ELSEVIER



# Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering

**Jens Schamberger, Michael Grimm, Andreas Steinmeyer and Alexander Hillisch**

Bayer HealthCare, Lead Generation and Optimization – Medicinal Chemistry Berlin and Wuppertal, Aprather Weg 18a, D-42096 Wuppertal, Germany

**Here, we compare the entire compound collections of Bayer HealthCare and Schering AG with respect to structural identities, similarities and physico-chemical properties. We discuss possible consequences stemming from unexpected findings in light of new collaborative models in pharmaceutical research.**

## Introduction

It is generally perceived that pharmaceutical companies worldwide are pursuing drug discovery programs aimed at very similar sets of targets and synthesizing small molecule compounds that are, at least to some extent, fairly similar in structure. As a consequence the compound collections of traditional pharmaceutical companies could be expected to show significant overlap in structural identity or similarity. Reports on close races for chemical patents cultivate this notion.

Working at both Bayer HealthCare AG (BHC) and Schering AG (SAG), we were presented with a situation that enabled us to challenge this paradigm. In March 2006, Bayer AG, Leverkusen, made a public takeover offer to the stockholders of SAG Berlin. This process led to the acquisition of SAG and the founding of a new company, Bayer Schering Pharma AG (BSP), which was completed by the end of 2006. As a result of the takeover, the pharmaceutical businesses and R&D activities of two long-standing international companies with headquarters in Germany, which had each been active for approximately 140 years, were merged.

With high throughput screening (HTS) being the most common method of lead finding in the pharmaceutical industry, compound collections are the cornerstones of early drug discovery. Thus, in BSP drug discovery research, the merging of the screening compound libraries of BHC and SAG was one of the first larger projects to be initiated.

As one measure of synergism the amount the compound collections would complement each other, in terms of structural overlap and their physico-chemical properties, was prudent to investigate. The similarity overlap of both libraries was examined using a straightforward clustering scheme, which led to easily interpretable results. Owing to the clarity of the results, this clustering method was preferred over other chemoinformatic approaches [1–4], such as artificial neural networks.

Here, we compare the small molecule compound libraries of BHC and SAG. We focus on a comparison of their structural identity and physico-chemical properties and discuss consequences of our results.

## The BHC corporate compound collection

Before the takeover, the BHC compound collection consisted of 2 357 206 unique structures as of November 2006. It represented, in part, a collection of compounds synthesized in medicinal chemistry projects over the past few decades. In particular, the library was expanded between 2000 and 2006 to enrich it with attractive starting points for lead optimization.

Combinatorial libraries were designed internally based on feasible chemistry and proprietary building blocks. Proposals for these libraries were based on medicinal chemistry experience, favorable physico-chemical properties and awareness of undesirable chemical groups. The resulting libraries were synthesized using external companies as well as in-house combinatorial chemistry. The BHC compound collection thus consisted of nearly two-thirds combinatorial chemistry compounds and over one-third medicinal chemistry compounds.

## The SAG corporate compound collection

As of November 2006, the SAG screening library contained 874 759 unique structures. It also represented a collection of medicinal chemistry compounds, reflecting project activities over the past few decades. In 2001, SAG began to develop a new screening library from scratch owing to quality issues with the

*Corresponding author:*. Hillisch, A. (alexander.hillisch@bayer.com)



1359-6446/06/$ - see front matter  doi:10.1016/j.drudis.2011.04.005

old library. Two major purchasing campaigns with four vendors occurred between 2002 and 2005, adding about 350 000 commercially available compounds to the library. This library clean-up process was grounded on stringent library design aiming at a collection of favorable diversity and containing compounds which fulfill certain chemical and lead-likeness rules. [5]. The new library consisted of approximately one-third commercial and two-thirds medicinal chemistry and combinatorial chemistry compounds.

