Teaser An assessment of 16 million commercially available compounds, (properties and quality), comparing vendors offerings and how they have evolved to meet modern physiochemical requirements. A selection of 500,000 lead-like compounds for high throughput screening.

Evolution of commercially available compounds for HTS

Dmitriy M. Volochnyuk[1], **Sergey V. Ryabukhin**[2], **Yurii S. Moroz**[3,7], **Olena Savych**[4], **Alexander Chuprina**[3], **Dragos Horvath**[5], **Yuliana Zabolotna**[5], **Alexandre Varnek**[5] and **Duncan B. Judd**[6]

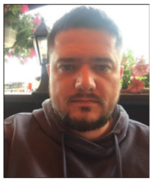
[1] Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine
[2] The Institute of High Technologies, Kyiv National Taras Shevchenko University, 64 Volodymyrska Street, Kyiv 01601, Ukraine
[3] ChemBioCenter, Kyiv National Taras Shevchenko University, 61 Chervonotkatska Street, Kyiv 02094, Ukraine
[4] Institute of Bioorganic Chemistry and Petrochemistry, National Academy of Sciences of Ukraine, Kyiv 02094, Ukraine
[5] Laboratoire de Chemoinformatique, 4, rue B. Pascal, Strasbourg 67081, France
[6] Awridian Ltd, Gunnelswood Road, Stevenage SG1 2FX, UK
[7] Chemspace, ilukstes iela 38-5, Riga, LV-1082, Latvia[8]

Over recent years, an industry of compound suppliers has grown to provide drug discovery with screening compounds: it is estimated that there are over 16 million compounds available from these sources. Here, we review the chemical space covered by suppliers' compound libraries (SCL) in terms of compound physicochemical properties, novelty, diversity, and quality. We examine the feasibility of compiling high-quality vendor-based libraries avoiding complicated, expensive compound management activity, and compare the resulting libraries to the ChEMBL data set. We also consider how vendors have responded to the evolving requirements for drug discovery.

Introduction

A growing body of evidence from clinical outcomes, along with scientific and technological advances over the past decades, has resulted in shaping the strategies of early-stage drug discovery [1]. High-throughput screening (HTS) has evolved since its introduction during the early 1990s. Initially, many pharmaceutical companies were screening hundreds of thousands of compounds against hundreds of targets per year. Today, HTS is often complemented with fragment-based lead discovery (FBLD) [2], encoded library technologies [3], and phenotypic approaches [4] to form a comprehensive screening toolbox and an opportunity to combine knowledge from each

Dmitriy Volochnyuk shares his time as head of the Biologically Active Compounds Department at the Institute of Organic Chemistry of the NAS of Ukraine and as a professor in the Institute of High Technology, Kiev National University. He received his PhD in Organic Chemistry in 2005 and his DSc in organic and organomettalic chemistry in 2011. He has 10+ years' experience in managing chemical outsourcing projects having previously worked in contract research organizations. Dr Volochnyuk is an expert in fluoroorganic, organophosphorus, heterocyclic, combinatorial,and medicinal chemistry. He is also an author on over 120 scientific papers.



Sergey Ryabukhin is an associate professor in the Institute of High Technology, Kiev National Taras Shevchenko University. He was awarded his PhD by Kiev National University in 2008. He has 10+ years' experience in managing combinatorial chemistry departments as well as chemical outsourcing projects having previously worked in contract research organizations. Dr Ryabukhin is an expert in combinatorial methods in organic chemistry, organosilicon, and organoboron chemistry. He is an author on over 50 scientific papers.

Duncan B. Judd is consultant at Awridian Ltd, currently working with a range of organizations including international companies. He is an accomplished medicinal chemist with extensive outsourcing experience and a 39-year proven track record with a blue-chip pharmaceutical company. Duncan has made significant contributions to numerous drug discovery projects, and is cited on many patents and publications. He has extensive outsourcing experience and has published and presented on open innovation in drug discovery, for which he is a strong advocate.

Corresponding author: Judd, D.B. (duncan.b.judd@awridian.co.uk)

[8] chem-space.com.

https://doi.org/10.1016/j.drudis.2018.10.016

approach to successfully identify new lead molecules. Despite these industry-changing 'paradigm shifts', the number of new drugs approved per US$1 billion spent on research and development (R&D) has been halving every 9 years since 1950 [5], and now an estimate of R&D spending per new product exceeds US$2 billion [6].

There has been much speculation in the literature and in the industry around the quality of HTS data derived from random screening, both in terms of sample purity and the physicochemical properties of HTS screening decks. Many consider the classical approaches used by James Whyte Black during the 1960s–1970s [5,7] as being a preferred alternative. However, further studies have clearly shown that HTS is a valuable part of a proven scientific toolkit, and the wide use of the method is essential for the discovery of new chemotypes [8]. Furthermore, the modern HTS is on the 'Plateau of Productivity' phase in the Gartner Hype Cycle, and is now integral in lead discovery along with a combination of different approaches.[*] Moreover, the content, size, and quality of a compound collection used in HTS campaigns are all fundamental to the success of a project: the most advanced screening technologies and the most physiologically relevant assays were thought to be compromised by the low quality of compound collections [9].

At a time when the HTS technology had achieved its 'Peak of Inflated Expectations' and ultra HTS (uHTS) had evolved, it became apparent that large numbers of screening compounds were required. In response, big pharmaceutical companies ('Big Pharma') started enhancing their compound collections, launching file enrichment programs during the early 2000s. However, many of the early combinatorial libraries are now considered far from the optimal chemical space appropriate to initiate a successful drug discovery project [10]. This activity, as well as mergers and acquisitions (M&A), have led to an increase in the size of their respective corporate libraries some to several million compounds: (Pfizer, 4 million [11]; BHC, 2.7 million [12]; AZ, 1.7 million, own collection[*] and 4 million, accessed through collaborations [13]; Novartis, 1.7 million [14]; GSK, 2 million (1.8 million diversity set) [15,16]; Sanofi in collaboration with Evotec, 1.7 million [17]; and Roche, 1.2 million [18]). Moreover, AstraZeneca (AZ) and Bayer have made their collections available to one another for specific HTS campaigns. The overlap for the combined AZ-Bayer set is minimal (~3.5% of the combined library size) and that is attributed to compounds being purchased from chemical vendors [19].

During this period, several companies emerged to meet the demand for more compounds. Furthermore, advances in cheminformatics tools have enabled the design of development libraries, such as the elimination of compounds with inappropriate parameters. Starting from the Lipinski Rule of 5 (Ro5) coined in 1997 [20], many related drug-like criteria have been proposed [21]. In 1999, Teague *et al*. [22] observed that, during optimization, the molecular weight (MW) of the lead molecule increased by 200 Da, whereas logP increased by 0.5–4, which yielded another key concept of lead-likeness. The latter was further developed in 2008 by Pfizer's researchers revealing the Rule of 3/75 (Ro3.75) [23], and the current list of filters is more stringent than the original drug-likeness philosophy. Finally, the Rule of 3 (Ro3) proposed by Congreve *et al*. in 2003 [24] has found a wide application in FBLD.

The aforementioned physicochemical guidelines in combination with the structural filters (reactive compounds [25], REOS [26], PAINS [27], Eli Lilly Rules [28] etc.) and diversity selection methodologies [29,30] have resulted in improvement in the quality of subsequent hits. In addition, the concept of lead-oriented synthesis introduced by Churcher *et al*. in 2012 [31] focused on appropriate chemical space. Despite criticism [32], the current trends in compound set design include filtering of databases before a screening campaign based on chemical structure, calculated properties, rule-based criteria, or the binding efficiency predictions. These filters are routinely combined to form an efficient triage [33] that effectively shrinks chemical space created during the 1990s and early 21st century to make it more appropriate for high-quality HTS. These filtering approaches combined with the synthetic methods have allowed the creation of large drug-like, lead-like, and fragment-like compound collections, which have been aligned with the current paradigm within the industry. Furthermore, it is Big Pharma, with their substantial financial and infrastructure resources, that have developed their collections, which have become 'family jewels' and, therefore, until recently had been inaccessible to those outside the companies, such as academic users and small biotechs.

Despite these challenges, there have been several initiatives to explore HTS outside the pharmaceutical industry [34]. In 2004, the US National Institutes of Health (NIH) and the European Union Innovative Medicines Initiative (EU IMI) both initiated projects to enhance their respective compound collections with the aim of making high-quality compound libraries accessible to the wider scientific community [35]). In the main, these initiatives relied on buying appropriate compounds from chemical vendors. In some cases, pharmaceutical companies have broken new ground by opening their technologies and resources in HTS to selected academics and external institutions [36].

Many outside of Big Pharma have the capabilities to select and order compounds, but the logistics of compound handling tend to get overlooked, such as in the consolidation of libraries from different vendors. Automated production of assay-ready compound plates for screening requires specialized formatting facilities, which could cost US$7 million [37], thus being unaffordable for smaller organizations. There are two approaches to overcome the above-mentioned issues: (i) ordering from companies that specialize in consolidating and formatting libraries; or (ii) purchasing a preformatted library ready to use from a limited number of vendors. To the best of our knowledge, there is only one study evaluating SCL from the user's standpoint, published in 2013 [38]. The main conclusion of that study was that the available screening compounds appeared small and was, at that time, represented by fewer than 350 000 compounds [38].

