Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets

E. Alexis Flores-Padilla,[a] K. Eurídice Juárez-Mercado,[a] José J. Naveja,[b] Taewon D. Kim,[c]
Ramón Alain Miranda-Quintana,[c,d] José L. Medina-Franco[a],*

[a] DIFACQUIM Research Group, Department of Pharmacy, National Autonomous University of Mexico, Mexico City 04510, Mexico

[b] Instituto de Quimica, National Autonomous University of Mexico, Mexico City 04510, Mexico

[c] Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States

[d] Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States

Abstract

The importance of epigenetic drug and probe discovery is on the rise. This is not only paramount to identify and develop therapeutic treatments associated with epigenetic processes but also to understand the underlying epigenetic mechanisms involved in biological processes. To this end, chemical vendors have been developing synthetic compound libraries focused on epigenetic targets to increase the probabilities of identifying promising starting points for drug or probe candidates. However, the chemical contents of these data sets, the distribution of their physicochemical properties, and diversity remain unknown. To fill this gap and make this information available to the scientific community, we report a comprehensive analysis of eleven libraries focused on epigenetic targets containing more than 50,000 compounds. We used well-validated chemoinformatics approaches to characterize these sets, including novel methods such as automated detection of analog series and visual representations of the chemical space based on Constellation Plots and Extended Chemical Space Networks. This work will guide the efforts of experimental groups working on high-throughput and medium-throughput screening of epigenetic-focused libraries. The outcome of this work can also be used as a reference to design and describe novel focused epigenetic libraries.

Keywords: analog series; cheminformatics; Constellation Plots; drug discovery; Extended Chemical Space Networks.

Introduction

Epigenetic drug and probe discovery continue to be relevant for a plethora of potential therapeutic applications, including cancer therapy.[1,2] Probes for epigenetic targets would be key components to understand epigenome data that can be the basis to develop personalized medicine.[3] Similarly, chemical probes targeting epigenetic processes are fundamental in basic chemical biology associated with epigenetic processes. Computational and experimental screening efforts of chemical libraries have been ongoing to identify potential hits for various epigenetic targets. In addition to the traditional general screening libraries available for high-throughput screening, chemical companies have been assembling libraries focused on epigenetic targets[4] and the chemical structures are available in the public domain. Indeed, in epigenetic drug discovery, there are recent and increased efforts to design and analyze focused libraries.[5,6] By design, epigenetic-focused libraries have the potential to increase the epigenetic relevant chemical space, which has recently been revised.[7]

Depending on the library developer, the compounds in the compound data sets are selected following a multi-step procedure, typically through computational approaches (although the specific methodological details remain undisclosed to the public). Many providers emphasize the diversity of the libraries, yet it remains hidden to the user what their contents and coverage of the chemical space are. We propose that a content analysis of a focused library would be informative before screening it for hits. To this end, validated chemoinformatic approaches provide a systematic and rigorous manner to characterize the contents, chemical diversity, and, in general, coverage of the compound libraries in chemical space.[8]

The goal of this study is to rigorously characterize the chemical content, diversity, and drug-like properties of eleven epigenetic-focused libraries containing more than 50,000 compounds in total. All data sets are available in the public domain. To the best of our knowledge, this is the first systematic chemoinformatic analysis of epigenetic-focused compound libraries. As part of the analysis, we implemented a recently introduced and validated methodology to compute and identify core structures and analog series based on retrosynthetic rules. The analog series were the basis to generate Constellation Plots as a rich representation of the chemical space that combines substructure and fingerprint representations. For each epigenetic-focused data set, we identified the most frequent core structures. We also generate novel representations of the chemical space using the concept of Extended Chemical Space Networks (eCSNs)[9] and rapidly identify the most representative individual molecules (i.e., medoids) in a large data set: these chemical structures can be used as a "chemical structural marker" or "chemical diagnostic molecules."[10] Herein, we discuss the most representative chemical structures of the epigenetic libraries that can be used as a criterion for comparing the chemical libraries and prioritization for follow-up studies, including computational and experimental screening.

Methods

To conduct the chemoinformatic characterization of the 11 epigenetic-focused data sets, we calculated properties of pharmaceutical interest; molecular scaffolds using the classic and most common method by Bemis-Murcko,[11] and a novel approach to identify core structures and analog series automatically.[12,13] We also quantify the chemical diversity and synthetic accessibility using whole-molecule structural fingerprints. To explore the distribution of the chemical libraries in chemical space, we generated two complementary and recently introduced visual representations of the chemical space, namely, Constellation Plots[14] and eCSNs.[9] The details of the data sets, their preparation, calculation of the properties, scaffolds, and fingerprints are described in the following sections.

Data sets preparation

The structure files of the focused libraries were obtained from different chemical vendors. The 11 chemical libraries are summarized in Table 1 that briefly describes each set and the number of compounds before and after data curation. The number of compounds after data curation was 53,443, in total. The chemical structures were curated using the open-source cheminformatics toolkit RDKit, version 2021.03.2 (www.rdkit.org). Data curation was performed using an established protocol.[15] Briefly, compounds with valence errors or any chemical element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were deleted. Stereochemistry information was removed because not all compounds in datasets have it defined. Compounds with multiple components were split, and the largest component was retained. The remaining compounds were neutralized and reionized to generate a canonical tautomer subsequently. Duplicated compounds were deleted.

<Table 1 here >

Properties of pharmaceutical interest

For each chemical structure of the 11 curated libraries, six properties of pharmaceutical interest were computed with RDKit, version 2021.03.2: molecular weight (MW), number of acceptor and donor hydrogen bonds (HBA, HBD, respectively), number of rotatable bonds (RB), topological surface area (TPSA), and partition coefficient octanol/water (SlogP). These properties are associated with compound size (MW), polarity (HBA, HBD, TSPA, SlogP), and flexibility (RB). The six properties are the basis of the well-known empirical rules of Lipinski[16] and Veber[17] (i.e., MW≤500, HBD≤5, HBA≤10, SlogP≤5, TPSA≤140, RB≤10) that help to guide the suitability of a compound to be orally absorbed (the preferred use of drug administration in most cases). The properties are not associated directly with biological activity but are commonly used to profile compound screening libraries in drug discovery projects.

Scaffolds: core structures and analog series

We used two approaches to systematically generate and analyze molecular scaffolds, exemplified in Figure 1: Bemis-Murcko scaffolds automatically generated with RDKit, version 2021.03.2; and core structures and analog series using open-source code we previously reported available at https://github.com/navejaromero/analog-series.[12,13]

< Figure 1 here >

To compute the Bemis-Murcko scaffolds[11] the side chains are removed, as illustrated in Figure 1. More specifically, based on the graph representation of the chemical structures of all vertices of degree one. The concept of core scaffolds and analog series is also illustrated in Figure 1 and discussed in detail elsewhere.[12,13] Briefly, core scaffolds and analog series apply a series of fragmentation rules based on retrosynthetic feasibility. Two molecules are considered analogs if the fragmentation rules can map them to the same core fragment, and that fragment is a significant part of the molecule (usually, it contains at least two-thirds of the total number of heavy atoms of each molecule).

