ACADEMIA ROMÂNĂ
Revue Roumaine de Chimie
http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2018, *63*(7-8), 749-757

Dedicated to Professor Bogdan C. Simionescu
on the occasion of his 70th anniversary

CHEMICAL REACTIVITY IN BIOLOGICAL PROMISCUOUS COMPOUNDS

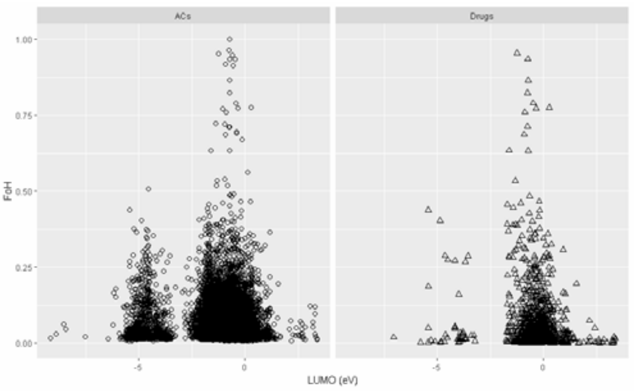
Ramona CURPĂN,[a,*] Sorin AVRAM,[a] Alina BORA[a], Liliana HALIP[a] and Cristian BOLOGA[b]

[a]Institute of Chemistry Timisoara of the Roumanian Academy, No. 24, Mihai Viteazul Av., 300223 Timişoara, Roumania
[b]Department of Internal Medicine, University of New Mexico, Health Sciences Center, Albuquerque NM 87131, NM, USA

Received July 31, 2017

Large scale high throughput screening (HTS) is one of the major sources of new chemical entities for drug discovery. The analysis of HTS outcomes has shown that many of the actives turned out to be dead-ends, false hits, due to their promiscuous biological profile, non-specific and/or apparent activity expressed against different unrelated proteins. For this reason, early labeling of these structures as promiscuous is a challenging task in drug discovery.

Here, we present the extension of our previous works describing the biological promiscuous compounds through chemical reactivity descriptors. We report a comparative analysis between the MLSMR (NIH Molecular Libraries Small Molecule Repository) set and drugs (DrugCentral; http://drugcentral.org/), in terms of reactivity indexes, computed with semiempirical methods AM1, PM3, and PM6, and frequency of hit scores (FoH), encoding different levels of promiscuity based upon assays.



INTRODUCTION

Modern drug discovery heavily depends upon high throughput screening (HTS) or large scale biological testing, as a major source of new chemicals, starting points for lead to drug optimization. In HTS large chemical libraries, hundreds of thousands of compounds are screened against one or multiple biological targets in short periods of time, days to weeks, to identify new biologically active compounds, hits. In this way, large and diverse parts of the chemical space are rapidly evaluated by biologically relevant assays and new classes of chemical compounds addressing different targets are discovered[1]. However, the retrospective analysis of the actives or hit lists originating for multiple unrelated HTS campaigns has revealed the existence of problematic classes of structures that have not been confirmed in subsequent follow-ups.[2,3] The recorded biological activity is only apparent as a consequence of the physico-chemical properties of the compounds: they interfere with the assay detection signal, *e.g.* fluorescent compounds, or interact with assay components, *e.g.* inhibitors of reporter enzyme, luciferase.[4] Depending on the chemical structure and assay conditions, small-molecules can form colloidal aggregates which sequester the protein-target leading to nonspecific inhibition.[5,6]

Furthermore, compounds can irreversibly (covalently) bind many biological targets which

[*] Corresponding author: ramo.curpan@gmail.ro
[**] Supplementary Information on http://web.icf.ro/rrch/ or http://revroum.lew.ro

might be undesirable in future drugs.[4,7] Much effort has been dedicated to the development of different methodologies to characterize and predict problematic compounds in HTS.[8-10]

In this respect, we have previously described biological promiscuous compounds through chemical reactivity descriptors[11] and the influence of the environment over the computed reactivity descriptors.[12] In this paper, we are presenting a comparative analysis between the MLSMR (NIH Molecular Libraries Small Molecule Repository) dataset, specially designed for HTS, and drugs available through DrugCentral (http://drugcentral.org/), in terms of reactivity indexes, computed with semiempirical methods AM1, PM3, and PM6, and frequency of hit scores (FoH), encoding different levels of promiscuity.

METHODS

1. Data sets preparation

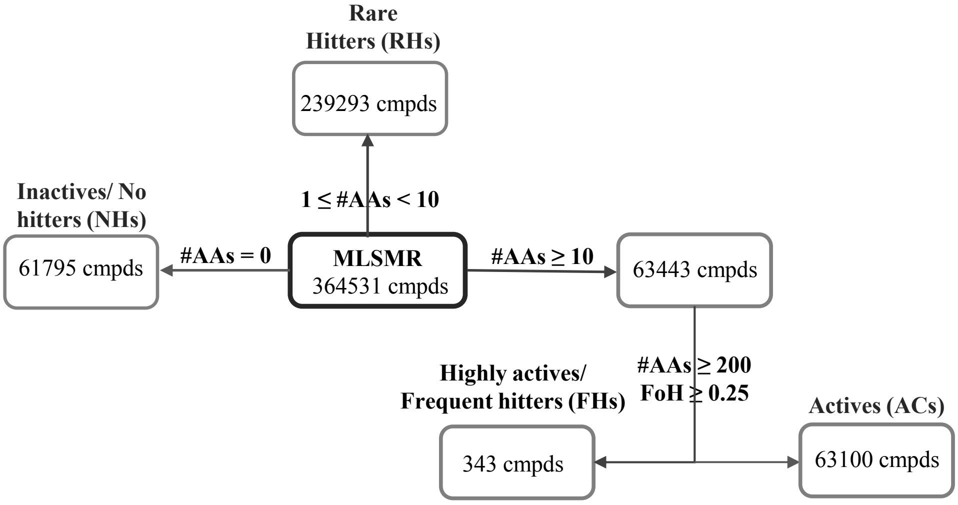
The PubChem Bioassay database[13] is the largest public repository of chemical structures and associated biological activity data, mainly originated from HTS campaigns. It comprises more than 2,350,000 compounds tested in over 1,250,000 assays covering 10,340 biological targets. The MLSMR set, specially designed for HTS, comprising 406,090 substances, was downloaded from PubChem Compound database.

