Exploring Structural Relationships between Bioactive and Commercial Chemical Space and Developing Target Hypotheses for Compound Acquisition

Carmen Cerchia,†,‡ Dilyana Dimova,† Antonio Lavecchia,‡ and Jürgen Bajorath*†

†Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany
‡Department of Pharmacy, ”Drug Discovery” Laboratory, University of Naples Federico II, 80131 Naples, Italy

ABSTRACT: Analog series were systematically extracted from more than 650 000 bioactive compounds originating from medicinal chemistry and screening sources and more than 3.6 million commercial compounds that were not biologically annotated. Then, analog series-based (ASB) scaffolds were generated. For each scaffold from a bioactive series, a target profile was derived and ASB scaffolds shared by bioactive and commercial compounds were determined. On the basis of our analysis, large segments of commercial chemical space were not yet explored biologically. Shared ASB scaffolds established structural relationships between bioactive and commercial chemical space, and the target profiles of these scaffolds were transferred to commercially available analogs of active compounds. This made it possible to derive target hypotheses for more than 37 000 compounds without biological annotations covering more than 1000 different targets. For many molecules, alternative target assignments were available. Target hypotheses for these compounds should be of interest, for example, for hit expansion, acquisition of compounds to design or further extend focused libraries for drug discovery, or testing of expanded analog series on different targets. They can also be used to search for analogs and complement compound series during target-directed optimization. Therefore, all of the commercial molecules with new target hypotheses as well as key scaffolds identified in our analysis and their target profiles are made freely available.

1. INTRODUCTION

Volumes of publicly available bioactive compounds and other small molecules increase rapidly, as exemplified by three major databases that collect bioactive compounds from different sources and small molecules from chemistry vendors. For example, release 22 of ChEMBL, the major public repository of active compounds from the medicinal chemistry literature and patents contains nearly 1.7 million compounds with activity against more than 11 000 targets. In addition, PubChem, the major public resource of screening data, lists more than 93 million compound entries and 1.25 million assays. Taken together, these two major databases contain more compounds and activity data than could have been imagined just a few years ago. Moreover, release 15 of ZINC, which collects compounds from chemistry vendors worldwide, reports an astonishing number of more than 200 million entries. Thus, in terms of compounds, structures, and activity data, chemistry is without doubt entering the “big data” era. ZINC compounds are not biologically annotated but represent a vast sample of chemical space that is relevant for drug discovery. Accordingly, the large numbers of bioactive compounds and vendor molecules that are available provide an unprecedented opportunity to systematically compare structural relationships between active compounds and other small molecules that are currently without biological annotations and attempt to derive target hypotheses for molecules that are available commercially.

Therefore, we have carried out a large-scale comparison of compound subsets selected from ChEMBL and PubChem on the basis of well-defined activity criteria and target annotations and ZINC molecules that are most likely available from vendor sources. The analysis was facilitated using the recently introduced concept of analog series-based (ASB) scaffolds. Different from other scaffold definitions, the ASB scaffold formalism first systematically extracts analog series from compound collections and then isolates conserved core structures from series that distinguish between analogs on the basis of a single substitution site (if available). Thus, ASB scaffolds carry series information and account for structural relationships. Furthermore, ASB scaffolds can be annotated with targets of compounds they represent. Biologically annotated ASB scaffolds can then be used to develop target hypotheses for other compounds containing the same scaffold. This approach was recently successfully applied to target deconvolution of chemical cancer cell line screens.

Herein, we report our systematic comparison of bioactive and commercial compounds and derivation of target hypotheses for nonannotated molecules. The large number of target hypotheses that were obtained, frequently including multiple
assignments for compounds, should be of interest for a variety of practical applications as discussed below.