## Comparison of BHC and SAG compound libraries

### *Structural identity*

All library comparisons were based on canonical SMILES with Pipeline Pilot (PP) software (http://accelrys.com/products/pipeline-pilot/). Stereochemistry and tautomeric forms were used as found in both corporate databases; salt-stripping was archived by keeping the largest fragment only. The standardization functionality of PP regarding stereochemistry and charges was applied before translation to canonical SMILES. Overlapping structures were identified by means of a case-sensitive string comparison of canonical SMILES.

Comparison of all structures of both libraries resulted in 48 194 duplicates. This represented only 1.5% of all structures that are available in both databases. If protomers were treated consistently, in addition, stereochemistry was ignored, the number of duplicates increased to 48 328 and 50 396, respectively.

In parallel, the analysis was also carried out using Tripos Sybyl software (http://tripos.com/). All structures were encoded in unique SYBYL Line Notation (SLNs) and stored in Tripos Unity hitlists. Salt stripping and a check for duplicate structures were carried out using the programs 'saltstrip' and 'dbhitlist' within Sybyl.

Both chemoinformatics packages (PP and Tripos) led to similar numbers of overlapping structures (±480). A slight deviation in overlap between the two methods was expected as methods of salt treatment and unification of structures are different. As both chemoinformatics packages performed equally well, PP was selected for further investigations owing to its ease of use.

Most of the duplicate structures can be traced back to external vendors, showing that BHC and SAG had purchased some of the same compounds. However, the companies used different suppliers for some of these compounds. Approximately 2000 overlapping compounds were synthesized in-house. Of these, approximately 800 could be identified as parts of combinatorial libraries. The remaining duplicates were contributed by medicinal chemistry programs. Even synthesis intermediates, which exhibit small fragment-like structures, are found in this group. The mean molecular weight of the approximately 1200 duplicate compounds was 300 g/mol, which was significantly less than in the case of the overall BHC or SAG libraries (432 and 407 g/mol, respectively). Of the duplicate structures, only 2.7% was accounted for by medicinal chemistry-derived compounds, which, in turn, means that only 0.04% of the combined database is represented by duplicate medicinal chemistry structures. The results of these investigations, particularly the small degree of overlap of the corporate libraries, were not expected. BHC and SAG are both traditional German pharmaceutical companies with various historical parallels. One might therefore speculate that their compound libraries would somehow reflect these parallel developments.

There might be several reasons for the low structural overlap. First, the indication areas of BHC and SAG overlapped only partly. If we consider the past 20 years, over which most of the compounds included in this analysis were synthesized or acquired, BHC was active in cardiovascular, oncology, anti-infective, central nervous system (CNS), chronic pulmonary disease (COPD) and metabolic disease indications, whereas SAG actively pursued women's and men's healthcare, oncology, diagnostic imaging, dermatology, CNS and cardiovascular projects. Therefore, there were only three common indication areas (oncology, CNS and cardiovascular research) in which both companies invested significant research resources. Second, the library build up and extension strategies of BHC and SAG were both different. BHC focused on proprietary compounds through exclusive collaborations with certain partners, whereas SAG acquired large numbers of commercial compounds, with stringent focus on physico-chemical properties.

There are only few publications dealing with an overlap analysis of pharmaceutical compound libraries. Johnson & Johnson (JNJ) published an analysis from their takeover of 3-Dimensional Pharmaceuticals Inc (3DP) [6]. Their library merger of approximately 700 000 and 400 000 compounds produced 27 000 duplicates, which represented a structural overlap of 2.5%. However, the situation and history of both companies were different from those of BHC and SAG. 3DP was a small startup company with a mainly combinatorial chemistry-based library that was built up over a short period of time and formed the heart of the DirectedDiversity technology platform [7]. The low structural overlap with the historically grown compound collection of JNJ is thus not surprising.