We used for this study a set of drugs extracted from DrugCentral, an open-access online drug database which integrates chemical structures, comprehensive bioactivity data, regulatory, pharmacologic actions and indications for active pharmaceutical ingredients approved by FDA and other regulatory agencies.[14] The current version of DrugCentral contains 93,084 pharmaceuticals and 4,509 drugs. Out of these a list of 3,845 small-molecule drugs were extracted. All chemical structures were standardized by removing salts, and disconnected fragments, followed by 3D geometry generation using ChemAxon JChem API package.[15]

2. Frequency of hits scores (FoH)

Frequencies of hit scores (FoH) were computed to evaluate the degree of promiscuity of the MLSMR compounds. Shortly, the HTS assays with specified target, in which compounds of the MLSMR set were tested, were extracted from PubChem Bioassay. This led to a subset of 364,531 MLSMR compounds with activity outcomes. Subsequently, this set was classified based on the number of assays in which a compound was found active (#AAs) as summarized in Scheme 1, in order to identify the extreme points on the biological activity scale, highly actives or frequent hitters (FHs) and completely inactive structures or no hitters (NHs), respectively. In between there are commonly active compounds and rare hitters (RHs).

Simple FoH scores were computed for each compound having a positive biological activity outcome (AA ≥ 1), as the ratio between the number of assays in which a compound was declared active relative to the number of assays in which it was evaluated (eq. 1).



Scheme 1 – Subsets of active and inactive compounds.

$$FoH = \frac{AA}{TA} \quad \text{(eq. 1)}$$

Data analysis and graphs generation were performed with statistical software package R.

3. Descriptors calculations

Geometry optimizations for all standardized structures were performed with AM1, PM3 and PM6 semiempiric methods as implemented in MOPAC2012,[16,17] using the default settings. The values of LUMO and HOMO electronic descriptors were extracted from the outputs of MOPAC2012 using an in-house script. The following chemical reactivity descriptors were computed based on Koopmans' theorem using HOMO and LUMO energies: ionization potential (IP), electron affinity (EA), electronegativity (χ), hardness (η), softness (S) and electrophilicity index (ω). The detailed description and equations used in calculations are presented in references.[11,12]

RESULTS AND DISCUSSION

Datasets biological profile

The MLSMR set was extracted from PubChem Compounds database. It comprises 406,090 substances corresponding to 390,697 unique compounds, trackable with the Compound Identifier (CID). The biological data associated with the prepared MLSMR set were extracted from PubChem Bioassay. Confirmatory and primary assays evaluated in HTS and having specified a protein target were considered. Setting these criteria, a list of 7,006 biological tests was compiled. Compounds with attributes "inconclusive" and "unspecified" were filtered out from the MLSMR set. Also, compounds showing ambiguous bioactivity data, *e.g.* designated as "active" and "inactive" in the same assay, were removed. Finally, 364,531 qualifying compounds were identified.

To identify potentially promiscuous compounds, the prepared MLSMR set was analyzed and classified based on the number of assays in which the compounds were annotated as actives (as shown in Scheme 1). Accordingly, a subset of 61,795 compounds was found to be inactive in all assays and denominated no hitters (NHs), 239,932 active compounds or rare hitters (RHs) represents compounds annotated active in a range of one up to 9 assays and 63,443 compounds were identified to be active in more than 10 assays representing the dataset of active compounds. Approximately 25% of the hits of the latter subset were declared active in more than 25 assays and tested in more than 675 assays, hence providing a solid basis to identify promiscuous compounds or frequent hitters (FHs).

The NHs subset comprises 61,795 compounds declared inactive in all assays, although they have been tested in hundreds of assays. Roughly 70% of the molecules were screened in more than 400 assays, showing that NHs were extensively tested (Fig. 1A), but they are consistently inactive structures despite the very large number of assays in which they were screened.

The largest set of molecules, RHs, was assembled with compounds that were found active in less than 10 assays. Overall, these compounds were evaluated in slightly larger number of assays than NHs. Thus, almost 85% of them were screened in more than 400 assays (Fig. 1B) but were found active in less than 10 assays.

Finally, the remaining 63,443 compounds were denominated the actives set because these compounds show consistent biological activity. Thus, almost 91% of them were evaluated in the range of 400 up to 900 assays (Fig. 1C) and ~93% of the compounds are found active in 10 up to 50 assays (Fig. 2B). On average, these compounds were tested in 604 assays per compound, in a total of 140 to 2,414 assays and reported to be active with the median frequency of 16 assays per compound. Hence, this is the most tested and active set of compounds, being suitable for the identification of promiscuous compounds or frequent hitters (FHs). Detailed statistics of these datasets are presented in Table S1 of Supporting Material.

