Towards the Bioassay Activity Landscape Modeling in Compound Databases

José Luis Medina-Franco* and Jacob Waddell

Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, FL 34987, USA. jmedina@tpims.org

Abstract. Public compound databases annotated with biological activity are increasingly being used in drug discovery programs. A prominent example is of such databases is PubChem. Herein, we introduce an approach to systematically characterize the structure-bioassay activity relationships in PubChem using the concept of *bioassay activity landscape*. This strategy is general and can be applied to any data set screened across multiple bioassays. We also present a visual representation of the chemical space of an in-house data set using a recently developed web-based public tool.

Keywords: Chemical space, chemoinformatics, drug discovery, molecular databases, Structure multiple Activity Similarity (SmAS) maps.

Resumen. Programas de investigación de descubrimiento de fármacos utilizan con mayor frecuencia bases de datos moleculares públicas que contienen información de actividad biológica. Un ejemplo de estas bases de datos es PubChem. En este trabajo se presenta una estrategia para caracterizar en forma sistemática relaciones estructura-actividad biológica en bioensayos disponibles en PubChem utilizando el concepto de panorama de actividad en bioensayos. Esta estrategia es general y puede aplicarse a cualquier grupo de compuestos que se han evaluado en bioensayos diversos. También se presenta una representación visual del espacio químico de una colección de compuestos institucional utilizando una herramienta en línea disponible públicamente.

Palabras clave: Bases de datos moleculares, descubrimiento de fármacos, espacio químico, mapas de Similitud Estructura-Actividad Múltiple, quimionformática.

Abbreviations: bAPS, Bioassay activity profile similarity; PCA, principal component analysis; SAR, structure-activity relationships; SmAS, Structure multiple Activity Similarity maps.

Introduction

High-throughput screening (HTS), combinatorial chemistry, parallel synthesis as well as the access to web-based bioinformatic tools have given rise to a new area of data-rich environment for life-science studies dealing with biomolecular targets and molecular ligands [1]. As such, several compound libraries in the public domain are annotated with biological activities and these can be used for structure-activity relationships (SAR) studies for specific targets [2] and to study quantitatively the *polypharmacology* of bioactive compounds [3]. Notable examples of such libraries in the public domain are: PubChem [4], ChEMBL [5], and BindingDB [6]. These databases are particularly useful because, in addition to be openly accessible to academic groups, non-for profit and other research institutions, they contain information of vast collections of chemical compounds that have been screened across multiple biological endpoints.

Mining the information included in annotated chemical databases is not an easy task and there is a continued effort to develop novel methods to retrieve the SAR for molecules tested against single or diverse biological endpoints. For example, PubChem has implemented several chemoinformatic tools to analyze the screening data for individual bioassays. However, a systematic analysis of the bioactivity profile of compounds screened *across several bioassays* is not straightforward. One of the major challenges in PubChem is that there are collections of compounds that were partially screened across a set of biological assays. Therefore, not all compounds in the collection have activity data across all the biological endpoints.

In order to characterize the bioactive profile of data sets in PubChem, herein we *introduce* a novel approach that systematically navigates through the structure-bioassay activity profile relationships of a set of compounds using the principles of 'activity landscape modeling' (vide infra) [7, 8]. To *illustrate* this approach, we used in this work a data set of 618 compounds obtained from an in-house collection and deposited in PubChem. We also report a visual analysis of the chemical space [9] of this data set using a web-based tool.

Results and Discussion

Bioassay activity landscape modeling

The analysis of structure-activity relationships (SAR) of a set of compounds with measured biological activity is a central topic in drug discovery [10,11]. Although SAR studies of small-to-medium size data sets can be performed with no need for computational approaches, the systematic study of the SAR of large data sets requires the application of automated methods. Systematic descriptions of the SAR of compound data sets using the emerging concept of *activity landscape* have been designed to access, visualize, and to help understand the data generated

