JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim
Article

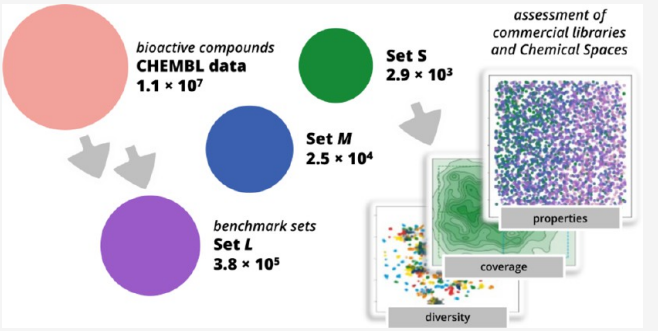
A Benchmark Set of Bioactive Molecules for Diversity Analysis of Compound Libraries and Combinatorial Chemical Spaces

Published as part of Journal of Chemical Information and Modeling special issue "Chemical Compound Space Exploration by Multiscale High-Throughput Screening and Machine Learning".

Alexander Neumann* and Raphael Klein

Cite This: J. Chem. Inf. Model. 2025, 65, 9097−9124

ABSTRACT: Sources for commercially available compounds have been experiencing continuous growth for several years, reaching their peak in billion- to trillion-sized combinatorial Chemical Spaces. To assess the quality of a compound collection to provide relevant chemistry, a benchmark set of pharmaceutically relevant structures is required that enables an unbiased comparison. For this purpose, the ChEMBL database was mined for molecules displaying biological activity, and three benchmark sets of successive orders of magnitude were created by systematic filtering and processing: Set *L* ("large-sized," 379k), Set *M* ("medium-sized," 25k), and Set *S* ("small-sized," 3k). Tailored for broad coverage of the physicochemical and topological landscape, the benchmark Set *S* was then employed to analyze the chemical diversity capacities of commercial combinatorial Chemical Spaces and enumerated compound libraries. Among the three utilized search methods—FTrees (pharmacophore features), SpaceLight (molecular fingerprints), and SpaceMACS (maximum common substructure)—eXplore and REAL Space consistently performed best. In general, each Chemical Space was able to provide a larger number of compounds more similar to the respective query molecule than the enumerated libraries, while also individually offering unique scaffolds for each method.



INTRODUCTION

The accessibility to small molecules of interest has been significantly expanded in recent years through the continuous growth of commercial compound libraries.[1−3] Where promising compounds and analogs previously had to be synthesized in-house, it has become common practice to screen vendor catalogs for structurally related substances that can be purchased in small quantities for biological testing. This allows for accelerated insights into structure−activity relationships (SARs) and ultimately a speedup of the whole lead optimization process. Beyond the obvious aspects of convenience and flexibility in choosing from one of the many vendors, this approach also enables savings in personnel, material, and operational costs. Vendors, for economic reasons, have largely optimized their processes to offer screening compounds at competitive prices, making it economically unfeasible for in-house synthesis to compete. For example, compounds from Enamine's REAL Space are officially listed in the price range of 163−245 USD per 1 mg, which can hardly be competed against with the costs of an experienced chemist, required materials, purification, and storage in the Western world.[4]

However, it cannot be denied that the continuous growth of compound libraries brings well-known challenges in processing larger data sets.[3,5−7] This becomes particularly evident in the context of combinatorial compound collections, the so-called Chemical Spaces (written with capital letters in this study to distinguish them from the concept of chemical space). Nowadays, Chemical Spaces encompass billions to trillions of commercially accessible compounds of which only the smallest fraction has already been listed somewhere, displaying vast potential for novel intellectual property.[8−10] After all, setting up a combinatorial Chemical Space focusing on just six functional groups can already result in a substantial size of 2×10^8.[11]

Understandably, these numbers are not physically kept in stock on shelves but are synthesized and delivered upon request for the customer. Alternatively, the option is also available to order the corresponding building blocks and synthesize the desired product oneself. In this case, the Chemical Spaces serve as a hunting ground for all possible synthesis options resulting

Received: March 31, 2025
Revised: May 21, 2025
Accepted: June 30, 2025
Published: August 20, 2025

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

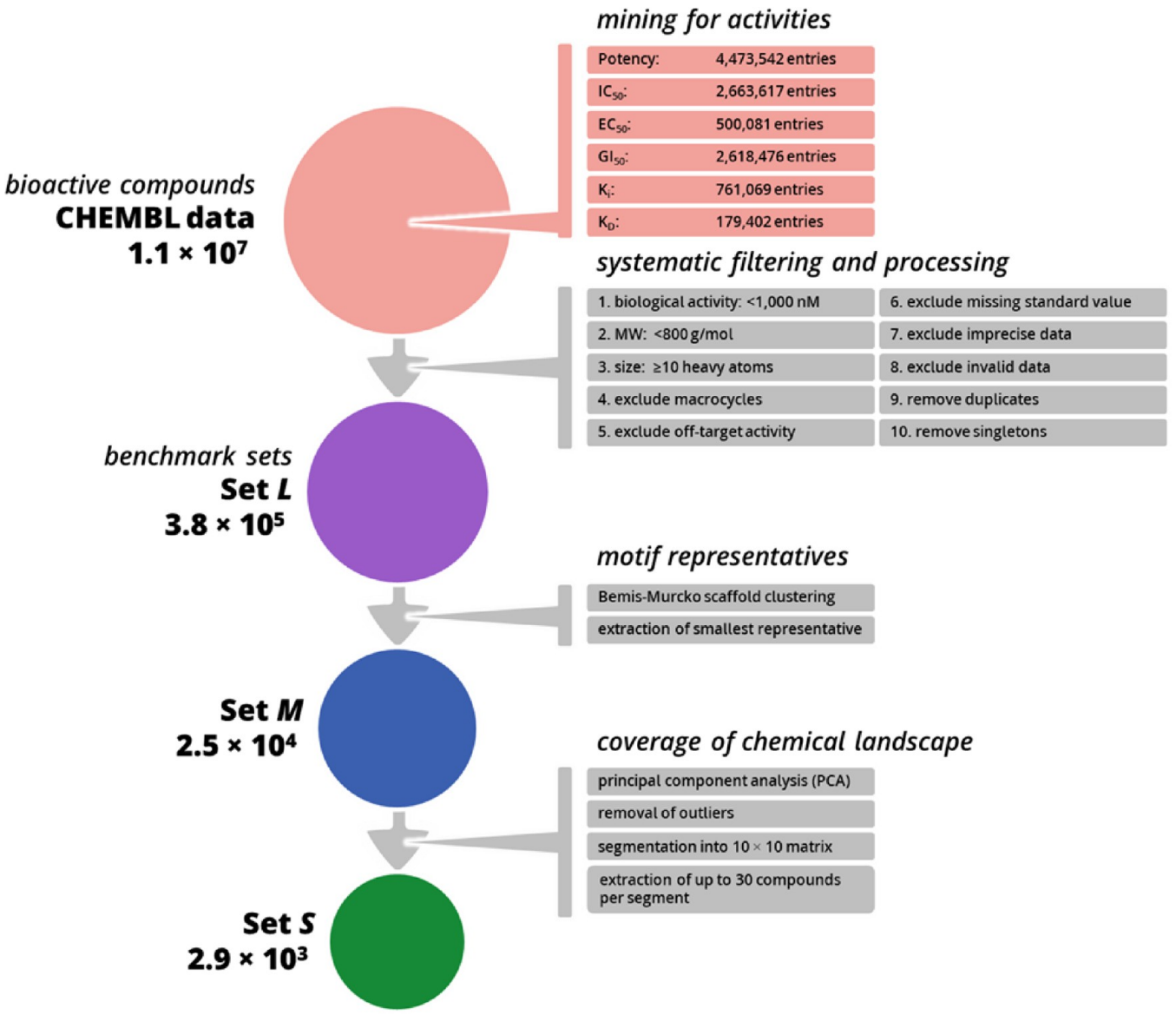


Figure 1. Systematic processing and filtering of the ChEMBL data with subsequent processing steps to achieve three benchmark sets: "large-sized" Set *L*, "medium-sized" Set *M*, and "small-sized" Set *S*.

from the utilized building blocks. Unlike conventional enumerated molecule libraries, where each compound entry is explicitly listed, Chemical Spaces consist of building blocks and encoded chemical reactions, which can be searched for relevant chemistry using algorithms specifically designed to operate in their combinatorial architecture.[12−14] Similarity searches with these algorithms enumerate the desired number of top-ranking results in common standard formats (SMILES, SDF). The associated sheer size of a compound inside the Chemical Spaces therefore makes it more difficult to conveniently assess the data.

While several efforts in studying combinatorial Chemical Spaces have already been conducted,[15−22] a holistic understanding of their capacities and blind spots has yet to be achieved. Although investigations of vendor libraries have already been conducted, no focus was placed on the applicable, direct relevance of the results in the context of modern drug discovery.[23−25] This leads to gaps in understanding their utility and relevance for modern drug discovery challenges from academic and industry perspectives.[26]

Against this background, the same questions arise as with smaller substance libraries: Despite the growing number of entries, what are the blind spots of commercial libraries regarding small molecule drug discovery? How can the diversity of a compound collection be assessed, and consequently, how should a data set be designed to address these questions?