Molecular fingerprints

All compound data sets were analyzed with RDKit and MACCS keys (166-bits) fingerprints,[18] both computed with the RDKit toolkit (www.rdkit.org).

Structural diversity

The structural diversity of each compound library was computed using two distinct representations: scaffolds and fingerprints. Each type of representation is complementary and, as discussed elsewhere,[19] provides complementary information towards a more comprehensive assessment of the structural diversity of compound data sets.

Scaffold-based with analog series analysis

The scaffold diversity was also quantified based on the number of cores and analog series, and their ratio compared with the total number of molecules in the dataset. We have previously used the number of identifiable cores and analog series as a measure of chemical diversity in a dataset.[12] However, the 11 data sets presented here are too variable in their size. Therefore, we propose two new indicators: the fraction of molecules in the constellation plot (only analog series with at least three compounds are included there) and the average size of the analog series, measured as the number of unique compounds in the dataset divided by the number of analog series represented in it.

Fingerprint-based

The diversity of the compound libraries was computed with the extended similarity indices as recently reported.[20,21] Briefly, instead of only measuring the similarity between pairs of molecules, the extended (e.g., n-ary) indices allow us to calculate the similarity of any number of molecules simultaneously.[20] This has two key advantages when it comes to the analysis of molecular libraries. First, the n-ary indices provide a truly global description of the correlation between the molecules in the set (as exemplified by their superior performance in estimating the compactness of a set.[21] Second, the extended indices are dramatically more efficient, requiring only O(N) operations to calculate the similarity of N molecules (opposed to the quadratic scaling of the standard binary similarity indices). Extended similarity has been successfully used in numerous applications, including compound diversity analysis,[21] comparison of nucleotide and protein sequences,[22] and, more recently, analysis of molecular dynamics simulations.[10] Herein, we used this approach as a novel and efficient manner to quantify and compare the fingerprint-based diversity of the 11 compound libraries. The extended similarity indices were computed with a fractional weight function, with various coincidence thresholds. The Python code to conduct the extended similarity indices calculations is freely available at https://github.com/ramirandaq/MultipleComparisons.

Chemical space: visual representation

Chemical space has been defined as the set of molecular descriptors in which molecules will be represented. Visual representation of the chemical space helps to better understand the mutual relationships between compounds in that multi-dimensional descriptor coordinates.[8] Among the several methods available (examples of the most common include principal component analysis and t-distributed stochastic neighbor embedding), herein we used two novel representations detailed in this section.

Constellation Plots

Constellation plots are useful to depict the chemical space of chemical libraries containing several analog series.[14] They are helpful at concisely depicting the structure-activity (property) relationships - SA(P)R - in a summarized representation of the data set, as analog series can be represented in fewer data points than individual compounds.[23,24] Of note, only a fraction of the total data is presented in the constellation plot: the compounds forming analog series; in this case, we included only analog series consisting of at least three compounds. Recently, constellation plots have been used to describe a library of antidiabetic natural products[25] and a collection of tubulin inhibitors.[26]

Extended Chemical Space Networks

The chemical space networks (CSNs), proposed and developed by Maggiora and Bajorath, start by measuring the pairwise similarity between the molecules in a data set (using a given similarity coefficient and a compound representation).[27,28] Then, the molecules are represented by nodes which are connected if the similarity is larger than an established threshold. A limitation of this approach is that it is difficult to visualize networks for large compound data sets. Moreover, this approach also requires $O(N^2)$ operations, so it is not well-suited to represent large sections of the chemical space. To overcome this issue, the eCSNs have been recently proposed (*vide supra*). This is a natural generalization of the CSNs, in which any given molecular set can be taken as a node in the network (in this study, the nodes will be the 11 libraries). Then, the relations between these nodes are established via the extended similarity calculated for the union of the corresponding libraries. This coarse-grained representation is markedly more efficient since it exploits the more favorable computational scaling of the n-ary indices.

Synthetic accessibility

The complexity of the compounds generated was estimated using the synthetic accessibility (SA) score previously published.[29] Briefly, the SA score implemented in this study is the difference between a fragment score and a complexity penalty. The fragment score captures common structural features in a large number of already synthesized molecules (934,046 representative molecules from the PubChem). Molecules are fragmented using extended connectivity fragments, and the fragment score is calculated as a sum of contributions of all fragments in the molecule divided by the number of fragments in the molecule. The fragment frequency is related to their synthetic accessibility and hence easy-to-prepare substructures are present in molecules quite often. The complexity score is calculated as the sum of ring complexity (i.e., ring bridge atoms and spiro atoms), the number of stereocenters, large rings (i.e., ring size greater than eight, molecular complexity increases), and molecule size. The SA score was calculated for all epigenetic-focused libraries.

Results and Discussion

Properties of pharmaceutical relevance

Figure S1 in the Supporting Information shows box plots and summary statistics of the distribution of the six calculated properties of pharmaceutical relevance. The profiling of the six properties indicated that, in general, all 11 compound libraries are within the Lipinski and Veber parameters. Based on this criterion, the libraries are acceptable candidate compounds for drug discovery and development programs (in particular, to be administered orally). All 11 compound data sets have a comparable distribution of the six properties, as shown in Figure S1. The outcome of the profiling might be anticipated since it is likely that the chemical vendors (developers) filter or consider the so-called "drug-like" properties during the assembly of the focused libraries. However, the profiling disclosed in this work is relevant and encourages the experimental screening of the 11 compound libraries for drug discovery projects.

Scaffold content: core structures and analog series

The relevance of analyzing the main core scaffold of a chemical compounds, in the the context of drug discovery, is particularly relevant because the central element drives the main molecule shape, arrange the substituents in their specific positions and take part of the biological activity itself.[30] For this reason, systematic profiling of scaffold content of synthetic organic compounds for drug discovery is of utmost relevance.

Generating automatically and consistently the main scaffold or core structure of large data sets can be done in several ways as recently reviewed.[24] In general, it is desirable to generate the scaffolds rapidly, consistently, and interpretable, in particular for an organic or medicinal chemist working on chemical synthesis. As detailed in the Methods section, in this work we implemented two methodologies generating the Bemis-Murcko scaffolds and analog series based on core scaffolds (Figure 1).

Figure S2 in the Supporting Information shows the most frequent Bemis-Murcko scaffolds in the 11 epigenetic-focused libraries. In addition to the ubiquitous benzene ring (present in 1.04% in all databases), which is the most common Bemis-Murcko scaffold in several screening compound libraries, *N*-phenyl-benzenesulfonamide, 1*H*-indol, and *N*-phenylbenzamide, and 1*H*-benzimidazol were the most frequent scaffolds in the 11 epigenetic focused libraries (percentages of 0.45 %, 0.22 %, 0.20 %, and 0.15 %, respectively).

