Expanding the medicinally relevant chemical space with compound libraries

Fabian López-Vallejo, Marc A. Giulianotti, Richard A. Houghten and José L. Medina-Franco

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

Analysis of marketed drugs and commercial vendor libraries used in high-throughput screening suggests that the medicinally relevant chemical space may be expanded to unexplored regions. Novel regions of the chemical space can be conveniently explored with structurally unique molecules with increased complexity and balanced physicochemical properties. As a case study, we discuss the chemoinformatic profile of natural products in the Traditional Chinese Medicine (TCM) database and a large collection assembled from 30 small-molecule combinatorial libraries with emphasis on assessing molecular complexity. The herein surveyed combinatorial libraries have been successfully used over the past 20 years to identify novel bioactive compounds across different therapeutic areas. Combinatorial libraries and natural products are suitable sources to expand the traditional relevant medicinal chemistry space.

Many drug discovery efforts focus on compound libraries commonly filtered using classical semi-empirical rules. A prominent example is the seminal Lipinski's Rule-of-Five (RO5) [1] that has been revised over the past decade. Nowadays it is largely recognized that 'new molecular entities are moving away from the traditional drug space' [2,3] and that 'as new targets emerge and optimization tools advance, the oral drug-like space might expand' [4]. Commercial vendor libraries, which are the current major source of small molecules for high-throughput screening (HTS), have led to the identification of novel leads for traditional drug targets, such as kinases, G protein-coupled receptors, and ligand-gated ion channels [5]. However, such libraries have failed for many classes of biological targets. This fact has been highlighted by Dandapani and Marcaurelle in their excellent commentary [6], pointing out that such libraries interrogate a narrow range of medicinally relevant chemical space. Emerging therapeutic targets, such as DNA methyltransferases [7,8], or unidentified targets could be successfully explored with larger and novel areas of chemical space [2,6,9]. Such spaces can be appropriate to identify potential lead compounds for the so-called 'undruggable' targets (i.e. targets that lack small molecule starting points in the traditional property space) [2].

Molecular complexity and chemical space

Molecular complexity is an attractive criterion to guide the selection of chemical libraries to experimentally explore largely neglected chemical space [6,10]. A recent study showed that compounds with a greater fraction of saturated carbons, which is an intuitive measure of complexity, have a higher success rate in the drug discovery process [11]. In that study the authors concluded that 'more highly complex molecules, as measured by saturation, have the capacity to access greater chemical space' [11]. In the same work, the authors also suggested that compounds with increased complexity as measured by the degree of saturation, might give rise to improved selectivity. This hypothesis has been supported experimentally by Clemons *et al.* [12] who screened across 100 diverse proteins, commercial compounds, natural products, and synthetic compounds from academic groups. Clemons *et al.* concluded that increasing the content of sp^3-hydridized and stereogenic atoms relative to compounds from commercial sources, improves selectivity and frequency of binding [12]. These results are in accord with the 'complexity model' of Hann *et al.* [13] which has been recently reviewed [14]. In an

Corresponding author:. Medina-Franco, J.L. (jmedina@tpims.org)

1359-6446/06/$ - see front matter © 2012 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.drudis.2012.04.001

independent study, analysis of the chemical complexity of the compounds in the Molecular Libraries Small Molecule Repository (MLSMR) using the fraction of saturated carbons as the measure of complexity, led to the conclusion that natural products and collections from academic and other research institutes are more complex than libraries from commercial vendors [6]. Using the same measures of complexity, Chen *et al.* showed the increased complexity of natural products over other collections including drugs, clinical candidates and bioactive molecules [15]. These results further support the rationale to develop more complex, 'natural product-like' libraries for improved drug discovery [16].

Combinatorial libraries as promising sources of complex molecules

In addition to natural products, combinatorial libraries also represent a rich source of complex molecules. Combinatorial chemistry, combined with HTS and other screening methodologies, continues to have a key role in drug discovery [17–19]. Several groups in pharmaceutical companies, universities and research institutes, such as the University of Cambridge, the University of Pittsburgh, the Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), the Howard Hughes Medical Institute, to name a few, have made significant contributions to expand the chemical space with combinatorial libraries. Among these groups, the Torrey Pines Institute for Molecular Studies (TPIMS) has developed over the past two decades a set of small-molecule combinatorial libraries assembled in a so-called 'scaffold ranking plate' which contains more than 35 libraries. The compound collections have been prepared over the years to introduce diversity in structures and chemical properties. Experimental evidence has shown that this plate is a valuable tool to rapidly select the most promising scaffolds for further screening and identifying novel lead compounds [17,18,20]. Experiments across several targets suggest that the small-molecule libraries contain compounds not present in other compound collections frequently used for lead identification [17,18,20]. As part of a continued effort to combine combinatorial chemistry with computational approaches for accelerated drug discovery [21], chemoinformatic analysis of combinatorial libraries has revealed that some of these libraries occupy the chemical space of marketed drugs and the MLSMR library, whereas other libraries will explore new regions of chemical space [22–24]. These libraries have been characterized and described quantitatively by means of molecular scaffolds, molecular properties, and structural fingerprints [22,23].

Expanding the chemical space with combinatorial libraries and natural products

As discussed above, combinatorial libraries and natural products represent promising sources to interrogate novel regions of medicinally relevant chemical space. As a case study, we surveyed the molecular complexity and computed physicochemical characteristics of 30 small-molecule combinatorial libraries present in the scaffold raking plate (Figure S1 in the Supplementary information). We also surveyed the complexity and profile of physicochemical properties of TCM which is a large collection of natural products available in the public domain [25]. The profile of these collections was compared with approved drugs and general screening collections including a diverse set from the National Cancer Institute (NCI) database, a set of natural products from commercial vendors and a diverse commercial vendor library.