### *Structural similarity*

The low overlap in identical structures prompted us to investigate similarities between the libraries of SAG and BHC. We combined both libraries for structural clustering. 'Functional Connectivity Fingerprints' (FCFP_4) as provided by the PP software were used as molecular descriptors for clustering. Structural clustering was carried out using the hierarchical maximum dissimilarity method of PP [8] based on Tanimoto distances. A comparison using MACCS fingerprints [9] resulted in only slightly more compounds appearing in non-exclusive clusters (data not shown), but led to the same conclusions. In total, 139 673 clusters and 23 258 singletons with an average 19.8 structures per cluster and an intra-cluster Tanimoto distance of 0.78 were obtained. The proportion of BHC compounds in each cluster was calculated. All clusters were categorized in ten bins with proportions ranging from 0% to 100% and a bin size of 10%. Results are shown in Fig. 1, with mean proportion marking the bin on the *x*-axis.

Owing to its sheer size, the BHC library was expected to dominate the clustering. With the BHC library being three times larger than its SAG counterpart, it was even more surprising that two-thirds of the resulting clusters, and the compounds within, were exclusive to either BHC or SAG. Only one-third of clusters were populated with compounds from both libraries to varying degrees (Fig. 1).

Clusters of the first three bins, covering 0–30% of BHC structures, contained 679 160 structures (Fig. 1a). Of these first three bins, 635 612 structures were derived from the SAG library,

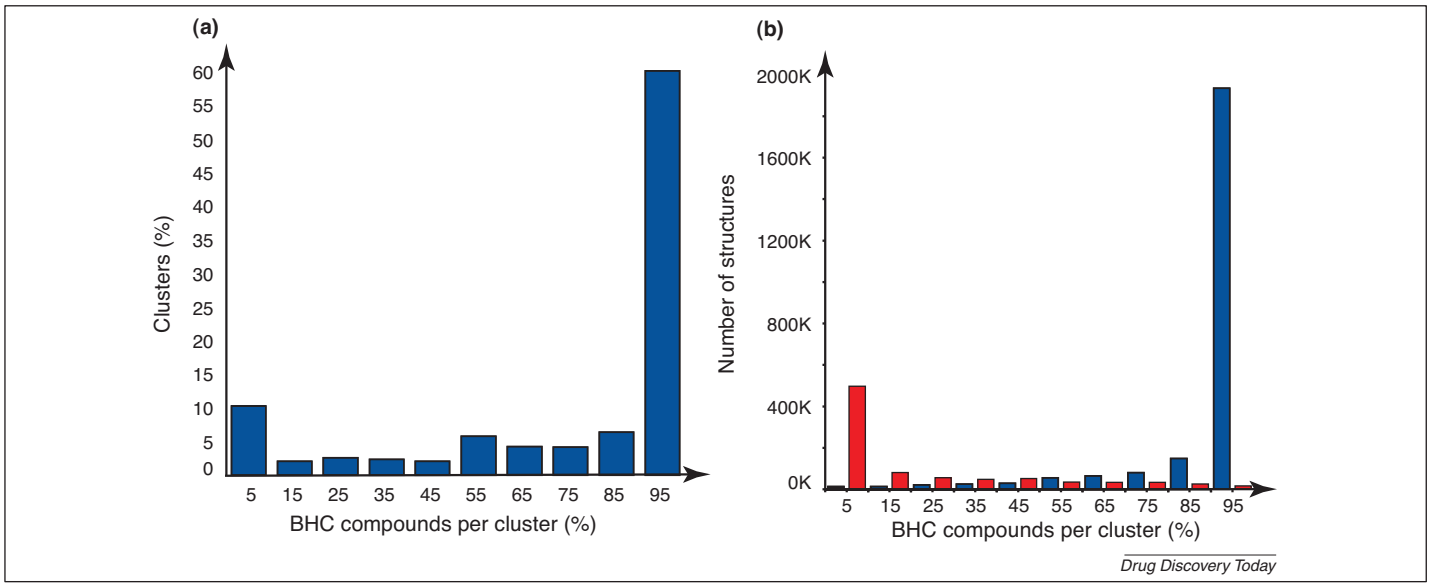


**FIGURE 1**

Structural clustering by FCFP_4 fingerprint similarity and Tanimoto distances. The proportion of BHC compounds in each cluster was calculated. All clusters were categorized in ten bins with proportions ranging from 0% to 100% and a 10% variance, with the mean proportion marking the bin on the *x*-axis. In **(a)**, the heights of the bins are defined by the number of clusters they contain, whereas in **(b)**, the heights of the bins are defined by the number of structures they contain. In **(b)**, each bin is split according to structures originating from SAG (red) and from BHC (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

whereas only 43 548 structures originated from BHC. This suggests that approximately three-quarters of the SAG library occupied previously unpopulated chemical space of the BHC collection.