Next, we identified the FHs by extracting the extreme activity points of the actives set, compounds declared actives in more than 200 assays and having a FoH $\geq$ 0.25 (Scheme 1). A subset of 343 complying molecules was compiled. Out of these, 242 compounds were evaluated in the range of 250 up to 1,000 assays (Fig. 1D) and 191 compounds were declared active in a total of 50-200 assays (Fig. 2C). Consequently, we have identified a small set of extensively tested and highly active compounds, promiscuous structures or FHs that we have used to analyze comparatively with the NHs, RHs and drugs.

Chemical reactivity descriptors

The chemical space of MLSMR-based sets and drugs has been characterized by global chemical reactivity descriptors computed using different semiempirical methods: AM1, PM3 and PM6.

To calculate global reactivity indexes the values of HOMO and LUMO descriptors were extracted from MOPAC output files (for statistics, see Table S2 of Supporting material). Pairwise comparisons were performed for the LUMO descriptor.

The results indicate high correlations of the LUMO values obtained via AM1, PM3 and PM6 methods for MLSMR-based sets (Fig. 3A) and drugs (Fig. 3B), with Pearson correlations ≥ 0.85 for all pairs. These results demonstrate that for large and diverse sets, *i.e.*, MLSMR chemical library, the semiempirical methods AM1, PM3 and PM6 produce similar results and can be used equally well for high-capacity calculations. Based on these data, the HOMO and LUMO descriptors computed with AM1 method were used to calculate the reactivity indexes: ionization potential (IP), electron affinity (EA), electronegativity (χ), hardness (η), softness (S) and electrophilicity index (ω).

Comparisons between MLSMR-based sets and drugs

In this study we identified sets of compounds showing extreme biological profile, *i.e.* FHs compounds declared actives in tens to hundreds of assays, *versus* NHs completely inactive structures, no biological activity detected in hundreds of assays, and we were interested to see if this biological profile is related with the chemical profile of the molecules described by several global chemical reactivity descriptors (see Materials and Methods section).

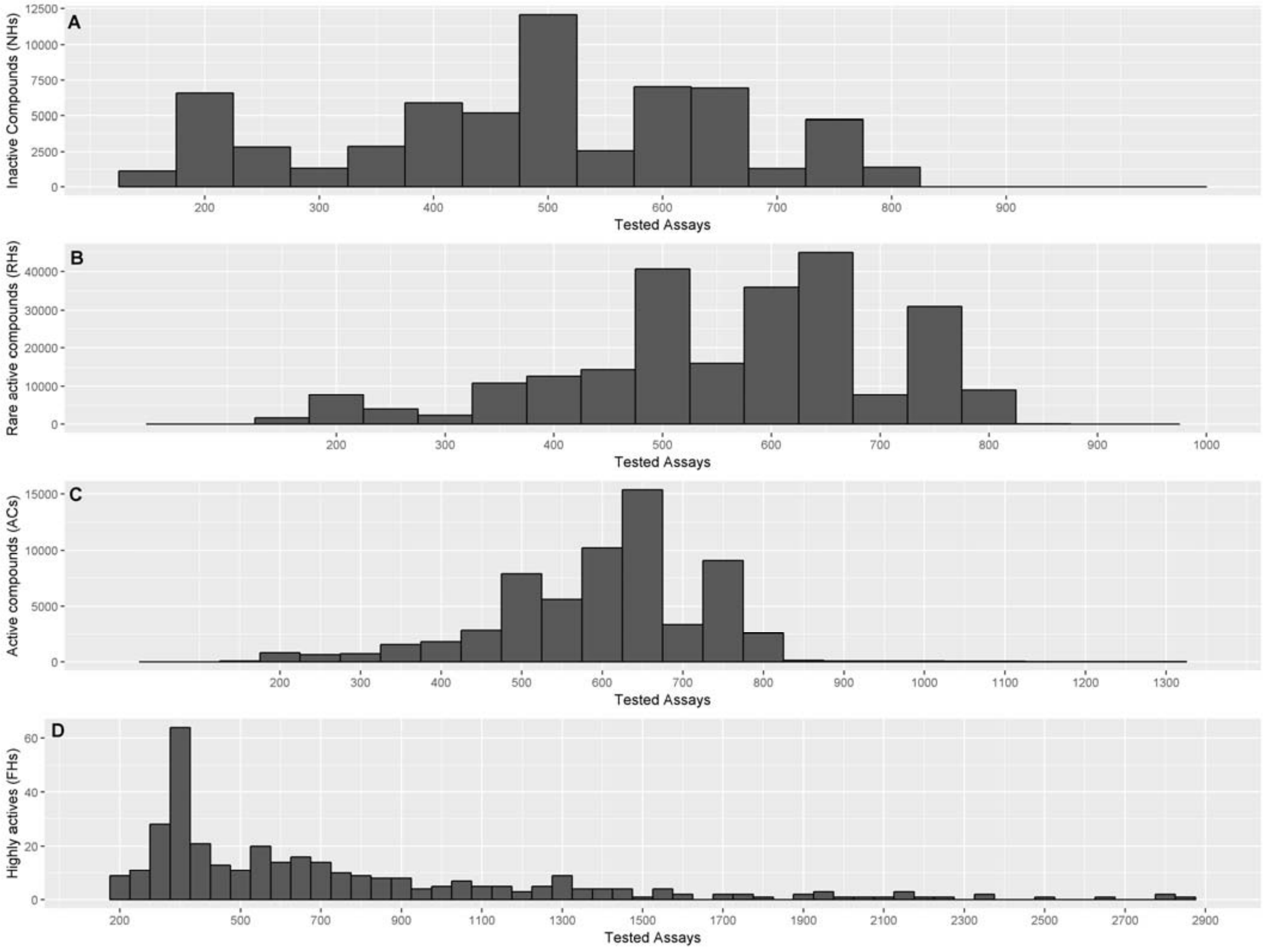


Fig. 1 – Tested assays frequency. The distribution of compounds in tested assays for the following sets: (A) inactives (NHs), (B) rare actives (RHs), (C) actives (ACs) and (D) frequent hitters (FHs).