2. MATERIALS AND METHODS

2.1. Compounds and Activity Data. Commercially available compounds were collected from ZINC version 15 (accessed March 5, 2017). To narrow down commercial chemical space and focus our analysis on compounds desirable for medicinal chemistry, the druglike subset was selected from ZINC using the “tranche browser”. Reactivity was set to “anodyne” (i.e., compounds that do not contain problematic functionalities, also excluding pan-assay interference compounds) and availability to “in-stock” (i.e., ready for delivery). The resulting compound subset was downloaded in two dimensional SMILES format and standardized with inhouse Python scripts implemented in OpenEye chemistry toolkit. These selection procedures resulted in a set of 3 658 425 unique compounds from ZINC.

Bioactive compounds were assembled from ChEMBL, version 22 and the current release of PubChem BioAssays. From ChEMBL, only compounds with high-confidence activity data were considered. Therefore, compounds involved in direct interactions (target relationship type “D”) with human targets at the highest confidence level (target confidence score 9) were selected. As potency measurements, only assay-independent equilibrium constants ($K_i$ values) and assay-dependent $IC_{50}$ values were considered. Approximate measurements associated with “>”, “<”, or “~” were discarded. If multiple $K_i$ or $IC_{50}$ values were available for the same compound, the geometric mean of the values was calculated as the final potency annotation, provided all of the values fell within the same order of magnitude. Otherwise, the measurements were discarded. Given these selection criteria, a total of 224 532 unique compounds were obtained with activity against 1687 human targets.

From PubChem BioAssays, a subset of 426 294 unique compounds tested both in primary assays (percentage of inhibition from a single dose) and confirmatory assays (dose–response assays yielding $IC_{50}$ values) was selected. Only compounds classified as active or inactive were taken, whereas compounds with designations such as unspecified or inconclusive were discarded. From primary assays, RNA interference screens were removed, whereas all of the chemical screens were retained, including cell-based assays for which no individual screens were removed, whereas all of the chemical screens were retained, including cell-based assays for which no individual target was specified. For confirmatory assays, a series of selection criteria was applied with the aim to collect chemical compounds tested against single target proteins, yielding a total of 299 021 unique compounds with activity against 772 targets.

2.2. Scaffold Analysis. ASB scaffolds were identified following a previously described two-stage procedure. First, the analog series was identified by applying the matched molecular pair (MMP) formalism. An MMP is defined as a pair of compounds that differ only by a structural change (chemical transformation) at a single site. Thus, an MMP consists of a shared core and a pair of exchanged substituents. Specifically, MMPs were generated using an adaptation of Hussain and Rea method, which systematically fragments molecules at exocyclic single bonds. Instead of random fragmentation, we applied retrosynthetic combinatorial analysis procedure (RECAP) fragmentation rules, according to which, exchanged fragments conformed with chemical reactions, yielding RECAP-MMPs. Transformations in RECAP-MMPs were confined to limited size because previously established fragment size restrictions were applied to limit RECAP-MMPs to pairs of compounds representing typical analogs. Specifically, the size of the conserved RECAP-MMP core was required to be at least twice the size of the larger exchanged fragment that was allowed to consist of at most 13 heavy atoms. These size restrictions ensured that substituents were limited to maximally a condensed two-ring system with no more than three additional atoms conforming with largest substitutions typically seen in the analog series from medicinal chemistry.

Second, a network was generated in which the nodes represented compounds and edges pairwise RECAP-MMP relationships. In this network, each separate cluster represented a unique series of analogs from which ASB scaffolds were extracted. Hence, ASB scaffold generation combines analog search and scaffold extraction algorithms. Because analog series often contained multiple possible RECAP-MMP cores, a computational search was carried out for a core that captured all of the MMP relationships within the series. If more than one core met this criterion, the largest one was selected. The qualifying RECAP-MMP core represented the ASB scaffold of the series. ASB scaffold analysis and compound statistics are summarized in Table 1. All of the calculations were carried out using inhouse Perl and Python scripts with the aid of the OpenEye chemistry toolkit and KNIME protocols.