from general screening. In this context, activity landscape has been conceptualized as the multidimensional space resulting from the addition of biological activity as another dimension to the chemical space of a compound data set [7]. SmAS maps [12], dual and triple activity-difference (DAD/TAD) maps [13, 14], a graph representation [15], and a novel approach using self-organizing maps [16], are examples of methods recently designed by our and other's research groups to characterize the activity landscape of compound data sets screened across several biological endpoints, i.e., multi-target activity landscapes. In this work, the SmAS maps the authors recently proposed [12] were adapted to systematically explore the relationship between the structure and bioactivity profiles of the 618 compounds i.e., the *bioassay activity profile landscape*. A major difference with the analysis of multi-target activity landscapes previously presented is that this work is not focused on a particular set of targets. Instead, this work is focused on the analysis of the activity profile of a set of compounds tested across different bioassays available in PubChem. We want to emphasize that each bioassay has its own quantitative definition of active/inactive/inconclusive. This fact represents a challenge to perform traditional quantitative structure-activity relationships studies [17] and it is required to categorize the activity data to perform quantitative comparisons (see the Methods section).

The distribution of the pairwise *bioassay activity profile similarities*, calculated with Eq. 1 described in the Methods is summarized in Table 1. Overall, the compounds have low activity profile similarity as indicated by the low median (0.30) and mean (0.39) and other statistics of the bioassay activity profile similarities (Table 1). The low similarities are associated with the different activities of the compounds across the tested assays (e.g., active, inactive or inconclusive) and with the different set of assays the compounds were screened across. As pointed out before, not all 618 compounds were screened across the 244 confirmatory assays.

Figure 1A shows the SmAS-like map that depicts the relationship between the bioassay activity profile similarity and MACCS keys/Tanimoto similarity for the 618 compounds (see the prototype plot in Figure 3). The plot contains 190,653 data points that represent a pairwise comparison. Data points are further distinguished by the maximum number of 'active' bioassays of the molecules of the pair using a continuous scale

from green (less active, zero assays) to purple (more active, 27 assays) (vide infra). Figure 1B depicts in color 6,125 pairs where at least one compound in the pair showed activity in at least 12 bioassays i.e., ~10% (12/128) of the maximum activity (thresholds different from 12 can be used to filter the data points for visualization). All other pairs are displayed in light gray for reference. Here, we focused on the compounds that showed activity in several bioassays as an indicative of 'bioactive compounds' (as opposed to compounds that were active in few or none assays). Of note, multiple activities across different bioassays may be a suggestion of promiscuity or polypharmacology [18] that can be associated with positive or negative effects [19].

Pairs of compounds with similar structure and similar bioactivity profile are found in the top-right region of the SmAS-like maps (*roughly* in region II of the prototype map in Figure 3). A representative example is the pair of compounds **26670058_26669932** (identified in the SmAS map in Figure 1A, B). Figure 1C shows a side-by-side comparison of the chemical structures. This figure clearly shows that this pair of molecules has very similar chemical structures (e.g., MACCS/Tanimoto similarity of 0.938) and high bioassay activity profile similarity (0.849). Compound **26670058** was active in 4 (out of 105) bioassays and compound **26669932** was active in 13 (out of 112) bioassays. The high bioassay activity profile similarity indicates that both compounds were tested in similar assays and that both showed similar activity profiles across those assays. Mining the bioassay information available in PubChem for these two compounds revealed that **26670058** and **26669932** were screened in 105 common bioassays. Similar conclusions can be derived from the pair of compounds **26670057_26669686** (Figure 1A, B) and other examples that can be found in region II of the plot.

The SmAS-like map readily identifies in the lower-right region of the plot *bioassay activity profile cliffs* i.e., pairs of compounds with high structure similarity but very different bioassay activity profiles (*roughly* in region IV of the prototype map in Figure 3). A representative example is the pair of compounds **85272523_26669932** with MACCS/Tanimoto similarity of 1.0 but low bioassay activity profile similarity of 0.183 (Figure 1C). The only structural difference between these compounds is the stereochemistry. The low bioassay activity

Table 1. Summary of the distribution of the 190,653 pairwise structure similarities of the 618 compounds calculated with four molecular representations and distribution of the bioassay activity profile similarities.