With the advent of machine learning (ML) and artificial intelligence (AI) methods, the principle of "the more, the merrier" toward data set sizes has been embraced to provide models with sufficient data points to improve the quality of results.[27−30] What the resulting benchmark sets have in common are several million data points. While highly relevant for training models, these numbers lead to long computational times during sequential processing in descriptor-driven approaches (e.g., molecular fingerprint screens, substructure searching, and ensemble docking). It would therefore be advantageous to have a representative set of bioactive molecules that is several orders of magnitude smaller than the published benchmark sets and suitable for routine applications. This data set should possess the following characteristics: (i) a size that allows the

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

completion of multiple chemical informatics tasks within a manageable time frame and can be processed even with standard hardware in most setups, (ii) contain molecules with high relevance for drug discovery, and (iii) a meaningful and broad coverage of chemical space.

Resorting to FDA-approved drugs to assess and compare compound collections for their diversity and chemical space coverage can introduce certain biases. After all, the FDA-approved drug list represents the crème de la crème of therapeutics—the result of years of campaigns, countless optimization iterations, and extensive human trials. This inevitably leads to certain target classes being overrepresented, which can impact the frequency of specific molecular motifs within the set. On the other hand, possible underexplored intervention points, or those whose first-in-class candidates are still in clinical phases, are not covered by the list, potentially limiting the exploration of appropriate chemical diversity for drug discovery. Another drawback of FDA-approved drugs is their prior optimization history: it is quite common for five or more synthesis steps to be required to obtain the desired product.[31] While this may be a necessity due to the lack of commercially available building blocks, on an industrial scale, it can also be driven by the need to avoid unwanted byproducts. Ultimately, there is no way around it when a specific compound with the desired decorations is needed because it possesses the best physicochemical and pharmacological and safest toxicological properties.

Nevertheless, this highlights an incompatibility with the approach of conveniently obtaining desired compounds in one or two steps with a high success rate—an expectation often set for initial hit molecules.

Other published data sets or publicly accessible collections come with different issues, such as blurred or missing activity data, low relevance for modern drug discovery purposes, or limitations in their design.[32−34] There is no question that using raw data without some form of scrutiny, filtering, or validation is neither effective nor useful.

For these reasons, the study presented here focuses on the creation of multiple benchmark sets of relevant bioactive molecules that reflect the current coverage of the chemical landscape. After mining the ChEMBL database for compounds with reported biological activity, the data were systematically filtered and processed for relevant entries. Chemical landscape coverage-focused extraction of representative molecules ultimately enabled a scaling down to three orders of magnitude: sets of 379k, 25k, and 3k structures.

Subsequently, the smallest set was used to examine commercial compound sources, namely, combinatorial Chemical Spaces and enumerated vendor libraries, for their chemical diversity and relevance for hit identification and expansion in the early stages of drug discovery.

RESULTS AND DISCUSSION

Data sets were extracted from ChEMBL, containing data points for the following commonly measured parameters: IC_{50}, GI_{50}, K_i, EC_{50}, and K_D, as they allow for a numerical categorization based on potency, facilitating the identification of promising hits. Compounds in the nanomolar range are typically of the highest interest as good starting points for subsequent lead optimization. Consequently, the following entry numbers were obtained (accessed in January 2024): potency 4,473,542, IC_{50} 2,663,617, GI_{50} 2,618,475, K_i 761,069, EC_{50} 500,081, and K_D 179,402, totaling 11,196,186 entries (including duplicates). The following parameters were excluded because they do not provide a numerical value for straightforward compound categorization: percent effect, activity, MIC, and inhibition. We took into account that the resulting data set may display a target-focused bias in regard to the chemical diversity of compounds, and, for example, data points for antibacterials or compounds with an unknown target structure may be excluded.

This raw data set of ChEMBL compounds of over 11 million data points with reported biological data was the starting point for setting up the benchmark set. From the perspective that the data set should cover relevant compounds for small molecule drug discovery, the following filters were applied (a summary is presented in Figure 1):

(1) Biological activity: values in the nanomolar range (<1000 nM). As described above, the purpose of the benchmark set is to represent relevant chemistry for drug discovery, which also entails favorable biological activity. Although the definition of an initial hit can vary depending on the project, and values between 10 and 100 μM are often tolerated, especially for targets without any known binders, our focus was on more potent compounds, as these are considered better starting points.[35−37] Artificial values of 0 nM were excluded.

(2) Molecular weight (MW): In recent years, a shift in trend has been observed regarding beyond-rule-of-five (bRo5) compounds.[38−42] Therefore, the MW cutoff was set at 800 g/mol to ensure coverage of the growing compound sizes of bioactive compounds.

(3) Size: Only compounds containing at least 10 heavy atoms were included. While it is certainly possible that some potent ultralow-molecular-weight compounds[43] may be excluded as a result, the typical potency of fragments lies in the micromolar to millimolar range,[44] meaning that they would be excluded by filter (1). Furthermore, compound vendors often make an explicit distinction between fragment and compound libraries. By applying a filter of at least 10 heavy atoms, the benchmark set focused on drug- and lead-like compounds. This setting also has the added benefit of removing "trivial" compounds, such as ChEMBL20936 (benzamidine) with 952 nM potency, ChEMBL1551365 (ethyl nitrite) with 944 nM, and ChEMBL1200471 (pyrithione) with 907 nM.

(4) No macrocycles: Macrocyclic structures (compounds containing more than nine atoms in a ring) were excluded. While macrocyclic structures are gaining increasing popularity, their chemical background is often decoupled from that of small molecules.[45] Compound vendors may offer dedicated macrocycle-focused libraries, which are based on different synthetic strategies, such as combinatorial approach (e.g., DNA-encoded libraries (DELs)), or cyclization at an advanced stage of the synthesis route.[46−48] Due to these pronounced differences in physicochemical properties and the differentiation of libraries practiced by the vendors, a decision was made here in favor of small molecules. It should be mentioned that the investigated Chemical Spaces do not contain macrocycles as building blocks or reactions that encode them as a product.

(5) No off-target activity: Data for the following targets and target families were excluded from the benchmark set: HERG (human Ether-à-go-go-Related Gene), Kir (inward-rectifier potassium channel), Kv (voltage-gated

