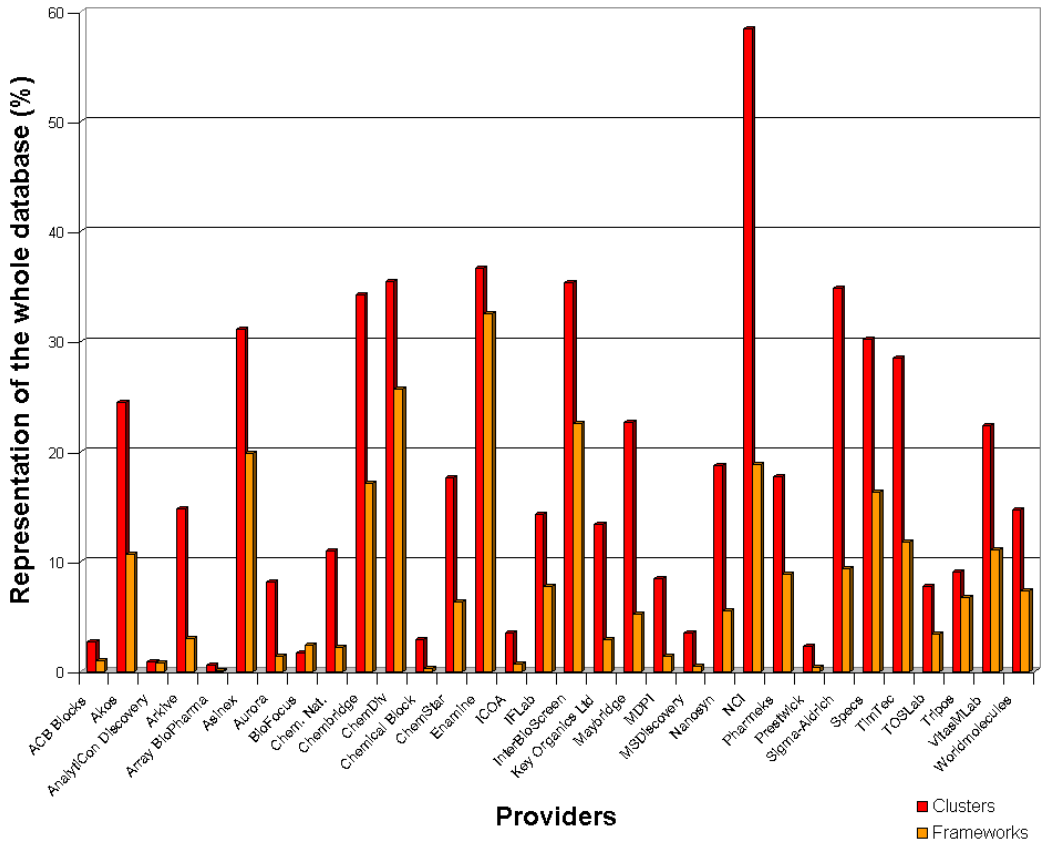

Figure 3. Diversity of each database compared to the diversity of the global database. Diversity is estimated by counting the clusters generated by the SCA algorithm (red) and the frameworks (orange).