	Max	Q3[a]	Median	Q1[b]	Min	Mean	StdDev
MACCS	1.00	0.64	0.55	0.47	0.24	0.57	0.14
	1.00	0.43	0.34	0.24	0.00	0.33	0.17
TGD	1.00	0.70	0.59	0.49	0.27	0.61	0.15
Atompairs	1.00	0.26	0.19	0.14	0.02	0.22	0.11
Radial	0.84	0.09	0.06	0.05	0.01	0.08	0.05
bAPS[c]	1.00	0.62	0.30	0.17	0.00	0.39	0.29

[a]Third quartile. [b]First quartile. [c]Bioassay activity profile similarity (calculated with Eq. 1).
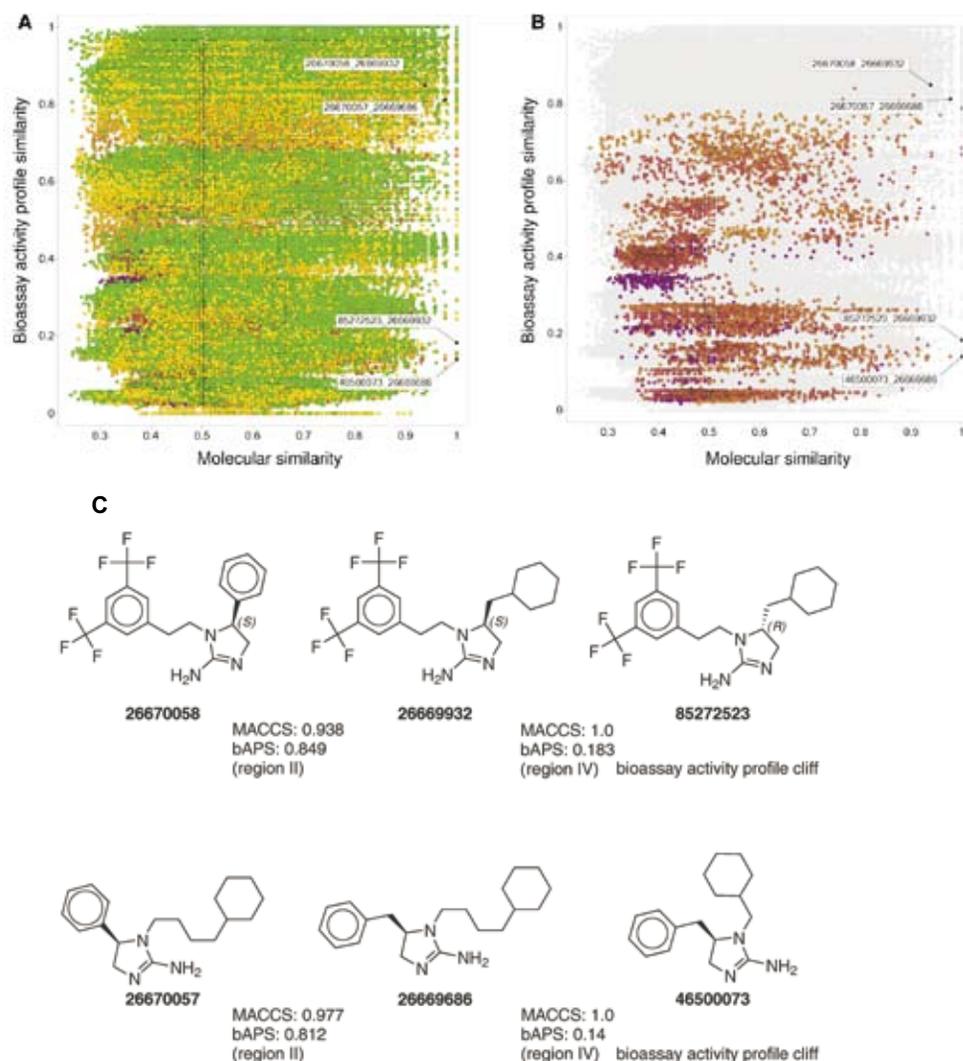
Fig. 1. Modeling of the bioassay activity landscape of the 618 compounds analyzed in this work. (A) The SmAS-like plot shows in color 190,653 pairwise comparisons. Data points are color-coded by the maximum number of confirmatory assays of the most bioactive compound in the pair using a continuous scale from purple (more active) to green (less active); (B) In color, 6,125 points with at least one compound in the pair showed activity in at least 12 bioassays (the remaining pairs of compounds are displayed in light gray); (C) Selected pairs of compounds from region II and IV (see prototype plot in Figure 3). Compounds are labeled with the corresponding SID number in PubChem.