Despite several analyses of the chemical space covered by SCL published in 2004 [39], 2005 [40], 2006 [41], and 2015 [42]

[*] Mayr, L.M. and Wigglesworth, M. High-Throughput Screening: Challenges & Opportunities 8th ELRIG Drug Discovery Conference Manchester/UK 2014, September 2–3.http://elrig.org/downloads/dd14/20140904_Mayr_ELRIG2014.pdf.

[*] Mayr, L.M. and Wigglesworth, M. High-Throughput Screening: Challenges & Opportunities 8th ELRIG Drug Discovery Conference Manchester/UK 2014, September 2–3. http://elrig.org/downloads/dd14/20140904_Mayr_ELRIG2014.pdf.

(including our studies in 2011 [43], 2012 [44]) the question remains as to whether the available purchasable chemical space could enable the creation of a high-quality compound library for HTS projects that are comparable to Big Pharma's proprietary repositories. Thus, the goals of present study were: (i) to provide a critical view from a user's standpoint on the existing SCL offerings and to clarify whether they are comparable to Big Pharma's collections in terms of 'compound novelty, diversity, and quality'; (ii) to examine the feasibility of facile compiling a high-quality compound library via a limited number of vendors, hence avoiding complicated and expensive compound management; and (iii) to include in the analysis a comparison of vendor's offerings.

A preferable supplier can be identified using the following criteria: (i) cost effective and timely delivery of quality compounds; (ii) a wide range of compounds with appropriate physical and chemical properties [Ro5, Ro3, with limited undesirable functionality: no 'PAINS', stable, no hot functionality (except covalent libraries)]; (iii) possibility of provision of analogs for hit follow-up in a time- and cost-effective manner (except for NP and metabolites); (iv) the SCL represents numerous and/or original chemotypes, as defined by Bemis-Murcko, Tanimoto, and so on; and (v) the vendor updates the catalog regularly, and is clear about pricing with transparent and prompt communication throughout the purchasing process.

However, a comprehensive analysis of the vendors fulfilling the above-mentioned criteria limited to the information extractable from open sources because most companies prefer not to share their analysis of various vendors. Therefore, we used cheminformatic approaches to compare the SCLs found in open platforms. As an indirect indicator of the vendor's activity in the field, we analyzed the dynamics of the reshaping and growth of their collections over a set time period.

Results and discussion

Collection of the data and characteristics of the data sets

The starting point of the current study was the creation of the chemical space covered by purchasable screening compounds using the ZINC database.[†] To create this space, we performed standardization of SMILES for all the sets involved in our search using RDKit nodes for the KNIME analytics platform.[‡] This space was defined as the union of standardized SMILES strings of all sets prepared, as mentioned earlier. Duplicates were deleted from the newly created large set. After removal of duplicates, the standardized space comprised 16 902 208 unique structures, including stereoisomers (all stereochemical features mentioned by vendors were included). As illustrated by Fig. 1 and Fig. S1 in the Supplementary information online, the impact of the vendors on the space differed significantly by the number of structures as well as by percentage of unique compounds. From 33 sets, eight showed a high fraction of unique compounds (80% and more): Abamachem, AnalytiCon Discovery, BCH Research, Enamine, FCH Group, Intermed, Selenachem, and UORSY; all these sets, except for AnalytiCon Discovery, contained more than 1 million molecules. Eight sets contained a medium number of unique compounds (40–80%), and three of these sets were of 1 million or more molecules (Asischem, ChemBridge, and ChemDiv). Even though Princeton Biomolecular Research and Vitas-M contained 1.2 million and 1.4 million molecules, respectively, the fraction of unique compounds was <10% for both databases.

Compound-level analysis (for the 16 902 208 set)

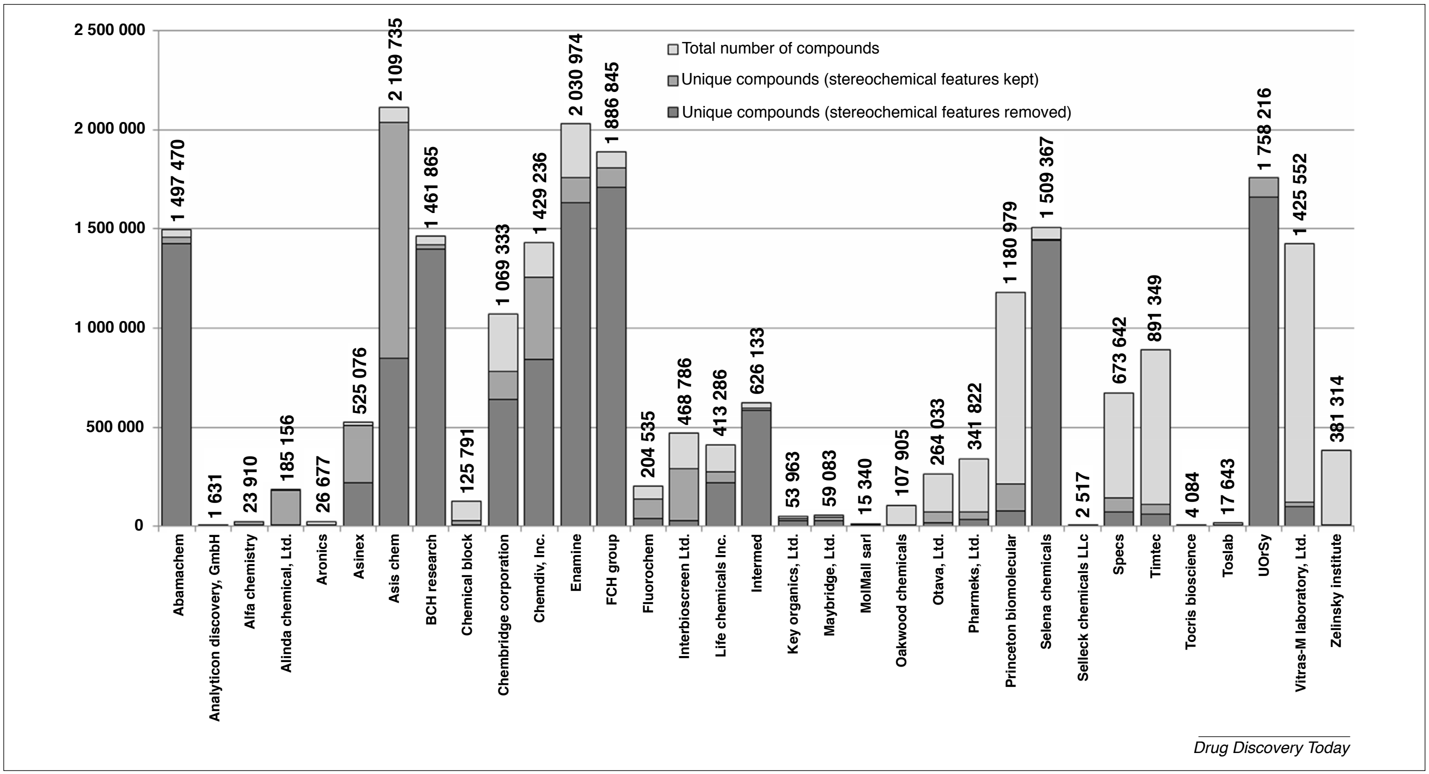
For the preliminary evaluation of the quality of the purchasable chemical space as well as the set from each vendor, ten selected molecular properties were chosen: MW, logP, heavy atom (HA) count, number of hydrogen bond donors (HBDs), number of hydrogen bond acceptors (HBAs), polar surface area (PSA), number of rotatable bonds (ROTB), Fsp^3, number of rings, and number of aromatic rings. The mean values of these parameters are detailed in Table 1. We also compared these values with the corresponding data from our previous analysis from 2011 [44]. The data showed that, during the past 7 years, the mean values of the six parameters mentioned in our previous paper significantly shifted from drug-likeness to lead-likeness, which accords with general trends of the screening libraries criteria. The mean MW ($\Delta = -26$), logP ($\Delta = -0.67$), PSA ($\Delta = -22.4$), HBA ($\Delta = -1.57$), and ROTB ($\Delta = -0.47$) significantly decreased whereas mean HBD slightly increased ($\Delta = +0.20$). Given the impact of historical compounds from the collections of the main players in the field, which strongly affected the mean values, we compared the mean value of the compounds appearing from 2010 to 2017[§]; encouragingly, these results were the closest to the lead-oriented synthesis concept.[¶] Comparison of the characteristics of the 'new compounds' set from the SCL 2010–2017 with the European Lead Factory[**] (ELF) library [45] (mean values, calculated on the basis of the data from two publications [46,47] showed that parameters of the SCL 2010–2017[‡‡] set were stricter [mean MW (SCL 2011–2017) = 340, MW (ELF) = 425; logP (SCL 2011–2017) = 2.38, logP (ELF) = 3.1] and closer to DrugBank mean values (MW = 315, logP = 2.4) than were those of ELF (Table 1).

In addition to the mean values, we analyzed the distribution of the aforementioned parameters for all purchasable chemical space as well as for each vendor collection (for exact information on vendors, see mmc3.xlsx in the Supplementary information online). To simplify the visualization of the distributions of each vendor compared with the space, we divided the distributions into several areas. The distributions that were difficult to assign to the areas are marked in the figures as 'outliers'. The representative examples of such simplifications are shown in Fig. S2 in the Supplementary information online.