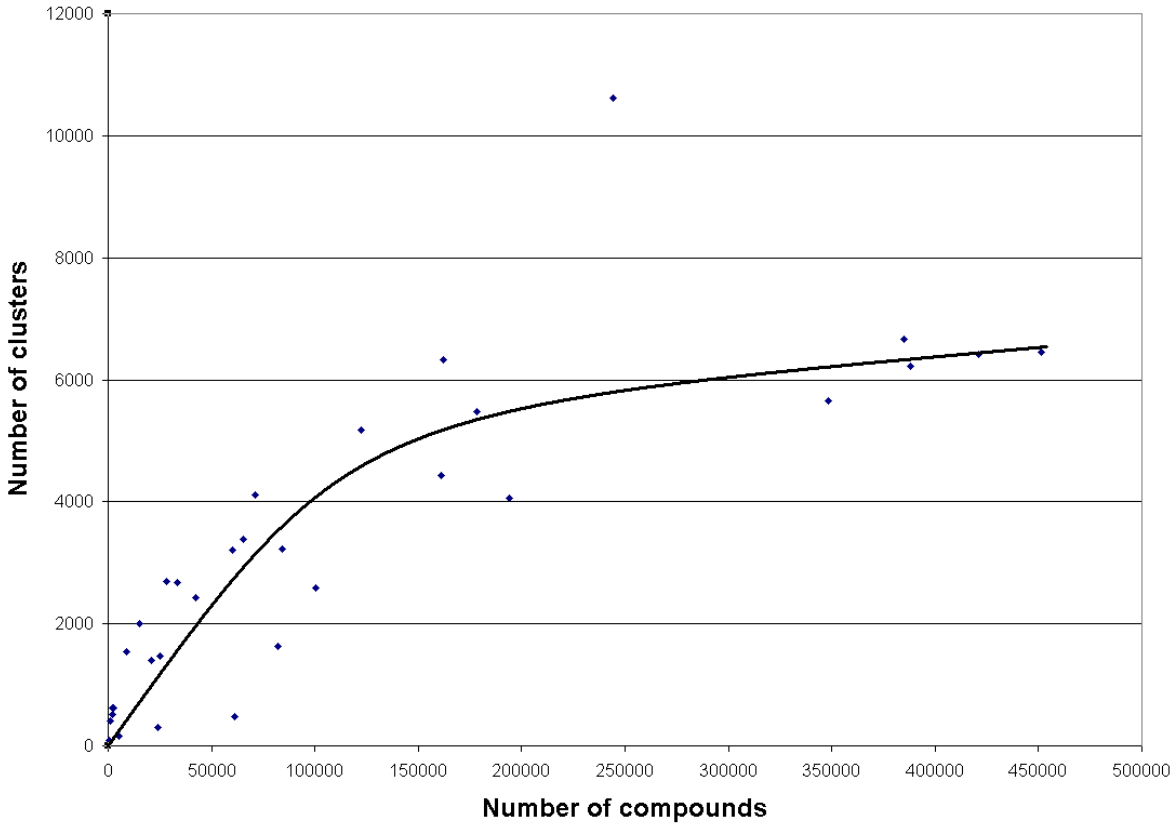

Figure 4. Increase of the diversity with the size of the databases.

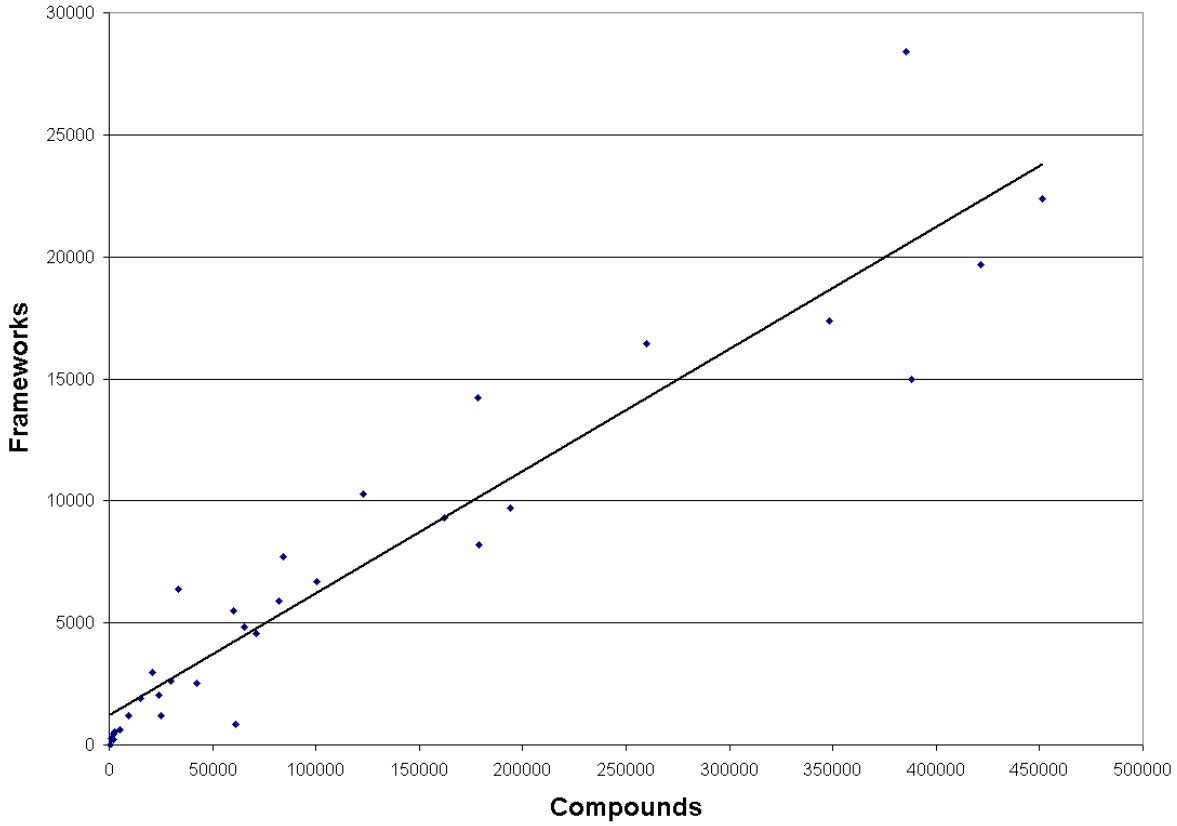

Figure 5. Relation between the number of frameworks and the number of compounds (linear $r^2 = 0.89$).