Concerning the core scaffolds and analog series, the most frequent are shown in Figure 2. The obtained cores overlap very little with the 201 substructures that have been annotated as epigenetic bioactive rings in a recent publication: of the 4016 cores matching at least three molecules from any epigenetic data set, only 19 contained at least one of the epigenetic rings. This highlights the structural novelty of the studied libraries[30] and it's potential to expand the epigenetic relevant chemical space.

< Figure 2 here >

After the compound screening, the most frequent scaffolds, core structures, and analog series are potentially privileged or enriched towards epigenetic targets.

Structural diversity

Based on scaffolds

Figure 3A shows the percentage of unique Bemis-Murcko scaffolds in each of the 11 libraries. The analysis revealed that TocrisScreen is the set with the largest percentage of unique scaffolds (85.6 %), indicating a large scaffold diversity considering the definition of the scaffold of Bemis-Murcko. The library with the second largest percentage of unique scaffolds was Targetmol (82.3 %). Figure 3B shows the percentage of scaffolds with a frequency of at least two per library. Clearly, ChemDiv was the data set with the largest proportion of non-unique scaffolds suggesting the lowest scaffold diversity.

< Figure 3 here >

To further quantify scaffold diversity based on the Bemis-Murcko scaffolds, we used cyclic system recovery curves. As documented elsewhere,[31] based on the scaffolds count, the fraction of cyclic systems is plotted against the cumulative fraction of the database. A diagonal represents maximum scaffold diversity, i.e., each compound will have its chemical scaffold. A vertical line represents the minimum scaffold diversity (all compounds have the same scaffold). Figure 3C shows the cyclic system recovery curves for all data sets. The curves can be further characterized by the area under the curve, AUC (maximum diversity: AUC = 0.5; minimum diversity: AUC = 1.0). Table S1 in the Supporting Information summarizes the AUC values for the 11 data sets. The results show that ChemDiv is the least diverse, followed (AUC = 0.87) by OTAVA DNMT3b. In contrast, TocrisScreen and Targetmol are the most diverse (AUC <= 0.56).

The analog series analysis suggested that ChemDiv, Asinex, Life Chemicals, and OTAVA DNMT3b are the least diverse data sets, as they have a larger average of compounds per series. All the other libraries seem to be more diverse, and it is hard to point out at the least diverse from this perspective.

<Table 2 here >

Based on fingerprints

Results of the fingerprint analysis are shown in Figure 4. The figure shows the similarity of the databases computed with RDKit fingerprints and the extended Tanimoto similarity coefficient, at different coincidence thresholds. The analysis revealed that the least diverse set is ChemDiv followed by Asinex. Notably, ChemDiv is the compound data set with the largest number of compounds (27,543) meaning that the larger data set is not necessarily the one with the largest structural diversity, as clearly shown here. Similar results have been obtained for other data sets.[31] These results further emphasize the need to quantify the structural diversity. Similar conclusions regarding the relative diversity of the compound libraries were obtained with other extended similarity indices (in addition to Tanimoto) and MACCS keys, as shown in Figure S3 in the Supporting Information.

< Figure 4 here >

Chemical space: visual representation

Constellation plots

We mapped all compounds into the same chemical space, regardless of the database they came from. We were only interested in analog series having no less than three compounds, even if all of them belonged to different data sets. Afterwards, we highlighted the cores (points) represented in each database. Plots of representative epigenetic libraries are shown in Figure 5, while other libraries are depicted in Figure S4 of the Supporting Information.

< Figure 5 here >

Extended Chemical Space Networks

The eCSNs have been used to visualize the chemical space of 19 large data sets of organic compounds, including natural products, drugs approved for clinical use and other compound libraries, with more than 18 million molecules.[9] As discussed in that paper, this novel representation of the chemical space based on molecular fingerprints is an efficient method to compare the structural relationship among compound libraries. Figure 6 shows a visual representation of the chemical space of the 11 compound libraries using RDKit fingerprints. The network shows that ChemDiv and Asinex (which happen to be the least diverse libraries based on RDKit fingerprints) are at the center of the representation with several connections (similarities) with other databases. In particular, ChemDiv (identified with the number 2 in this network: ID, 2) has the most number of connections, and these connections are closer (visualized by darker linkers between the nodes) with other libraries, such as ApeXBio (ID, 0), SelleckChem (ID, 8), and Targetmol (ID, 9).

< Figure 6 here >

Most representative compound based on fingerprints

Figure 7 shows the three most representative chemical structures per library as calculated with RDKit fingerprints. The twenty more representatives (medoids) per library (including the three in Figure 7) are listed in Table S2 in the Supporting Information. The medoids are calculated using the algorithm described recently.[10] In short, we calculate the complementary similarity of every molecule in a library, that is, the similarity of all but the selected molecule (which we can do in O(N) for the whole set). Then, we can rank all the molecules from more (e.g., medoid-like) to less (e.g., outlier-like) representative by simply ordering them according to the increasing value of their complementary similarity. The reasoning behind this is very simple: by removing a molecule that is closely related to all of the rest we leave behind a more ‘disorganized’ set, which will have a lower complementary similarity. In this context, the medoids could be interpreted as chemical structural markers or “signatures” of the compounds libraries and contribute to profile the chemical contents of each data set.

<Figure 7 here >

Synthetic accessibility

We also profiled the chemical libraries using a validated in silico approach to estimate the synthetic accessibility, as described in the Methods section.[29] Results presented in Figure S5 in the Supporting Information (box plots and summary statistics) indicated that all compound libraries have a comparable profile and, in general, it is expected that the vast majority of the libraries are synthetically accessible. Of note, all focused epigenetic libraries analyzed in this work are commercially available from the chemical providers listed in Table 1.

Conclusions

Herein we report the first comprehensive chemoinformatic analysis of 11 compound libraries focused on epigenetic targets commercially available for screening. Different vendors have previously selected the molecular libraries, but their profile of properties, scaffold contents, and structural diversity was unknown. Profiling of the six properties of pharmaceutical relevance: MW. LogP, HBD, HBA, TSPA, and RB, revealed that all 11 compound libraries are suitable to be screened in drug discovery campaigns to identify molecules that eventually could be orally administered. It was found that, other than benzene, *N*-phenyl-benzenesulfonamide, 1*H*-indol, and *N*-phenylbenzamide, and 1*H*-benzimidazol were the most prevalent. The results of the fingerprint-based diversity indicated that SelleckChem is among the most diverse libraries. In contrast, ChemDiv and Asinex are the least diverse, relative to all other data sets. Regarding the Bemis-Murcko scaffolds and analog series, ChemDiv was also the least diverse, while TocrisScreen, Targetmol, and SelleckChem were the most diverse. Taken together, based on the results of structural diversity, the most diverse library overall (TocrisScreen) should be prioritized for experimental medium-throughput screening. Interestingly, out of the 11 databases analyzed, TocrisScreen was the smallest data set (100 compounds), yet it is the most diverse. In sharp contrast, ChemDiv was the largest data set (27,543 compounds) but is the least structurally diverse. The scaffold content and analog series, analyzed in the context of rings present in currently known compounds with activity against epigenetic targets revealed that the focused libraries have a large potential to expand the epigenetic relevant chemical space. Results of the calculated synthetic accessibility showed that all compound data sets are, in general, feasible to make. For practical applications, the libraries could be acquired first by the chemical vendors but, if needed, could be synthesized in-house. We anticipate that the results of the chemoinformatic characterization discussed in this work will assist research teams in the decision-making process and prioritize what libraries move forward to experimental screening in, for example, a high-throughput screening setting.