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

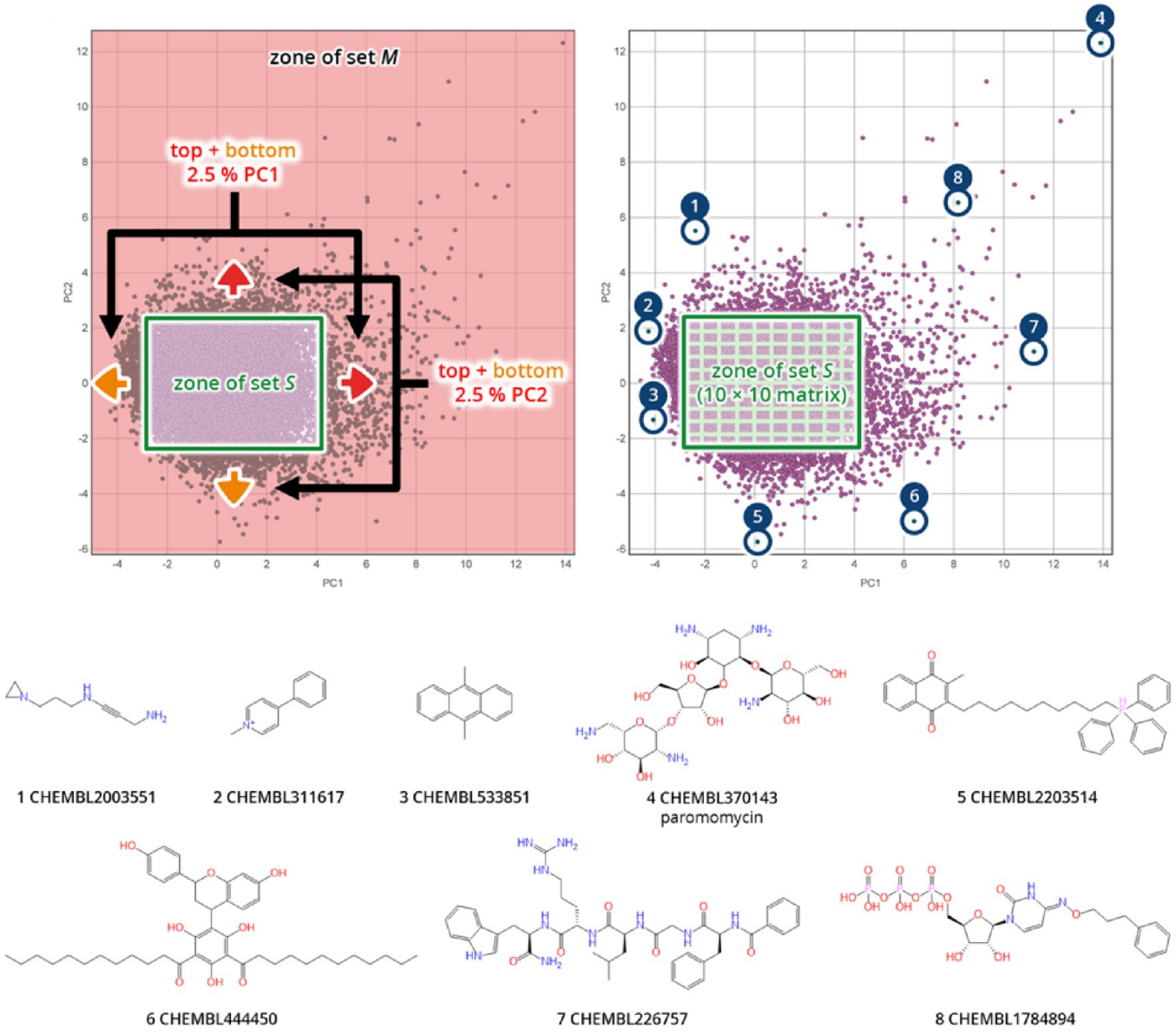


Figure 2. Visualization of the trimming of Set *M* by outlier removal to achieve Set *S*. Examples of excluded outliers from the sparsely populated regions of the PCA of Set *M* feature fragments, polycyclic aromatic hydrocarbons (PAHs), structures bearing long aliphatic chains, peptides, and charged molecule classes such as nucleotides and aminoglycosides.

potassium channels), and cytochrome P450 enzymes. Although there are numerous data points for biological activity on these structures available, they are generally collected during assessments of off-target compound toxicity.[49,50] By eliminating these data points, any misleading bias regarding toxically active compounds can be minimized.

(6) Activity: Compounds missing a standard value were removed.

(7) Data interpretability: Imprecise or semiquantitative values for standard relations ("≫" and ">"), which do not allow for a comparison, were excluded.

(8) Data validity: Compounds with the following parameters for "Data Validity Comment" were removed: "outside range", "potential missing data", "potential transcription error", "outside", and "error".

(9) Duplicates: Duplicates were removed.

(10) Singular events: In the final step, singletons were removed. In this study, we defined singletons as compounds that have fewer than five molecules with the same Bemis−Murcko scaffold within the raw data set. The reason for removing the singletons is to eliminate random, not further investigated hits. Compounds that, for instance, emerged from high-throughput screening (HTS) provide limited data, which in turn makes it difficult to verify them as relevant scaffolds. In contrast, compounds that are part of a series or exhibit biological activity against multiple biological targets indicate their potential as drug candidates and privileged structures.

After filtering, a Set *L* ("L" denoting "large-sized") consisting of 379,169 compounds was obtained. Characteristics of this set include drug- to lead-like structures with relevant reported biological activities. Additionally, activity singularities were removed, resulting in an accumulation of privileged scaffolds. To enable the most versatile options for various drug discovery

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

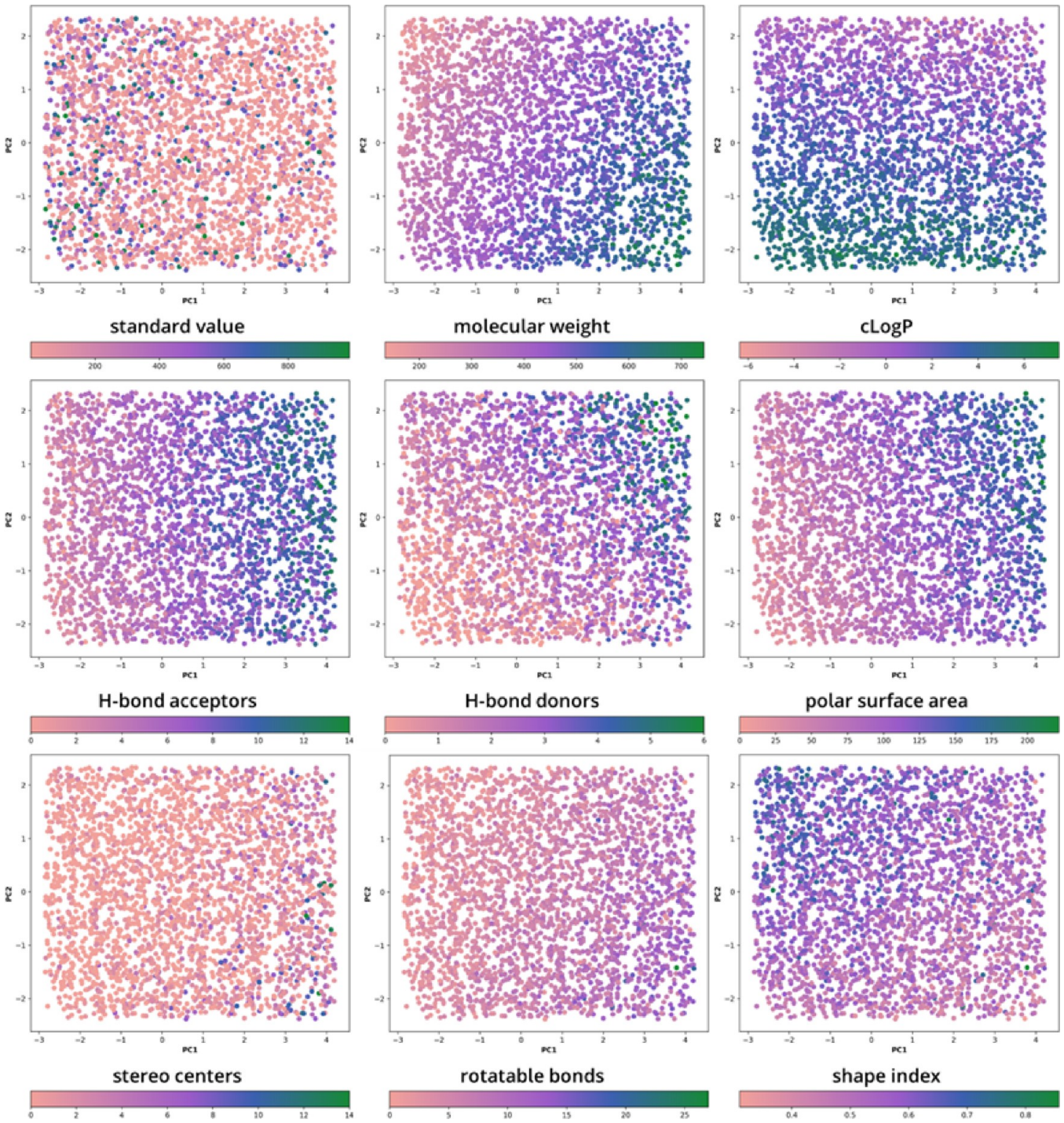


Figure 3. Overview of the distribution of compounds in Set *S* based on biological activity, physicochemical properties, and topological attributes. Lower numerical values are consistently represented in rosé, higher numerical values in green, and the middle of the range in blue. Less complex compounds can be found in the upper left region of the first quadrant, with increasing complexity distributed progressively moving to the other quadrants.

workflows and scenarios, three different data set sizes were generated: "large-sized" (on the order of hundreds of thousands, corresponding to Set *L*), "medium-sized" (on the order of tens of thousands, corresponding to Set *M*), and "small-sized" (on the order of thousands, corresponding to Set *S*).

To achieve a manageable five-digit size for the data set, suitable for utilization in a standard hardware setup, Set *L* was further scaled down in the next step. The compound with the lowest weight from each Bemis−Murcko scaffold cluster was extracted, and the rest were removed. The goal was to identify the most undecorated motif to increase the chances of finding related structures using different methods (e.g., maximum common substructure (MCS) search). After removal, 25,234 unique compounds remained, resulting in Set *M* ("M" denoting medium-sized).

Arguably, the size of Set *M* is already quite comfortably usable for many purposes. Nevertheless, in addition to this data set, which can involve intensive computations depending on its use

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

case, we decided to reduce the size further by another order of magnitude to the four-digit range. This smaller size can be employed in most daily applications without fully consuming computational resources. For this purpose, chemically diverse representatives were extracted to ensure coverage of the investigated chemical space. Another aspect that needed to be addressed was translating the complexity of the chemical space into an interpretable 2D landscape, aimed at enabling a straightforward visual understanding of the results. To achieve this, a principal component analysis (PCA) was performed using DataWarrior (version 6.0).[51] In this context, we prioritized PCA over uniform manifold approximation and projection (UMAP) because PCA focuses on the interpretability of the results and is also long-term reproducible. Additionally, it allows for the categorization of new compounds within the generated 2D landscape, enabling other groups to benefit from this groundwork. After the compound sources were assessed, however, a UMAP was also performed for the 2D visualization of the results (see below). The following eight parameters were selected to perform a 2D PCA: atomic polarization, clogP, number of H-bond acceptors, number of H-bond donors, polar surface area (PSA), shape index, stereocenters, and rotatable bonds. Further insights on the PCA, as well as the eigenvalues of the parameters, can be found in the Supporting Information (see Table S1). Subsequently, outliers in the sparsely populated regions were removed. The rationale behind this is that compounds in the outlier zone often exhibit extreme physicochemical properties such as high logP, MW, a large number of stereocenters, or violations of the rule of five, which would typically disqualify them from being considered as drug candidates. Additionally, they belong to molecular categories (e.g., fragment-like, peptides, nucleotides) that rely less frequently on the commonly used medicinal chemistry transformations employed in commercial libraries.[22,52−55] To eliminate the outliers, compounds with principal component (PC) values in the top or bottom 2.5% of the entire range were removed, leaving a total of 22,862 compounds. Visualization of the outlier trimming and compound examples is presented in Figure 2. In the next step, the ranges of PC1 and PC2 were each divided into 10% segments, creating a 10 × 10 matrix that encompassed all remaining compounds. From each of the hundred segments, up to 30 random representatives were selected (if an area contained fewer than 30 compounds, all were included) and merged into a new data set. This final data set, Set *S* ("S" denoting small-sized), contained 2,917 compounds, which cover the condensed chemical space of Set *M*.