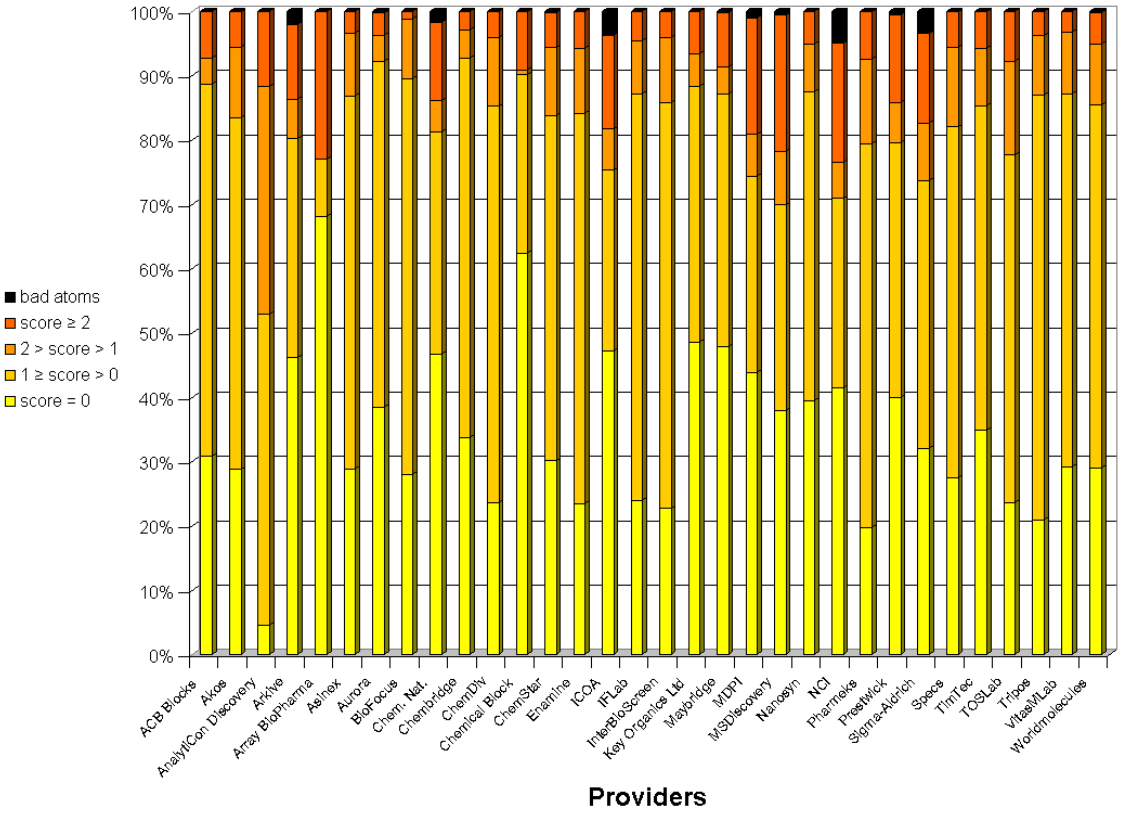

Figure 6. Percentage of ‘drug-like’ scores for each provider’s database.