TABLE 1

Compound libraries considered in this study

Database (abbreviation)	Size[a]
30 combinatorial libraries (TPISR)	30,000
Traditional Chinese Medicine (TCM)	28,277
Natural products in ZINC (NP)	89,032
Approved drugs (DrugBank)	1731
Commercial vendor library (Maybridge)	14,400
NCI diversity (NCI)	1817

[a] Number of unique compounds considered in this study.

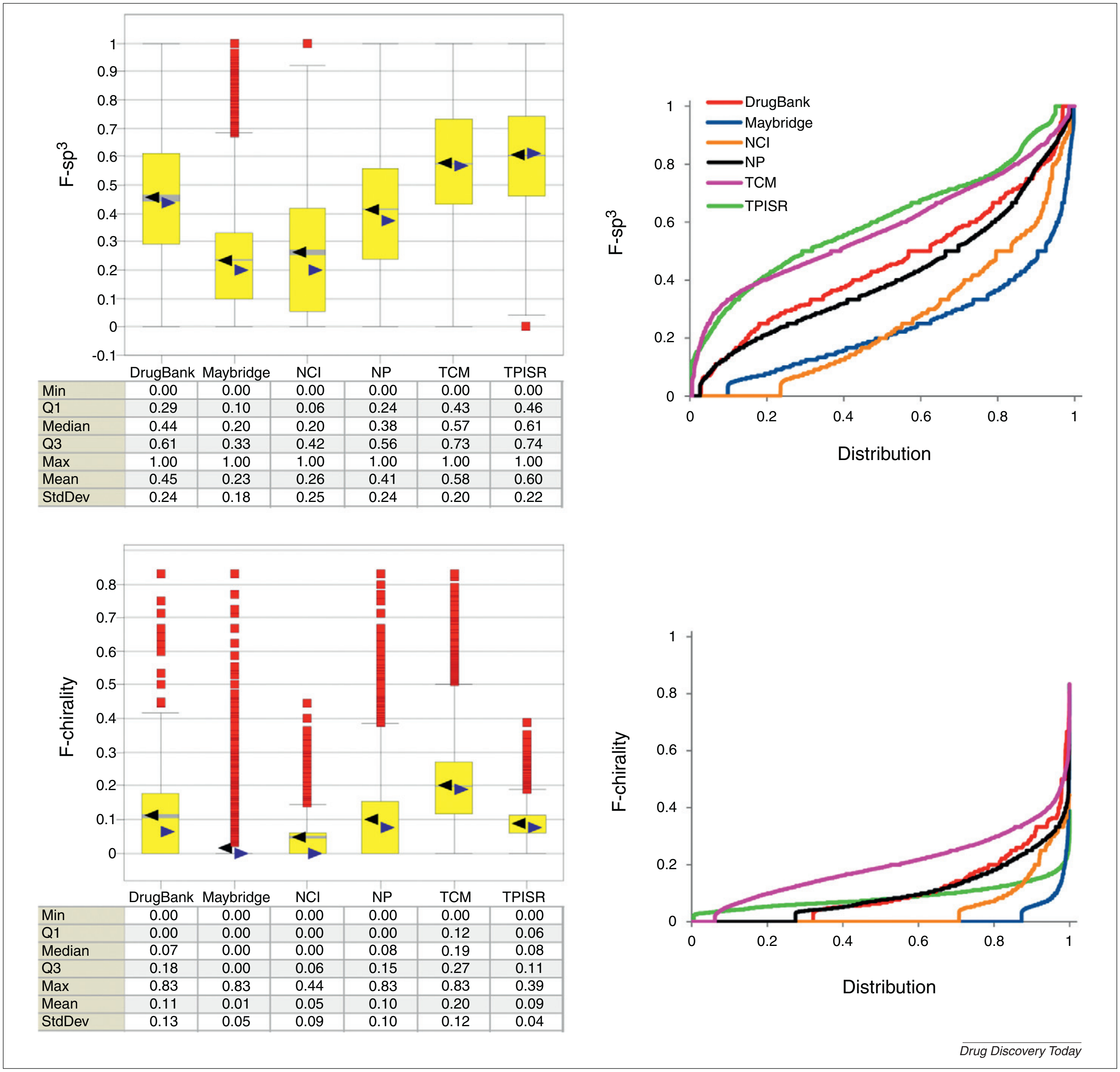
The core-scaffolds of the 30 combinatorial libraries [17,18,20] are depicted in Figure S1. Random subsets of 1000 compounds per combinatorial library were assembled in a single data set containing 30,000 compounds referred in this work as the TPIMS Scaffold Ranking Library (TPISR). It has been shown that random samples of 1000 molecules are representative of the molecular diversity [22,26]. The combinatorial libraries were enumerated with Molecular Operating Environment (MOE) [27]. TCM, which was assembled by the China Medical University and Asia University in Taiwan, and the MIT in USA, was retrieved from the collection's website (downloaded on October 2011) [25]. Natural products from commercial sources (denoted as 'NP' onwards) were obtained from the ZINC database (downloaded on March 2011) [28]. The collection of drugs from DrugBank [29] and the National Cancer Institute Diversity II (denoted 'NCI' onwards) were retrieved from ZINC (downloaded on March 2011). The Maybridge HitFinder^{TM} database which represents the drug-like diversity of more than 56,000 organic compounds from the Maybridge Screening Collection was used as the commercial vendor library [Maybridge: http://www.maybridge.com]. All databases were prepared with MOE by disconnecting group I metals in simple salts and keeping the largest fragments. Unique compounds were selected. To avoid bias in the comparison, all molecules with molecular weight (MW) over 1100 in the reference collections were excluded. In total, we analyzed 28,277 compounds from TCM; 89,032 compounds from NP; 1731 compounds from DrugBank; 14,400 compounds from the commercial vendor library (Maydridge); and 1817 compounds from NCI. Table 1 summarizes the compound libraries considered in this work.