For the performed PCA and the resulting Set *S* (and consequently for Set *M*), the following distribution with regard to physicochemical properties and topological chemistry trends was observed (see Figure 3): More potent biological activity was observed in quadrants 2 (Q2) and 4 (Q4), while triple-digit nanomolar values were more frequently found in quadrants 1 (Q1) and 3 (Q3), which are also reflected in the larger standard deviation for both (see Table S2 for all statistics). The range of MW was evenly distributed across all four quadrants, with the smallest compounds accumulating in Q1. Based on the clogP parameter, it can be inferred that the lipophilicity of the molecules increases from Q1 to Q3 and Q4, while more hydrophilic molecules are concentrated in Q2. With average values of 4.65 and 4.14 H-bond acceptors and 1.75 and 0.80 H-bond donors for Q1 and Q3, these quadrants exhibit the lowest number of heteroatoms besides carbon, indicating a higher prevalence of aliphatic chains and carbon ring systems. In contrast, Q2 and Q4 show significantly higher numbers of interaction atoms with mean H-bond acceptor counts of 8.46 and 7.88 and mean H-bond donor counts of 3.30 and 2.23, respectively. PSA increases from Q1 to Q2 and Q4 and decreases from Q1 to Q3. Accordingly, the compounds in Q3 have the smallest PSA, while those in Q2 have the largest. To assess the molecular complexity of the compounds, the numbers of stereocenters and rotatable bond parameters were examined. Q1 exhibits both the smallest standard deviation and the lowest mean of stereocenters among all quadrants with an average of 0.49. Q3 shows a slightly increased mean of 0.70 stereocenters, while Q2 and Q4 display significantly higher averages of 1.68 and 1.94, respectively, compared to Q1. The standard deviations increase from Q2 to Q4, indicating a broader distribution and greater heterogeneity of the results, which can be attributed to the presence of natural-like compounds. Particularly in Q2 and Q4, which have the highest standard deviations of 1.97 and 2.43, respectively, it can be inferred that these quadrants contain compounds with a potentially higher synthetic complexity. The number of rotatable bonds can provide insights into a molecule's rigidity, which in turn can influence solubility and cell permeability. With an average of 3.35, Q1 contains compounds with the fewest rotatable bonds. The average increases to 4.48 in Q3 and almost doubles in Q2 to 7.10 and in Q4 to 8.7. Finally, the shape index of the compounds was analyzed as part of the topological evaluation. The shape index describes the shortest distance between any two heavy atoms of a molecule, taking the total number of heavy atoms into account. For example, linear undecorated alkanes have shape indexes of 1.0, cyclopentane an index of 0.6, and adamantane an index of 0.5. Consequently, the shape index allows for insights into the topology of a molecule: Higher shape index values indicate linear, flexible compounds without significant branching (e.g., aliphatic chains, alkyne groups, and biphenyl groups), while lower values suggest spherical shapes and more rigid compounds with branching decorations. Given the narrower range of the shape index, from 0.857 to 0.311, with a mean of 0.539 and a standard deviation of 0.080 for the whole set, only blurred boundaries for the compound characteristics can be drawn between the quadrants. For Q1−Q4, the calculated means are 0.60, 0.55, 0.52, and 0.49, respectively. This suggests that compounds with minimal branching tend to accumulate in Q1, while highly branched molecules and those with numerous fused ring systems (e.g., glucocorticoids and saturated ring systems composed of carbon and oxygen atoms) are primarily found in Q4. Peptides and peptidomimetics were mostly found in Q2 and Q4.

Subsequently, it can generally be stated that, due to the similar ranges of standard deviations for MW, clogP, number of H-bond acceptors and donors, PSA, and shape index, Set *S* exhibits a uniform distribution of compounds with regard to physicochemical properties across the quadrants, which in turn can be associated with good coverage of chemical diversity.

These findings allow the following assignments of the PCA: Less complex compounds, with many representatives of the fragment-like class, are primarily found in Q1. Consequently, these characteristics of the compounds decrease as they progress toward other quadrants. Given their size, it is not surprising that the standard values for potency in Q1 are higher than those in the other quadrants, which contain significantly larger molecules with more functionalities. These additional features enable further interactions with the target structure, potentially increasing the potency of the compounds. A similar argument for increased standard values can also be applied in the case of

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

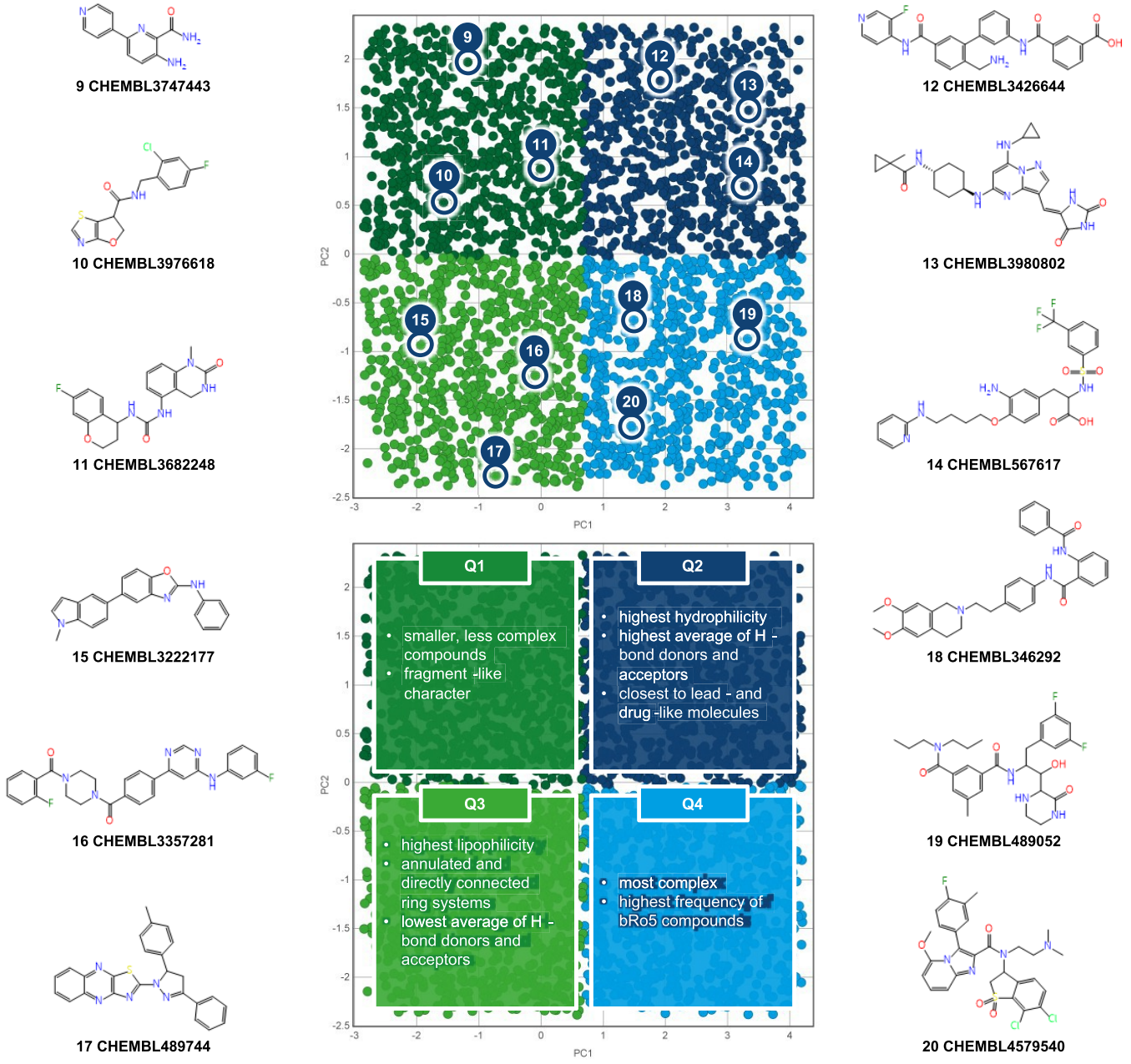


Figure 4. Overview of the properties of the compounds in Set *S* within the quadrants of the PCA. Quadrant Q1 is shaded in dark green, Q2 in dark blue, Q3 in light green, and Q4 in light blue. Three examples of molecules are shown for each quadrant. The boundaries between the quadrants are fluid: the characteristics of the molecules can be very similar, particularly in the transitions between quadrants.