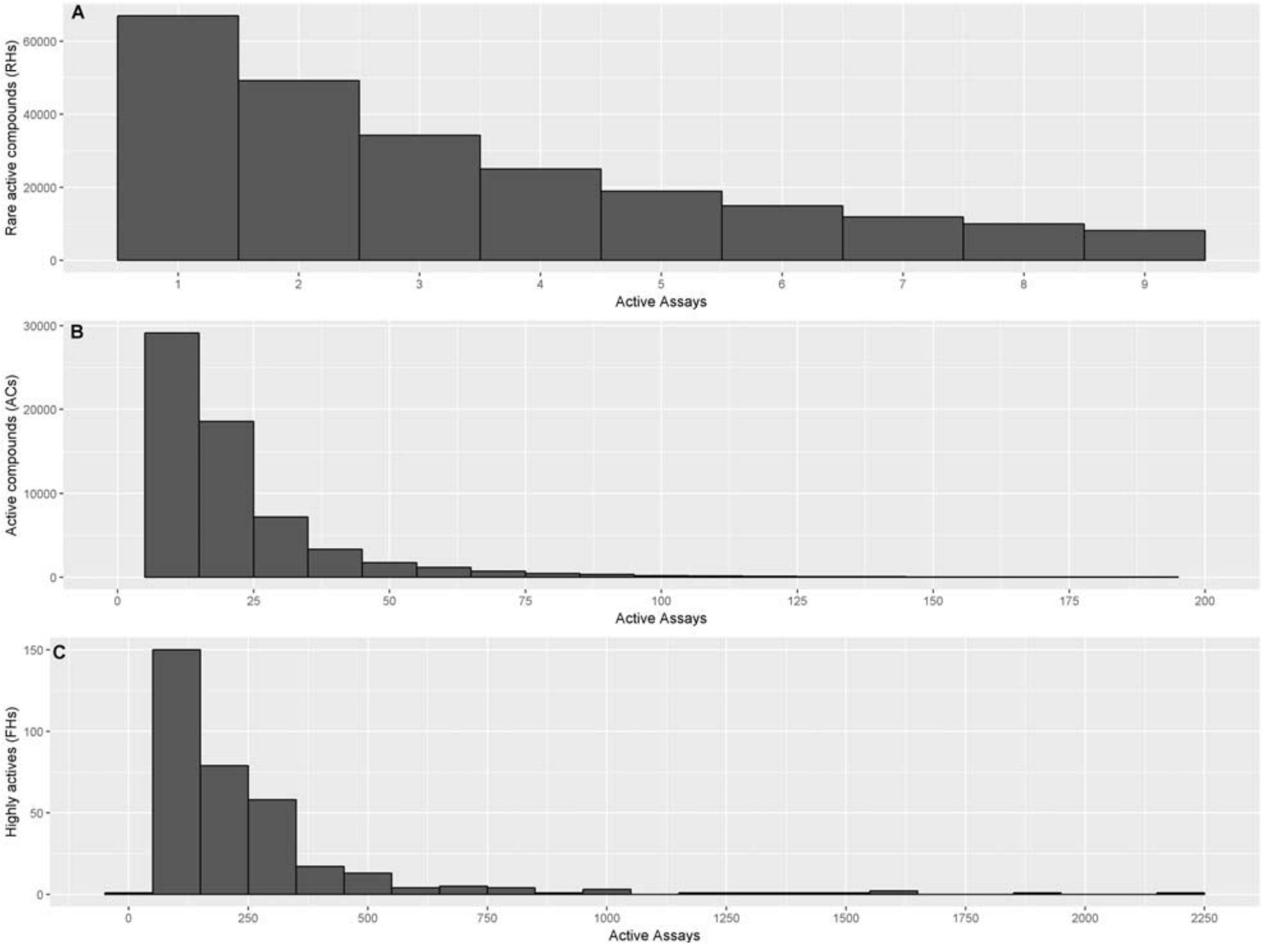


Fig. 2 – Active assays frequency. The distribution of compounds declared active in assays for the following sets: (A) rare actives (RHs), (B) actives (ACs) and (C) frequent hitters (FHs).

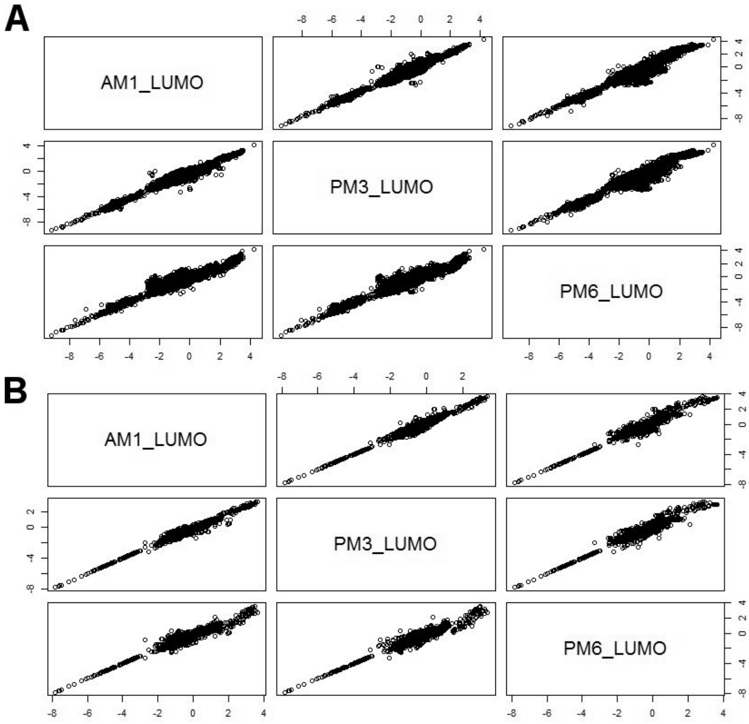


Fig. 3 – Correlation of LUMO values (eV) calculated with different semiempirical methods for MLSMR set (A) and drugs (B).