Molecular complexity

Several measures of complexity have been reported including MW [30–33]. In this survey, we focused on intuitive and well-established measures that are increasingly being used to compare the complexity of compound collections so that the results can be directly compared with other reports. Carbon bond saturation was defined by fraction sp^3 (Fsp^3) where Fsp^3 = (number of sp^3 hybridized carbons divided by total carbon count) [11]. Overall, a larger Fsp^3 value indicates that the molecule is more likely to have a 3D structure (i.e. the structure would be less flat) [11,15]. Stereochemical complexity was defined as the proportion of carbon atoms that are chiral [12], F-Chirality = (number of chiral carbon atoms divided by total carbon count). MOE was used to calculate the number of chiral carbon atoms and the total carbon count. Maestro 9.2 [34] was used to calculate the number of sp^3 hybridized carbons.

Carbon bond saturation

Figure 1 shows cumulative distribution function (CDF) curves and box plots summarizing the distribution of Fsp^3 values of TPISR, TCM, and other reference collections. Selected statistics from the CDFs are also shown in the figure. In agreement with previous reports, approved drugs have higher Fsp^3 values than the commercial vendor library (Maybridge) (e.g. compare the corresponding median and mean values). The calculated mean Fsp^3 value for drugs obtained from DrugBank, 0.45, is similar to the mean values reported for other collections of drugs (0.47 and 0.40) [11,15]. Similarly, low Fsp^3 values have been calculated for other commercial vendor libraries (e.g. mean of 0.30 for commercial compounds in MLSMR) [6]. The low Fsp^3 values for the commercial vendor library of this work were comparable to the low values of the NCI



	DrugBank	Maybridge	NCI	NP	TCM	TPISR
Min	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.29	0.10	0.06	0.24	0.43	0.46
Median	0.44	0.20	0.20	0.38	0.57	0.61
Q3	0.61	0.33	0.42	0.56	0.73	0.74
Max	1.00	1.00	1.00	1.00	1.00	1.00
Mean	0.45	0.23	0.26	0.41	0.58	0.60
StdDev	0.24	0.18	0.25	0.24	0.20	0.22

	DrugBank	Maybridge	NCI	NP	TCM	TPISR
Min	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	0.00	0.00	0.00	0.12	0.06
Median	0.07	0.00	0.00	0.08	0.19	0.08
Q3	0.18	0.00	0.06	0.15	0.27	0.11
Max	0.83	0.83	0.44	0.83	0.83	0.39
Mean	0.11	0.01	0.05	0.10	0.20	0.09
StdDev	0.13	0.05	0.09	0.10	0.12	0.04

FIGURE 1

Molecular complexity of chemical libraries. Cumulative distribution function and box plots of the molecular complexity measures of the compound libraries analyzed in this work. The yellow boxes enclose data points with values within the first and third quartile; the black and blue triangles denote the mean and median distributions, respectively; the lines above and below indicate the upper and lower adjacent values. The red squares indicate outliers. Selected statistics of each distribution are also shown. *Abbreviations*: NCI, National Cancer Institute; NP, Natural products; TCM, Traditional Chinese Medicine; TPISR, Torrey Pines Institute Scaffold Ranking Library.

data set. The higher values of Fsp^3 for NP as compared to commercial vendor libraries are in agreement with the higher values reported for natural products from different sources [6,15]. Figure 1 clearly shows that TCM and TPISR have the highest Fsp^3 values with mean and median values of approximately 0.59. These values are similar to the values reported for other natural product databases and metabolites (e.g. median values of 0.59 and 0.63, respectively) [15]. Based on previously published studies about molecular complexity [11], the results of this comparison support the hypothesis that TPISR and TCM have the capacity to explore novel areas of chemical space. As discussed earlier, databases with greater complexity than commercial databases, have a greater chance to identify compounds that better complement the spatial requirement of target proteins [11]. A potential caveat, however, is that screening the entire highly complex collections might not lead to a large number of hits.

Stereochemical complexity

Figure 1 also shows the distribution of F-Chirality values [12] for all the six databases. Similar to the trends observed with the Fsp^3 values, the set of drugs showed increased content of stereogenic atoms relative to the commercial vendor library and NCI. The median and mean values of F-Chirality for the commercial library (Maybridge) (0.00 and 0.01, respectively) are in agreement with the low values reported for other commercial vendor libraries [12]. NP showed similar F-Chirality values as compared to drugs. TCM showed the highest F-Chirality values (e.g. median and mean of 0.19 and 0.20, respectively) which are similar values reported for a different natural product database [12]. TPISR has a distribution of F-Chirality values comparable to NP and currently approved drugs (similar mean and medians) but with less standard deviation. In agreement with previously formulated hypotheses, the increased F-Chirality values of TPISR and TCM relative to compounds from commercial sources support the fact that these collections are attractive sources of high-value probes and drugs [12].

We want to point out that the discussion of the combinatorial libraries is focused on TPISR (i.e. a portion of the collection of small-molecules libraries present in the scaffold ranking plate as a whole). We followed this approach because the experimental screening of the entire scaffold ranking plate is the first step in the identification of lead compounds (see above). However, each library in TPISR was characterized in detail as shown in Figure S2 in the Supplementary information which shows the distributions of the Fsp^3 and F-Chirality values of each of the 30 combinatorial libraries.