Clusters containing 90–100% BHC structures amounted to approximately 1 951 404 structures. This indicates that two-thirds of the combined library (635 612 structures from SAG and 1 951 404 structures from BHC) were structurally dissimilar. This analysis further substantiated the structural complementarity of the two libraries.

We further investigated both libraries in terms of their structural overlap with regards to Murcko assemblies [10] as part of the PP software. Before calculation of the Murcko fragments, all compounds with a molecular weight above 800 g/mol and less than 2 rings or eight heavy atoms were filtered out. Canonical tautomers were calculated. In the BHC compound collection, 376 853 unique Murcko fragments were found, whereas 142 281 fragments were found in the SAG collection. The overlap of both individual sets of Murcko assemblies evaluated to 6.3% or 32 884 scaffolds.

To visualize the structural complementarity found between the compound collections, we compared them with the World Drug Index (WDI; http://www.daylight.com/products/wdi.html). Again FCFP_4 fingerprints were applied to all data sets. PP was used to build a principle components analysis (PCA) model of the WDI to visualize the drug-like space using the first two main components of this model (Fig. 2a) plotted on the *x* and *y* axes.

Despite the fact that the compound collections of both BHC and SAG are larger than the WDI (48 129 unique entries), all three databases peak in the same region, which underlines the principle drug likeness of the compounds in these collections. In Fig. 2a, contour levels indicated by the coloring scheme show that the BHC library obviously is the largest collection. However, the smaller SAG collection shows a second, larger area of high compound density around an *x*-value of −10 to −5, further emphasizing an additional chemical space.

A similar comparison was made between the BHC compound collection as underlying chemical space and all SAG compounds (Fig. 2b). SAG compounds showed a different distribution compared with the BHC compounds, indicated by the color distribution in Fig. 2b. It is obvious that the roughly three times larger BHC collection branches out much further than the SAG collection. However, regions can be determined where one collection alone dominates the space, which is consistent with the clustering results shown in Fig. 1.

### *Physico-chemical properties*

For measuring the quality of the libraries in terms of their structural excellence, we used computational measurements, namely the oral PhysChem Score (oPCS) [11] as well as an undesirable groups flagging algorithm [12]. The oPCS integrates several physico-chemical properties in one score to evaluate the oral drug-likeness of a compound. The *in silico* oral PhysChem Score is formed by summing the values for five *in silico* parameters: predicted aqueous solubility, molecular weight corrected for halogens, clogP, polar surface area and number of rotatable bonds. These five values are binned into three bins each, counting from 0 to 2. The score results from summing the bin numbers. It can adopt a minimum value of 0, for compounds that are always in the lowest bin, indicating that all of the calculated physico-chemical properties are in a favorable range. Its maximum value is 10, indicating that all of the calculated physico-chemical properties are in an unfavorable range. The lower its value, the more positive the *in*