Acknowledgments

Discussions and technical assistance of Ana L. Chávez-Hernández are greatly acknowledged. We thank support from DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación

Tecnológica (PAPIIT), grant IN201321. RAMQ acknowledges support from the University of Florida in the form of a startup grant. JJN is grateful to Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for a postdoctoral scholarship.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Ganesan, P. B. Arimondo, M. G. Rots, C. Jeronimo, M. Berdasco, *Clin. Epigenetics* **2019**, *11*, 174.

[2] Y. Wang, Q. Xie, H. Tan, M. Liao, S. Zhu, L.-L. Zheng, H. Huang, B. Liu, *Pharmacol. Res.* **2021**, 105702.

[3] G. De Riso, S. Cocozza, *Curr. Med. Chem.* **2020**, DOI 10.2174/0929867327666201117142006.

[4] Z. Sessions, N. Sánchez-Cruz, F. D. Prieto-Martínez, V. M. Alves, H. P. Santos, E. Muratov, A. Tropsha, J. L. Medina-Franco, *Drug Discov. Today* **2020**, *25*, 2268–2276.

[5] A. I. Green, G. M. Burslem, *J. Med. Chem.* **2021**, *64*, 7231–7240.

[6] A. I. Green, G. M. Burslem, *RSC Med. Chem.* **2021**, DOI 10.1039/D1MD00193K.

[7] D. L. Prado-Romero, J. L. Medina-Franco, *ACS Omega* **2021**, *6*, 22478–22486.

[8] J. L. Medina-Franco, N. Sánchez-Cruz, E. López-López, B. I. Díaz-Eufracio, *J. Comput. Aided Mol. Des.* **2021**, DOI 10.1007/s10822-021-00399-1.

[9] T. B. Dunn, G. M. Seabra, T. D. Kim, K. E. Juárez-Mercado, C. Li, J. L. Medina-Franco, R. A. Miranda-Quintana, *J Chem Inf Model* **2021**. Submitted.

[10] L. Chang, A. Perez, R. A. Miranda-Quintana, *BioRxiv* **2021**, DOI 10.1101/2021.08.08.455555.

[11] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

[12] J. J. Naveja, B. A. Pilón-Jiménez, J. Bajorath, J. L. Medina-Franco, *J. Cheminform.* **2019**, *11*, 61.

[13] J. J. Naveja, M. Vogt, D. Stumpfe, J. L. Medina-Franco, J. Bajorath, *ACS Omega* **2019**, *4*, 1027–1032.

[14] J. J. Naveja, J. L. Medina-Franco, *Front. Chem.* **2019**, *7*, 510.

[15] A. L. Chávez-Hernández, N. Sánchez-Cruz, J. L. Medina-Franco, *Mol. Inform.* **2020**, *39*, e2000050.

[16] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.

[17] D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615–2623.

[18] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

[19] M. González-Medina, F. D. Prieto-Martínez, J. R. Owen, J. L. Medina-Franco, *J. Cheminform.* **2016**, *8*, 63.

[20] R. A. Miranda-Quintana, D. Bajusz, A. Rácz, K. Héberger, *J. Cheminform.* **2021**, *13*, 32.

[21] R. A. Miranda-Quintana, A. Rácz, D. Bajusz, K. Héberger, *J. Cheminform.* **2021**, *13*, 33.

[22] D. Bajusz, R. A. Miranda-Quintana, A. Rácz, K. Héberger, *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3628–3639.

[23] J. L. Medina-Franco, J. J. Naveja, E. López-López, *Drug Discov. Today* **2019**, *24*, 2162–2169.

[24] J. J. Naveja, M. Vogt, *Molecules* **2021**, *26*, DOI 10.3390/molecules26175291.

[25] A. Madariaga-Mazón, J. J. Naveja, J. L. Medina-Franco, K. O. Noriega-Colima, K. Martinez-Mayorga, *RSC Adv.* **2021**, *11*, 5172–5178.

[26] E. López-López, C. M. Cerda-García-Rojas, J. L. Medina-Franco, *Molecules* **2021**, *26*, DOI 10.3390/molecules26092483.

[27] G. M. Maggiora, J. Bajorath, *J. Comput. Aided Mol. Des.* **2014**, *28*, 795–802.

[28] M. Zwierzyna, M. Vogt, G. M. Maggiora, J. Bajorath, *J. Comput. Aided Mol. Des.* **2015**, *29*, 113–125.

[29] P. Ertl, A. Schuffenhauer, *J. Cheminform.* **2009**, *1*, 8.

[30] P. Ertl, *J. Chem. Inf. Model.* **2021**, DOI 10.1021/acs.jcim.1c00761.

[31] J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender, T. Scior, *QSAR Comb. Sci.* **2009**, *28*, 1551–1560.

TABLES

Table 1. Epigenetic-focused screening libraries and reference data sets.

ID	Library	Size	Size (curated)[a]	General contents	Web site accessed June 2021
0	ApeXBio Discovery Probe Epigenetics Compound Library	328	310	Small molecules with activity against different epigenetic targets.[b]	https://www.apexbt.com/discoveryprobetm-epigenetics-compound-library.html
1	Asinex Epigenetic Library	5,391	5,313	Library focused on bromodomains and histone methyltransferase inhibitors. Compounds designed by a combination of structure and ligand-based methods.	http://www.asinex.com/Epigenetics/
2	ChemDiv Epigenetic Set	30,431	27,543	Library designed for modulation for all three classes of epi-targets. Molecules were designed based on a combination of structure-based (based on X-ray and nuclear magnetic resonance) and ligand-based approaches.	https://www.chemdiv.com/epigenetics-library/
3	Enamine Epigenetics Library	9,352	9,352	Library designed with a combination of ligand and structure-based methods. Compound optimization was made via pharmacophore profiling.	https://enamine.net/hit-finding/focused-libraries/view-all/epigenetics-libraries
4	Life Chemicals Epigenetics Screening Library	7,019	7,011	The library is focused on methylation-related epi-enzymes. Designed with similarity methods.	https://lifechemicals.com/screening-libraries/targeted-and-focused-screening-libraries/epigenetic-screening-libraries
5	MedChemExpress Epigenetics Library	700	650	Library designed for several epi-targets. The targets are not disclosed.[b]	https://www.medchemexpress.com/virtual-screening/epigenetics-library.html
6	OTAVA DNMT1 DNMTs Targeted Libraries	466	399	Drug-like compounds selected from virtual screening using docking and pharmacophore modeling.	https://www.otavachemicals.com/targets/dnmt1-and-dnmt3b-targeted-libraries
7	OTAVA DNMT3b DNMTs Targeted Libraries	1,261	1,230	Drug-like compounds selected from virtual screening using docking and pharmacophore modeling.	https://www.otavachemicals.com/targets/dnmt1-and-dnmt3b-targeted-libraries
8	SelleckChem Epigenetics Compound Library	699	677	The library contains inhibitors for several epi-targets. Focused on experimental tests and for HTS validation.[b]	https://www.selleckchem.com/screening/epigenetics-compound-library.html
9	Targetmol Epigenetics Compound Library	932	859	A set of epi-regulators whose primary focus is lead optimization and HTS.[b]	https://www.targetmol.com/compound-library/Epigenetics-Inhibitor-Library
10	TocrisScreen Epigenetics	101	99	Collection of small molecules covering more than 40 epigenetic targets (including readers, writers, erasers, and transcriptional modulators).[b]	https://www.tocris.com/products/tocriscreen-epigenetics-library_6801