Q3. Here, the reduced average number of H-bond acceptors and donors indicates the absence of potential functionalities. This assumption is further supported by the higher clogP and lower PSA compared with Q1, suggesting a greater prevalence of carbon-based systems. In these cases, lipophilic and π-interactions would make a significant contribution to the compound's affinity. In contrast, entries Q2 and Q4 exhibit significantly more H-bond donors and acceptors, which due to their increased interaction potential may contribute to the lower standard values. This is further supported by the fact that certain functional groups are more frequently found in these quadrants: 302 out of 338 (89%) compounds with a sulfonamide group, 199 out of 281 (71%) with a carboxylic acid, and 87 out of 121 (72%) with a carbamate functionality are located in Q2 or Q4 (see Figure S1). Due to the filtering step in which only the smallest scaffold was selected, it becomes apparent that the compounds in Set *S* predominantly consist of undecorated systems, offering opportunities for further functional extensions. Compounds in Q2 are characterized by the highest hydrophilicity (by consequence of having the lowest lipophilicity) and the highest average number of H-bond donors and acceptors, making the compounds in this quadrant the most similar to typical lead- and drug-like molecules. Q3 contains the most lipophilic compounds and the lowest averages for H-bond donors and acceptors, which can be attributed to a large number of compounds with fused systems of two or more rings as well as directly connected rings or rings linked through carbon-based linkers. Q4 covers mainly the bRo5 compounds. Of the 796 compounds in Set *S* with MW > 500, 552 (69%) are found in Q4. This corresponds to 76% of the total 727 compounds in Q4,

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

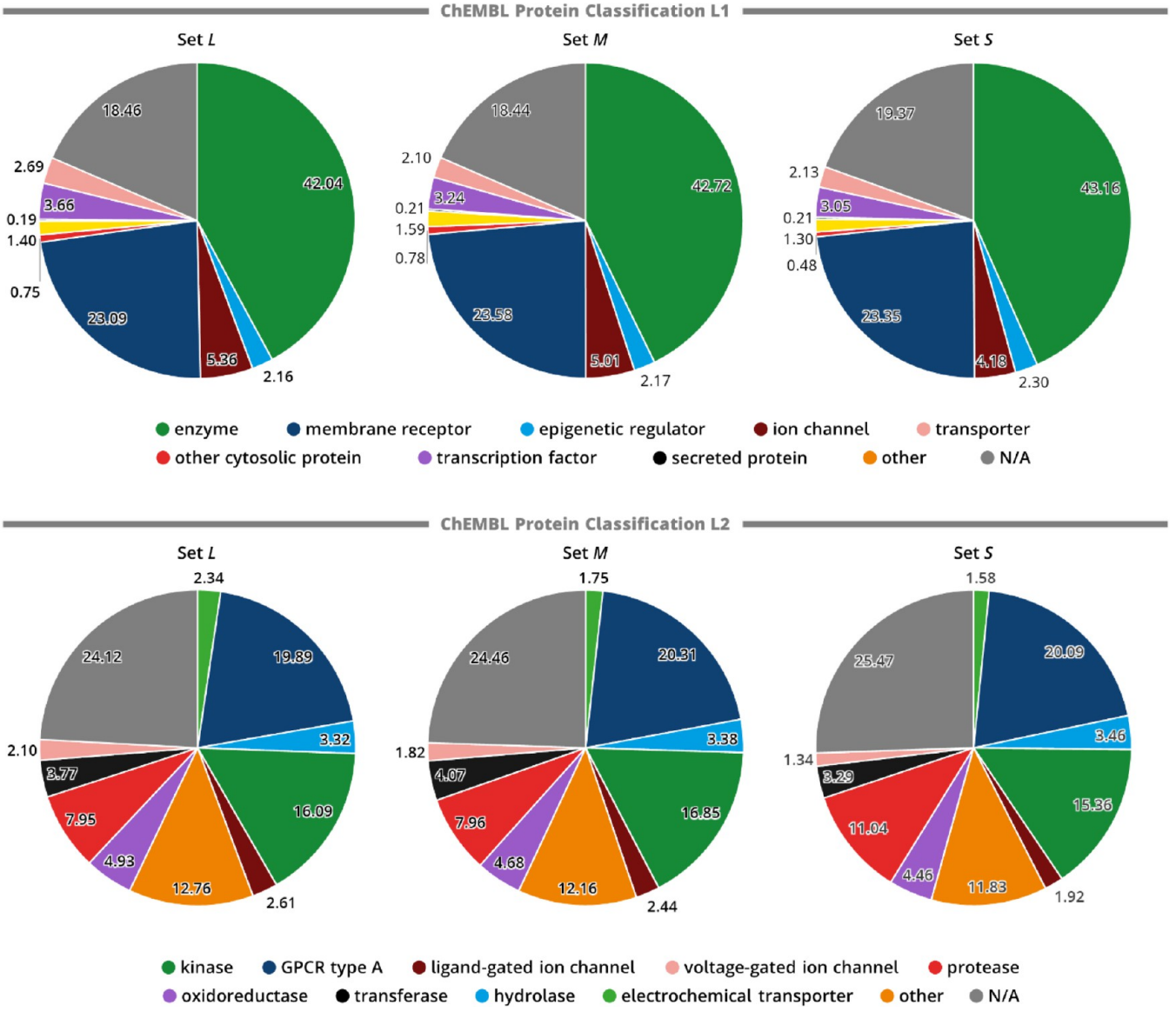


Figure 5. Target coverage is based on the ChEMBL protein classifications for the generated sets. The numbers represent the percentage share.

which is also reflected in the average MW of 547.94 for this cluster. The summarized observations are presented in Figure 4.

In the future, these findings can also be used for the approximate classification of new compounds based on their calculated parameters and corresponding eigenvalues for the calculation of PCs within the chemical landscape context. Additionally, Set *S* can be further filtered according to the requirements (e.g., for drug-like compounds/compliant with the rule of five) while maintaining an appropriate size. For example, the number of rule-of-five-compliant compounds is 1,894, and the number of rule-of-three-compliant compounds is 111, indicating that Set *S* focuses on more developed compounds.

The target coverage of the generated sets was further examined. For this, ChEMBL-specific protein classification was used. It should be noted in this context that the target associated with the potency of the compounds was used, meaning that other off-targets or activities against different macrostructure classes were not included in the assessment. Based on this, it was observed that the downsizing from one set size to the next one had no significant impact on the target class coverage (see Figure 5).

The distribution of all three analyzed data sets reflects the profile of previously reported molecular drug targets and current trends.[56−58] The most prominent targets—such as kinases, GPCRs, and ion channels—also represent the largest share of targets in the generated sets. Deviations from the proportions seen in FDA-approved drugs (e.g., two-thirds of all approved drugs target GPCRs[58]) can be attributed to the fact that the ChEMBL source set also includes a large number of data points for targets that have historically not been confirmed as therapeutically relevant. Successful targets, on the other hand, are pursued more intensively, leading to several approvals over time, which are reflected in their higher percentages of the total share. The absolute target counts can be found in Table S3 of the Supporting Information.

Assessment of Combinatorial Chemical Spaces. The generated Set *S* was subsequently used to analyze commercial compound libraries and Chemical Spaces. Here, we differentiate between enumerated molecule libraries, where each compound entry is explicitly listed, and combinatorial Chemical Spaces. As previously mentioned, Chemical Spaces contain information about the building blocks and chemical reaction rules for

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

Table 1. Overview of Commercial Combinatorial Chemical Spaces and Enumerated Compound Libraries Investigated in This Study[a]

				results with a similarity of 1.0			
					SpaceLight		
	vendor	compound set	size	FTrees	ECFP4	fCSFP4	SpaceMACS
Chemical Spaces	Ambinter	AMBrosia	1.1×10^{11}	51 (23)	20 (20)	24 (23)	23
	OTAVA	CHEMriya	1.2×10^{10}	38 (26)	27 (26)	28 (27)	27
	eMolecules	eXplore	5.0×10^{12}	255 (173)	155 (155)	180 (175)	174
	Chemspace	Freedom Space	5.1×10^{9}	142 (83)	84 (81)	87 (83)	83
	WuXi	GalaXi	1.2×10^{10}	47 (25)	27 (27)	28 (27)	27
	Enamine	REAL Space	7.0×10^{10}	288 (190)	184 (183)	203 (193)	191
libraries	ChemDiv	Representative Diversity Libraries	1.5×10^{5}	4 (2)	2 (2)	3 (2)	2
	Life Chemicals	HTS Compound Collection	5.7×10^{5}	21 (13)	15 (15)	15 (15)	15
	Mcule	Mcule Full	5.9×10^{6}	286 (228)	245 (242)	247 (242)	242
	Molport	Drug-Like Compounds Library	1.8×10^{6}	65 (35)	47 (47)	51 (47)	47

[a]The table includes the number of rank 1 results with a similarity of 1.0 that were retrieved for Set *S* from the corresponding compound sets. The number in parentheses represents the amount of identical molecules found in both Set *S* and the compound set. For the SpaceMACS algorithm, this number corresponds to the listed Sim = 1 value.

combining them. They, therefore, do not list every possible combination of building blocks but instead enable the on-the-fly generation of molecules according to the employed screening method and algorithm.