For example, in reviewing the results for MW, we believe there are three general categories of suppliers: Area 1: ten distribution curves (Abama Chemicals, BCH Research, Intermed Chemicals,

[†] Database released on March 2017 at http://zinc.docking.org/ was used.
[‡] www.knime.com/knime-analytics-platform.

[§] Comparison of the mean value of the compounds appearing during the period 2010–2017 was estimated by simple math approximation using the formula: <X>(2010)*F(cpd, 2010 in 2017) + <X>(2010–2017)*F(cpd, 2010–2017) = <X>(2017), where <X> – median values of the compounds number, and F(cpd) – fration of compounds of 'old' and 'new' appearance in the database of 2017.
[¶] GSK Novel Synthetic Methods Symposium, Stevenage, 24–25th May 2010.
[**] www.europeanleadfactory.eu/.
[‡‡] SCL 2010–2017: screening compounds libraries from the vendors for 2010–2017.



Drug Discovery Today

FIGURE 1

The chemical space of purchasable screening compounds represented by vendors.

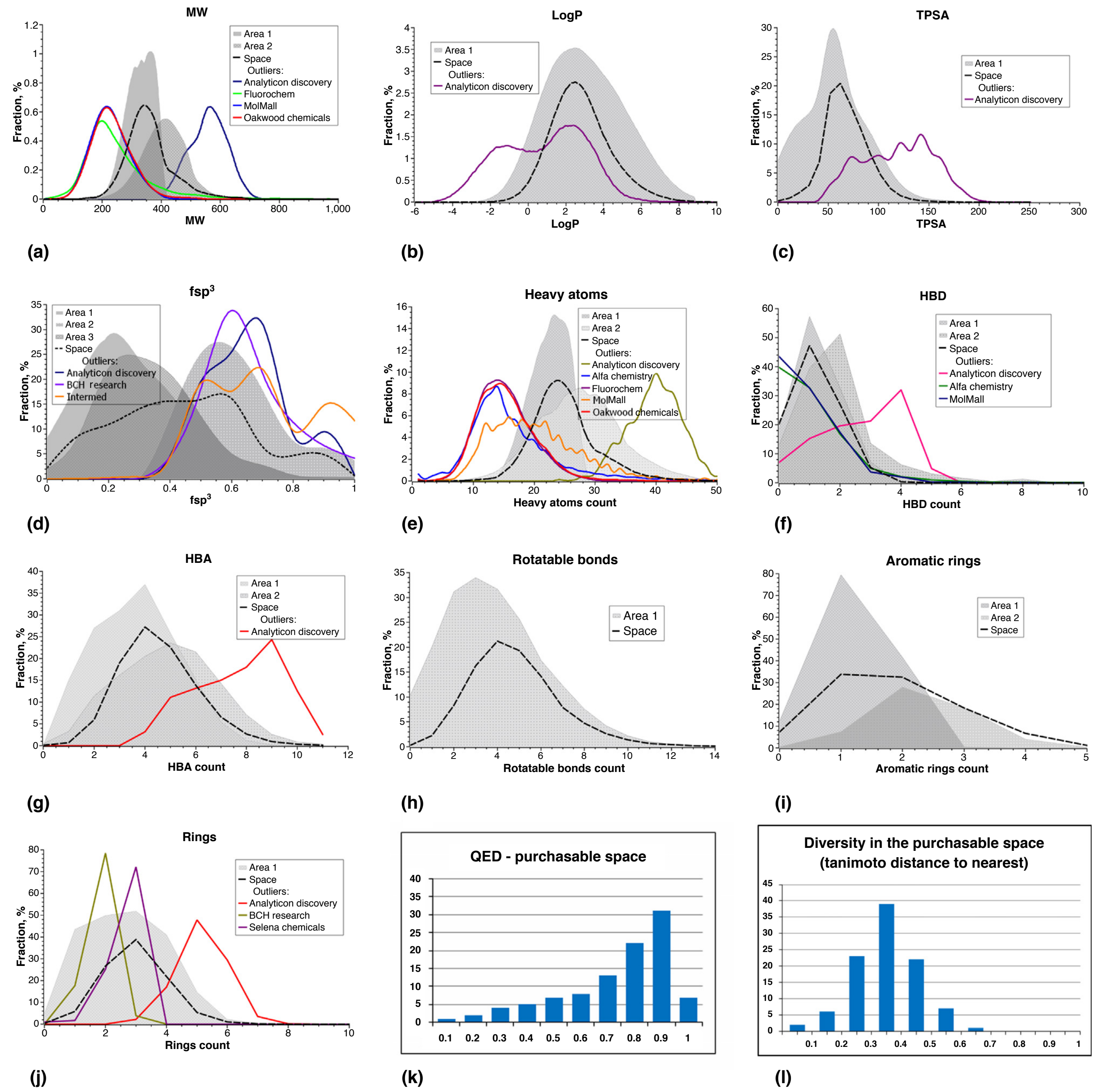
TABLE 1

Mean values of selected molecular properties of the purchasable chemical space in 2010, 2017, and the ELF library

Parameter (X)	2010	2017	Δ<X> (2010–2017)	<X>Δ (2010–2017)	ELF
MW	388.82	362.49	−26.33	339.59	425
logP	3.64	2.96	−0.67	2.38	3.1
Fsp^3	–	0.40	–	–	0.4
tPSA	94.23	71.84	−22.39	52.38	91
Heavy atoms	–	25.11	–	–	–
HBA	6.18	4.61	−1.57	3.25	–
HBD	0.96	1.16	0.20	1.33	–
ROTB	5.28	4.82	−0.47	4.41	–
Rings	–	3.02	–	–	–
Aromatic rings	–	2.03	–	–	–

Selena Chemicals, ChemBridge, Enamine, FCH Group, Key Organics, Maybridge, and UORSY) have narrow peaks with maxima between 300 and 400 Da; Area 2: 18 distribution curves (Alinda Chemicals, Asinex, ChemDiv, Aronis, Asischem, Chemical Block, InterBioScreen, Life Chemicals, Otava Chemicals, Pharmeks, Princeton Biomolecular Research, Selleck Chemicals, Specs, Timtec, Tocris, Toslab, Vitas-M Laboratory, and Zelinsky Institute) have wide peaks with a vertex at 400 Da. By contrast, five curves (AnalytiCon Discovery, Alfa Chemistry, Fluorochem, MolMall, and Oakwood Chemicals) were left as is and recognized as 'outliers'. Another representative example of simplification is the distribution of HBD number given in Fig. S2 in the Supplementary information online. Using such an approach, distributions of all above-mentioned parameters were calculated and are shown in Fig. 2.

Among the compound suppliers, AnalytiCon Discovery, Alfa Chemistry, Fluorochem, MolMall,and Oakwood Chemicals were identified as 'frequent outliers'. The main reason for this rests on the main business activity of these companies. AnalytiCon Discovery specializes on natural products and macrocycles; Fluorochem and Oakwood Chemicals are widely known as suppliers of building blocks and reagents; Alfa Chemistry is a contract research organization; and MolMall is a small collection of samples from different sources. All these companies are not 'classical' producers of the compounds for HTS. However, despite differences in the parameter distributions of each vendor, the cumulative distribu-



Drug Discovery Today

FIGURE 2

Distribution of the selected molecular properties of the purchasable chemical space with 'vendor areas' and outliers together with QED and ECFP4-based Tanimoto similarity profiles for the space. Please see main text for definitions of abbreviations.

tions of the parameters of purchasable space have one peak, which is usual for screening collection. An exception is the Fsp^3 distribution, which has a more complex character, unlike the curves of vendors. In this case, old historical collections and the newly synthesized compounds have significantly different Fsp^3 parameter values (Fig. S3.01 in the Supplementary information online). Nevertheless, the quantitative estimate of drug-likeness (QED) [48] histogram for the purchasable space revealed the quality of the compounds based on this parameter (see mmc4.xlsx in the Supplementary information online). The maximum QED accounted for 0.8–0.9 (Fig. 2).

The chemical diversity of the space and vendor collections was analyzed by ECFP4-based Tanimoto similarity of each compound with its nearest neighbor (for all vendors, see Figs. S3.01–3.10 in the Supplementary information online). For the purchasable space, the corresponding histogram is shown in Fig. 2. Its profile demonstrates a diverse set with a mean Tanimoto distance to nearest neighbor of 0.3. Notably, Tanimoto diversity for the purchasable space is worse than the data announced for the Joint European Compound Library (JECL): a mean Tanimoto distance of 0.4 to the nearest neighbor [47]. Deeper analysis of the contribution of each supplier to a joint diversity of the space showed that some sets represent completely different areas of chemical space, whereas others have a significant overlap. As an example, the AnalytiCon set has a low internal diversity but occupies a significantly different space from other vendors (median Tanimoto distance 0.18 within the set, but 0.55 against the full space).By contrast, the Vitas-M set is narrowly distributed (median Tanimoto distance 0.24 in set, and median Tanimoto distance in comparison with the full space 0.29). Selleck set had high internal diversity and differed from other vendors (median Tanimoto distance was 0.56 in the set but median Tanimoto distance in comparison with full space was 0.46). The corresponding histograms are shown in Figs. S4.01–4.33 in the Supplementary information online.

For the 3D-shape analysis of the purchasable space as well as vendor sets, the Plane of Best Fit (PBF) – Principal Moments of Inertia (PMI) approach was used [49]. Generation of coordinates and geometry optimization (mmff94, 100 iterations per molecule) along with subsequent PMI and PBF calculations, were performed using RDKit. Density plots were built in R Statistics using the hexbin package; the plot for the complete space is shown in Fig. 3a.