[a] Data curation was done with a standard protocol described in [16] to conduct a comparative chemoinformatic profile of the compound libraries. [b] Design approach not disclosed.

Table 2. Measures of scaffold diversity based on analysis of cores and analog series.

Library	Number of analogue series (AS)	Number of cores	Fraction of molecules in constellations plot	Average size of AS
ApeXBio	298	785	2.9% (9/310)	1.04
Asinex	3423	8795	33.4% (1775/5313)	1.55
ChemDiv	10,681	43,292	61.7% (17,017/27,543)	2.58
Enamine	8737	20,572	4.4% (409/9352)	1.07
Life Chemicals	5159	13,500	24.6% (1726/7011)	1.36
MedChemExpress	618	1636	2.8% (18/650)	1.05
OTAVA DNMT1	372	548	9.8% (39/399)	1.07
OTAVA DNMT3b	961	2855	24.5% (301/1230)	1.28
SelleckChem	649	1635	4.4% (30/677)	1.04
Targetmol	829	2045	3.6% (31/859)	1.04
TocrisScreen	95	242	7% (7/99)	1.04

FIGURES

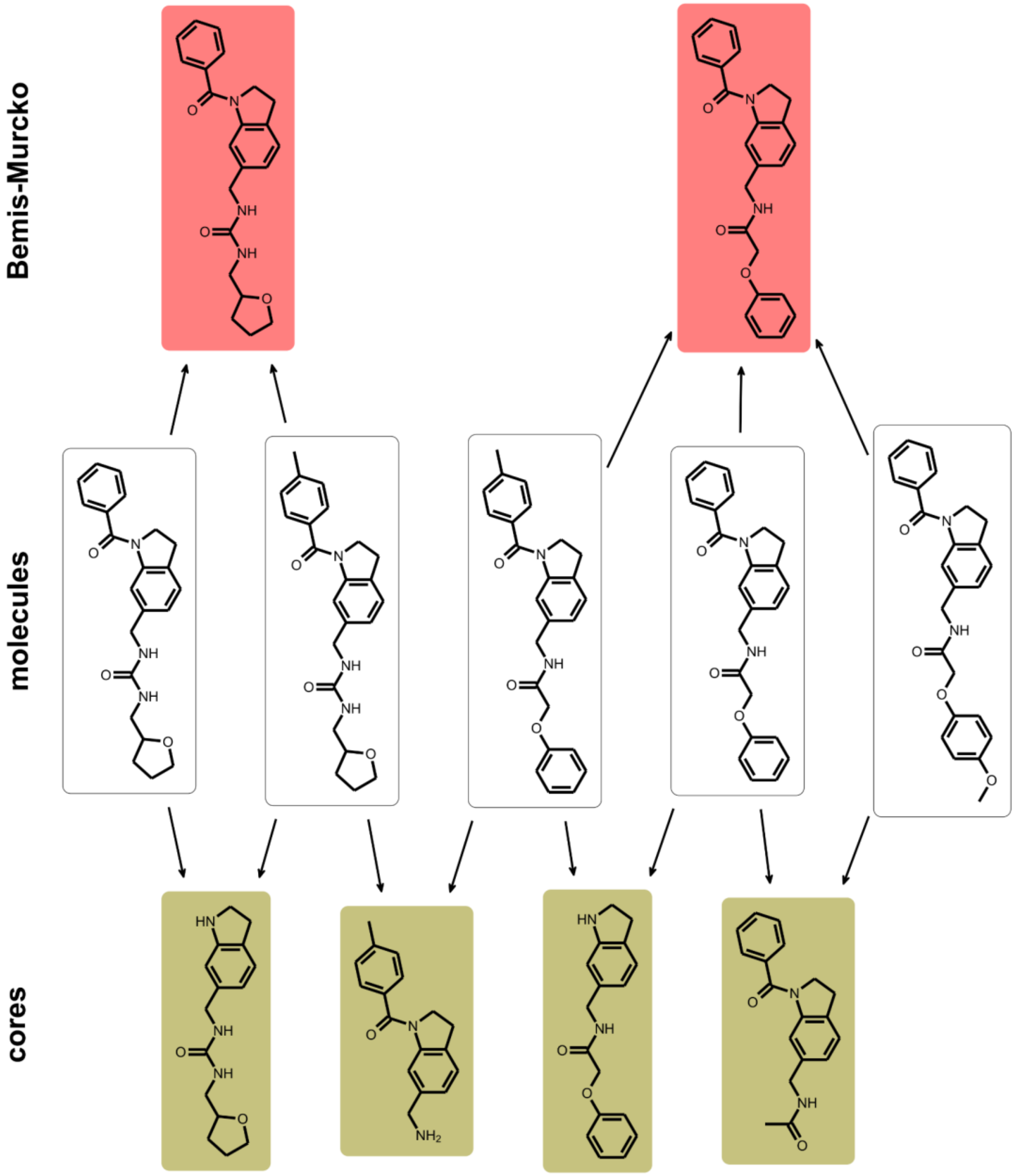

Figure 1. General approaches to compute molecular scaffolds. Note that the Bemis-Murcko approach maps every molecule to only one scaffold. Small changes in the scaffold results in a failure to identify analogs. On the other hand, the core approach is in many instances (but not always) able to identify such analogs. The molecules shown are only a small subset of an analog series consisting of over 400 compounds (AS7684).