2.3. Target Annotations. A pool of unique 1687 targets originating from selected ChEMBL compounds and 772 targets from selected PubChem compounds was used to develop plausible target hypotheses for commercially available ZINC molecules. For each analog series, the union of target annotations from bioactive compounds comprising the series was determined and assigned to ZINC molecules sharing the corresponding ASB scaffold. To identify unique targets and assign targets to families, the UniProt classification scheme was applied. Therefore, target GI numbers from PubChem and IDs from ChEMBL were mapped to UniProt IDs, hence ensuring consistency of target annotations.

3. RESULTS AND DISCUSSION

3.1. ASB Scaffold Analysis. ASB scaffolds were systematically extracted from the analog series originating from 224 532 ChEMBL compounds for which high-confidence activity data and target annotations were available, 426 294 most extensively assayed PubChem compounds, and 3 658 425 ZINC molecules that had the highest probability of commercial availability. As
reported in Table 1, for each database, around 60% of the compounds participated in the formation of analog series. ZINC, ChEMBL, and PubChem compounds yielded 264,496, 22,015, and 42,513 series, respectively, from which ASB scaffolds were isolated. Figure 1 shows the size distribution of analog series. In each case, more than 60% percent of all of the series were compound pairs, an additional 23–27% of the series consisted of 3–5 compounds, and less than 10% of 6–10. Thus, whereas some very large series were extracted from each database (Table 1), small series dominated the distribution. From more than 70% (ChEMBL, PubChem) to close to 80% (ZINC) of all analog series, ASB scaffolds were obtained. Figure 1 shows the size distribution of analog series. In each case, more than 60% percent of all of the series were compound pairs, an additional 23–27% of the series consisted of 3–5 compounds, and less than 10% of 6–10. Thus, whereas some very large series were extracted from each database (Table 1), small series dominated the distribution. From more than 70% (ChEMBL, PubChem) to close to 80% (ZINC) of all analog series, ASB scaffolds were obtained. Figure 1 shows the size distribution of analog series. In each case, more than 60% percent of all of the series were compound pairs, an additional 23–27% of the series consisted of 3–5 compounds, and less than 10% of 6–10. Thus, whereas some very large series were extracted from each database (Table 1), small series dominated the distribution. From more than 70% (ChEMBL, PubChem) to close to 80% (ZINC) of all analog series, ASB scaffolds were obtained.

ASB scaffolds from ChEMBL exclusively represented active compounds, whereas scaffolds from PubChem represented both active and consistently inactive screening compounds. Scaffolds from ZINC represented compounds without activity annotations. Importantly, the comparison revealed that ZINC contained many more analog series and ASB scaffolds than were available for bioactive compounds. In addition, more than 40% of the ZINC compounds did not participate in series. Taken together, these observations indicate that ZINC covers large chemical space that is most likely not yet explored biologically.

3.2. Scaffold Overlap. Next, the overlap between the independently derived ASB scaffold populations was determined. The results are shown in Figure 2. ChEMBL and ZINC shared 1216 ASB scaffolds, whereas PubChem and ZINC shared 11,270 scaffolds. For PubChem, only ASB scaffolds representing multiple active compounds were taken into consideration.
Figure 3. Distribution of targets over different families. The distribution of targets over different families assigned to ZINC compounds on the basis (a) ChEMBL and (b) PubChem annotations is shown. UniProt-based target family assignments were available for 549 of 610 PubChem targets.