According to the PBF = 0.6 and NPRsum = 1.1 cut-off filter, the number of ‘out-of-plane molecules’ in purchasable space was 8 668 016 (51%). The same calculations for each vendor set (Figs. S5.01–5.34 in the Supplementary information online) revealed that the fraction of compounds passed through the filter fell in a range of 36–47%, with exception for AnalytiCon (76%), Alfa Chemistry (20%), Alinda (33%), Aronis (26%), Fluorochem (20%), and Oakwood (21%).

Scaffold level analysis

Bemis–Murcko loose frameworks (scaffolds) analysis [50] was used to evaluate the 2D shape and topology of the compounds in the purchasable space and each vendor collection (Figs. S6.01–6.33 in the supplementary information online). This analysis gave 2 886 942 unique frameworks representing purchasable space. Cumulative scaffold frequency plots (CSFP) [51] were built for the space and vendor collections. As in the case of compound-level analysis, the main ‘area’ and outliers were identified. This time, UORSY appeared in outliers, the CSFP of which was close to those of Binding DB and DrugBank (Fig. 3b).

Equal distributions of compounds across molecular scaffolds were found in the Selleck and Tocris collections, mainly because of the main profiles of these companies: Selleck and Tocris are worldwide recognized suppliers of reference compounds, which are usually used as standards in different screening assays as well as in biomedical investigations. Our data are in slight disagreement with a recently published analysis of the libraries of the main players [52], but the CSFP curves obtained therein fit the ‘area’ in Fig. 3b.

SCL changes analysis

An important factor in the choice of compound vendor is the viability of the sample resupply and further opportunity for the hit follow-up support [38]. Another is how vendors have responded to the desire for more lead-like compounds. To address these issues, we focused on companies active in this field. Promotional materials of those companies do not give a true picture; therefore, we evaluated such companies by comparing the results of analyses carried out in 2010 and in the current paper. Initially, differences in compound numbers in collections were plotted (Fig. 4). Some vendors presented in 2010 (AMRI, ComGenex, Tripos, ART-CHEM, Nanosyn, SALOR, IVK Laboratories, ChemStar, Ufark, and Spectrum) were absent in 2017 in ZINC. Some of these companies had been sold (e.g., ComGenex[§§] or Tripos[¶¶]), whereas others, such as AMRI and Nanosyn, provided integrated MedChem solutions using in-house libraries. Moreover, all these vendors were not active participants in screening compound production. In 2017, 14 new vendors were present: AnalytiCon, Selleck, Tocris, MolMall, Alfa Chemistry, Aronis, Chemical Block, Alinda, Zelinsky Institute, Intermed, BCH research, Abamachem, Selena Chemicals, and FCH Group. The libraries of the latter four contain more than 1 million unique diverse compounds with good PhysChem properties (see mmc2.xlsx in the Supplementary information online), proving their activity on screening compounds market.

The vendors referred to in the analysis of 2010 could be divided into several categories (i) outgoing from the market: TOSLab, Maybridge (−9070 cpds/33% and −10 779 cpds/15%) and InterBioScreen (almost no changes in 7 years); (ii) not growing: Key Organics, Asinex (+6307 cpds/13% and +67 234 cpds/15%, respectively: <15% increase of the library size without significant qualitative changes) and Life Chemicals (−12 849 cpds/3% decrease in size but with considerable qualitative changes, $\Delta\langle MW\rangle$ (2010–2017) = −26; $\Delta\langle \log P\rangle$ (2010–2017) = −0.36); (iii) growing: ChemBridge (+328 157 cpds/44%); (iv) extremely growing: ChemDiv (+643 496 cpds/82%), Enamine (+809 017 cpds/66%), UORSY (+963 219 cpds/120%), and Asis Chem (+2 076 986 cpds/634%); and (v) companies that proposed building blocks in mg quantities: Oakwood and FluoroChem. The latest category appears to be growing, with seven vendors currently included: Otava Chemicals, Pharmeks, TimTec, Specs, Princeton Biomolecular Research, Vitas-M, and Zelinsky Institute. Despite the increased number of compounds, these collections include a few unique structures (Fig. S2 in the Supplementary information online). We carried out further analysis of cross-overlapping of these collections (Table 2) that revealed that the libraries of five vendors (Otava Chemicals, TimTec, Princeton, Vitas-M, and Zelinsky Institute) substantially overlap, which is an indirect proof of common source of these compounds and questions the production ability of these compounds.

At a cursory glance, the space was sufficiently diverse and covered significant PhysChem parameters for most screening campaigns; thus, it could deliver an appropriate HTS set. To verify this statement, several case studies were performed.

Case study: an ‘ideal’ million

Among the variety of screening paradigms that exist to identify hits [53], we chose an example comprising building a compound set to screen against a novel target with an unknown structure, with few known active chemotypes, or without existing small-molecule modulators. In this case, HTS is the method of choice for its potential to identify quality leads because it does not require

[§§] https://bbj.hu/business/albany-molecular-closes-comgenex-acquisition_9580.

[¶¶] www.thepharmaletter.com/article/tripos-to-sell-drug-discovery-business.

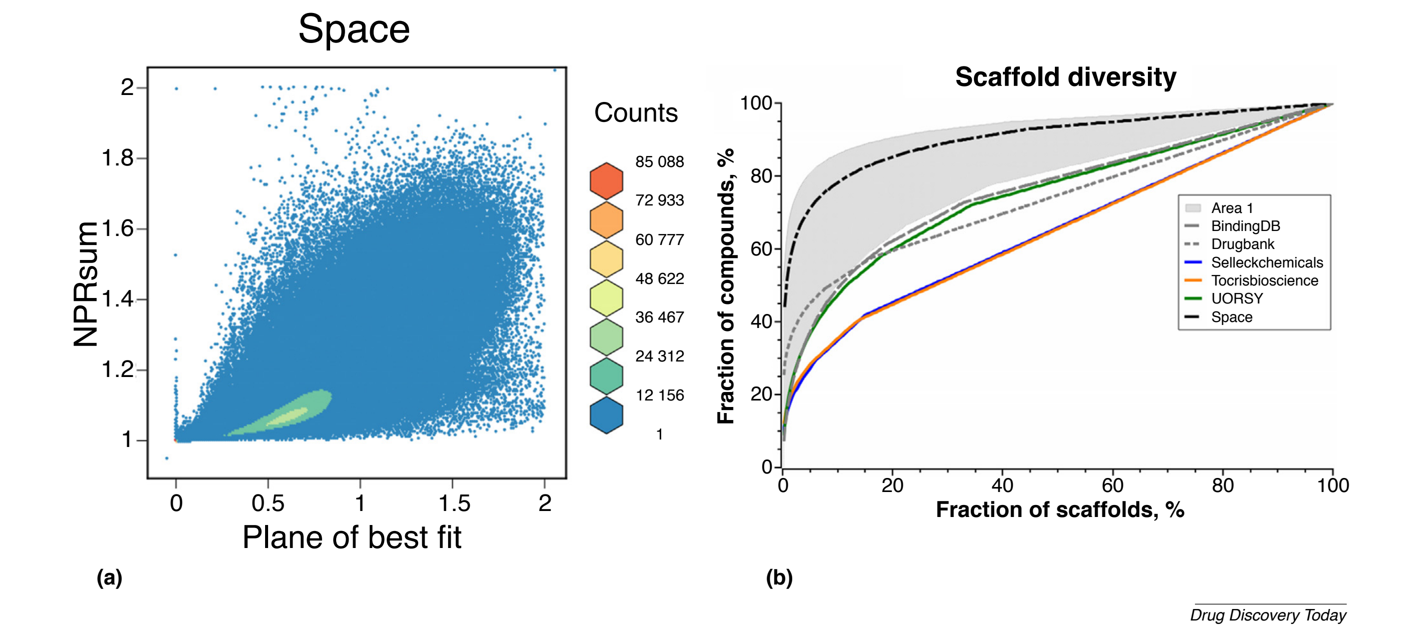


FIGURE 3

3D shape and scaffold diversity of the purchasable chemical space. **(a)** Density plot of Plane of Best Fit (PBF) score versus the sum of normalized principal moments of inertia (NPR). **(b)** Cumulative Scaffold Frequency Plots of the scaffold with 'vendor areas' and outliers compared with Binding DB and DrugBank.

information about the target. However, determining the optimal size of such a screening deck is problematic. Several studies have addressed this question but the optimal size of a screening collection [54,55] has remained undefined and varied.