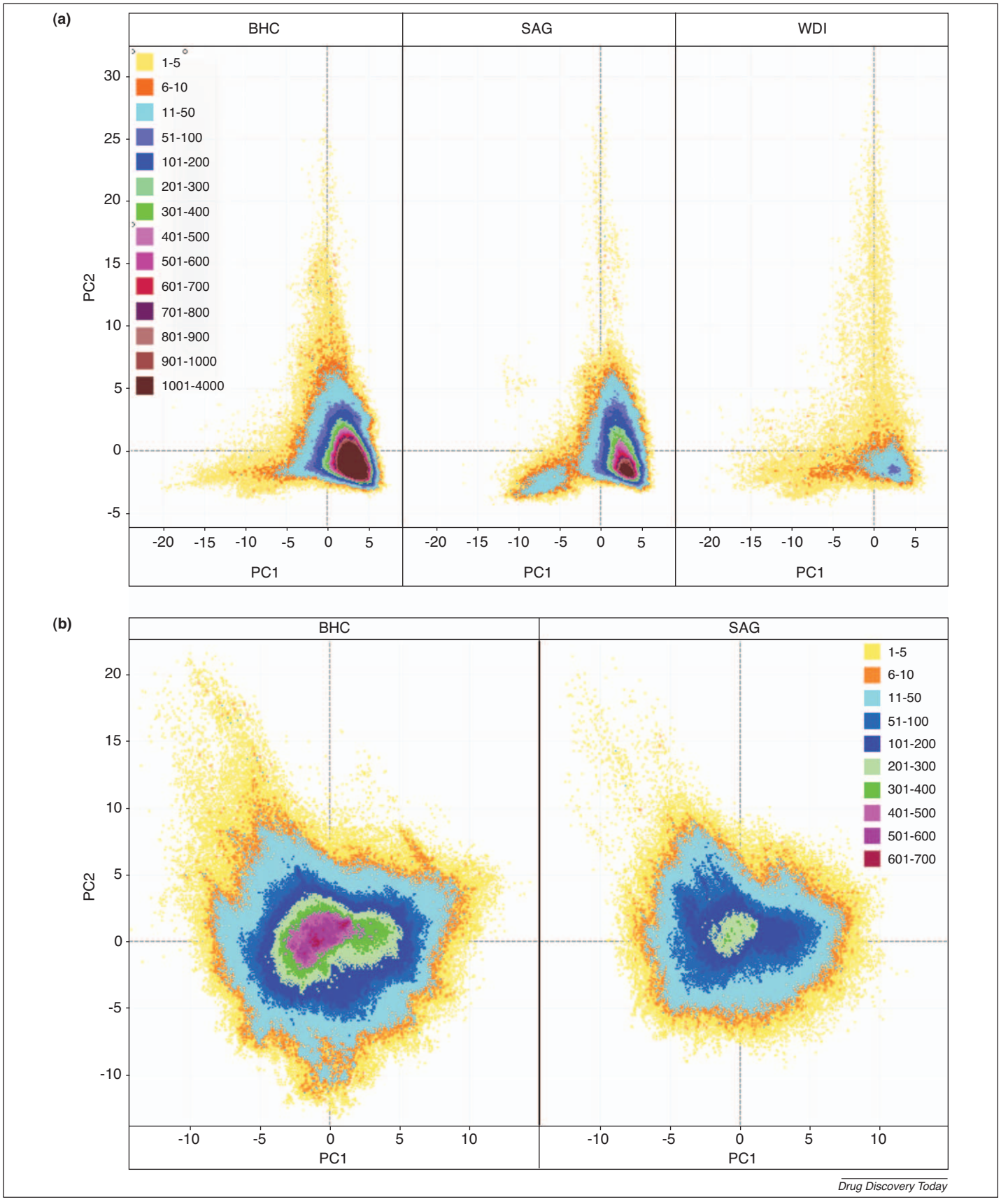


**FIGURE 2**

PCA models of **(a)** the WDI based on FCFP_4 fingerprints, applied to the BHC and SAG compound collections and to the WDI itself, and **(b)** the BHC library based on FCFP_4 fingerprints, applied to BHC and SAG compound collections. In both **(a)** and **(b)**, the two main components define the *x* and *y* axes, respectively. For a clearer visualization, the continuous PCA values have been binned. Each bin measures 0.1 *x*-units by 0.1 *y*-units; the number of all compounds per bin was built in order to visualize regions of high compound density. The coloring scheme above adds a third dimension to the map and indicates those high density regions.