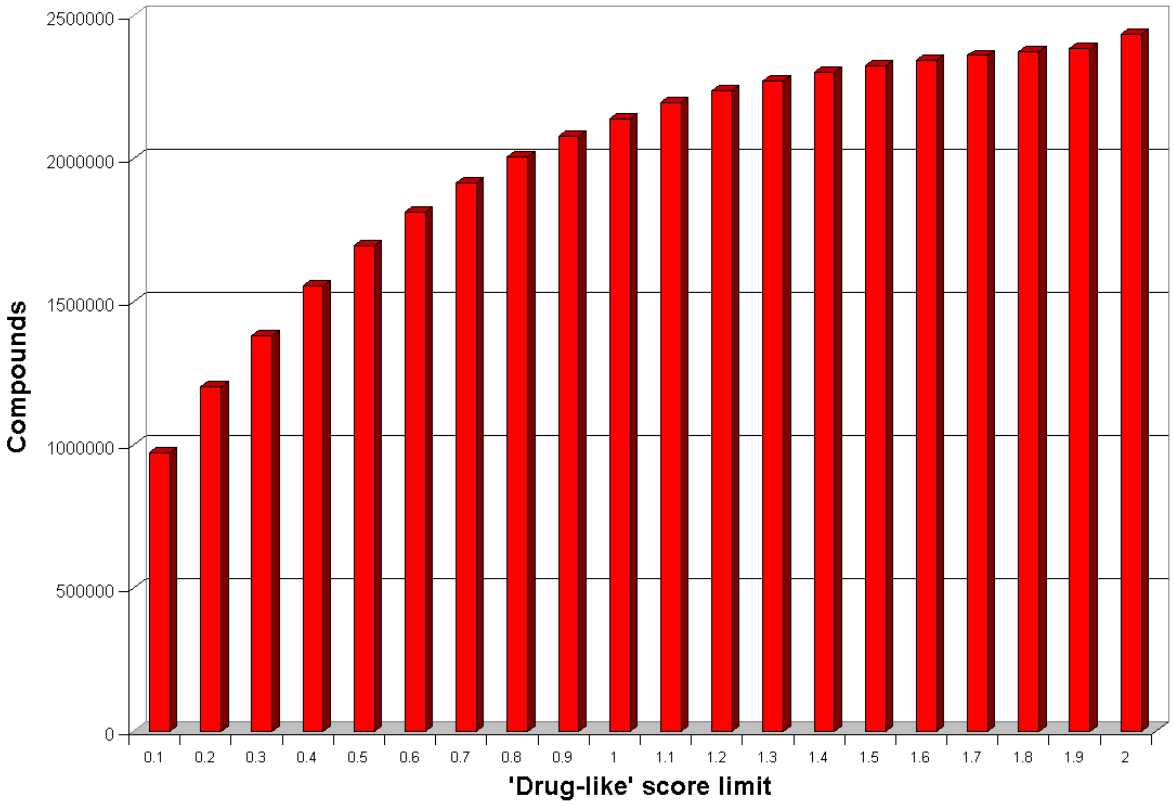

Figure 7. Cumulative ‘drug-like’ scores distribution.