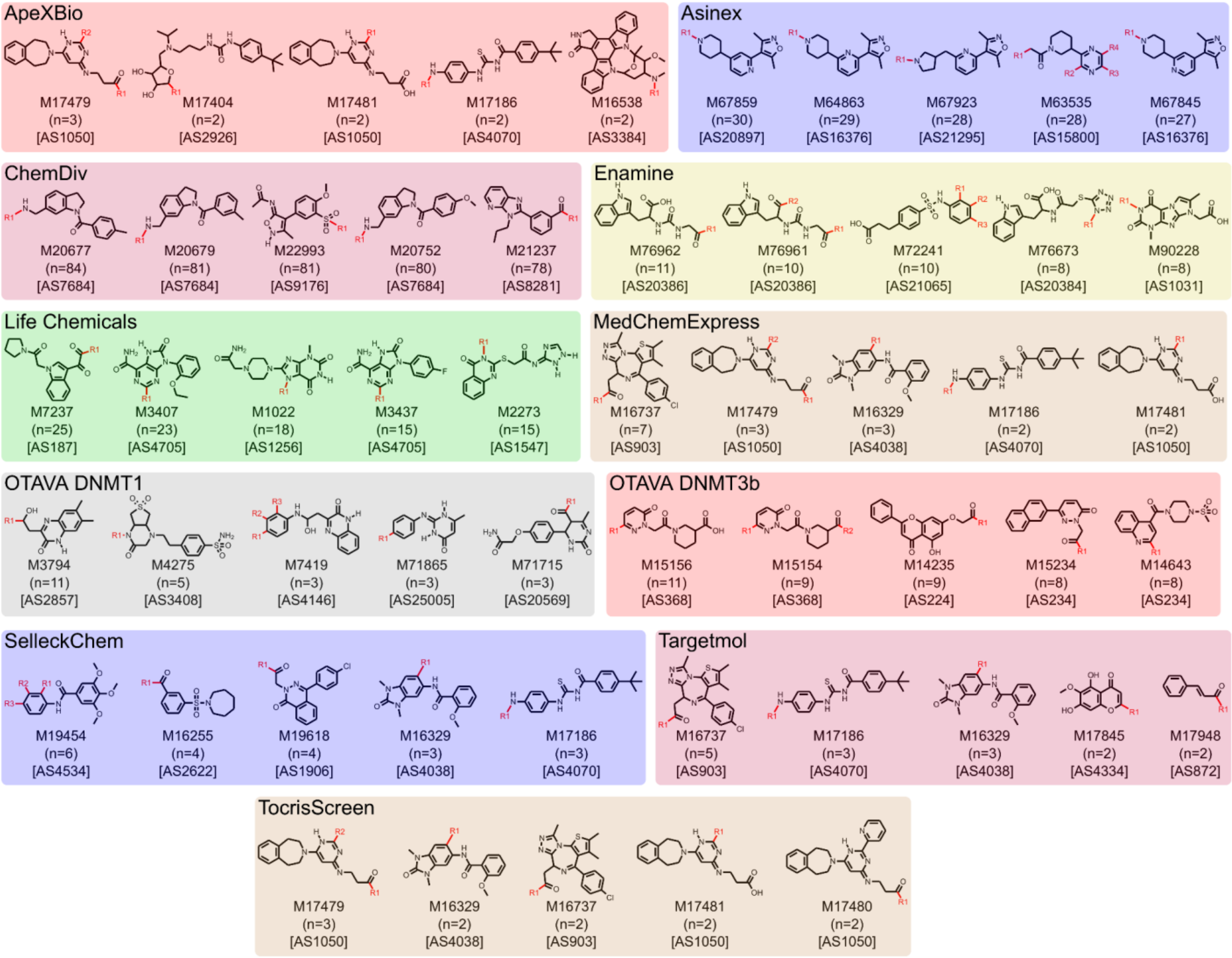

Figure 2. Chemical structures of the five most frequent core scaffolds of each library. At every instance, the substitution sites present in the complete database (all the libraries) are indicated, as well as the number of molecules represented in the specific library between brackets. The analog series ID (ASID) is indicated in square brackets.

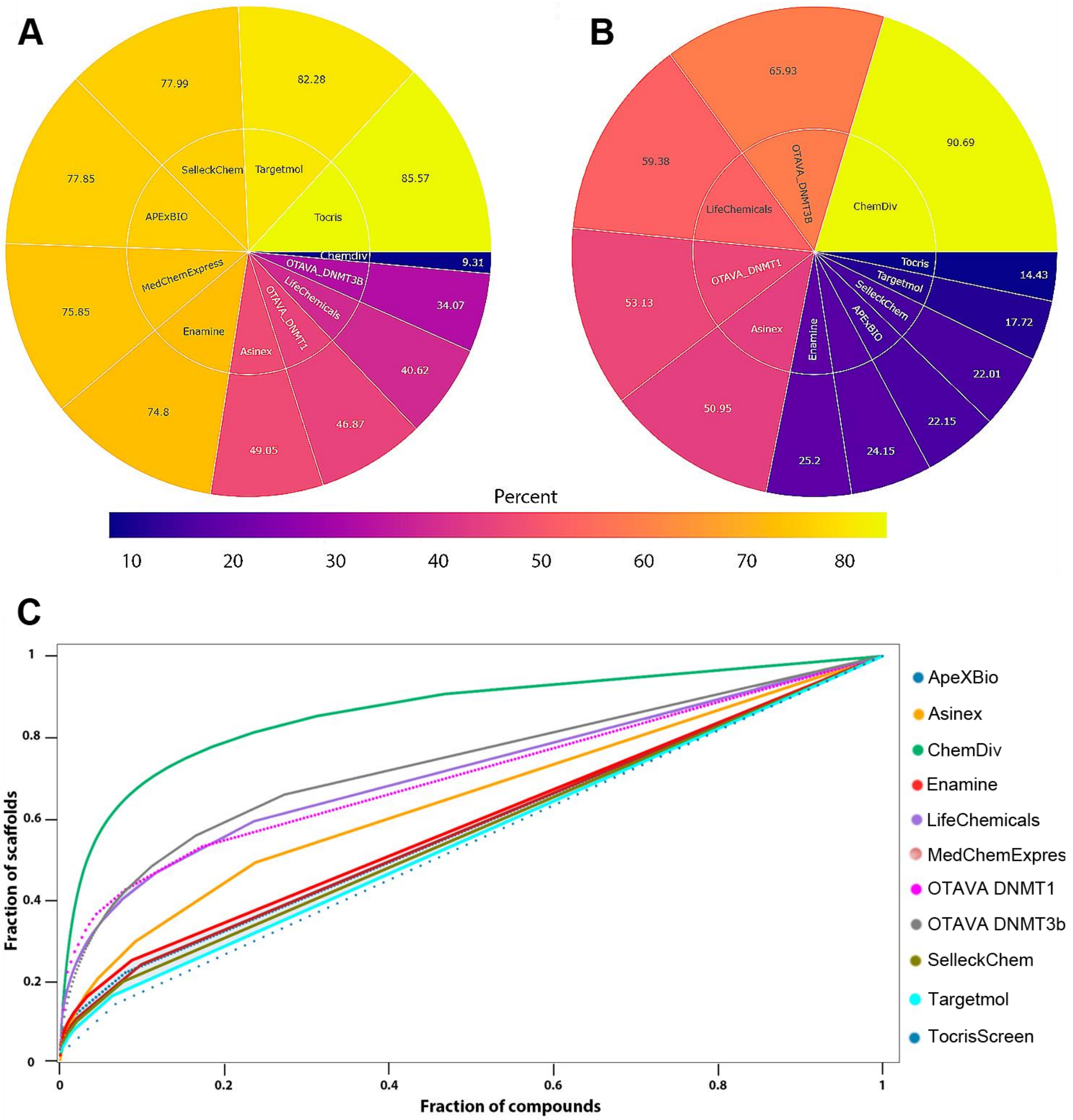

Figure 3. Quantification of scaffold diversity based on Bemis-Murcko. **A**) Percentage of unique scaffolds. **B**) Percentage of scaffolds with a frequency of at least two. **C**) Cyclic system recovery curves.