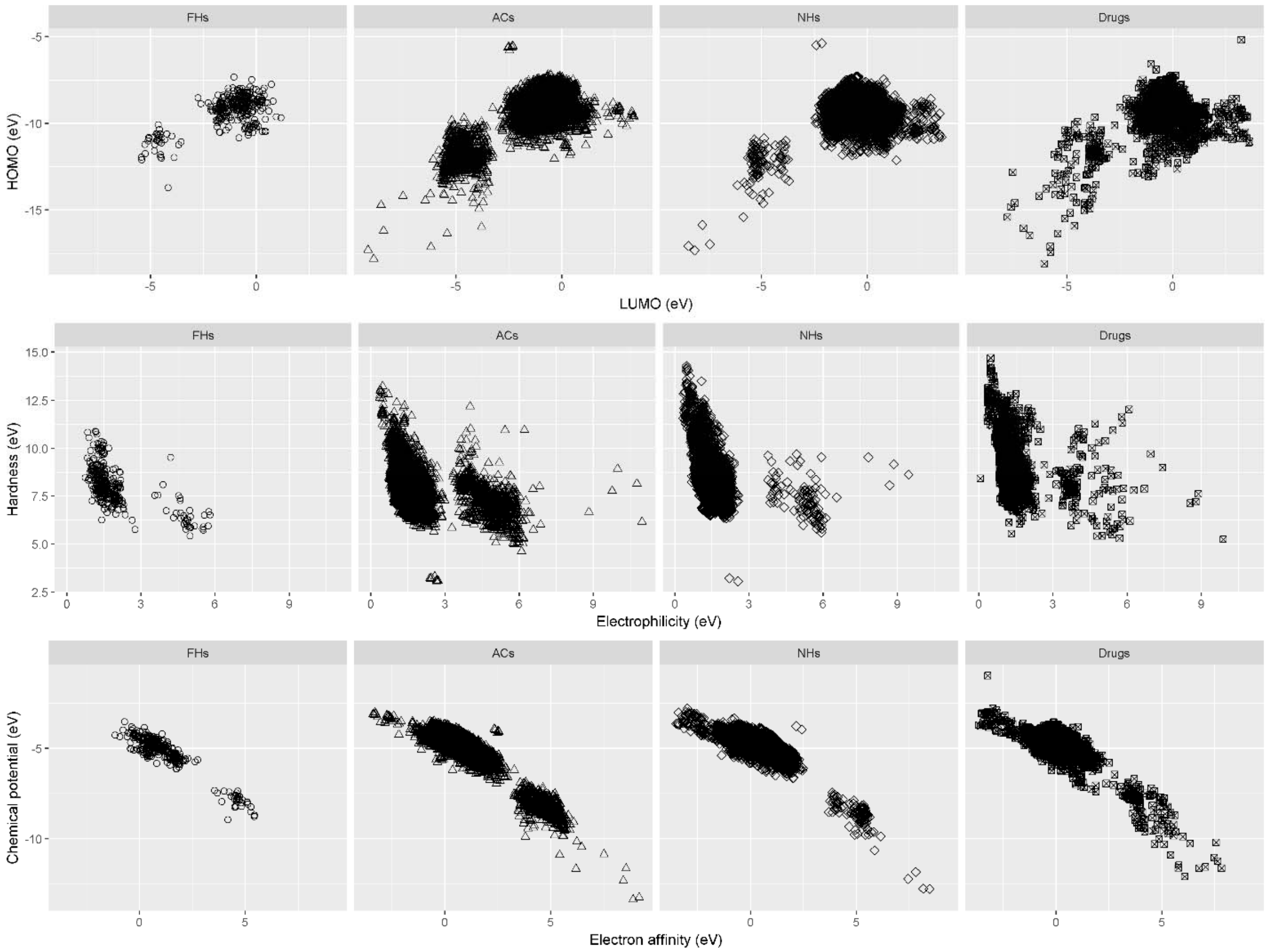


Fig. 4 – Dependencies between chemical reactivity descriptors presented comparatively for MLSMR-based datasets and drugs.

In this respect, we performed one-to-one comparisons between the chemical reactivity descriptors calculated at AM1 semiempirical level for MLSMR-based sets, *i.e.* FHs, ACs and NHs, and drugs, a high-confidence dataset in terms of the biological activities. Selected results for several descriptors, LUMO, HOMO, hardness and electrophilicity, are shown in Fig. 4, for the remaining descriptors the results are presented in Supporting Materials (Figs. S1 and S2).

A careful analysis of the data has shown no significant differences between chemical reactivity profiles of the investigated sets. Similar values have been computed for the selected descriptors, irrespective of the set, ranging in the same intervals, depending on the size of the set. However, a pattern has been noticed for LUMO and LUMO-based descriptors, *i.e.* electron affinity, electrophilicity, electronegativity and chemical potential. In all cases, the compounds are clearly split in two distinct groups around a particular value depending on descriptor (see Fig. 4 and Fig. S1 of SM). Specifically, the FHs, ACs, NHs sets are divided at roughly -3eV, 3eV and -6.8 eV for LUMO, electrophilicity and chemical potential, respectively. At this moment, we do not have a clear explanation of the observed patterns, more investigations are needed to shed light on this aspect.