Table 2. Promiscuity of Shared ASB Scaffolds

| # targets | # ASB scaffolds | # ChEMBL CPDs | # ZINC CPDs | # ASB scaffolds | # PubChem CPDs | # ZINC CPDs |
|-----------|----------------|---------------|-------------|----------------|----------------|-------------|
| 1         | 863            | 1154          | 1851        | 3091           | 3200           | 7832        |
| 2         | 193            | 285           | 417         | 2065           | 2473           | 5450        |
| 3         | 66             | 90            | 138         | 1420           | 1903           | 4010        |
| 4         | 52             | 94            | 139         | 1099           | 1518           | 2891        |
| 5         | 14             | 31            | 18          | 753            | 1224           | 2144        |
| 6–10      | 26             | 49            | 43          | 1823           | 3563           | 6467        |
| 11–15     | 1              | 1             | 605         | 1440           | 2248           |             |
| 16–20     | –              | –             | 240         | 738            | 889            |             |
| 21–25     | 1              | 1             | 117         | 360            | 426            |             |
| >25       | –              | –             | 147         | 582            | 605            |             |

The table reports the number of shared ASB scaffolds associated with single- and increasing multitarget activities. For each ASB scaffold, the total number of bioactive compounds (from ChEMBL and PubChem) and ZINC compounds forming the analog series is reported. ASB scaffolds shared by ChEMBL and ZINC. PubChem and ZINC compounds.

3.3. Target Transfer. Targets of a series of bioactive compounds sharing the same ASB scaffold were assembled and assigned to the scaffold. The union of unique targets then constituted the target profile of the scaffold, representing a form of metadata. ZINC molecules sharing the same scaffold were analogs of bioactive compounds from which the target profile was derived. Accordingly, a shared ASB scaffold established a direct structural link between compounds with and without target annotations, and its target profile provided the target hypotheses for molecules from ZINC. Underlying ideas included that structural analogs are likely to display similar or overlapping biological activities, and that ASB scaffolds can be used for activity assignments at a higher level of structural abstraction compared to individual analogs comprising a series. Activity assignments to scaffolds are often attempted in compound-based scaffold analysis. However, ASB scaffolds are more suitable for metalevel activity assignments on the basis of analogs comprising a series than compound-based scaffolds. Figure 2 shows 3 exclusive subsets of 1216, 11 270, and 581 ASB scaffolds representing compounds from ChEMBL and ZINC,PubChem and ZINC, and all of the 3 databases, respectively. ChEMBL and PubChem compounds sharing ASB scaffolds with ZINC molecules were active against a total of 543 and 586 unique targets, respectively. Thus, despite the availability of many more PubChem than CHEMBL compounds sharing ASB scaffolds with ZINC molecules, the target coverage of these compound sets was comparable. In addition, compounds from ChEMBL and PubChem represented by the 581 ASB scaffolds shared by all of the 3 databases were active against a total of 600 targets. In total, by combining these exclusive sets, 588 and 610 targets from ChEMBL and PubChem were obtained, which could be assigned to 37 763 ZINC compounds. The distribution of these targets over different families is shown in Figure 3a for ChEMBL and Figure 3b for PubChem. Although the distributions were partly similar, there were notable differences. For example, compounds from ChEMBL covered larger proportions of kinases and proteases than PubChem compounds, whereas PubChem covered a larger proportion of targets that did not belong to major therapeutically relevant families ("other targets"). The 581 ASB scaffolds that were common to ChEMBL, PubChem, and ZINC represented the compounds active against a large number of targets, thus assigning a key role to these conserved ASB scaffolds.