Physicochemical properties

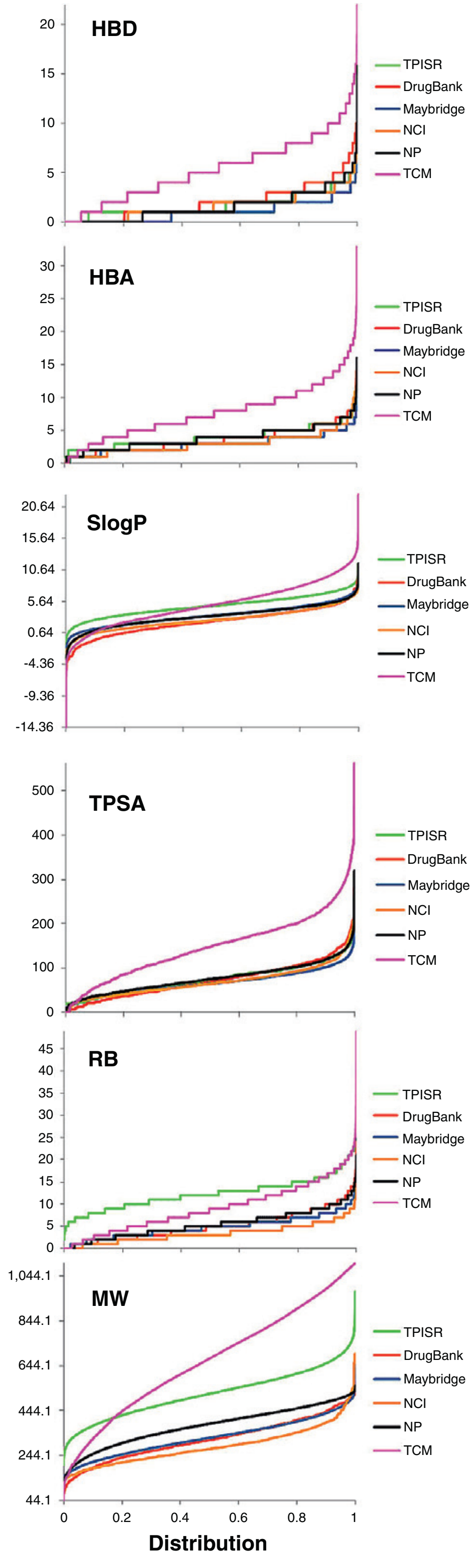
Physicochemical properties have been used extensively to define and compare the property space of approved drugs, bioactive compounds and other molecular databases. Several of these properties have served as the basis to establish classical semi-empirical rules, such as the seminal Lipinski's RO5 [1] or the Veber's rule-of-three [35]. Of note, the RO5 was derived from drug candidates that reached Phase II and studies with newly marketed drugs indicate that these classical rules are not necessarily valid [2,36]. The following properties were computed with the MOE program: hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), the octanol and/or water partition coefficient (SlogP), topological polar surface area (TPSA), number of rotatable bonds (RB), and MW.

Figure 2 summarizes the distribution of six physicochemical properties of the six compound collections. The three important molecular properties of polarity, flexibility and size are described by HBD, HBA, SlogP, and TPSA; RB; and MW, respectively. These descriptors have been extensively used to compare the property space of drugs, combinatorial libraries, natural products and other compound databases [22,23,37,38]. According to Fig. 2, the distribution of the properties of TPISR associated with polarity are, overall, comparable to drugs, NP, NCI, and the commercial vendor library but with less standard deviation. An exception is the SlogP values that are higher for TPISR as compared to drugs, NP and the commercial compounds (but lower than the SlogP values for TCM). Of note, recent analysis reviewed by Leach and Hann [14] suggests that off-target promiscuity increases with MW and, in particular with lipophilicity. However, the relationship between promiscuity and lipophilicity has not been clearly established due, in part, to the different definitions of 'promiscuity' employed in the studies [14]. TCM has the largest values of HBD, HBA, SlogP, and TPSA with respect to the other five collections as reflected by the corresponding mean and median values. Regarding flexibility, TPISR has higher values of RB as compared to the other compound collections, followed by TCM. NCI is the least flexible according to this measure. Concerning the size, in general, TCM has the largest molecules followed by TPISR. NP, the commercial vendor library, and currently approved drugs have similar distributions of MW whereas the NCI diversity set has the lowest MW values. Interestingly, studies comparing drugs launched before 1983 ('old' drugs) and drugs lunched between 1983 and 2002 ('new' drugs) reveal that the mean and median RB and MW values are larger in new drugs [2,3]. Table S1 in the Supplementary information summarizes the distributions of the property values for each of the 30 libraries in TPISR (Figure S1). Taken together, these results show that TPISR and TCM are able to sample and expand the property space of drugs and other currently available general screening collections. The overall high lipophilicity of TPISR and TCM suggest that these collections might be not as selective as indicated otherwise by the measures of molecular complexity discussed above.

Structure fingerprints

Inter-library similarity

TPISR, TCM and other selected libraries were compared using fingerprints including Molecular Access System (MACCS) keys (166 bits), graph-based three-point pharmacophore (GpiDAPH3) and typed graph distances (TGD) fingerprints implemented in MOE. Fingerprints with different design were used to reduce the dependence of chemical space with structure representation [39,40]. Inter-library similarity can be evaluated by means of multifusion similarity maps [41] which have been used to compare combinatorial libraries and other compounds collections [42,43] and nearest-neighbor curves. These curves represent the distribution of the maximum similarity values of molecules in a test set with respect to the molecules in the reference set [41]. In this work, random subsets of 1000 compounds from TPISR and TCM were used as test sets while the commercial vendor library and drugs were reference sets. Additional comparisons with TCM as the reference set were also performed.



Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	0	1	1	2	7	1.72	1.17
DrugBank	0	1	2	3	10	2.02	1.81
Maybridge	0	0	1	2	10	1.04	1.05
NCI	0	1	1	2	8	1.61	1.31
NP	0	0	1	2	16	1.56	1.47
TCM	0	3	5	7	22	5.36	3.27

Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	1	3	4	5	11	3.93	1.49
DrugBank	0	2	3	5	14	3.58	1.92
Maybridge	0	2	3	4	12	2.92	1.32
NCI	0	2	3	4	12	3.00	1.67
NP	0	3	4	5	16	3.82	1.66
TCM	0	5	7	10	33	7.67	3.97

Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	-2.54	3.78	4.92	6.04	11.62	4.89	1.70
DrugBank	-5.40	1.06	2.47	3.83	9.91	2.37	2.20
Maybridge	-4.59	2.24	3.36	4.51	9.46	3.39	1.68
NCI	-3.19	1.59	2.66	3.81	10.39	2.72	1.75
NP	-6.91	2.20	3.32	4.42	11.66	3.24	1.76
TCM	-14.36	2.78	5.02	7.40	22.61	5.11	3.47

Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	18.5	47.9	69.5	96.4	230.0	74.1	33.7
DrugBank	0	41.2	68.5	98.2	279.1	72.4	41.2
Maybridge	0	46.9	64.6	83.3	259.5	66.1	27.5
NCI	0	45.6	63.8	86.9	213.3	68.7	33.6
NP	0	52.6	74.7	97.4	319.5	76.4	33.1
TCM	0	97.3	147.7	191.7	562.8	147.8	72.9

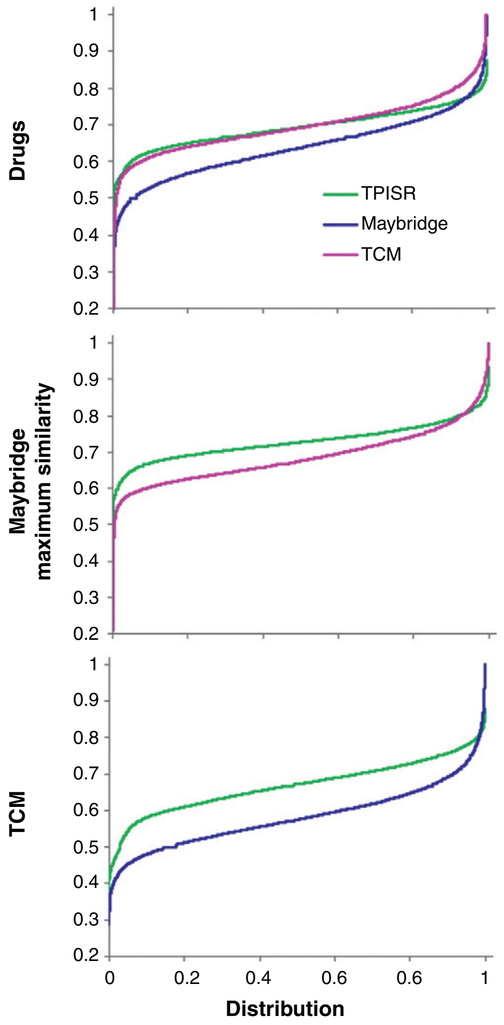
Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	2	10	12	14	26	12.31	3.47
DrugBank	0	3	5	7	24	5.05	3.29
Maybridge	0	3	5	6	19	4.83	2.44
NCI	0	2	3	5	19	3.41	2.14
NP	0	3	5	7	21	5.41	3.03
TCM	0	5	9	13	49	9.11	5.42

Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	203.3	441.7	518.8	590.8	974.6	517.6	106.6
DrugBank	75.1	249.5	313.8	384.4	537.6	315.2	93.5
Maybridge	123.2	263.3	322.4	380.5	656.5	323.5	81.5
NCI	114.1	224.4	273.1	330.2	697.1	284.1	84.6
NP	78.1	317.4	385.5	441.6	553.9	375.0	86.8
TCM	44.1	486.5	675.9	855.1	1100.0	663.9	243.1

Drug Discovery Today

FIGURE 2

Cumulative distribution function (CDF) of the distribution of molecular properties. The tables summarize statistics of the CDFs. *Abbreviations*: HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; MW, molecular weight; NCI, National Cancer Institute; NP, Natural products; Q1, first quartile; Q3, third quartile; RB, rotatable bonds; SD, standard deviation; TPSA, topological polar surface area; TCM, Traditional Chinese Medicine; TPISR, Torrey Pines Institute Scaffold Ranking Library.



Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	0.48	0.66	0.69	0.73	0.88	0.69	0.05
Maybridge	0.29	0.58	0.64	0.69	1.00	0.64	0.09
TCM	0.23	0.65	0.69	0.74	1.00	0.70	0.07

Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	0.52	0.70	0.73	0.76	0.93	0.73	0.05
TCM	0.24	0.63	0.68	0.73	1.00	0.69	0.07

Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	0.38	0.62	0.67	0.72	0.88	0.67	0.07
Maybridge	0.29	0.52	0.57	0.63	1.00	0.58	0.09

Drug Discovery Today

FIGURE 3

Inter-library similarity. Cumulative distribution function (CDF) of the maximum structure similarity computed with MACCS keys/Tanimoto comparing the scaffold ranking plate, TPISR (green), the Traditional Chinese Medicine (magenta), and the commercial vendor library (blue) with three reference collections (designated along the left-hand side of the figure). The tables summarize statistics of the CDFs. *Abbreviations*: Q1, first quartile; Q3, third quartile; SD, standard deviation; TCM, Traditional Chinese Medicine; TPISR, Torrey Pines Institute Scaffold Ranking Library.