The technical possibilities of modern HTS are almost unlimited. Nowadays, 384-well microtitre plates are the 'golden standard,' whereas 1536-well plates are increasing in popularity, and even 3456-well microtitre plates are used in some projects. Throughputs of ≥100 000 compounds screened per day are routine in leading HTS practitioner laboratories using *in vitro* biochemical, functional cell-based, reporter gene, and phenotypic assays [56]. According to reports on screening campaigns, the number of compounds used in an 'all-or-nothing' screening mode ranges from 50 000 to 1 500 000 [57]: a maximum mean value of 800 000 compounds per screen was reported in 2003, whereas this number had decreased to 500 000 in 2009 [58]. Despite a low true positive hit rate (<1% in 2010 [59]), in 2018, AZ concluded that increasing success could be achieved by gaining access to as many compounds as possible [13]. Moreover, choosing the 'relevant region' of the chemical space [28] would decrease further attrition and increase the true positive hit rate [60]. Support for the trend to use several million screening compound campaigns is the multiplexing of more than one compound per well during primary HTS to increase the capacity without compromising screening quality [61]. Thus, we assembled a screening deck of 1 million lead-like compounds, based on 50 000 scaffolds with 20 representatives each, belonging to clusters that were as diverse as possible for the first case study. We limited the number of the compounds to eliminate the molecular redundancy [62], but left a sufficient number of compounds per cluster to efficiently identify latent hit series and rapid preliminary structure–activity relationships (SARs), and to avoid any singletons [63]. Currently, there is controversy over the optimal size of compounds per cluster per scaffold. The first papers discussing the issue were published in early 2000, although their conclusions varied from 10 [64] to 50–100 [65] compounds per scaffold. By contrast, the 'Open Scaffolds' collection from Compounds Australia was build with ≤30 SAR-meaningful compounds per scaffold (avarage value 28) [66]. Nevertheless, a series of 5–20 compounds was most frequently used by Pfizer [67] during plate-based diversity subset generation 2 (PBDS2). Therefore, we selected a model value of 20 compounds per scaffold, also in agreement with the opinion of Bostwick.[***] For comparison, we also ran the study using 50 compounds per scaffold.

To build an 'ideal million' set, we initially subjected the purchasable chemical space of 16 902 208 compounds to structural filtering against PAINS (despite recent criticism [68], the filters are routinely used) and toxicology/reactive Eli Lilly Rules [27,28], which selected 15 968 338 compounds. Further application of the lead-likeness [69] and Ro3/75 [23] criteria resulted in two spaces with 6 544 044 and 3 705 803 compounds, respectively. Bemis–Murcko loose framework analysis of the sets gave only 39 101 and 22 162 scaffolds bearing more than 20 compounds per scaffold and 13 156 and 8006 scaffolds bearing more than 50 compounds per scaffold (Table 3). Given that the first model ideal million set (20 compounds per scaffold) would require 50 000 scaffolds and fewer than this were available from drug-like space, we targeted a 0.5 million set represented by 25 000 scaffolds with 20 compounds per scaffold and used the 6 544 044 set. From this set of 39 101 scaffolds, we extracted 25 000 of the most diverse using the MaxMin algorithm [70]. If the scaffolds had more than 20 compounds in the lead-like space, we selected the 20 most diverse structures using the above-mentioned MaxMin algorithm for compounds from overpopulated scaffolds [70]. In this 'ideal half million', the unique structures from all 33

[***] www.uab.edu/medicine/adda/images/BostwickHTS.pdf.

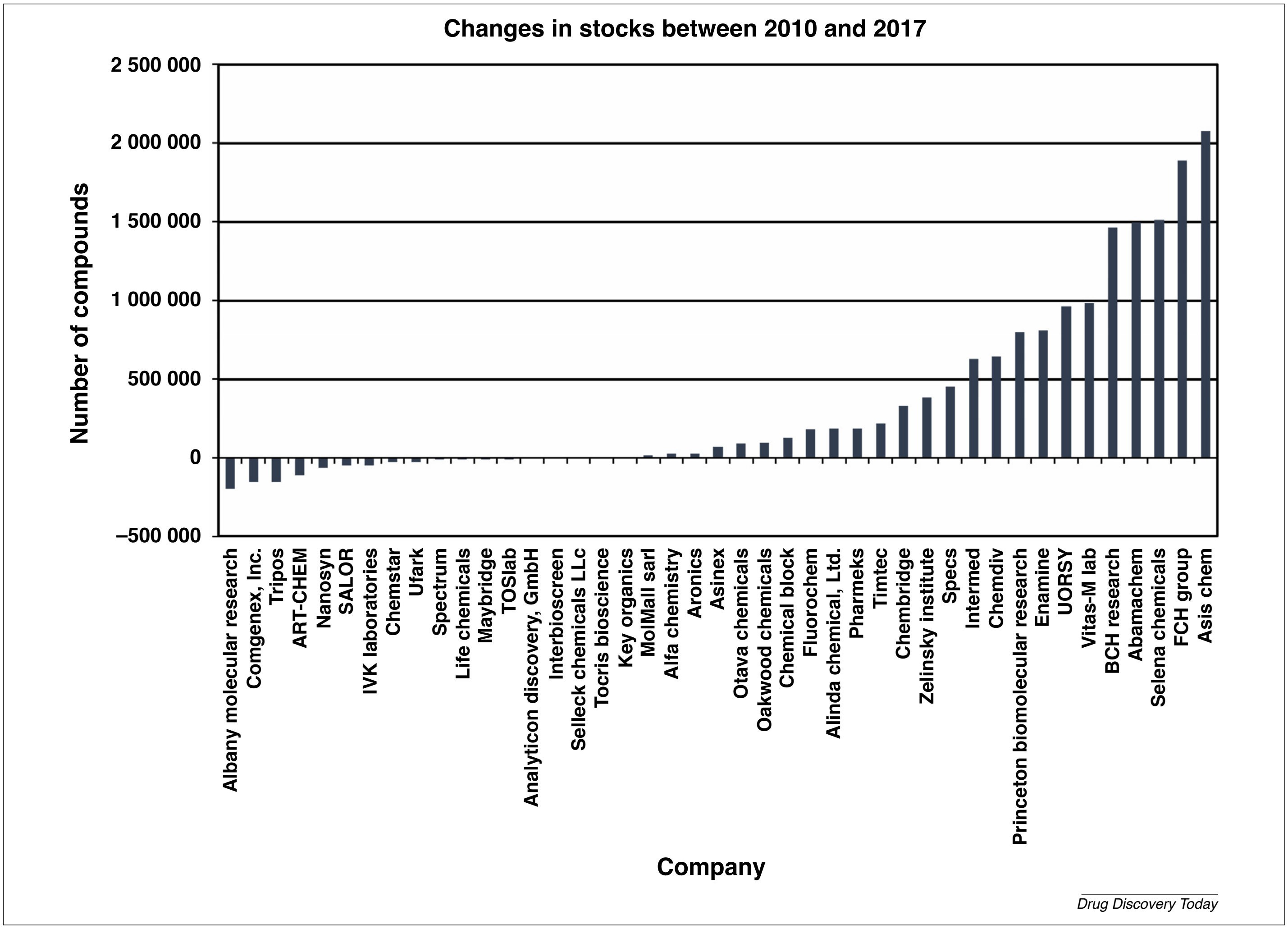


FIGURE 4

Changes in suppliers' compound libraries (SCL) size from 2010 to 2017.

TABLE 2

Cross-overlapping of the 'seemingly growing' vendors[a,b]

	Otava	Pharmek	Princeton	Specs	Timtec	Vitas-M	Zelinsky
Otava Chemicals		9	14	4	11	12	2
Pharmek	12		15	4	4	16	3
Princeton Biomolecular Research	62	52		37	51	64	86
Specs	10	7	21		24	21	33
Timtec	36	10	38	32		32	80
Vitas-M	66	69	77	44	51		84
Zelinsky Institute	3	3	28	19	34	23	

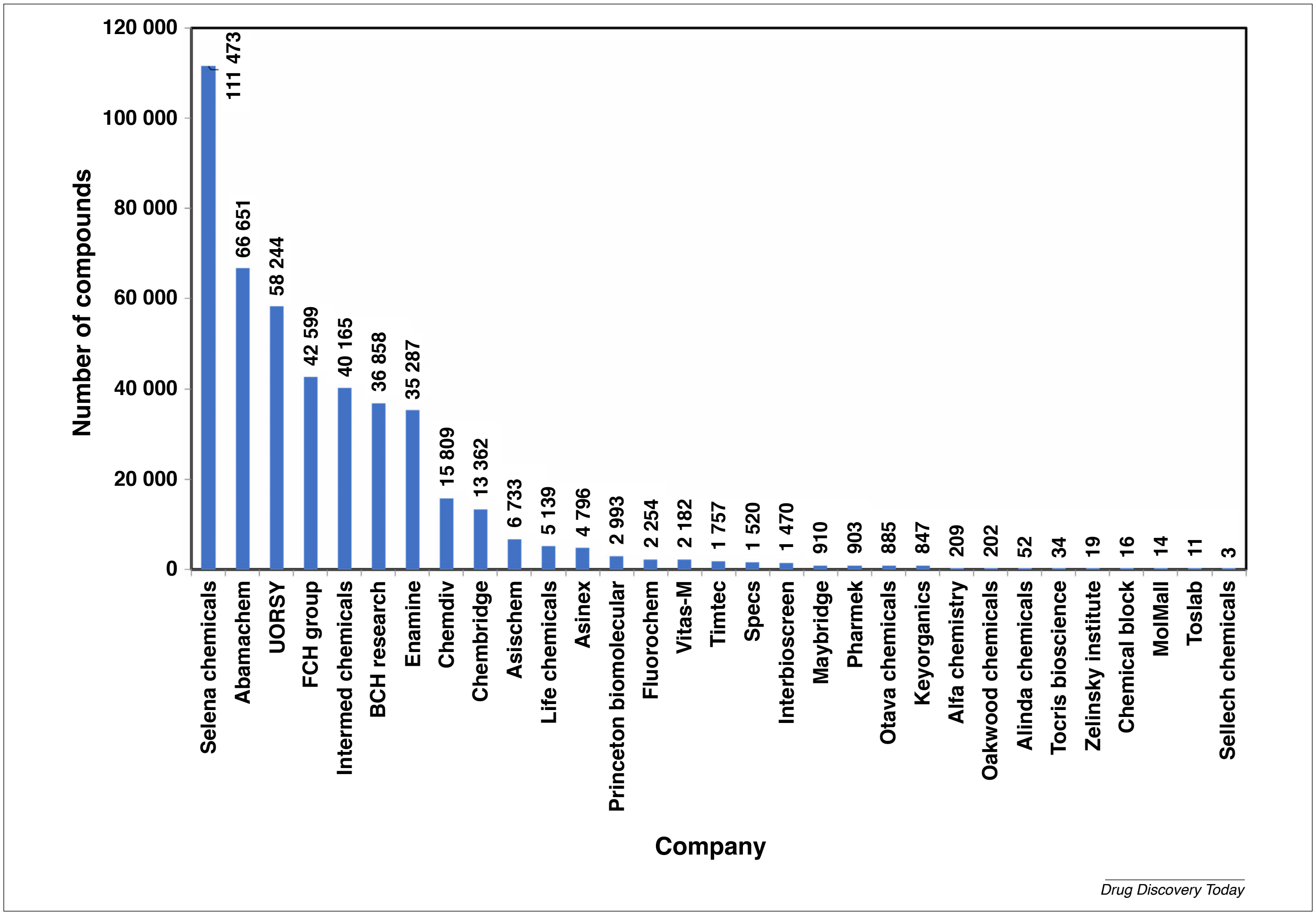
[a] The fraction (%) of vendor 1 compounds <in column> that are present in the vendor 2 database <in string>.
[b] XXXXX.