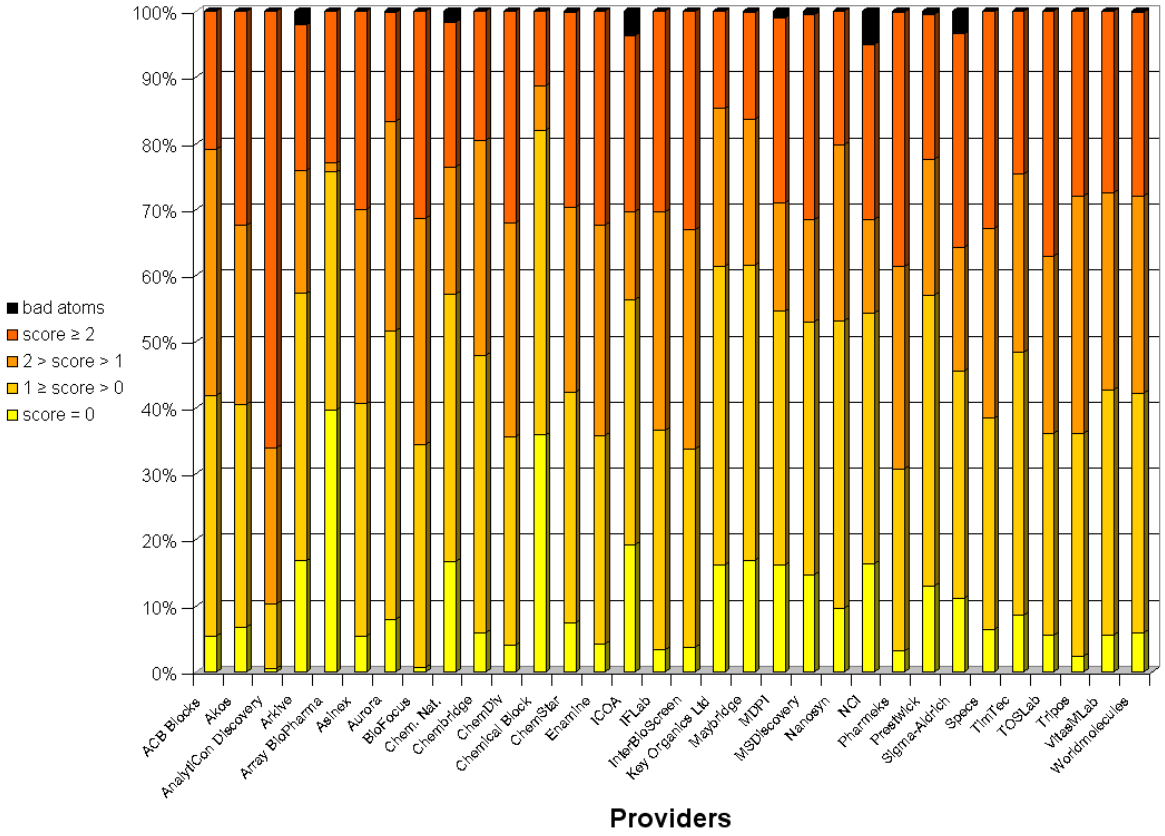

Figure 8. Percentage of 'lead-like' scores for each provider's database.

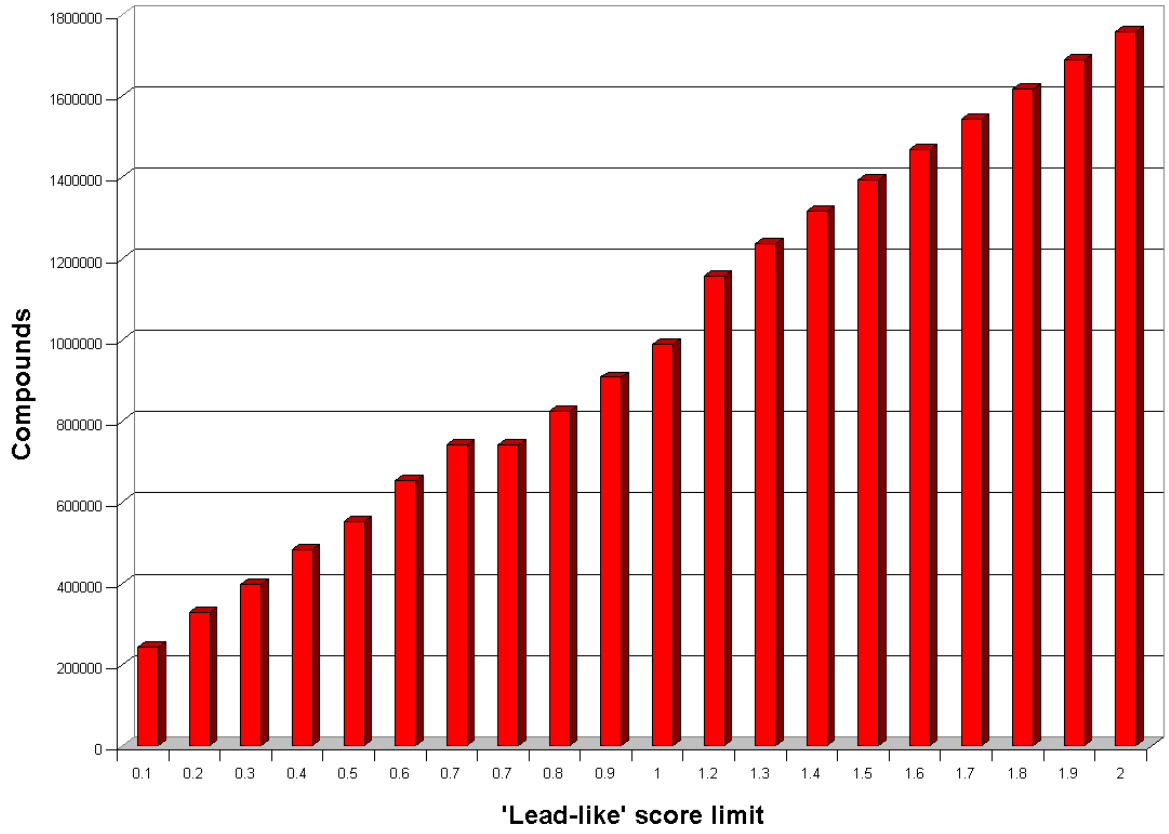

Figure 9. Cumulative ‘lead-like’ score distribution.