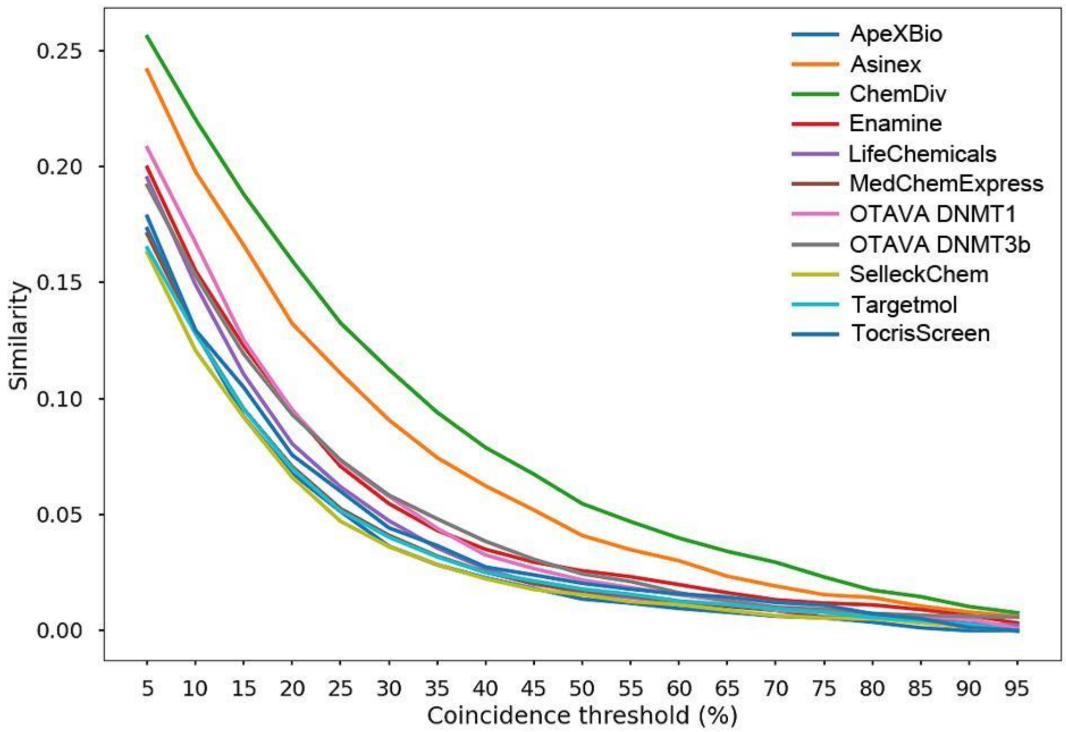

Figure 4. Fingerprint-based similarity of the eleven data sets calculated with RDKit fingerprints and the extended Jaccard-Tanimoto coefficient at different coincidence thresholds.