Next, we proposed to compare the set of actives, ACs, to which the FHs were added, with the 3,845 drug molecules via correlations between the FoH scores and chemical reactivity descriptors.

FoH scores were calculated for each compound of MLSMR declared active in at least one assay (see Methods section). For the purpose of this study, we used only the sets with significant FoH values, *i.e.* the 63443 compounds declared active in more than 10 assays, the extended ACs set which includes FHs (see Scheme 1). For drugs, we could not calculate FoH scores in a similar manner to those calculated for MLSMR-based sets. Nevertheless, we used the information available from MLSMR, a set of small molecules specially designed from HTS screenings that includes drugs. Accordingly, similarity based calculations were performed between the 3,845 drugs from Drugcentral and MLSMR compounds. We identified 1,194 drugs in the MLSMR subset. Out of these, 26 drugs have FoH = 0 because they were declared inactives in all assays from PubChem and for the remaining 1,164 drugs FoH values range from 0.001 up to 0.95, allowing the identification of promiscuous drugs.

To assess the influence of compounds' chemical reactivity on their promiscuity, we comparatively analyzed the ACs dataset and the 1,164 drugs in terms of FoH scores and computed reactivity descriptors. In Fig. 5 are presented correlations for representative descriptors. Surprisingly, the results are showing no significant correlations between compound's promiscuity quantified by FoH scores and none of the reactivity descriptors (Fig. 5 and S3). As previously observed, the sets are clearly classified in two classes, for LUMO and electrophilicity descriptors, around the values we have previously suggested as promiscuity cutoffs for the descriptors discriminating highly promiscuous compounds.

In Fig. 6 are shown 31 highly promiscuous compounds from ACs and drugs datasets that have FoHs ≥ 5. For each compound the structures are provided together with FoH value, compound identifier from PubChem (CID), generic name (if available) and biological indications. Out of these compounds, 13 are approved drugs, most of them neoplastic agents used in cancer treatment and 6 structures are evaluated in phase I-II clinical trials for similar indications. Of the MLSMR FHs, 13 compounds are extensively tested as kinase inhibitors, a well-known class of promiscuous compounds. Among drugs, 3 structures (marked with * in Fig. 6) were found on a list of 164 promiscuous drugs recently published in a study analyzing drugs promiscuity.[19]

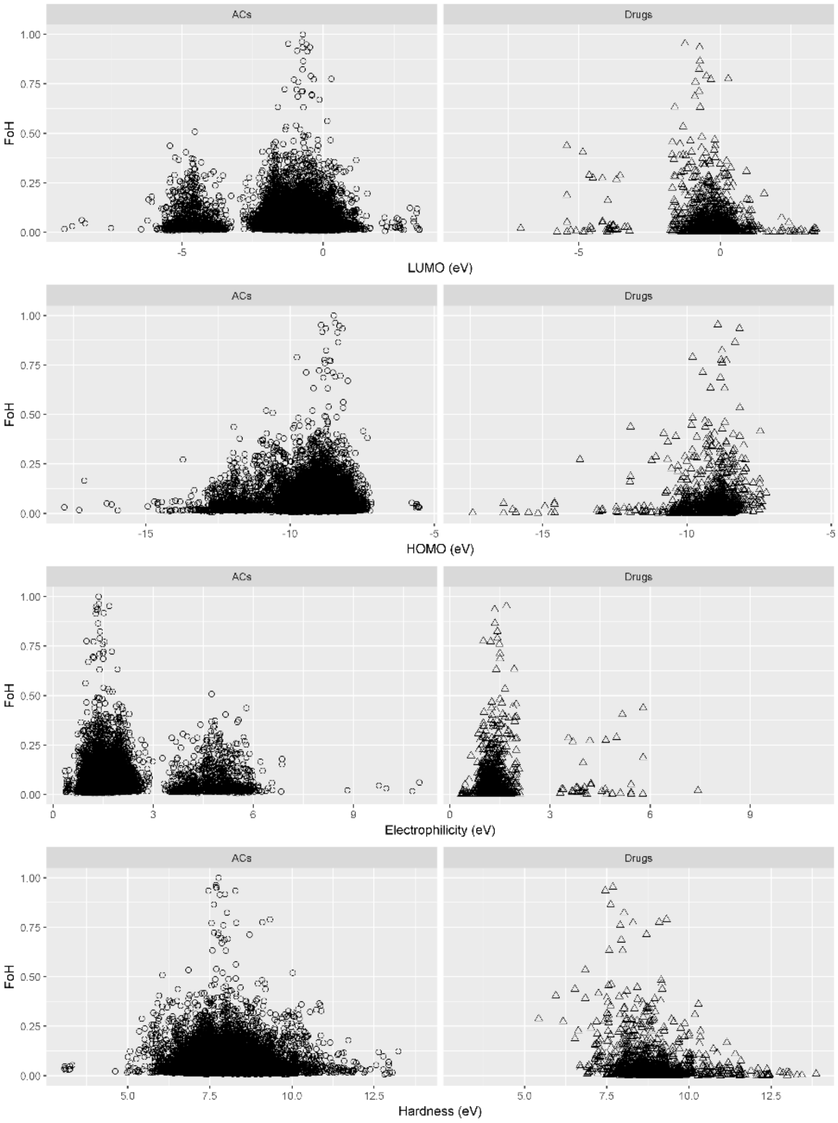


Fig. 5 – Chemical reactivity descriptors plotted against FoH for extended set of ACs and selected drugs with calculated FoH scores.