suppliers were presented, although the contribution of each supplier varied significantly (Fig. 5). To simplify compound management (as mentioned in the Introduction), we studied the dependence of the quality of the selected set on the number of suppliers. Based on the obtained data (Fig. 5), we selected 12, six, and three suppliers that contributed the most. The above-mentioned procedure for the 'ideal half million' selection was applied for the chemical space covered by these 12, six, and three suppliers, respectively. For the 12 and six suppliers, the generated space contained 0.5 million compounds, whereas for three suppliers, the size of the space decreased to 384 520 compounds based on 19 226 scaffolds. We then compared these three spaces with the initial space from 33 suppliers at the compound and scaffold levels. Diversity at the compound level as well as QED were similar for all the three spaces (Figs. S7.01 and S7.02 in the Supplementary information online). However, a similar analysis at the scaffold level showed a significant decrease in diversity from the 33 to the three supplier sets (Fig. 7a).

TABLE 3

Bemis–Murcko loose framework scaffolding of the prefiltered chemical space covering 15 968 338 compounds

Number of structures per scaffold	Number of scaffolds			Resulting number of structures		
	Lead-like	3/75 rule	Drug-like	Lead-like	3/75 rule	Drug-like
≥50	13 156	8006	28 815	657 800	400 300	1 440 750
≥20	39 101	22 162	78 756	782 020	443 240	1 575 120
≥10	88 155	47 375	169 072	881 550	473 750	1 690 720
≥5	198 649	102 369	365 419	993 245	511 845	1 827 095
Total number of structures	6 544 044	3 705 803	14 191 016			



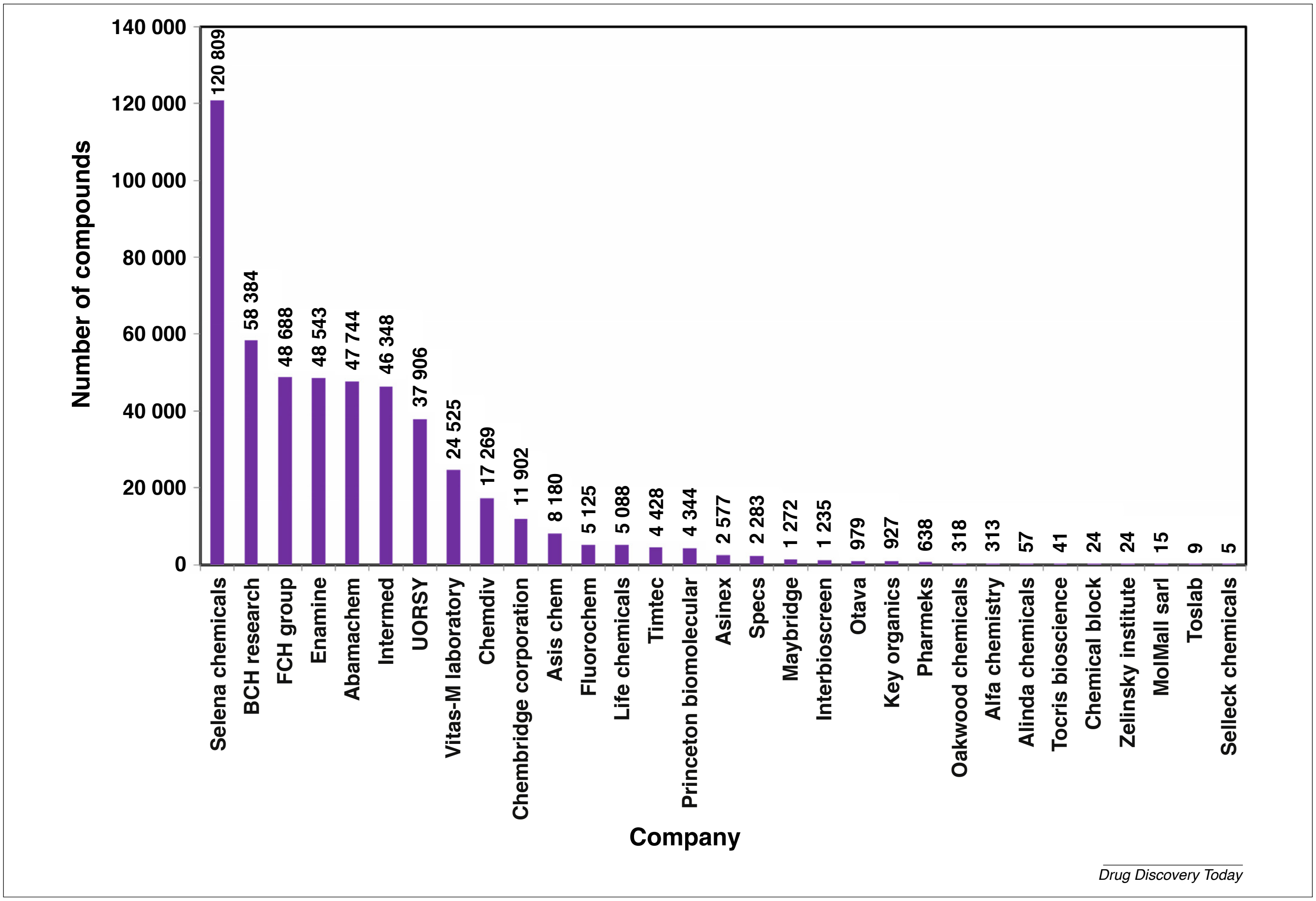
Drug Discovery Today

FIGURE 5

The contribution of vendors to the ‘ideal half million’ set.

The second model ‘ideal million’ set (50 compounds per scaffold) was collected using the above-mentioned algorithm. Similarly, for 50 compounds per scaffold set, only an ‘ideal half million’ could be generated. However, in contrast to the previous analysis, this resulted in a different level of contribution from each supplier (Fig. 6). We also analyzed the contribution from the top 12, six, and three suppliers. For 12 suppliers, applying the algorithm resulted in a 0.5 million compound set, whereas for six and three suppliers, the size of the r sets was 494 450 and 306 200 compounds based on 9889 and 6124 scaffolds, respectively. Compared with the 20 compounds per scaffold set analysis, decreasing the number of suppliers did not significantly influence the Tanimoto diversity at the compound level or the QED (Figs. S7.03 and S7.04 in the Supplementary information online), but did significantly decreased diversity at the scaffold level (Fig. 7b). In general, the comparison of the two sets (20 and 50 compounds per scaffolds) showed that the 50 compounds per scaffold set was significantly less diverse at the scaffold level. Therefore, the 20 compounds per scaffold set with the number of suppliers reduced to six or three subsets would be a pragmatic way to build a useful set of compounds for HTS screening campaigns based on compounds purchased from commercial sources.