profile similarity indicates that both compounds were tested in different assays and/or that both showed a different activity profile across those assays. For example, compound **26669932** (*S*) was active in 13 (out of 112) assays; in contrast **85272523** (*R*) showed activity in 6 (out of 33) assays. Searching the list of assays available in PubChem for each compound showed that **26669932** and **85272523** were tested only in 25 common bioassays. This result strongly suggests that both compounds are very promising and should be tested across the same set of assays. Similar conclusions can be obtained from the molecule pair **46500073_26669686** in Figure 1. Additional examples of *bioassay activity profile cliffs* can be found in region IV of the plot.

The examples of pairs of molecules discussed above are located in the same relative regions II and IV of the prototype

plot (Figure 3) generated with GpiDAPH3, atom pairs, and radial fingerprints (data not shown). Therefore, these pairs can be considered as *consensus data points* [8] in the bioassay landscape.

In this data set we did not identify remarkable examples of data points representative of *scaffold hopping* (region I of the prototype map), i.e., compounds with very high activity profile similarity (e.g., bAPS > 0.8) and low structure similarity. Although there are several data points in this area (Figure 1A), Figure 1B clearly shows that there are no pairs of compounds in which at least one molecule in the pair showed activity in a number of assays.

The authors have reported different ways to further quantify the contents of the SmAS and related maps [20-22]. Such quantification is beyond the scope of this work that is focused
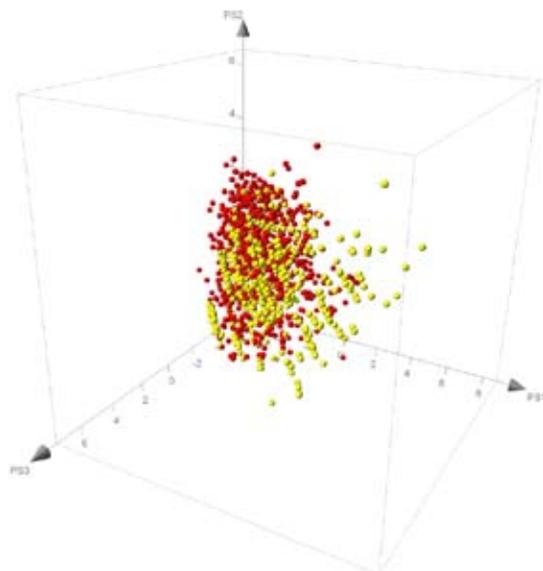
Fig. 2. Visual representation of the chemical space of a set of approved drugs (red) and compounds obtained from in-house collections and deposited in PubChem (yellow). The plot was generated using the ChemGPS-NP prediction scores calculated using the on-line tool ChemGPS-NP$_{Web}$.
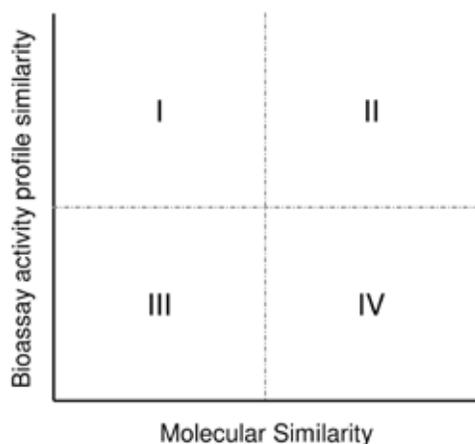

Fig. 3. Prototype SmAS map adapted to model bioassay activity landscapes. The bioassay activity profile similarity is plotted against molecular similarity. The plot can be roughly divided in four major regions. Region IV contains bioassays activity profile cliffs. See text for details.

on the *introduction* of a novel approach to conduct bioassay activity landscape modeling in compound databases.