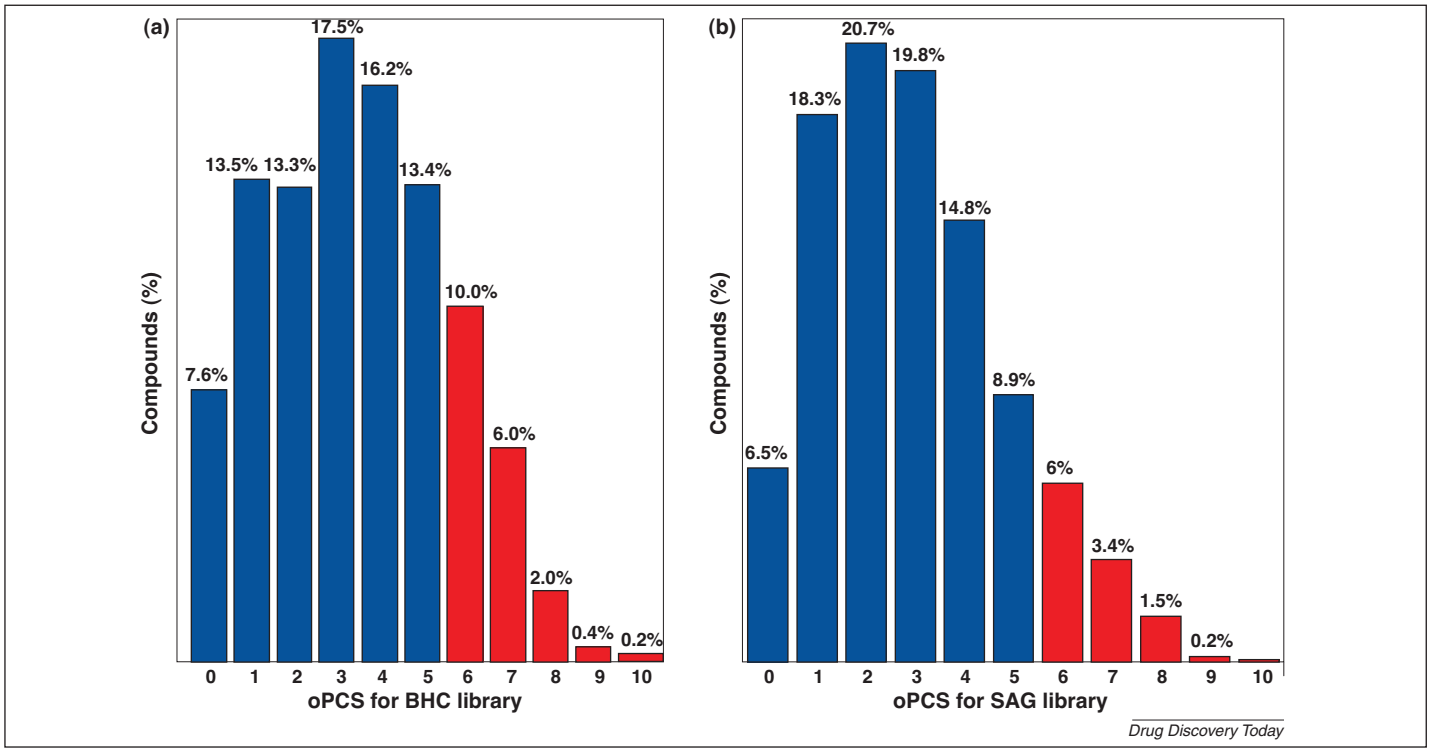


**FIGURE 3**

Oral PhysChem Score (oPCS) distribution for **(a)** BHC and **(b)** SAG compound collections. Blue bars indicate compounds with calculated physicochemical properties that resemble 94% of oral marketed drugs. Red bars indicate compounds outside this window. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

*silico* evaluation of a certain compound is with regard to favorable physico-chemical properties and, therefore, its use as a lead for the discovery of an orally administered drug. Marketed small molecule drugs show an average *in silico* oPCS of 1.9 [11].