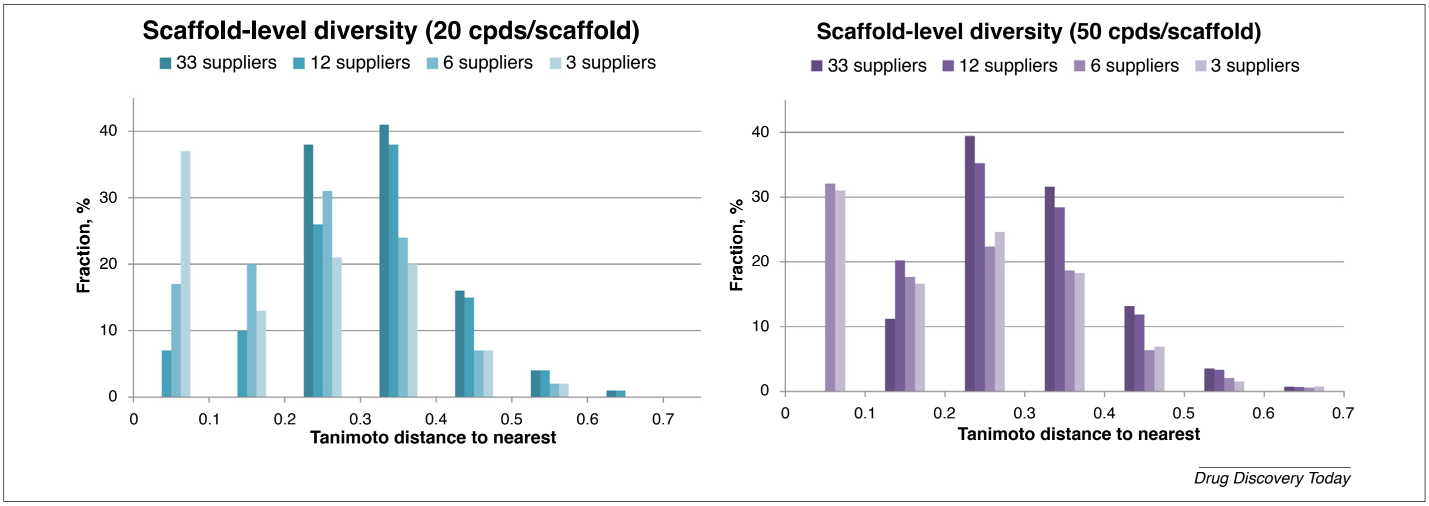
The last step of our investigation was to compare the results from 33, 12, six, and three suppliers (for the libraries bearing 20 compounds per scaffold). For this purpose, we utilized the recently developed Generative Topographic Mapping (GTM) [71,72] be-



Drug Discovery Today

FIGURE 6

The contribution of vendors to the ‘ideal half million 50 compounds per scaffold’ set.



Drug Discovery Today

FIGURE 7

Comparison of the scaffold diversity of the libraries collected from 33, 12, six, and three suppliers. **(a)** For 20 compounds per scaffold set; **(b)** For 50 compounds per scaffold set.

cause it is considered the most efficient tool among the published methods for multiple descriptor chemical space comparison. The 1.5-million ChEMBL compound data set was used as a reference database. The four compound sets corresponded to three, six, 12, and 33 suppliers. These were mapped against the background of ChEMBL compounds, with blue zones corresponding to chemical space areas dominated by supplier compounds, versus dark–red zones containing (almost) exclusively ChEMBL compounds, after applying Bayesian normalization to compensate for the initial imbalance of set size (300 000–500 000 for supplier sets, versus 1.5-million ChEMBL compounds). Intermediate colors, from light red through yellow and green, corresponded to chemical space zones in which supplier and ChEMBL compounds mingled (increasing relative density of supplier compounds corresponding to a 'blue shift'). Three maps were built on the basis of the aforementioned principles, shown in Fig. 8.

Map #1 was based on ISIDA [73] force-field-type colored atom sequence counts acting as molecular descriptors. The force field types assigned to atoms (the CVFF forcefield typing rules were applied) were specific to their chemical environment and, therefore, this class of ISIDA fragment descriptors provides a fine-grained analysis of chemical space. The three-supplier set dominated the 'north-eastern' chemical space zone, clearly separated by a ChEMBL-dominated central part from some secondary 'islands' in both the north-western and south-eastern regions. Increasing the number of suppliers resulted in a gradually growth of overlap with the ChEMBL set, by embracing more compounds in the central area, which remained dominated by ChEMBL compounds while also starting to be populated by supplier molecules. The extent of library overlaps, calculated as the Tanimoto score of the mean vectors responsible from the supplier and ChEMBL libraries, respectively, increased from 0.28 (three suppliers) to 0.33 (six suppliers) to 0.42 (12 suppliers) and remained constant when all suppliers were considered.

Map#2 relied on ISIDA pharmacophore-type colored atom sequence count descriptors (i.e., it monitors pharmacophore pattern diversity). Therefore, it ignored the precise chemical nature of the atoms, rendered as hydrophobes, aromatics, HBA and HBD, cations, and anions, respectively. The three-supplier set provided significant coverage of the chemical space, with the only ChEMBL-dominated area close to the 'south pole' of the map. The addition of compounds from further suppliers gradually filled this initial diversity hole. The degree of library overlap was generally higher than in the more fine-grained map #1, and gradually increased from 0.51 (three suppliers) to 0.54 (six suppliers), 0.63 (12 suppliers),and 0.65 (all suppliers).

Map#3 was based on plain ISIDA atom sequence counts. Similar to map#1, it also focused on chemical constitution and connectivity patterns, but was less fine-grained than the latter; thus, the libraries are strongly overlap. On this map, the three-supplier library appears as a core collection that gradually expands (in particular, into the north-west and south-west regions) as

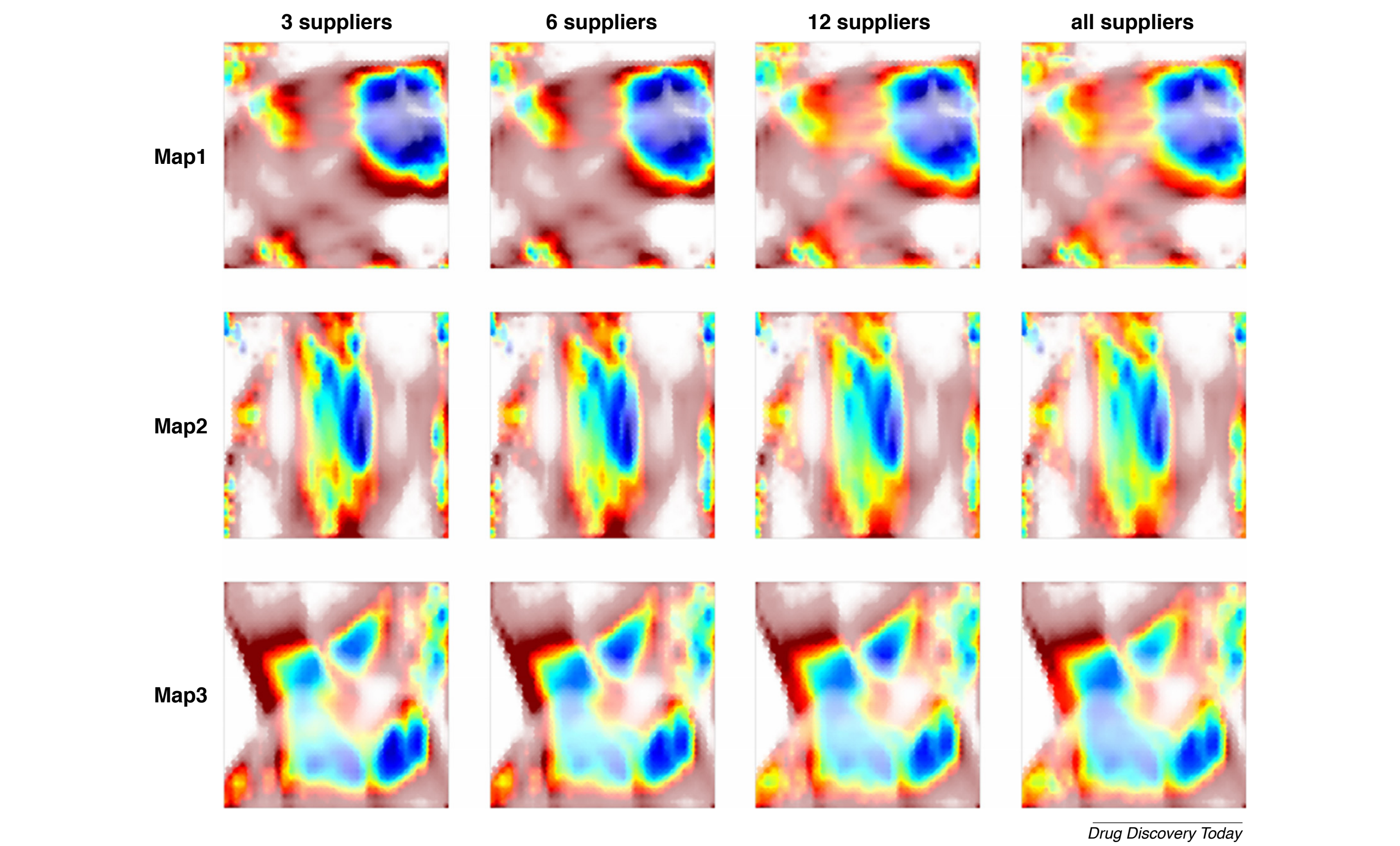


FIGURE 8

Generative Topographic Mapping (GTM) maps of four compound sets corresponding to three, six, 12, and 33 suppliers on the ChEMBL compounds background. See main text for key to colors.

compounds from further suppliers were added. Overlap degrees varied from 0.34 (three suppliers) to 0.40 (six suppliers), 0.47 (12 suppliers), and 0.49 (all suppliers).

Concluding remarks

As HTS has matured, our understanding of what features constitute a quality hit and lead has evolved. It is generally regarded that low lipophilic, and higher Fsp^3 properties are preferred. From our analysis, it appears that, over the past 10 years, the market has evolved to meet these demands, with new compounds from many suppliers having modern physiochemical properties. Currently, it is not possible to purchase an 'ideal' 1-million compound set (50 000 scaffolds, minimum of 20 compounds per scaffold). However, it appears that an 'ideal' 500 000 set can be purchased. If sample logistics is an issue, then we have shown that it is possble to purchase the 500 000 set from only six suppliers, with a 350 000 set available from just three suppliers. Many large companies have been through similar exercises and have built their screening decks accordingly. If you are considering building a screening deck *ab initio*, then it is possible to achieve this from purchasable space. In the interest of open innovation, we have made our data available online (www.awridian.co.uk/Resources). We are confident that, as new challenges in sample supply emerge, the market place will respond positively.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2018.10.016.