Chemical space

Several visualization methods have been used to represent chemical spaces [9]. In-house collections have been analyzed by the authors and other groups using several methods including PCA [23, 24], multi-fusion similarity plots [25] and Latent Trait Model [26]. Herein, we employed ChemGPS-NPWeb, a

web-based public tool based on global mapping onto a consistent, eight dimensional map over structure derived physicochemical characteristics for a reference set of compounds (see also Methods section) [27]. Figure 2 shows a visual representation of the chemical space of the 729 compounds obtained from in-house collections available in PubChem (*vide supra*). Figure 2 clearly shows that compounds from in-house collections share the same chemical space of approved drugs and can be regarded as drug-like in terms of physicochemical properties. Visual representations of the chemical space using fingerprint-based representations for in-house collections in PubChem are published elsewhere [8].

Conclusions and perspective

The large amount of bioactivity data available in public repositories such as PubChem represents a major opportunity to further advance drug discovery endeavors. We present a general method to systematically describe structure-bioactivity profile relationships using the concept *bioassay activity landscape modeling*. The analysis was based on the pairwise comparison of bioactivity profile similarity and molecular similarity using molecular fingerprint representations. Dislike current approaches to model multi-target activity landscapes which are focused on a particular set of related targets, the focus of this work is to model bioactivity profiles that may be obtained after screening a compound data set across different and/or unrelated biological targets. To illustrate the method, we used a collection of more than 600 compounds obtained from in-house libraries that have been screened across more than 200 bioassays in PubChem. This collection shares the same chemical space of approved drugs as demonstrated by an analysis of the chemical space herein presented. The approach to *bioassay activity landscape modeling* is general and can be applied to any other set of compounds available in PubChem screened across multiple bioassays or to any other chemical databases with annotated biological activity. A major perspective of this work is to apply this approach to model the *bioassay activity landscape* of other data sets.

Methods

Data sets and activity data

An initial set of 729 compounds derived from in-house collections were obtained from PubChem using the query "Torrey Pines" (accessed June 2011). All tested bioassays available were retrieved for each compound from PubChem. A final set of 618 compounds tested in any confirmatory assay (244 in total) was selected for the bioassay activity landscape modeling. The total number of confirmatory bioassays each compound was tested across, along with the number of 'active', 'inactive' or 'inconclusive' bioassays, were recorded. It is worth noting that not all compounds where tested in all the bioassays: 75%

of the compounds were tested in 101 or fewer bioassays and 50% of the compounds were screened in 37 or fewer bioassays. The total maximum number of confirmatory assays that a single compound was tested in was 128 and the minimum was six.

Bioassay landscape modeling

We investigated the relationship between bioassay activity profile and structure similarity using the principles of multi-target activity landscape modeling i.e., computational methods to explore the structure-activity relationships (SAR) of data sets with biological activity across different biological endpoints [12-15, 21, 28-30]. Herein, for each pair of compounds, the computed molecular similarity was compared with the *bioassay activity profile similarity* across multiple assays. Pairwise structure-bioassay activity profile relationships were visually depicted in 2D plots that are reminiscent of the Structure multiple Activity Similarity (SmAS) maps we recently reported (vide infra) [12].

Bioassay activity profile similarity

In PubChem each bioassay has its own quantitative definition of active/inactive/inconclusive. This is because, at least in part, the bioassays have a different nature and the assays can be performed by different screening centers. Since the main goal of this work is to obtain a general characterization of the bioactivity profile, herein we used a categorical classification of the activity data. For each of the 618 compounds tested in any of the 244 confirmatory assays, the bioassay activity profile was represented as a multiset fingerprint encoding of the activity data available in PubChem as follows: 'active' was set to '2'; 'inactive' as '1'; inconclusive or not tested as '0'. Then, the pairwise bioassay activity profile similarity (bAPS) across the 244 bioassays was calculated using the Tanimoto coefficient [31]:

$$
\text{bAPS}(i, j) = \frac{\sum_{k=1}^{n} \min[m_k(i), m_k(j)]}{\sum_{k=1}^{n} \max[m_k(i), m_k(j)]} \quad (1)
$$

where bAPS(*i,j*) is the bioassay activity profile similarity of the *i*th and *j*th molecules, $m_k(i)$ and $m_k(j)$ are the activity encodings of the *i*th and *j*th molecules, respectively, and *n* is the total number of assays that the molecules were screened across.