Figure 3 shows CDF curves of the maximum similarity of TPISR, TCM and the commercial vendor library with three reference libraries using MACCS keys. Selected statistics from the CDFs are also shown. The name of the reference dataset is designated along the left-hand side of the figure. The upper and middle panel of Fig. 3 shows a comparison of TPISR (green) and TCM (magenta) to drugs and the commercial vendor library, respectively. The corresponding reference libraries are not represented in the CDFs. The low values of the CDFs and the corresponding statistics clearly indicate that the combinatorial libraries (TPISR) and the natural products from TCM have compounds with different chemical structures as compared to drugs and the diverse collection of commercial compounds. We did not identify any compounds that were both in the TPISR and any of the reference libraries. This is in agreement with a previous analysis conducted with representative in-house libraries [22–24,42]. Similar conclusions were obtained with GpiDAPH3 and TGD representations.

The bottom panel of Fig. 3 shows a comparison of the TPISR combinatorial libraries (green) and commercial compounds (blue) to TCM which is used as a reference. The CDF shows that, on average, the chemical structures of the commercial vendor library are less similar to TCM as compared to the combinatorial libraries [e.g. the mean and median of the maximum similarity values of the commercial compounds (0.57 and 0.58) are lower than the corresponding similarity values of the combinatorial libraries (0.67)].

The similarity values obtained with TGD were, overall, higher than the similarity values obtained with MACCS keys and GpiDAPH3 (data not shown). Similar results have been obtained for several other data sets [39,44,45]. However, the overall relative similarity of the test sets with the reference data sets in Fig. 3 remains the same.

The distribution of the maximum similarity values of each combinatorial library in TPISR, TCM, drugs, and the commercial vendor library to different reference sets is presented in Figure S3 in

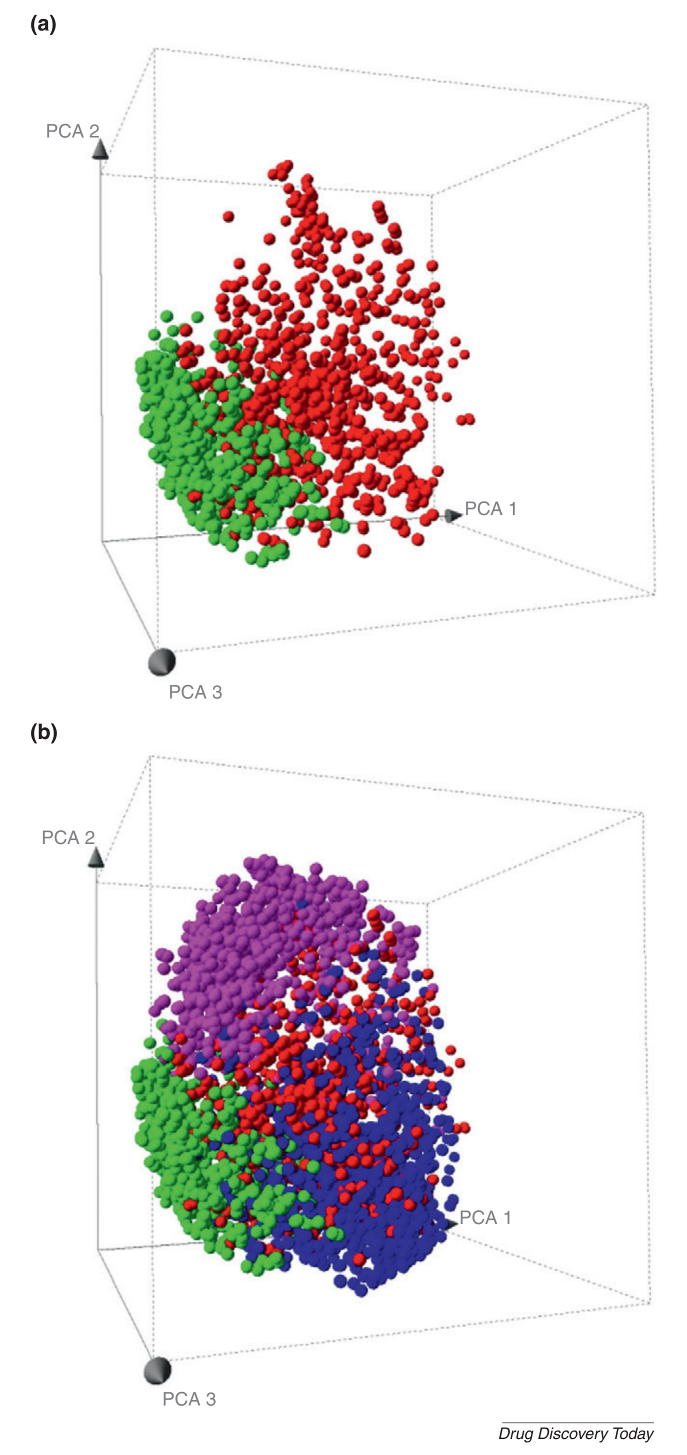


FIGURE 4

Visual representation of the chemical space of the scaffold ranking plate. TPISR (green) drugs (red), natural products from the Traditional Chinese Medicine (magenta) and a commercial vendor library, Maybridge (blue). The first three principal components account for 81.48% of the variance. For clarity, panels **a** and **b** display two or the four collections, respectively, in the same coordinate space.

the Supplementary information. As mentioned above, a detailed discussion of each of the 30 combinatorial libraries is beyond the scope of this manuscript that is focused on the discussion of TPISR as a whole, TCM and other compound collections.