To serve as a lead compound for drug discovery projects, oral PhysChem Scores should be less than 6. The oPCS was solely designed for compounds with an oral route of administration. The advantage of small molecules is the ability to apply them orally and, thus, we focused our analysis on physico-chemical properties associated with oral administration.

Our in-house 'Undesirable Groups' flagging indicates functional groups in a structure that are known to result in toxicity or pharmacokinetic problems. The flagging uses 153 substructures and other structural properties that often confer undesirable properties, such as high reactivity, mutagenicity or susceptibility to form reactive metabolites. This procedure was inspired by several schemes of functional group filtering, such as REOS introduced by Vertex [13,14]. Twelve experienced medicinal chemists from BHC selected these alerts and qualified their undesirability rating. Both algorithms have been tested extensively in-house in numerous lead optimization programs and have been found to be useful in screening compounds for adverse effects.

Comparison of both libraries according to the in-house oral PhysChem Score of BHC revealed a similar distribution (Fig. 3). With a peak value of 2, the SAG library had a slightly better oPCS maximum than did the BHC library, which peaked at an oPCS value of 3. Of the BHC library, 18.6% expressed an oPCS greater than 5, whereas only 11.1% of the SAG compounds were in that range. In particular, the BHC compounds were characterized by higher-than-average molecular weight (mean, 419 g/mol vs 396 g/mol; median, 420 g/mol vs 396 g/mol) and more rotatable bonds (mean, 6.9 vs 6.5; median, six in both collections). BHC compounds overall showed higher clogP values (http://www.biobyte.com/bb/prod/clogp40.html; mean, 4.0 vs 3.8; median, 4.0 vs 3.7).

The slightly better oPCS distribution of the SAG library was probably the result of the library clean-up process that occurred during 2002–2005. In terms of physico-chemical properties and undesirable groups, both libraries were similar and showed a desirable profile, which made them a perfect fit to each other.

As can be seen in Fig. 4, both libraries showed a green light for approximately half of their structures. Only 14% were flagged in red for the BHC library, and 16% in the SAG collection. Red flagging is not a definitive indicator for a toxic or undesirable compound. What we considered to be the worst offenders were marked in black, which amounted to 2% of BHC compounds and 1% for SAG compounds.

## Consequences for pharmaceutical research

Some general aspects can be deduced from this study. The structural overlap in compound collections of independent pharmaceutical companies can be rather low. Taking into account the many possible drug-like chemical compounds or, in other words, the 'chemical space' [15], it would indeed be surprising to find a significant number of identical compounds in unrelated corporate collections. We only found less than 1200 compounds (0.04% of

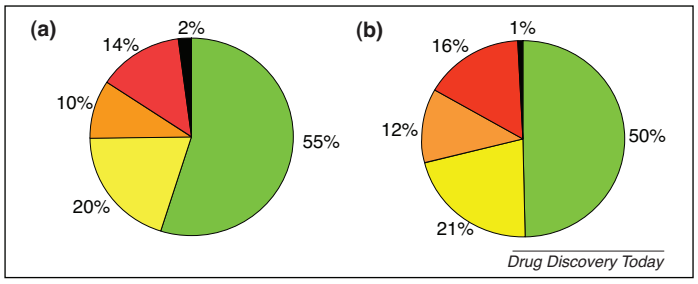


**FIGURE 4**

The traffic light distribution of undesirable groups for the **(a)** BHC and **(b)** SAG libraries.

the entire library) that were synthesized in-house independently by BHC and SAG. We speculate that an overlap of less than 10% also exists between historically grown compound collections of other pharmaceutical companies and that their libraries would therefore complement each other. This paves the way for screening consortia in which companies collaborate to screen their corporate collections mutually on selected targets in non-competing indication areas. It would not be necessary to exchange entire compound libraries, only hit lists and an easily manageable number of physically existing hits. Given the enormous investment that would otherwise be required to gain access to several hundred thousand medicinal chemistry-like screening compounds, this proposed approach appears attractive. This would be a strategy to replace expensive and protracted in-house investments in new compounds that complement internal compound collections based on, for example, chemogenomic approaches [16,17]. It might help to find leads for targets that are principally druggable but for which acceptable leads have so far not been identified by screening certain in-house libraries [18]. Indeed, we found examples of important in-house projects for which several screening approaches with the BHC library had not yielded new leads. Screening of the SAG library after the takeover resulted in promising starting points for lead optimization. Such a scenario has already been hypothesized in previous analyses [19].