For the assessment of their content, the following commercial Chemical Space were investigated: AMBrosia by Ambinter (1.1×10^{11} compounds), CHEMriya by OTAVA (1.2×10^{10} compounds), eXplore by eMolecules (5.0×10^{12} compounds), Freedom Space by Chemspace (5.1×10^{9} compounds), GalaXi by WuXi (1.2×10^{10} compounds), and REAL Space by Enamine (7.0×10^{10} compounds). An overview of the investigated compound collections is presented in Table 1.

For this study, we employed three different similarity search methods developed to retrieve relevant chemistry from Chemical Spaces: FTrees, SpaceLight, and SpaceMACS. FTrees screens for similar compounds based on fuzzy pharmacophore matching.[12] SpaceLight is a molecular fingerprint-based algorithm that can mine the closest analogs to a query compound using either extended-connectivity fingerprints (ECFPs) or connected subgraph fingerprints (CSFPs).[13] SpaceMACS is a substructure-driven algorithm focused on the longest connected substructure chain of heavy atoms and, therefore, the maximum common substructure (MCS).[14,59] All three methods aim to retrieve compounds fitting a drug discovery challenge, and while they were developed for combinatorial Chemical Spaces, they can also be applied to enumerated sets, making them suitable in this study for comparing Chemical Spaces and enumerated commercial compound libraries.

For all three search algorithms, the default settings were used: For each query compound of Set *S* (2,917 molecules), 100 results were requested from the corresponding Chemical Space. For SpaceLight, we chose the most commonly used molecular fingerprint variant, ECFP4,[60] as well as fCSFP4,[13] a fingerprint developed for screening combinatorial Chemical Spaces, to compare both methods in parallel. In the case of SpaceMACS, the MCS search was used. The corresponding software versions of FTrees, SpaceLight, and SpaceMACS were 6.13, 1.5, and 1.3, respectively. The results of all searches are summarized in Figure 6.

We also investigated whether the results were tied to the randomly selected 30 molecules per cluster. To this end, an additional counterset, Set *S*′, was generated. For this set, up to 30 different molecules from each cluster were selected, if available. If the number of alternative molecules in a cluster was exhausted, the remaining slots up to 30 were filled with molecules already present in Set *S*. As a result, the new selection again yielded 2,917 molecules drawn from 100 clusters, forming Set *S*′. This set was likewise used as a query in the four search runs. The subsequent analysis revealed no differences in the means compared to Set *S*, and the trends in ranking based on the standard deviation also remained unchanged (see Table S4 in the Supporting Information).

An analysis of the FTrees results is summarized in Figure 7. It should be noted that the FTrees algorithm is agnostic to the topology of the captured pharmacophore fragments during the calculation of the features: the decoration pattern (ortho-, meta-, para-), stereochemistry, and position of heavy atoms within a ring system are not considered as long as the pharmacophore profile does not change. This can result in the calculated total similarity of the molecule being 1.0 even though the molecule is a constitutional isomer or diastereomer of the query compound, which in turn requires an additional assessment of the results in the subsequent step.

Furthermore, we were also interested in how many similar compounds a Chemical Space can offer for a query from Set *S*.

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

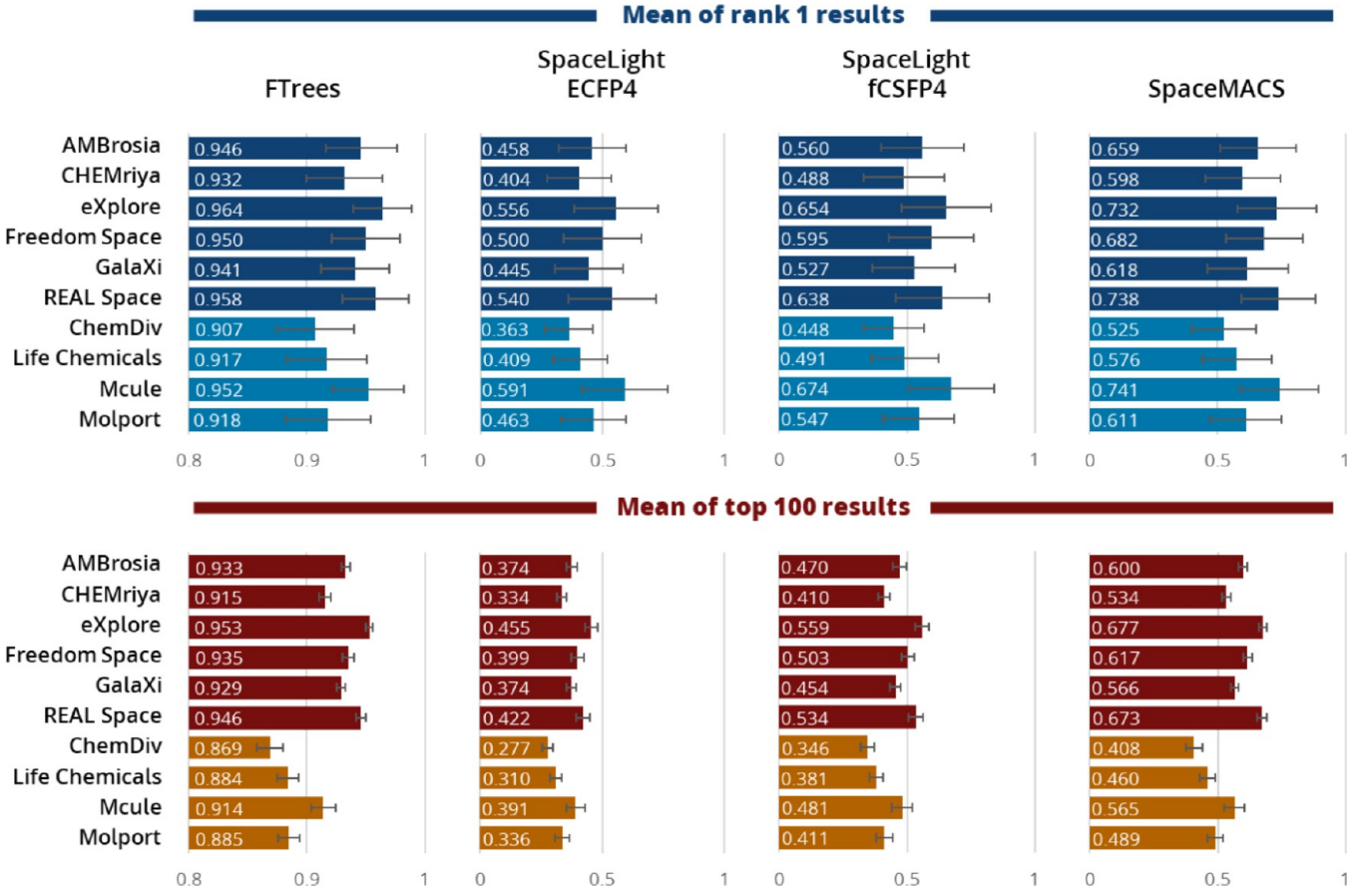


Figure 6. Summary of the results from the searches in the compound collections using Set *S* as queries. Displayed at the top are the mean values of the respective scores for all result molecules with rank 1. Displayed at the bottom are the mean values of the top 100 results: the average score of all 100 retrieved results for each query was calculated, and from the total of 2,917 averages, the overall mean per source was derived.

For this, we analyzed the mean and SD of all 100 results for each query and their distribution. The mean of all results for a query compound in this context indicates the average value of the similarity scores relative to the query compound. Consequently, a high value (maximum 1.0) signifies that the top 100 ranking results are very similar to those of the query compound. Lower values suggest that fewer closely related substances are present among the top 100 ranking results. The SD in this context reflects the spread of the results. A low SD indicates that, based on the FTrees similarity score, many results with similar scores are present (e.g., numerous constitutional isomers with the same structural motifs in different arrangements). In contrast, a higher SD suggests there are gaps within the results in the similarity score (e.g., many heterogeneous molecular motifs, where the total similarity score as a whole matches the query compound best compared to its structural analogs). Thus, the mean answers, on the one hand, which queries have many similar compounds and, on the other hand, what the average value of a search query with Set *S* within a Chemical Space is. The SD value can provide insight into how the results for a query compound are distributed within the chemical landscape.