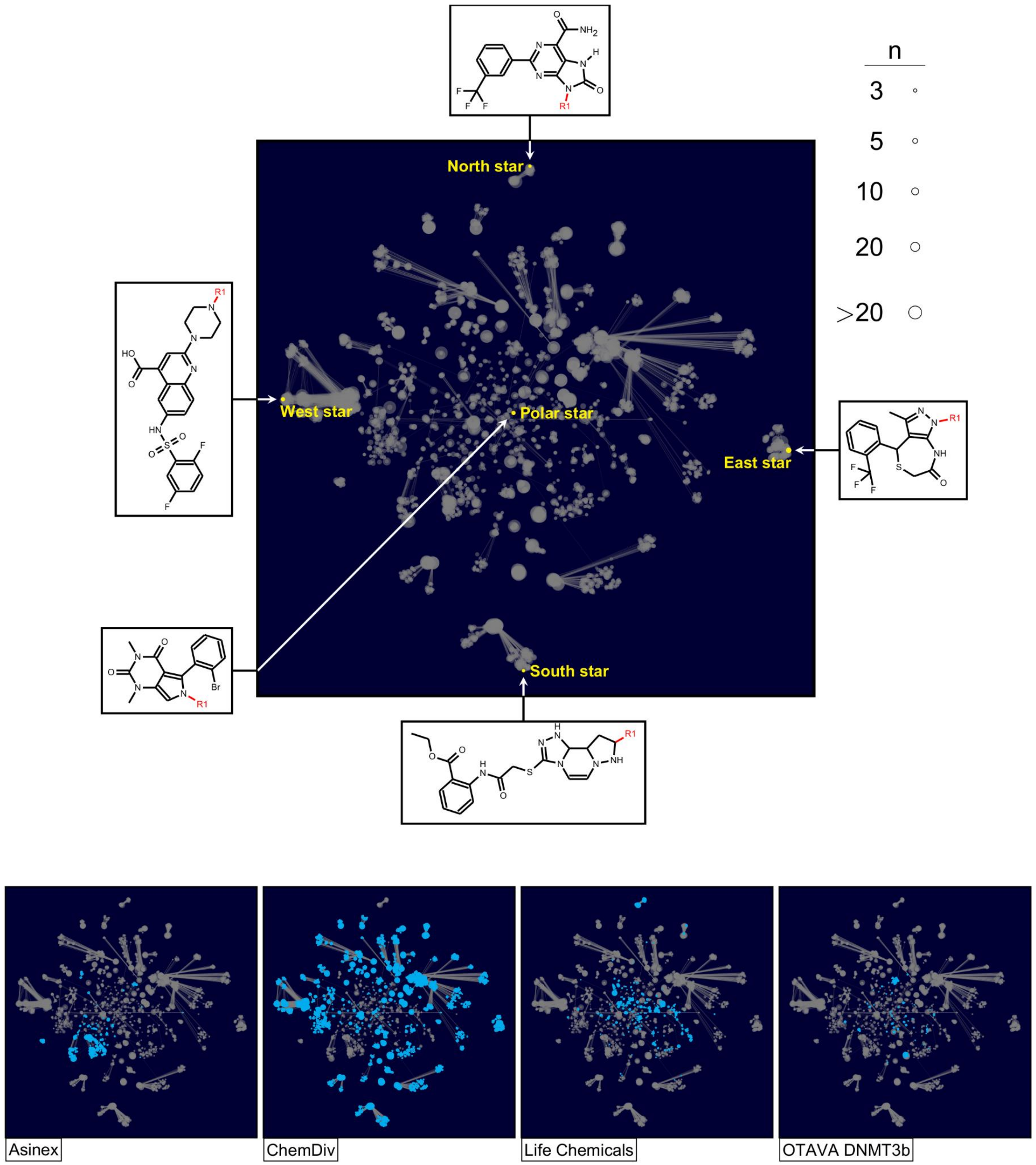

Figure 5. Constellation Plots for selected data sets. Every dot represents a core structure, and analog series consisting of more than a single core are connected by lines. The mapping in the chemical space is the same for all libraries. Above, a few reference cores (stars) are shown to illustrate the fact that constellations represent the chemical space of analog series. Below, the distribution of selected data sets is shown by coloring cores represented in the corresponding library. For inspecting the distribution of every data set see the Supplementary Information, Figure S4.

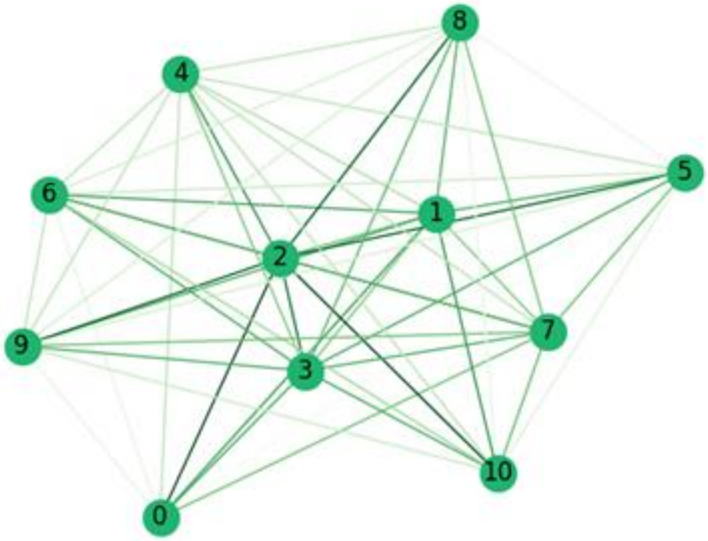

Figure 6. Visual representation of the chemical space of the 11 libraries using extended Chemical Space Networks using RDKit fingerprints at coincidence threshold of 90. Code: **0**: ApeXBio; **1**, Asinex; **2**, ChemDiv; **3**, Enamine; **4**, LifeChemicals; **5**, MedChemExpress; **6**, OTAVA DNMT1; **7**, OTAVA DNMT3b; **8**, SelleckChem; **9**, Targetmol; **10**, TocrisScreen.

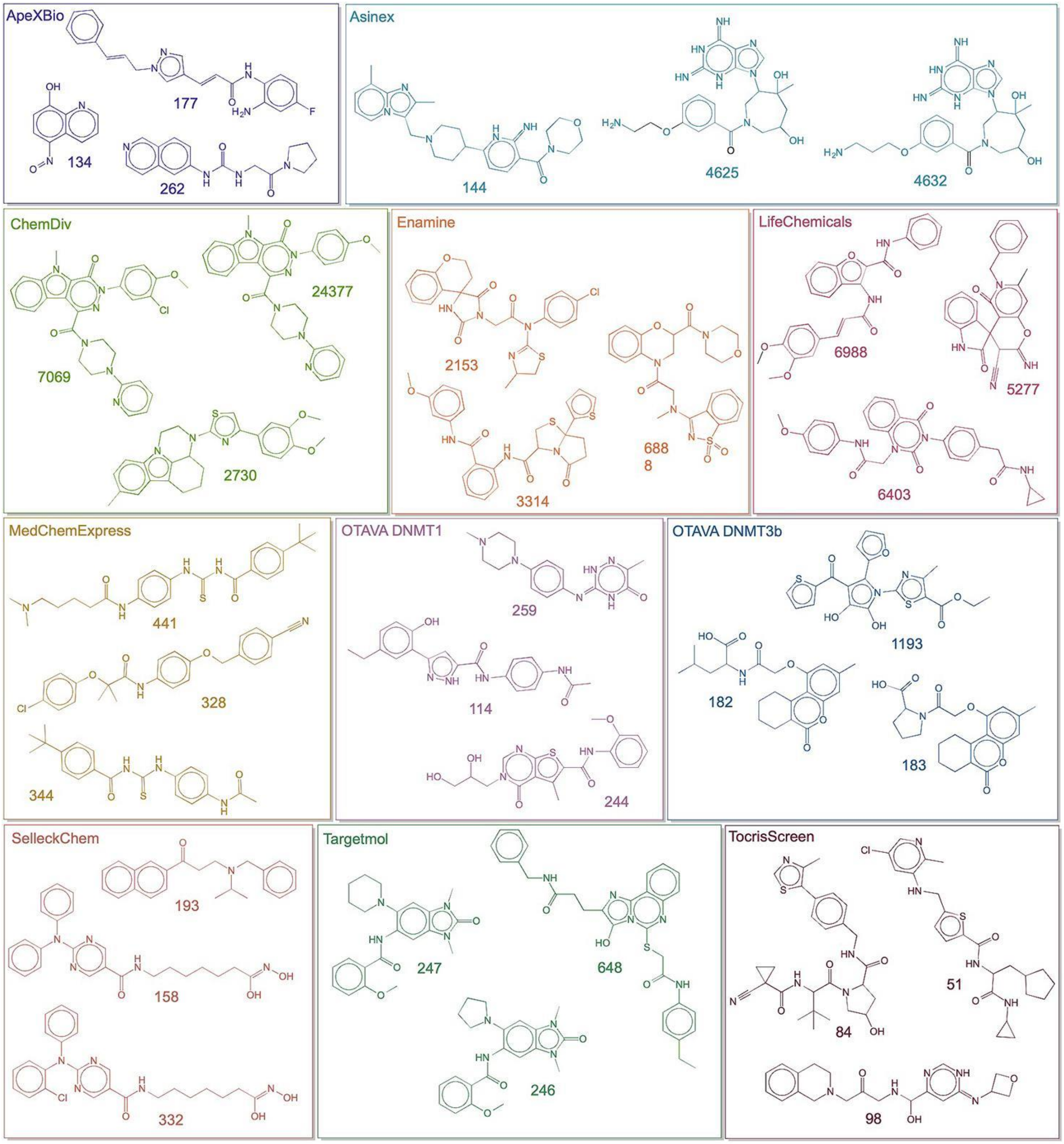

Figure 7. Three most representative compounds (medoid) as calculated with RDKit fingerprints.