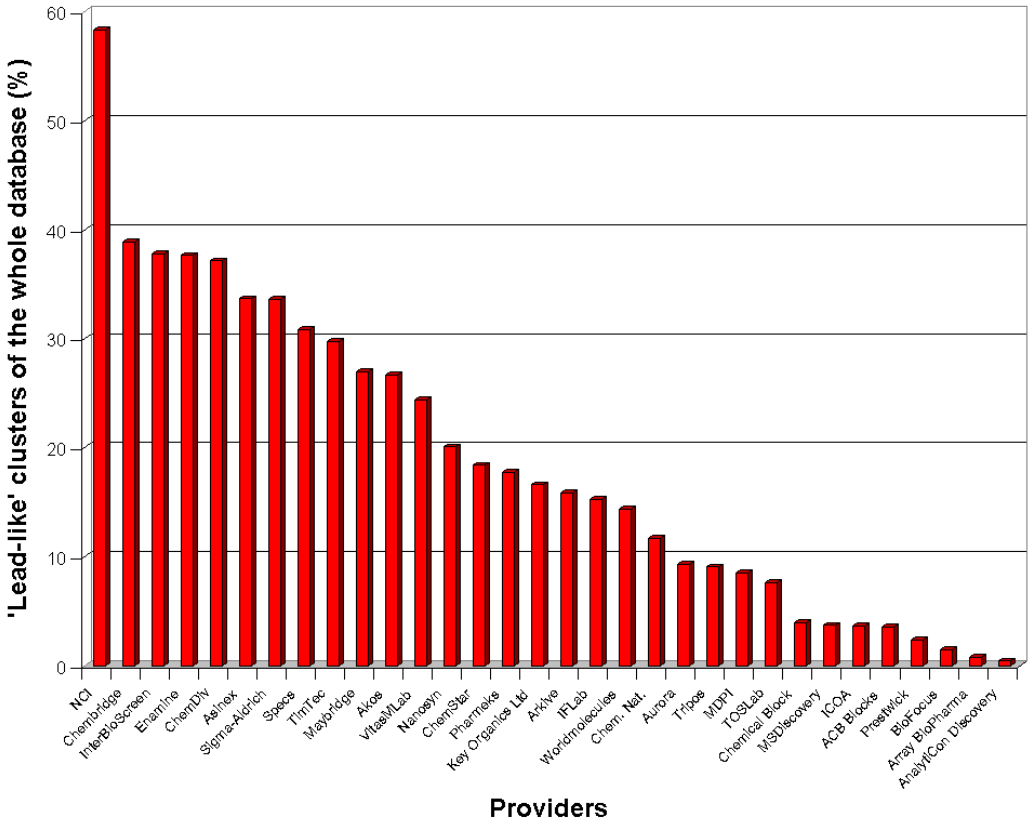

Figure 10. 'Lead-like' space of the whole database covered by each provider.

TABLES

Functionality	InChI	MOE	OEChem	Marvin
Sp3 stereoisomerism	X	X	X	X
Sp2 stereoisomerism	X	X	X	X
Simple tautomerism C(O)-[NH]	X			
Keto-enol tautomerism				
NO2 representation: N(=O)=O and [N+](=O)-[O-]	X	X		
Moveable positive charge detection *	X			

* example taken from InChI documentation:

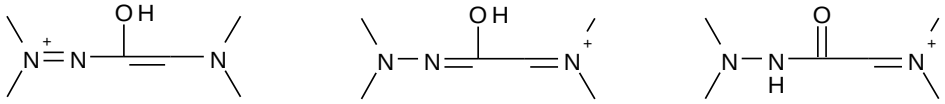

Table 1. Comparison of the unique codes of four softwares.