Visualization of the chemical space

A visual representation of the chemical space of random subsets of 1000 compounds selected from TPISR, TCM, the commercial vendor library and drugs was obtained performing principal component analysis (PCA) of the similarity matrix of the compound databases (4000 compounds total) computed with MACCS keys (166 bits) and the Tanimoto coefficient. Other representations can be used [40]. Figure 4 shows a visual representation of the chemical space of TPISR, TCM, drugs, and the commercial vendor library. The first three principal components account for 81.48% of the variance. Figure 4a displays TPISR (green) and drugs (red) and Fig. 4b also shows the position in chemical space of TCM (magenta) and the commercial vendor library (blue) within the same coordinates. Figure 4a clearly shows that several compounds from TPISR share the same structural space of drugs. Also, in agreement with previous results for representative combinatorial libraries, it is concluded that there are several compounds from TPISR that cover sparsely and unexplored regions of the chemical space of drugs as represented by MACCS keys [22]. The results of the visual comparison of the chemical spaces are in agreement with the quantitative similarity values captured in the CDFs of the maximum similarity values discussed above.

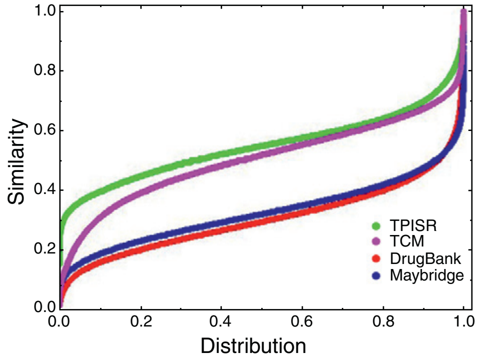
Figure 4b shows that the commercial compounds share the same area of chemical space as currently approved drugs. By contrast, the natural products from TCM cover a different and unexplored region of the chemical space of the combinatorial libraries, drugs, and the commercial vendor library. Similar conclusions were obtained in a recent analysis of the chemical space of a specific implementation of TCM in ZINC [46]. The similarity matrix that served as the basis to generate the chemical space in Fig. 4 can also be visualized as a heat map [23,46] which is presented and discussed in Figure S4 in the Supporting information.

Intra-library similarity

The intra-library similarity of the compound collections was measured using MACCS keys. This fingerprint representation was selected because of its widespread use for comparing compound collections. In addition, MACCS keys gave similar trends in similarity, although different absolute values, as compared to GpiDAPH3 and TGD fingerprints (see above). Figure 5 shows the distribution of the pairwise similarity values computed with MACCS keys/Tanimoto using CDF curves. The CDFs and the corresponding statistics summarized in Fig. 5 indicate that drugs are the database with the least similarity (highest diversity) followed by the commercial compounds, TCM and TPISR. The latter two have comparable intra-library similarities (e.g. median and mean values of 0.5). Also, drugs and the commercial vendor library have comparable similarities (e.g. median and mean values of 0.3). Similar conclusions can be obtained from the heat map in Figure S4.

Concluding remarks

Comparisons of new versus old drugs and the lack of suitable starting points for novel targets in the traditional medicinal chemistry space is increasing the awareness to explore novel regions of chemical space. A suitable way to expand the medicinally relevant chemical space is screening libraries with structurally



Dataset	Min	Q1	Median	Q3	Max	Mean	SD
TPISR	0.15	0.47	0.55	0.62	1.00	0.55	0.12
TCM	0.00	0.42	0.52	0.61	1.00	0.50	0.14
DrugBank	0.00	0.22	0.30	0.38	1.00	0.31	0.12
Maybridge	0.00	0.25	0.32	0.40	1.00	0.33	0.11

Drug Discovery Today

FIGURE 5

Intra-library similarity. umulative distribution function (CDF) of the intra-library similarity computed with MACCS keys and/or Tanimoto for four selected datasets. The table summarizes statistics of the CDFs. *Abbreviations*: Q1, first quartile; Q3, third quartile; SD, standard deviation; TCM, Traditional Chinese Medicine; TPISR, Torrey Pines Institute Scaffold Ranking Library.

unique and complex compounds which also maintain balanced physicochemical properties. It has been shown that compounds with greater *sp*^{3} centers and higher stereochemical complexity have higher success rates at different stages of the drug development process. Moreover, compounds with increased complexity as measured by the degree of saturation, may give rise to improved selectivity although screening highly complex libraries may not lead to a significant number of hits. Over the years, pharmaceutical companies and several groups in different universities and research institutes have developed small-molecule combinatorial libraries and natural product-like libraries. Screening these collections has given rise to the identification of novel lead compounds. The molecular complexity, physicochemical profile and structural diversity of compound collections can be readily assessed using chemoinformatic approaches. As a case study herein we surveyed the complexity, structural diversity and properties profile of 30 small-molecule combinatorial libraries and a large collection of natural products assembled in the TCM database. The high molecular complexity and structural uniqueness of these collections suggest that, overall, these databases are suitable to interrogate novel regions of the neglected chemical space. Characterization of the chemical space of the compound libraries generated with other representations, such as ADMETox-related properties and aromaticity [47] is warranted. This survey encourages the systematic computational and experimental characterization of additional combinatorial libraries and natural products collections to explore their potential to expand the medicinally relevant chemical space.

Conflict of interest

The authors declare that they do not have any conflict of interest related to this manuscript.

Acknowledgements

Authors thank Jacob Waddell for his assistance in performing part of the similarity calculations. This work was funded by the Multiple Sclerosis National Research Institute (J.L.M-F. and M.G.) and partially funded by the State of Florida, Executive Office of the Governor's Office of Tourism, Trade, and Economic Development.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.drudis.2012.04.001.