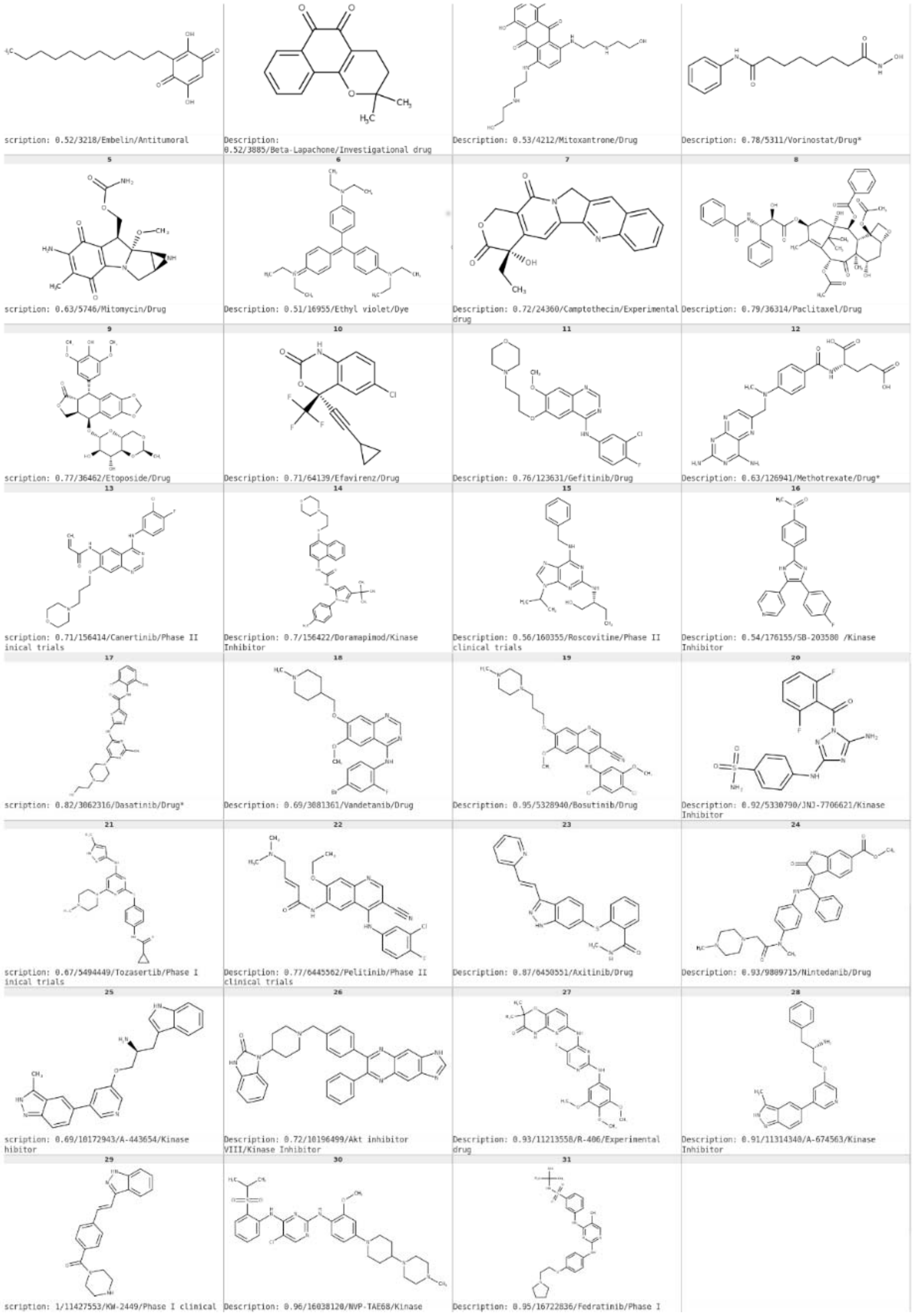

scription: 0.52/3218/Embelin/Antitumoral
Description: 0.52/3885/Beta-Lapachone/Investigational drug
Description: 0.53/4212/Mitoxantrone/Drug
Description: 0.78/5311/Vorinostat/Drug*
5
6
7
8
scription: 0.63/5746/Mitomycin/Drug
Description: 0.51/16955/Ethyl violet/Dye
Description: 0.72/24360/Camptothecin/Experimental drug
Description: 0.79/36314/Paclitaxel/Drug
9
10
11
12
scription: 0.77/36462/Etoposide/Drug
Description: 0.71/64139/Efavirenz/Drug
Description: 0.76/123631/Gefitinib/Drug
Description: 0.63/126941/Methotrexate/Drug*
13
14
15
16
scription: 0.71/156414/Canertinib/Phase II inical trials
Description: 0.7/156422/Doramapimod/Kinase Inhibitor
Description: 0.56/160355/Roscovitine/Phase II clinical trials
Description: 0.54/176155/SB-203580 /Kinase Inhibitor
17
18
19
20
scription: 0.82/3062316/Dasatinib/Drug*
Description: 0.69/3081361/Vandetanib/Drug
Description: 0.95/5328940/Bosutinib/Drug
Description: 0.92/5330790/JNJ-7706621/Kinase Inhibitor
21
22
23
24
scription: 0.67/5494449/Tozasertib/Phase I inical trials
Description: 0.77/6445562/Pelitinib/Phase II clinical trials
Description: 0.87/6450551/Axitinib/Drug
Description: 0.93/9809715/Nintedanib/Drug
25
26
27
28
scription: 0.69/10172943/A-443654/Kinase hibitor
Description: 0.72/10196499/Akt inhibitor VIII/Kinase Inhibitor
Description: 0.93/11213558/R-406/Experimental drug
Description: 0.91/11314340/A-674563/Kinase Inhibitor
29
30
31
scription: 1/11427553/KW-2449/Phase I clinical
Description: 0.96/16038120/NVP-TAE68/Kinase
Description: 0.95/16722836/Fedratinib/Phase I

Fig. 6 – Examples of promiscuous compounds. There are shown 31 structures FoH ≥ 0.5. For each compound the following data are provided: FoH/CID (compound identifier extracted from PubChem)/Generic name/Compound information, with * are marked promiscuous drugs.