Provider	Web site	Imported compounds	Origin
ACB Blocks	http://www.acbblocks.com	61 237	Organic synthesis
Akos	http://www.akosgmbh.com	161 316	Organic synthesis
AnalytiCon Discovery	http://www.ac-discovery.com	5 438	Pure and semi-synthetic natural products
Arkive	http://ark.chem.ufl.edu/pages/arkive.htm	28 504	Organic synthesis
Array BioPharma	http://www.arraybiopharma.com	517	Organic synthesis focused primarily in cancer and inflammatory disease
Asinex	http://www.asinex.com	348 203	Organic synthesis
Aurora	http://www.aurora-feinchemie.com	25 295	Organic synthesis
BioFocus	http://www.biofocus.com	23 836	Organic synthesis focused primarily in kinase, GPCR and ion channel.
Chembridge	http://www.chembridge.com	387 859	Organic synthesis
ChemDiv	http://www.chemdiv.com	451 205	Organic synthesis and combinatorial chemistry
Chemical Block	http://www.chemical-block.com	1 993	Organic synthesis
ChemStar	http://www.chemstar.ru	60 066	Organic synthesis
Chem. Nat.	http://chimiotheque-nationale.enscm.fr/	14 946	Organic synthesis and natural products
Enamine	http://www.enamine.relc.com	385 175	Organic synthesis
ICOA	http://www.univ-orleans.fr/icoa/	2 811	Organic synthesis
IFLab	http://www.iflab.kiev.ua	100 392	Organic synthesis
InterBioScreen	http://www.ibscreen.com	421 058	Organic synthesis
Key Organics Ltd	http://www.keyorganics.ltd.uk	42 414	Organic synthesis
Maybridge	http://www.maybridge.com	71 041	Organic synthesis
MDPI	http://www.mdpi.org	8 853	Organic synthesis
MSDiscovery	http://www.msdiscovery.com	1 982	Known drugs, experimental bioactives, and pure natural products
Nanosyn	http://www.nanosyn.com	65 184	Organic synthesis
NCI	http://dtp.nci.nih.gov	244 321	Organic synthesis and natural products focused in anticancer and anti-AIDS
Pharmeks	http://www.pharmeks.com	83 992	Organic Synthesis
Prestwick	http://www.prestwickchemical.com	876	Marketed drugs and others
Sigma-Aldrich	http://www.sigma-aldrich.com	162 171	Organic synthesis and natural products
Specs	http://www.specs.net	178 492	Organic synthesis and natural products
TimTec	http://www.timtec.net	122 238	Organic synthesis and natural products
TOSLab	http://www.toslab.com	21 004	Organic synthesis

			and semi-natural compounds
Tripos	http://leadquest.tripos.com	82 370	Organic synthesis
VitasMLab	http://www.vitasmlab.com/	193 993	Organic synthesis
Worldmolecules	http://www.worldmolecules.com/	33 259	Organic synthesis

Table 2. List of providers with number of molecules present in the database (gathered before March 2005, internal duplicates are not evaluated).

	‘Drug-like’ penalties	‘Lead-like’ penalties
HBD	≤ 3.5: 0 > 3.5 and < 6.5: 0.3333 * P – 1.1667 ≥ 6.5: 1	-
HBA	≤ 7: 0 > 7 and < 13: 0.1667 * P – 1.1667 ≥ 13: 1	≤ 6.3: 0 > 6.3 and < 11.7: 0.1852 * P – 1.1667 ≥ 11.7: 1
Rotatable bonds	≤ 10.5: 0 > 10.5 and < 19.5: 0.1111 * P – 1.1667 ≥ 18.5: 1	≤ 7: 0 > 7 and < 13: 0.1667 * P – 1.1667 ≥ 13: 1
Number of SSSR	≤ 4.2: 0 > 4.2 and < 7.8: 0.2778 * P – 1.1667 ≥ 7.8: 1	≤ 2.8: 0 > 2.8 and < 5.2: 0.4167 * P – 1.1667 ≥ 5.2: 1
Maximum ring size	≤ 6: 0 > 6 and < 9.1: 0.3226 * P – 1.9355 ≥ 9.1: 1	-
Number of halogens	≤ 4.9: 0 > 4.9 and < 9.1: 0.2381 * P – 1.1667 ≥ 9.1: 1	-
MW	≤ 100: 1 > 100 and < 150: -0.02 * P + 3 ≥ 150 and ≤ 350: 0 > 350 and < 800: 0.0022 * P – 0.7778 ≥ 800: 1	≤ 100: 1 > 100 and < 150: -0.02 * P + 3 ≥ 150 and ≤ 322: 0 > 322 and < 588: 0.0038 * P – 1.2105 ≥ 588: 1
LogP	≤ -5: 1 > -5 and < -1.5: -0.2857 * P – 0.4286 ≥ -1.5 and ≤ 4.5: 0 > 4.5 and < 7.5: 0.3333 * P – 1.5 ≥ 7.5: 1	≤ -5: 1 > -5 and < -1.5: -0.2857 * P – 0.4286 ≥ -1.5 and ≤ 2.94: 0 > 2.94 and < 5.46: 0.3968 * P – 1.667 ≥ 5.46: 1

P is the considered property; - means that the ‘lead-like’ penalty is equal to the ‘drug-like’ penalty.

Table 3. Functions used in ‘drug-like’ and ‘lead-like’ scores.