Furthermore, an argument that might support mergers and acquisitions (M&A) in the pharmaceutical sector can be harvested from this analysis. Currently, M&As in this industry are driven by product portfolios rather than by drug discovery competencies. With the current need for innovative drugs, R&D skills of pharmaceutical companies might again become more important. The technological complementarity of two companies is often quoted as an important factor for successful M&As in the long term [20,21]. If compound libraries are regarded as a kind of company knowledge-base, then a high degree of complementarity is clearly desirable and would improve drug discovery skills. Based on our data, the libraries of BHC and SAG are structurally complementary and fit together well in terms of their physico-chemical properties. However, it remains to be proven if this leads to additional innovative products.

## Summary

To the best of our knowledge, this analysis constitutes the first detailed comparison of larger compound collections of competing pharmaceutical companies. Although the overall physico-chemical properties of both libraries are strikingly similar, their overlap in terms of structural identity and similarity is surprisingly low. Approximately three-quarters of all SAG compounds occupy previously completely unpopulated chemical space of the BHC library. Thus, both libraries complement each other to a degree that was originally not expected.

## Acknowledgments

We would like to thank the following colleagues for fruitful discussions: M. Bauser, U. Bömer, T. Flessner, A. Göller, H. Haning, J. Hüser, D. Nguyen, M. Koppitz, J. Köbberling, H. Meier, S. Mundt, E. Ottow, B. Riedl and H. Wild.

## References

1 Southan, C. *et al.* (2009) Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminform.* 1, 10
2 Muresan, S. and Sadowski, J. (2005) In-house likeness': comparison of large compound collections using artificial neural networks. *J. Chem. Inf. Model.* 45, 888–893
3 Turner, D.B. *et al.* (1997) Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* 37, 18–22
4 Cummins, D.J. *et al.* (1996) Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* 36, 750–763
5 Steinmeyer, A. (2006) The hit-to-lead process at Schering AG: strategic aspects. *ChemMedChem* 1, 31–36
6 Engels, M.F. *et al.* (2006) A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J. Chem. Inf. Model.* 46, 2651–2660
7 Agrafiotis, D.K. *et al.* System and Method for Automatically Generating Chemical Compounds with Desired Properties. (US Patent 5,574,564). 1996.
8 Hassan, M. *et al.* (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers.* 2, 64–74
9 Durant, J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280
10 Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893
11 Lobell, M. *et al.* (2006) In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* 1, 1229–1236
12 Wunberg, T. *et al.* (2006) Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* 11, 175–180
13 Walters, W.P. *et al.* (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 3, 384–387
14 Charifson, P.S. and Walters, W.P. (2002) Filtering databases and chemical libraries. *Mol. Divers.* 5, 185–197
15 Dobson, C.M. (2004) Chemical space and biology. *Nature* 432, 824–828
16 Jacoby, E. and Mozzarelli, A. (2009) Chemogenomic strategies to expand the bioactive chemical space. *Curr. Med. Chem.* 16, 4374–4381
17 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
18 Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today* 11, 277–279
19 Shibayama, S. *et al.* (2008) Effect of mergers and acquisitions on drug discovery: perspective from a case study of a Japanese pharmaceutical company. *Drug Discov. Today* 13, 86–93
20 Cloodt, M. *et al.* (2006) Mergers and acquisitions: Their effect on the innovative performance of companies in high-tech industries. *Research Policy* 35, 642–654
21 Prabhu, J.C. *et al.* (2005) The impact of acquisitions on innovation: poison pill, placebo, or tonic? *J. Market.* 69, 114–130