Overall, these findings might indicate that biological promiscuity is not significantly influenced by the chemical reactivity of the compounds and it is not an indicative of promiscuity. Our current results suggest that reactive structures have a low impact on the HTS testing and highly promiscuous compounds identified from these screening are originating from other sources of errors. However, these results are limited by the size and characteristics of the datasets investigated in this study. More studies against different datasets are needed to confirm these results.

CONCLUSIONS

In this study, we have performed a comparative analysis of a dataset of small molecules specially designed for HTS, MLSMR from PubChem database and a dataset of drugs from DrugCentral to investigate the impact of chemical reactivity upon biologically promiscuous compounds or frequent hitters.

A total of 364,531 compounds tested in 7,006 HTS assays and 3,845 drugs provided the basis for our analysis. Global reactivity indexes have been calculated to describe the chemical profile of the datasets using different semiempirical methods. Irrespective of the methods used the correlations were high suggesting that regardless of the method of choice the impact upon results is negligible.

Compounds classification, based upon the biological information and FoH scores, measure of promiscuity, has led to the assembly of the working sets: highly actives or promiscuous compounds and consistently inactives or no hitters together with moderate active compounds or rare hitters and actives. Comparisons of the chemical reactivity profiles for these datasets have revealed no significant differences between them, indicating that promiscuous compounds can have chemical reactivity similar to that of NHs or RHs. Correlations between FoH scores and chemical reactivity descriptors have been insignificant suggesting that biological promiscuity is affected by the chemical reactivity of compounds to a lesser extent than expected. These findings are limited by the characteristics and size of the datasets used. More studies, employing large datasets coming from other bioactivity databases, e.g. CHEMBL, are needed to confirm or deny these observations.

Acknowledgements. This work was supported by a grant of the Roumanian National Authority for Scientific Research and Innovation, CNCS−UEFISCDI, project number PN-II-RU-TE-2014-4-0422.

REFERENCES

1. J. Inglese and D. S. Auld, “High Throughput Screening (HTS) Techniques: Applications in Chemical Biology” in “Wiley Encyclopedia of Chemical Biology”, John Wiley & Sons Inc., 2008.
2. N. Thorne, D. S. Auld and J. Inglese, *Curr. Opin. Chem. Biol.*, **2010**, *14*, 315-24.
3. G. M. Rishton, *Drug Discov. Today*, **2003**, *8*, 86-96.
4. A. Jadhav, R. S. Ferreira, C. Klumpp, B. T. Mott, C. P. Austin, J. Inglese, C. J. Thomas, D. J. Maloney, B. K. Shoichet and A. Simeonov, *J. Med. Chem.*, **2010**, *53*, 37-51.
5. K. E. Coan, D. A. Maltby, A. L. Burlingame and B. K. Shoichet, *J. Med. Chem.*, **2009**, *52*, 2067-2075.
6. S. L. McGovern, B. T. Helfand, B. Feng and B. K. Shoichet, *J. Med. Chem.*, **2003**, *46*, 4265-4272.
7. G. M. Rishton, *Drug Discov. Today,* **1997**, *2*, 382-384.
8. J. B. Baell and G. A. Holloway, *J. Med. Chem.,* **2010**, *53*, 2719-2740.
9. N. Y. Mok, S. Maxe and R. Brenk, *J. Chem. Inf. Model.*, **2013**, *53*, 534-544.
10. J. Che, F. J. King, B. Zhou and Y. Zhou, *J. Chem. Inf. Model.,* **2012**, *52*, 913-926.
11. R. Curpăn, S. Avram, R. Vianello and C. Bologa, *Bioorg. Med. Chem.*, **2014**, 22, 2461-2468.
12. R. Curpăn, S. Avram, L. Halip and C. Bologa, *Rev. Roum. Chim.*, **2015**, *60*, 205-211.
13. The Pubchem Project, http://pubchem.ncbi.nlm.nih.gov/; Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant, *Nucleic Acids Res.*, **2012**, *40*, D400-D412.
14. DrugCentral, http://drugcentral.org; O. Ursu, J. Holmes, J. Knockel, C. G. Bologa, J. J. Yang, S. L. Mathias, S. J. Nelson and T. I. Oprea, **2017**, *45*, D932–D939.
15. J. Chem. API package, version 6.1.0., Chemaxon, 2013; http://www.chemaxon.com.
16. MOPAC2012, J.J.P. Stewart, Stewart Computational Chemistry; Colorado Springs, CO, USA.
17. J. J. P. Stewart, *J. Mol. Model.*, **2013**, *19*, 1-32.
18. R. Development Core Team. R: A Language and Environment for 708 Statistical Computing, version 3.1.1; The R Foundation for Statistical 709 Computing: Vienna, Austria, 2014; http://www.R-project.org/.
19. V. J. Haupt, S. Daminelli and M. Schroeder, *PLoS ONE*, **2013**, *8*, e65894, doi:10.1371/journal.pone.0065894.