Provider	Comp[a]	Dup[b] (%)	Uniq[c] (%)	Clusters[d] (%)	Number of clusters	Fw[e] (%)	Number of fw[f]	'Lead-like' clusters[g] (%)
ACB Blocks	61 237	0.0	97.9	2.6	480	0.9	829	3.6
Akos	161 316	0.6	19.5	24.4	4 437	10.6	9 306	26.7
AnalytiCon Discovery	5 438	0.0	100.0	0.9	164	0.7	633	0.5
Arkive	28 504	3.6	72.3	14.8	2 691	3.0	2 597	15.9
Array BioPharma	517	0.2	79.3	0.5	94	0.0	15	0.8
Asinex	348 203	0.0	51.1	31.1	5 659	19.9	17 371	33.7
Aurora	25 295	0.2	21.4	8.1	1 477	1.4	1 213	9.3
BioFocus	23 836	0.0	100.0	1.7	306	2.3	2 039	1.5
Chem. Nat.	14 946	2.3	87.3	11.0	1 997	2.2	1 886	11.7
Chembridge	387 859	0.0	36.6	34.3	6 224	17.1	14 956	38.9
ChemDiv	451 205	0.0	38.3	35.5	6 447	25.6	22 400	37.1
Chemical Block	1 993	0.5	22.6	2.9	520	0.2	213	3.9
ChemStar	60 066	0.3	32.3	17.6	3 206	6.3	5 496	18.4
Enamine	385 175	0.1	84.6	36.7	6 663	32.5	28 423	37.6
ICOA	2 811	1.3		3.4	626	0.6	539	3.7
IFLab	100 392	0.1	31.7	14.3	2 593	7.7	6 704	15.3
InterBioScreen	421 058	0.1	56.2	35.3	6 418	22.5	19 684	37.8
Key Organics Ltd	42 414	0.0	89.2	13.4	2 433	2.9	2 528	16.6
Maybridge	71 041	0.2	77.2	22.6	4 108	5.2	4 572	27.0
MDPI	8 853	5.3	74.2	8.5	1 538	1.4	1 190	8.5
MSDiscovery	1 982	0.9	46.1	3.5	628	0.5	437	3.7
Nanosyn	65 184	0.1	15.8	18.7	3 393	5.5	4 833	20.1
NCI	244 321	6.0	85.1	58.5	10 623	18.8	16 428	58.3
Pharmeks	83 992	0.2	41.7	17.7	3 223	8.8	7 714	17.7
Prestwick	876	0.5	25.6	2.2	406	0.3	288	2.4
Sigma-Aldrich	162 171	9.3	51.0	34.8	6 328	9.4	8 180	33.6
Specs	178 492	0.0	36.2	30.2	5 480	16.2	14 203	30.9
TimTec	122 238	0.4	17.5	28.5	5 172	11.8	10 293	29.8
TOSLab	21 004	0.1	50.2	7.7	1 401	3.4	2 966	7.6
Tripos	82 370	0.0	97.0	9.0	1 639	6.7	5 875	9.1
VitasMLab	193 993	0.1	26.6	22.4	4 064	11.1	9 686	24.4
Worldmolecules	33 259	0.2	27.8	14.7	2 672	7.3	6 372	14.3

[a] Number of compounds. [b] Number of duplicates. [c] Number of unique compounds (not present in any other database), ICOA is not analyzed because it is contain in Chem. Nat.. [d] Coverage of the chemical space of the total database (fingerprint diversity clusters). [e] Percentage of the frameworks of the total database represented. [f] number of frameworks of the database. [g] Coverage of the 'lead-like' space of the total database (fingerprint diversity clusters).

Table 4. Analysis of the 32 databases.

Filters	Accepted compounds
total	2 582 944
‘drug-like’	2 141 031
no reactive functions	2 477 248
rotatable bonds ≤ 15	2 504 385
SlogP ≤ 7	2 510 652
HBA ≤ 10	2 537 713
SSSR ≤ 6	2 563 999
no single chains	2 565 839
O and N atoms ≥ 1	2 569 747
max ring size ≤ 7	2 571 725
100 ≤ MW ≤ 800	2 574 611
HBD ≤ 5	2 575 045
halogens ≤ 7	2 578 167
no perflurinated chain	2 581 301

Table 5. Influence of ‘drug-like’ filters.

Filters	Accepted compounds
total	2 582 944
‘lead-like’	994 154
$-4 \leq \text{SlogP} \leq 4.2$	1 498 794
rotatable bonds ≤ 10	2 038 011
$\text{MW} \leq 460$	2 047 891
$\text{SSSR} \leq 4$	2 241 841
‘drug-like’	2 141 031
$\text{Hba} \leq 9$	2 486 377

Table 6. Influence of ‘lead-like’ filters.