The highest mean scores for rank 1 were calculated for eXplore, REAL Space, and Mcule, delivering the most similar compounds for Set *S* according to the FTrees score. The lowest SD was observed for eXplore, indicating that this source delivered the narrowest range of similarity values for the results of each query. In turn, this means that the retrieved compounds from eXplore had the highest degree of relatedness to the query compound among all investigated sources and therefore offered the highest quantity of similar structures within each query's results.

Figure 7 also features different coloring ranges of the mean FTrees similarity scores for the retrieved results. To provide a better understanding of how the similarity is distributed for the ChEMBL results, three different ranges were colored: 0.95 to 1.0 for very similar compounds, 0.9 to 1.0 for related compounds, and 0.85 to 1.0 to spot more distantly related structures. This depiction is aimed at illustrating the distribution and categorization of compounds for which a broader range of similar compounds is available. The visual assessment further highlights the higher prevalence of very similar results for Q1 and Q3, while the degree of relatedness decreases toward Q2 and Q4.

When the distribution of rank 1 results with an FTrees similarity of 1.0 across the quadrants was examined, higher content percentages were observed for all Chemical Spaces in Q1. Accordingly, the coverage of the relatively simpler query compounds from Set *S* was proportionally the best for all Chemical Spaces. Given the premise of applying robust chemical reactions[61] for higher synthesis success rates during the generation of Chemical Spaces, it seems logical that more complex compounds, primarily found in Q2−Q4, cannot always be synthesized with just one- to two-step reactions. Therefore, Q4, which features an increased proportion of bRo5 compounds, consistently yielded the fewest results, with an FTrees similarity score of 1.0. The distribution is summarized in

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

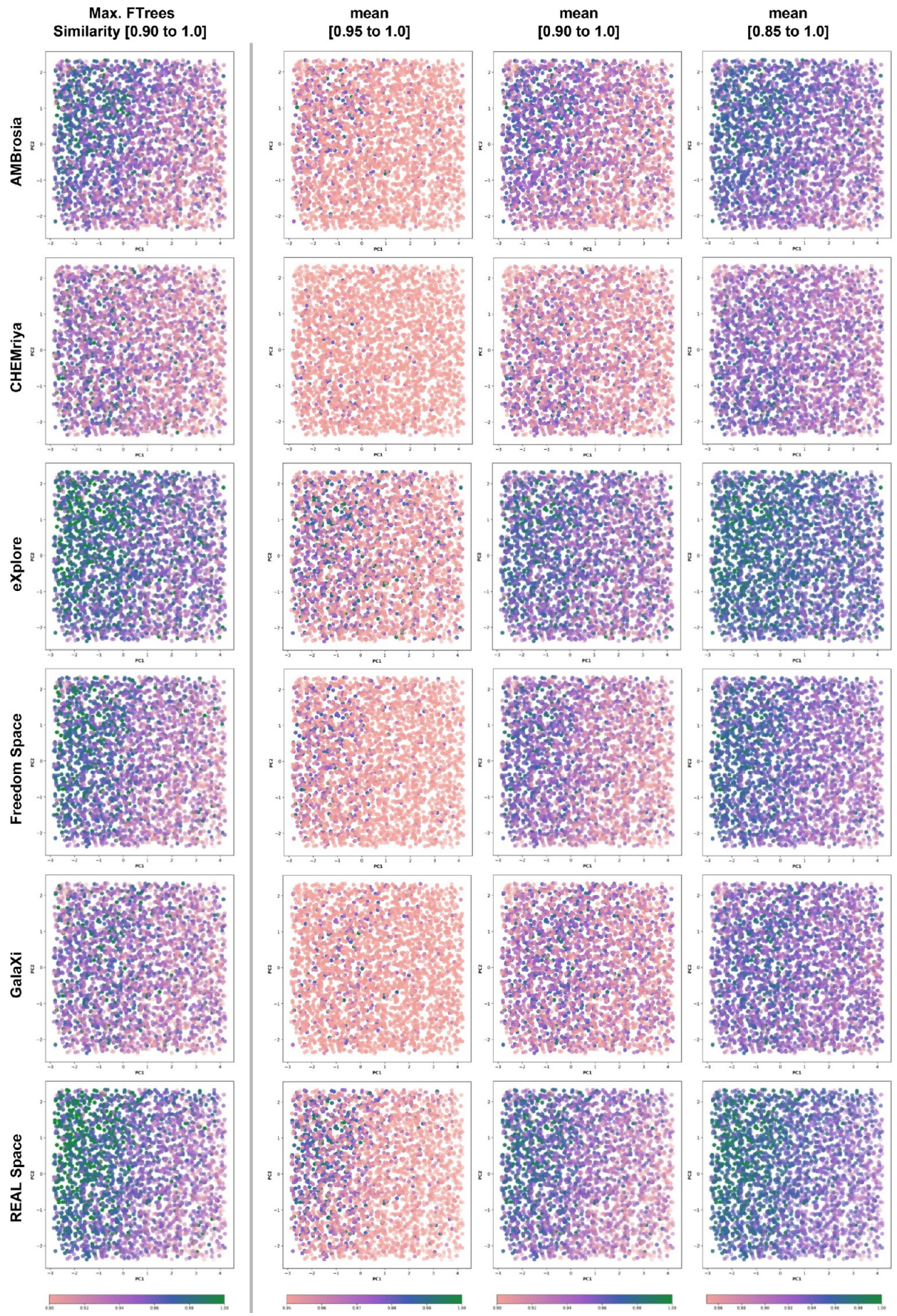


Figure 7. Overview of the FTrees assessment of the commercial Chemical Spaces. Point coordinates correspond to those of the associated Set *S* query compound. The FTrees similarity score of the highest-ranking compound in the corresponding Chemical Spaces is shown on the left. For color coding,

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

Figure 7. continued

0.90 was set as the lower cutoff (pink) and 1.0 as the upper cutoff (dark green). Values between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided. To better assess the distribution of the mean values, the lower cutoff (colored pink) was shifted by 0.05 units each time.

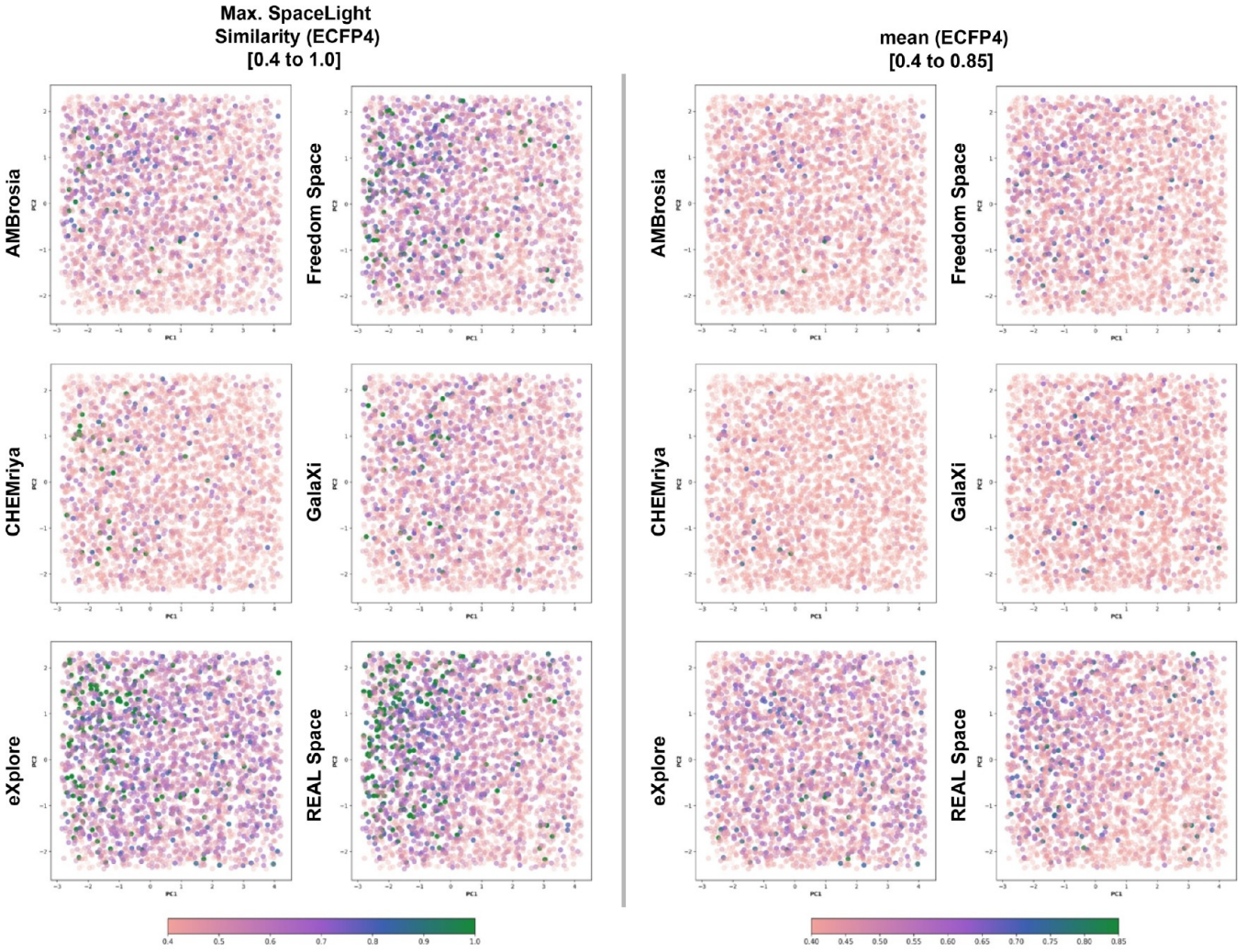


Figure 8. Overview of the ECFP4 SpaceLight assessment of commercial Chemical Spaces. Point coordinates correspond to those of the associated Set*S* query compound. The molecular fingerprint similarity of the highest-ranking compound in the corresponding Chemical Spaces is shown on the left. For color coding, 0.40 was set as the lower cutoff (pink), and 1.0 was set as the upper cutoff (dark green). Values in between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided. Here, 0.40 was selected as the lower cutoff (pink) and 0.85 as the upper cutoff (dark green).