3.4. Promiscuity of Shared Scaffolds. On the basis of target profiles, the promiscuity of scaffolds was determined, also representing an analysis at the level of metadata (i.e., scaffold promiscuity was inferred from unique targets of compounds
Figure 4. Analog series and shared ASB scaffolds. Shown is exemplary analog series with shared ASB scaffolds for (a) ChEMBL (blue box), (b) PubChem (green), (c) ChEMBL and PubChem, and ZINC compounds (orange) in an R-group table format. Substituents ($R_1$) in analogs are in red. For each series, the union of targets (assigned targets) from bioactive compounds is provided that can be potentially assigned to ZINC compounds containing the same scaffold. In (a) eight analogs from ChEMBL that share an ASB scaffold with six analogs from ZINC are shown that are active against a total of seven targets including protein-tyrosine phosphatase LC-PTP (T1), induced myeloid leukemia cell differentiation protein Mcl-1 (T2), carboxy-terminal domain RNA polymerase II polypeptide A small phosphatase 1 (T3), nuclear receptor subfamily 4 group A member 1 (T4), ...
protein tyrosine kinase 2 β (T5), estrogen receptor β (T6), and apoptotic protease-activating factor 1 (T7). In (b), six analogs from PubChem that share an ASB scaffold with five analogs from ZINC are shown that are active against a total of three targets including TDP1 protein (T1), thioredoxin glutathione reductase (T2), and dopamine receptor D3 (T3). In (c), 5 PubChem, 4 ChEMBL, and 6 ZINC analogs containing 1 of the S81 conserved ASB scaffolds are shown. PubChem and ChEMBL analogs were active against a total of five unique targets including dopamine D1 receptor (T1), TDP1 protein (T2), adenosine 5′-triphosphate-dependent Clp protease proteolytic subunit (T3), serotonin 5 (5-HT6) receptor (T4), and urea transporter 1 (T5).

represented by a given ASB scaffold). Table 2 reports the distribution of the 1216 and 11270 ASB scaffolds and the corresponding compounds from ChEMBL and PubChem over increasing numbers of targets (i.e., increasing promiscuity levels). The majority of ASB scaffolds shared by ChEMBL and ZINC compounds (i.e., 863) were only annotated with a single target. Hence, these scaffolds were classified as nonpromiscuous. In addition, there were more than 300 scaffolds with 2–4 targets available and smaller numbers with up to 25 targets (promiscuous scaffolds). In addition, for 3091 ASB scaffolds shared by PubChem and ZINC, only a single target was available. However, in this case, there were 4494 scaffolds with 2–4, 754 with 5, and 1822 scaffolds with 6–10 targets. Furthermore, more than 1100 ASB scaffolds with 11–25 or more targets were available. Thus, ASB scaffolds shared by PubChem and ZINC compounds were overall much more promiscuous than those shared by ChEMBL and ZINC.

3.5. Target Hypotheses. For a total of 37 763 commercial ZINC molecules, target hypotheses were derived on the basis of shared ASB scaffolds, covering a total of 1051 unique targets. Figure 4 shows the representative examples. Figure 4a shows a series comprising eight analogs from ChEMBL that were active against varying numbers of seven different targets. On the basis of the shared ASB scaffold, this series was combined with a corresponding series from ZINC containing six analogs, hence yielding a total of seven target hypotheses for these ZINC molecules. In addition, in Figure 4b, a series of six PubChem analogs is shown. Each of these compounds was active against one of three targets. Five analogs from ZINC were found to contain the same ASB scaffold, thus providing three target hypotheses for these molecules. Furthermore, Figure 4c shows 5 PubChem and 4 ChEMBL analogs containing 1 of the S81 conserved ASB scaffolds found in all of the 3 databases. ChEMBL and PubChem analogs were active against a total of five different targets. The conserved ASB scaffold was also present in six ZINC molecules. Thus, for these analogs from ZINC, five plausible target hypotheses were available.

Because more than 37 000 biologically unannotated compounds were linked to activities against more than 1000 targets via the ASB scaffold approach, often including multiple target hypotheses ASB scaffolds and resulting compounds, several practical applications can be considered. For example, these compounds provide a pool for target-directed hit expansion (“analog-by-catalog”) and for the design or extension of focused libraries used in drug discovery. To these ends, it is often attempted to acquire new compounds from vendor sources. Taken together, the results of your analysis assign priorities to commercial compounds for many targets. Furthermore, analog series are often evaluated on closely related targets in the search for selective compounds. For this purpose, analog series can also be further extended with commercial compounds linked to activities of interest. On the other hand, ZINC compounds containing promiscuous ASB scaffolds, as described above, can also be selected as candidates for exploring multitarget activities, another current topic of interest in drug discovery research. Thus, the pool of commercial compounds with analog relationships to known bioactives should have utility for a variety of applications.