References

1 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

2 Faller, B. *et al.* (2011) Evolution of the physicochemical properties of marketed drugs: can history foretell the future? *Drug Discov. Today* 16, 976–984

3 Leeson, P.D. and Davis, A.M. (2004) Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* 47, 6338–6348

4 Zhao, H. (2011) Lead optimization in the nondrug-like space. *Drug Discov. Today* 16, 158–163

5 Hert, J. *et al.* (2009) Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* 5, 479–483

6 Dandapani, S. and Marcaurelle, L.A. (2010) Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.* 6, 861–863

7 Foulks, J.M. *et al.* (2012) Epigenetic drug discovery. *J. Biomol. Screening* 17, 2–17

8 Medina-Franco, J.L. and Caulfield, T. (2011) Advances in the computational development of DNA methyltransferase inhibitors. *Drug Discov. Today* 16, 418–425

9 Drewry, D.H. and Macarron, R. (2010) Enhancements of screening collections to address areas of unmet medical need: an industry perspective. *Curr. Opin. Chem. Biol.* 14, 289–298

10 Selzer, P. *et al.* (2005) Complex molecules: do they add value? *Curr. Opin. Chem. Biol.* 9, 310–316

11 Lovering, F. *et al.* (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52, 6752–6756

12 Clemons, P.A. *et al.* (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U. S. A.* 107, 18787–18792

13 Hann, M.M. *et al.* (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41, 856–864

14 Leach, A.R. and Hann, M.M. (2011) Molecular complexity and fragment-based drug discovery: ten years on. *Curr. Opin. Chem. Biol.* 15, 489–496

15 Chen, H. *et al.* (2012) A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *MedChemComm* 3, 312–321

16 Thomas, G.L. and Johannes, C.W. (2011) Natural product-like synthetic libraries. *Curr. Opin. Chem. Biol.* 15, 516–522

17 Houghten, R.A. *et al.* (1999) Mixture-based synthetic combinatorial libraries. *J. Med. Chem.* 42, 3743–3778

18 Houghten, R.A. *et al.* (2008) Strategies for the use of mixture-based synthetic combinatorial libraries: scaffold ranking, direct testing, *in vivo*, and enhanced deconvolution by computational methods. *J. Comb. Chem.* 10, 3–19

19 Kennedy, J.P. *et al.* (2008) Application of combinatorial chemistry science on modern drug discovery. *J. Comb. Chem.* 10, 345–354

20 Pinilla, C. *et al.* (2003) Advances in the use of synthetic combinatorial chemistry: mixture-based libraries. *Nat. Med.* 9, 118–122

21 López-Vallejo, F. *et al.* (2011) Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. *Comb. Chem. High Throughput Screen.* 14, 475–487

22 Singh, N. *et al.* (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.* 49, 1010–1024

23 López-Vallejo, F. *et al.* (2011) Increased diversity of libraries from libraries: chemoinformatic analysis of bis-diazacyclic libraries. *Chem. Biol. Drug Des.* 77, 328–342

24 Owen, J.R. *et al.* (2011) Visualization of molecular fingerprints. *J. Chem. Inf. Model.* 51, 1552–1563

25 Chen, C.Y-C. (2011) TCM Database@Taiwan: the world's largest traditional chinese medicine database for drug screening in silico. *PLoS ONE* 6, 862–868

26 Agrafiotis, D.K. (2001) A constant time algorithm for estimating the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci.* 41, 159–167

27 *Molecular Operating Environment (MOE), version 2010.10*; Chemical Computing Group Inc., Montreal, Quebec, Canada

28 Irwin, J.J. and Shoichet, B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182

29 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D910

30 Bertz, S.H. (1981) The 1st general index of molecular complexity. *J. Am. Chem. Soc.* 103, 3599–3601

31 Barone, R. and Chanon, M. (2001) A new and simple approach to chemical complexity. Application to the synthesis of natural products. *J. Chem. Inf. Comput. Sci.* 41, 269–272

32 Allu, T.K. and Oprea, T.I. (2005) Rapid evaluation of synthetic and molecular complexity for in silico chemistry. *J. Chem. Inf. Model.* 45, 1237–1243

33 Schuffenhauer, A. *et al.* (2006) Relationships between molecular complexity, biological activity, and structural diversity. *J. Chem. Inf. Model.* 46, 525–535

34 *Maestro, version 9.2.* Schrödinger, LLC

35 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623

36 Ganesan, A. (2008) The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.* 12, 306–317

37 Fink, T. and Reymond, J-L. (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* 47, 342–353

38 López-Vallejo, F. *et al.* (2011) Benzotriazoles and indazoles are scaffolds with biological activity against *Entamoeba histolytica*. *J. Biomol. Screening* 16, 862–868

39 Yongye, A. *et al.* (2011) Consensus models of activity landscapes with multiple chemical, conformer and property representations. *J. Chem. Inf. Model.* 51, 1259–1270

40 Medina-Franco, J.L. *et al.* (2012) Consensus models of activity landscapes. In *Statistical Modeling of Molecular Descriptors in QSAR/QSPR* (Matthias, D. *et al.* eds), p. 307L 326, Wiley-VCH

41 Medina-Franco, J.L. *et al.* (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Des.* 70, 393–412

42 Medina-Franco, J.L. *et al.* (2008) Visualization of the chemical space in drug discovery. *Curr. Comput.-Aided Drug Des.* 4, 322–333

43 Akella, L.B. and DeCaprio, D. (2010) Cheminformatics approaches to analyse diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* 14, 325–330

44 Medina-Franco, J.L. *et al.* (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* 49, 477–491

45 Pérez-Villanueva, J. *et al.* (2010) Towards a systematic characterization of the antiprotozoal activity landscape of benzimidazole derivatives. *Biorg. Med. Chem.* 18, 7380–7391

46 Yoo, J. and Medina-Franco, J.L. (2011) Chemoinformatic approaches for inhibitors of DNA methyltransferases: comprehensive characterization of screening libraries. *Comput. Mol. Biosci.* 1, 7–16

47 Ritchie, T.J. *et al.* (2011) The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discov. Today* 16, 164–171