Structure similarity

MACCS keys (166 bits) were computed with Molecular Operating Environment (MOE) [32]. In order to address the well-known dependence of chemical space on structure representation [8] other four 2D fingerprints were also explored; two fingerprints implemented in MOE: pharmacophore graph triangle (i.e., graph-based three point pharmacophores) Gpi-

DAPH3 and typed graph distances (TGD); and two fingerprints implemented in Canvas [33], namely atom pairs and radial fingerprints (equivalent to the extended connectivity fingerprints, ECFPs [34]). Structure similarities were computed with the Tanimoto coefficient, which has been successfully used to model the activity landscape of several data sets [13, 14, 20-22]. A summary of the distribution of the 190,653 pairwise structure similarities calculated with the five molecular representations is displayed in Table 1. Overall, the data set is structurally diverse as indicated by the low MACCS/Tanimoto similarity (e.g., median of 0.55; mean of 0.57) and the distribution of the other fingerprints.

Structure multiple Activity Similarity (SmAS) maps

For each pair of compounds, their bioassay activity similarity was plotted against their structural similarity generating SmAS-like maps. These maps, which are an extension of the SAS maps initially proposed for single targets [35], were recently developed in our group and represent a general approach to systematically explore the activity landscapes of data sets tested across multiple biological endpoints [12]. A prototype SmAS map adapted to model bioassay activity landscapes is depicted in Figure 3. The molecular and bioassay activity profile similarities are represented on the *X*- and *Y*-axes, respectively. Four major regions can be *roughly* distinguished in the plot. In this study, pairs of compounds that fall in region I have low structural similarity, but the bioassay activity profile is very similar (although the tested bioassays are not necessarily the same). Region II denotes pairs of compounds with both high structure similarity and high bioassay activity profile similarity. Compounds in region IV have high structure similarity, but low bioassay activity profile similarity and therefore correspond to *bioassay activity profile cliffs* (vide supra). Region III is the least interesting, containing pairs of molecules with low molecular similarity and low bioassay activity profile similarity.

Chemical space

Several visualization methods of the chemical space are available [9, 36]. In this work, we employed the recently developed web-based public tool ChemGPS-NPWeb [27, 37]. ChemGPS-NP [37, 38] is a principal component analysis (PCA) based global chemical positioning system [39] tuned for exploration of biologically relevant chemical space. The first four dimensions of the ChemGPS-NP map capture 77% of data variance. The first dimension (principal component one, PC1) represents size, shape and polarizability (main contribution is size); PC2 is associated with aromatic and conjugation related properties (main influence is aromaticity); PC3 describes lipophilicity, polarity, and H-bond capacity (major contribution is lipophilicity); and PC4 expresses flexibility and rigidity. Chemical compounds can be positioned onto this map using interpolation in terms of PCA score prediction. Further details of this method are provided elsewhere [27].