4. CONCLUSIONS

Our study was designed to systematically compare bioactive and commercial chemical space on the basis of analog series, establish structural links via ASB scaffolds, and develop target hypotheses for commercial molecules. To these ends, a total of more than 4.2 million carefully selected compounds were analyzed. For ASB scaffolds, target profiles were generated on the basis of active compounds they represented, and scaffolds shared by bioactive compounds were determined. A subset of S81 conserved ASB scaffolds was identified whose profiles covered 600 different targets. The scaffolds played a central role in target exploration and transfer across different databases. Shared scaffolds and their target profiles made it possible to derive target hypotheses for 37 763 compounds from ZINC having a high probability of commercial availability. These compounds represent interesting candidates for acquisition and focused library design or other applications in compound optimization, as discussed above. Therefore, as a part of our study, key ASB scaffolds and ZINC compounds with new target annotations will be made freely available on an open access platform.20

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-2699-306.

ORCID
Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions
The study was carried out and the manuscript written with contributions of all of the authors. All of the authors have approved the final version of the manuscript.

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Martin Vogt for helpful discussions.

■ REFERENCES

(1) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100–D1107.
(2) Wang, Y.; Xiao, J.; Sunek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem’s BioAssay Database. Nucleic Acids Res. 2012, 40, D400–D412.
(3) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. J. Chem. Inf. Model. 2015, 55, 2324–2337.
(4) Hu, Y.; Bajorath, J. Entering the ‘Big Data’ Era in Medicinal Chemistry: Molecular Promiscuity Analysis Revisited. Future Sci. OA 2017, 3, No. FSO179.

(5) Dimova, D.; Stumpfè, D.; Hu, Y.; Bajorath, J. Analog Series-Based Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. Future Sci. OA 2016, 2, No. FSO149.

(6) Hu, Y.; Stumpfè, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. J. Med. Chem. 2016, 59, 4062–4076.

(7) Kunimoto, R.; Dimova, D.; Bajorath, J. Application of a New Scaffold Concept for Computational Target Deconvolution of Chemical Cancer Cell Line Screens. ACS Omega 2017, 2, 1463–1468.

(8) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for their Exclusion in Bioassays. J. Med. Chem. 2010, 53, 2719–2740.

(9) OEChem, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

(10) Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. PLoS One 2016, 11, No. e0153873.

(11) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In Chemoinformatics in Drug Discovery; Opdea, T. L., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.

(12) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J. Chem. Inf. Model. 2010, 50, 339–348.

(13) Lowell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. J. Chem. Inf. Comput. Sci. 1998, 38, 511–522.

(14) de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. Med. Chem. Commun. 2014, 5, 64–67.

(15) Hu, X.; Hu, Y.; Vogt, M.; Stumpfè, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. J. Chem. Inf. Model. 2012, 52, 1138–1145.

(16) Stumpfè, D.; Dimova, D.; Bajorath, J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. J. Med. Chem. 2016, 59, 7667–7676.

(17) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinh, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In Studies in Classification, Data Analysis, and Knowledge Organization; Preissach, C.; Burkhardt, H.; Schmidt Thieme, L.; Decker, R., Eds.; Springer: Berlin, Germany, 2008; pp 319–326.

(18) The UniProt ConsortiumUniProt: A Hub for Protein Information. Nucleic Acids Res. 2015, 43 D204 D212. 10.1093/nar/gku989.

(19) Stumpfè, D.; Dimova, D.; Bajorath, J. Systematic Analysis of Structural and Activity Relationships between Conventional Hierarchical and Analog Series-Based Scaffolds. RSC Adv. 2017, 7, 18718–18723.

(20) https://zenodo.org/deposit/1043537.