References

1 Winquist, R.J. *et al.* (2014) The fall and rise of pharmacology – (re-)defining the discipline? *Biochem. Pharmacol.* 87, 4–24

2 Erlanson, D.A. *et al.* (2016) Twenty years on: the impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* 15, 605–619

3 Goodnow, R.A., Jr *et al.* (2017) DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* 16, 131–147

4 Moffat, J.G. *et al.* (2017) Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discov.* 16, 531–543

5 Scannell, J.W. *et al.* (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11, 191–200

6 Prasad, V. and Mailankody, S. (2017) Research and development spending to bring a single cancer drug to market and revenues after approval. *JAMA Intern. Med.* 177, 1569–1575

7 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–716

8 Macarron, R. *et al.* (2011) Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* 10, 188–195

9 Peakman, M.-C. *et al.* (2015) Experimental Screening Strategies to Reduce Attrition Risk. In *Attrition in the Pharmaceutical Industry: Reasons, Implications, and Pathways Forward* (Alex, A., ed.), pp. 180–214, John Wiley & Sons

10 Feher, M. and Schmidt, J.M. (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 43, 218–227

11 Bakken, G.A. *et al.* (2012) Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *J. Chem. Inf. Model.* 52, 2937–2949

12 Koge, T. *et al.* (2013) Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. *Drug Discov. Today* 18, 1014–1024

13 Morgan, P. *et al.* (2018) Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat. Rev. Drug Discov.* 17, 167–181

14 Njoroge, M. *et al.* (2014) Recent approaches to chemical discovery and development against malaria and the neglected tropical diseases human African trypanosomiasis and schistosomiasis. *Chem. Rev.* 114, 11138–11163

15 Cooper, C.B. (2013) Development of *Mycobacterium tuberculosis* whole cell screening hits as potential antituberculosis agents. *J. Med. Chem.* 56, 7755–7760

16 Peña, I. *et al.* (2015) New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource. *Sci. Rep* 5, 8771

17 Scott, A. (2015) Sanofi off-loads R&D activities in France to Evotec. *C@EN* 93, 6

18 Cabrera, A.C. *et al.* (2016) Aggregated compound biological signatures facilitate phenotypic drug discovery and target elucidation. *ACS Chem. Biol.* 11, 3024–3034

19 Anon (2012) AstraZeneca and Bayer share their entire compound libraries. *Nat. Rev. Drug Discov.* 11, 739

20 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.* 46, 3–26

21 Todeschini, R. and Consonni, V., eds (2009) *Molecular Descriptors for Chemoinformatics*, Wiley-VCH Verlag GmbH & Co

22 Teague, S.J. *et al.* (1999) design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* 38, 3743–3748

23 Hughes, J.D. *et al.* (2008) Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* 18, 4872–4875

24 Congreve, M. *et al.* (2003) A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877

25 Jadhav, A. *et al.* (2010) Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* 53, 37–51

26 Walters, W.P. and Namchuk, M. (2003) A guide to drug discovery: designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* 2, 259–266

27 Baell, J.B. and Holloway, G.A. (2010) Compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740

28 Bruns, R.F. and Watson, I.A. (2012) Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* 55, 9763–9772

29 Gorse, A.-D. (2006) Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* 6, 3–18

30 Gillet, V.J. (2008) New directions in library design and analysis. *Curr. Opin. Chem. Biol.* 12, 372–378

31 Nadin, A. *et al.* (2012) Lead-oriented synthesis: a new opportunity for synthetic chemistry. *Angew. Chem. Int. Ed. Engl.* 51, 1114–1122

32 Senger, M.R. (2016) Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discov. Today* 21, 868–872

33 Kitchen, D.B. and Decornez, H.Y. (2015) Computational Techniques to Support Hit Triage. In *Small Molecule Medicinal Chemistry: Strategies and Technologies* (Czechtizky, W. and Hamley, P., eds), pp. 191–210, John Wiley & Sons

34 Janzen, W.P. (2014) Screening technologies for small molecule discovery: the state of the art. *Chem. Biol.* 21, 1162–1170

35 Mullard, A. (2013) European lead factory opens for business. *Nat. Rev. Drug Discov.* 12, 173–175

36 Schuhmacher, A. *et al.* (2016) Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* 14, 105

37 Green, C. and Taylor, D. (2016) Consolidating a distributed compound management capability into a single installation: the application of overall equipment effectiveness to determine capacity utilization. *J. Lab. Automat.* 21, 811–816

38 Baell, J.B. (2013) Broad coverage of commercially available lead-like screening space with fewer than 350,000 compounds. *J. Chem. Inf. Model.* 53, 39–55

39 Baurin, N. *et al.* (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* 44, 643–651

40 Siroisa, S. *et al.* (2005) Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* 29, 55–67

41 Verheij, H.J. (2006) Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Diver.* 10, 377–388

42 Lucas, X. *et al.* (2015) The purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.* 55, 915–924

43 Chuprina, A. *et al.* (2010) Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* 50, 470–479

44 Petrova, T. *et al.* (2012) Structural enrichment of HTS compounds from available commercial libraries. *Med. Chem. Commun.* 3, 571–579

45 Wigglesworth, M.J. *et al.* (2015) Increasing the delivery of next generation therapeutics from high throughput screening libraries. *Curr. Opin. Chem. Biol.* 26, 104–110

46 Karawajczyk, A. *et al.* (2015) Expansion of chemical space for collaborative lead generation and drug discovery: the European Lead Factory Perspective. *Drug Discov. Today* 20, 1310–1316

47 Besnard, J. *et al.* (2015) The Joint European Compound Library: boosting precompetitive research. *Drug Discov. Today* 20, 181–186

48 Bickerton, G.R. *et al.* (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.* 4, 90–98

49 Firth, N.C. *et al.* (2012) A novel method to characterize the three-dimensionality of molecules. *J. Chem. Inf. Model.* 52, 2516–2525

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

51 Langdon, S.R. *et al.* (2011) Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* 51, 2174–2185

52 Shang, J. *et al.* (2017) Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *J. Cheminform.* 9, 25

53 Hughes, J.P. *et al.* (2011) Principles of early drug discovery. *Br. J. Pharmacol.* 162, 1239–1249

54 Lipkin, M.J. *et al.* (2008) How large does a compound screening collection need to be? *Comb. Chem. High Throughput Screen.* 11, 482–493

55 Renner, S. *et al.* (2011) Recent trends and observations in the design of high-quality screening collections. *Future Med. Chem.* 3, 751–766

56 An, W.F. and Tolliday, N. (2010) Cell-based assays for high-throughput screening. *Mol. Biotechnol.* 45, 180–186

57 Mayr, L.M. and Fuerst, P. (2008) The future of high-throughput screening. *J. Biomol. Screen.* 13, 443–448

58 Downey, W. *et al.* (2010) Compound profiling: size impact on primary screening libraries. *Drug Discov. World Spring* 81–86

59 Glaser, V. (2010) High throughput screening retools for the future. *Bio-IT World Mag.* 8, 20–24

60 Hansson, M. *et al.* (2014) On the relationship between molecular hit rates in high-throughput screening and molecular descriptors. *J. Biomol. Screen.* 19, 727–737

61 Elkin, L.L. *et al.* (2015) Just-in-time compound pooling increases primary screening capacity without compromising screening quality. *J. Biomol. Screen.* 20, 577–587

62 Bakken, G.A. *et al.* (2012) Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *J. Chem. Inf. Model.* 52, 2937–2949

63 Kitchen, D.B. and Decornez, H.Y. (2015) Computational techniques to support hit triage. In *Small Molecule Medicinal Chemistry: Strategies and Technologies* (Czechtizky, W. and Hamley, P., eds), pp. 211–214, John Wiley & Sons

64 Harper, G. *et al.* (2004) Design of a compound screening collection for use in high throughput screening. *Comb. Chem. High Throughput Screen.* 7, 63–70

65 Nilakantan, R. *et al.* (2002) A novel approach to combinatorial library design. *Comb. Chem. High Throughput Screen.* 5, 105–110

66 Preston, S. *et al.* (2017) Screening of the 'Open Scaffolds' collection from Compounds Australia identifies a new chemical entity with anthelmintic activities against different developmental stages of the barber's pole worm and other parasitic nematodes. *Int. J. Parasitol. Drugs Drug. Resist.* 7, 286–294

67 Bell, A.S. *et al.* (2016) Plate-based diversity subset screening generation 2: an improved paradigm for high-throughput screening of large compound files. *Mol. Divers.* 20, 789–803

68 Chakravorty, S.J. *et al.* (2018) Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discov.* 23, 532–545

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

70 Ashton, M. *et al.* (2002) Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant. Struct. Act. Relat.* 21, 598–604

71 Horvath, D. *et al.* (2017) Generative topographic mapping approach to chemical space analysis. In *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences* (Roy, K., ed.), pp. 167–199, Springer

72 Gaspar, H.A. *et al.* (2015) Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J. Chem. Inf. Model.* 55, 84–94

73 Ruggiu, F. *et al.* (2010) ISIDA property-labelled fragment descriptors. *Mol. Inf.* 29, 855–868