Acknowledgements

References

1. Scior, T.; Bernard, P.; Medina-Franco, J. L.; Maggiora, G. M. *Mini-Rev. Med. Chem.* **2007**, *7*, 851-860.
2. Bender, A. *Nat. Chem. Biol.* **2010**, *6*, 309-309.
3. Hopkins, A. L. *Nat. Chem. Biol.* **2008**, *4*, 682-690.
4. PubChem. Available at http://pubchem.ncbi.nlm.nih.gov/
5. ChEMBL. Available at http://www.ebi.ac.uk/chembldb/index.php
6. Liu, T. Q.; Lin, Y. M.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucleic Acids Res.* **2007**, *35*, D198-D201.
7. Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. *Drug Discovery Today* **2009**, *14*, 698-705.
8. Medina-Franco, J. L.; Yongye, A. B.; López-Vallejo, F., in: *Statistical Modeling of Molecular Descriptors in QSAR/QSPR*; Matthias, D., Kurt, V., Danail, B., Eds.; Wiley-VCH: **2012**, 307-326.
9. Medina-Franco, J. L.; Martínez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322-333.
10. Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. *Drug Discovery Today* **2010**, *15*, 630-639.
11. Medina-Franco, J. L.; López-Vallejo, F.; Castillo, R. *Educación Química* **2006**, *17*, 452-457.
12. Waddell, J.; Medina-Franco, J. L. *Bioorg. Med. Chem.* **(2012)**, in press. doi:10.1016/j.bmc.2011.1011.1051.
13. Pérez-Villanueva, J.; Santos, R.; Hernández-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. *Med. Chem. Comm.* **2011**, *2*, 44-49.
14. Medina-Franco, J. L.; Yongye, A. B.; Pérez-Villanueva, J.; Houghten, R. A.; Martínez-Mayorga, K. *J. Chem. Inf. Model.* **2011**, *51*, 2427-2439.
15. Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 258-266.
16. Iyer, P.; Bajorath, J. *Chem. Biol. Drug Des.* **2011**, *78*, 778-786.
17. Medina-Franco, J. L.; Golbraikh, A.; Oloff, S.; Castillo, R.; Tropsha, A. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 229-242.
18. Merino, A.; Bronowska, A. K.; Jackson, D. B.; Cahill, D. J. *Drug Discovery Today* **2010**, *15*, 749-756.
19. Peters, J. U.; Schnider, P.; Mattei, P.; Kansy, M. *ChemMedChem* **2009**, *4*, 680-686.
20. Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. *J. Chem. Inf. Model.* **2009**, *49*, 477-491.
21. Pérez-Villanueva, J.; Santos, R.; Hernández-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. *Biorg. Med. Chem.* **2010**, *18*, 7380-7391.
22. Yongye, A.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. *J. Chem. Inf. Model.* **2011**, *51*, 1259-1270.
23. López-Vallejo, F.; Nefzi, A.; Bender, A.; Owen, J. R.; Nabney, I. T.; Houghten, R. A.; Medina-Franco, J. L. *Chem. Biol. Drug Des.* **2011**, *77*, 328-342.
24. López-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. *Drug Discovery Today* **2012**, *17*, 718-726.
25. Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. *Chem. Biol. Drug Des.* **2007**, *70*, 393-412.
26. Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; López-Vallejo, F. *J. Chem. Inf. Model.* **2011**, *51*, 1552-1563.
27. Rosen, J.; Lovgren, A.; Kogej, T.; Muresan, S.; Gottfries, J.; Backlund, A. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 253-259.
28. Peltason, L.; Hu, Y.; Bajorath, J. *ChemMedChem* **2009**, *4*, 1864-1873.
29. Wassermann, A. M.; Peltason, L.; Bajorath, J. *ChemMedChem* **2010**, *5*, 847-858.
30. Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. *Bioorg. Med. Chem.* **2012**, *20*, 3523–3532.
31. Maggiora, G. M.; Shanmugasundaram, V., in: *Chemoinformatics and Computational Chemical Biology, Methods in Molecular Biology,* Vol. 672, Bajorath, J., Ed.; Springer, New York, **2011**, 39-100.
32. Molecular Operating Environment (MOE), version 2010; Chemical Computing Group Inc., Montreal, Quebec, Canada.
33. Canvas, version 1.3; Schrödinger, LLC, New York, NY, 2010.
34. Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742-754.
35. Shanmugasundaram, V.; Maggiora, G. M. In *222nd ACS National Meeting, Chicago, IL, United States*; American Chemical Society, Washington, D. C: Chicago, IL, United States, **2001**.
36. Le Guilloux, V.; Colliandre, L.; Bourg, S.; Guenegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. *J. Chem. Inf. Model.* **2011**, *51*, 1762-1774.
37. Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. *J. Nat. Prod.* **2007**, *70*, 789-794.
38. Larsson, J.; Gottfries, J.; Bohlin, L.; Backlund, A. *J. Nat. Prod.* **2005**, *68*, 985-991.
39. Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157-166.