Table S5 of the Supporting Information. An important aspect required to contextualize the results of FTrees and the other two algorithms is the design of the Chemical Spaces. A fundamental decision for the combinatorial architecture could involve explicitly applying filters on the building blocks for the molecular size, rotatable bonds, a maximum number of H-bond donors/acceptors, or other parameters, to guide the properties of the results to comply with drug-like filters, thereby excluding larger molecules from the Chemical Spaces.[20] This may lead to better coverage of Q2 and Q4 in the case of eXplore, which contains more compounds with MW > 500.[22]

The SpaceLight ECFP4 search results are presented in Figure 8. Compared to FTrees, it is noticeable that the trend in the quality of the results continues in such a way that eXplore, Freedom Space, and REAL Space still provide results with higher similarity scores, which predominantly remain in Q1 and Q3. This is also evident in the means of the results, which were calculated analogously to FTrees (the mean of the average values of all results for a Set *S* query).

Interestingly, there were changes in the ranking for the SD of the ECFP4 fingerprint similarity. The lowest SD was displayed by GalaXi and CHEMriya, while the highest SD was observed for REAL Space. A closer inspection of results with high SD from eXplore, Freedom Space, and REAL Space shows that this observation may result from the presence of some very good results with high fingerprint similarity, which are accompanied by results with significantly decreasing fingerprint similarity in the lower ranks. Based on the sensitivity of the fingerprint method, even small structural variations or decorations can lead to a significant drop in the score, which contributes to the

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

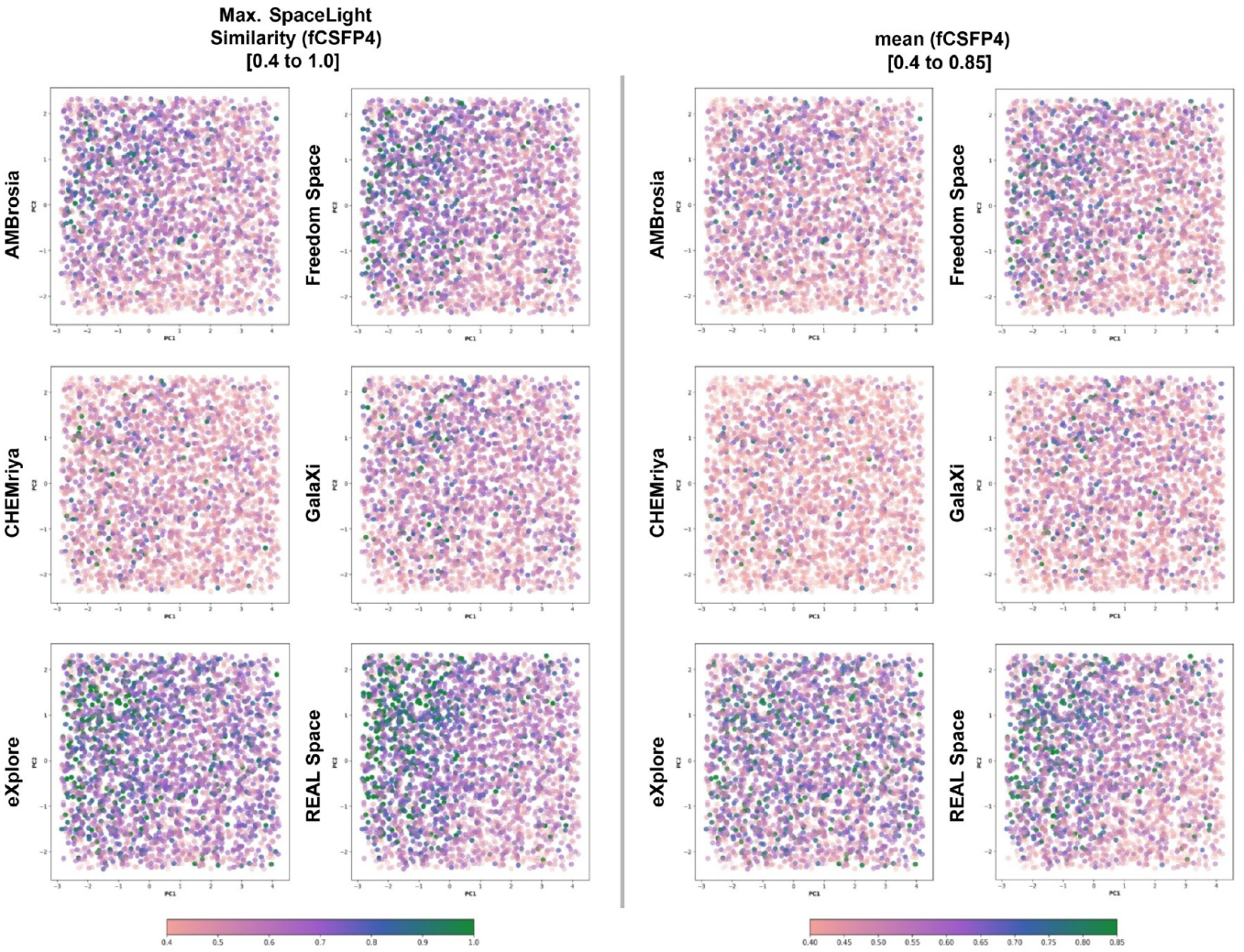


Figure 9. Overview of the fCSFP4 SpaceLight assessment of commercial Chemical Spaces. Point coordinates correspond to those of the associated Set*S* query compound. The molecular fingerprint similarity of the highest-ranking compound in the corresponding Chemical Spaces is shown on the left. For color coding, 0.40 was set as the lower cutoff (pink) and 1.0 as the upper cutoff (dark green). Values in between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided; 0.40 was selected as the lower cutoff (pink) and 0.85 as the upper cutoff (dark green). No significant changes compared to ECFP4 were observed in the distribution across the quadrants (see Table S5 of the Supporting Information). The molecular landscapes of result categorization by similar compounds (range 0.85 to 1.0), related compounds (range 0.65 to 0.85), and more distantly related structures (range 0.45 to 0.65) can be found in Figure S3 of the Supporting Information. Ranges were kept consistent with those of the above-mentioned ECFP4 definitions.

increased standard deviations.[62−66] In the case of the other Chemical Spaces, a lower average molecular similarity leads to more compounds with a correspondingly lower score, which is reflected in a smaller standard deviation. Translated to this scenario, it means that the Chemical Spaces with a low SD (GalaXi, CHEMriya, and AMBrosia) provide results with a lower similarity score in general, but all of which are equally similar to the query. Chemical Spaces with a higher SD (eXplore, Freedom Space, and REAL Space) provide more results with higher similarity, but the score declines more along the top 100 ranks. Further assessment of the data is provided in the fCSFP4 SD result section below.

Furthermore, trends similar to those of FTrees were observed regarding the quadrant distributions of rank 1 ECFP4 SpaceLight results with a fingerprint similarity of 1.0 (see Table S5 of the Supporting Information).

Similarly to the evaluation of the FTrees results, three categories for the means of all results for a query compound were introduced to visualize the distribution across the quadrants. For very similar compounds, a range of 0.85 to 1.0 was defined; for related compounds, a range of 0.65 to 0.85; and for more distantly related structures, a range of 0.45 to 0.65. The molecular landscapes can be found in Figure S2 of the Supporting Information. We acknowledge that the ECFP4 boundaries may vary depending on the project and compound classes. In this case, they serve merely as a rough categorization of the results for an easier visual assessment.

The overall distribution of fCSFP4 scores per query compound is depicted in Figure 9. Looking at the means for the results per Set *S* query, eXplore, REAL Space, and Freedom Space once again show the best results followed by AMBrosia, GalaXi, and CHEMriya. Analogous to the ECFP4 molecular fingerprint results, the lowest SD values were observed for GalaXi, CHEMriya, and AMBrosia. Higher SD values were recorded for Freedom Space, eXplore, and REAL Space.

As a third method to mine for relevant chemistry from commercial Chemical Spaces, SpaceMACS was applied. The overview of the chemical space coverage is presented in Figure

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

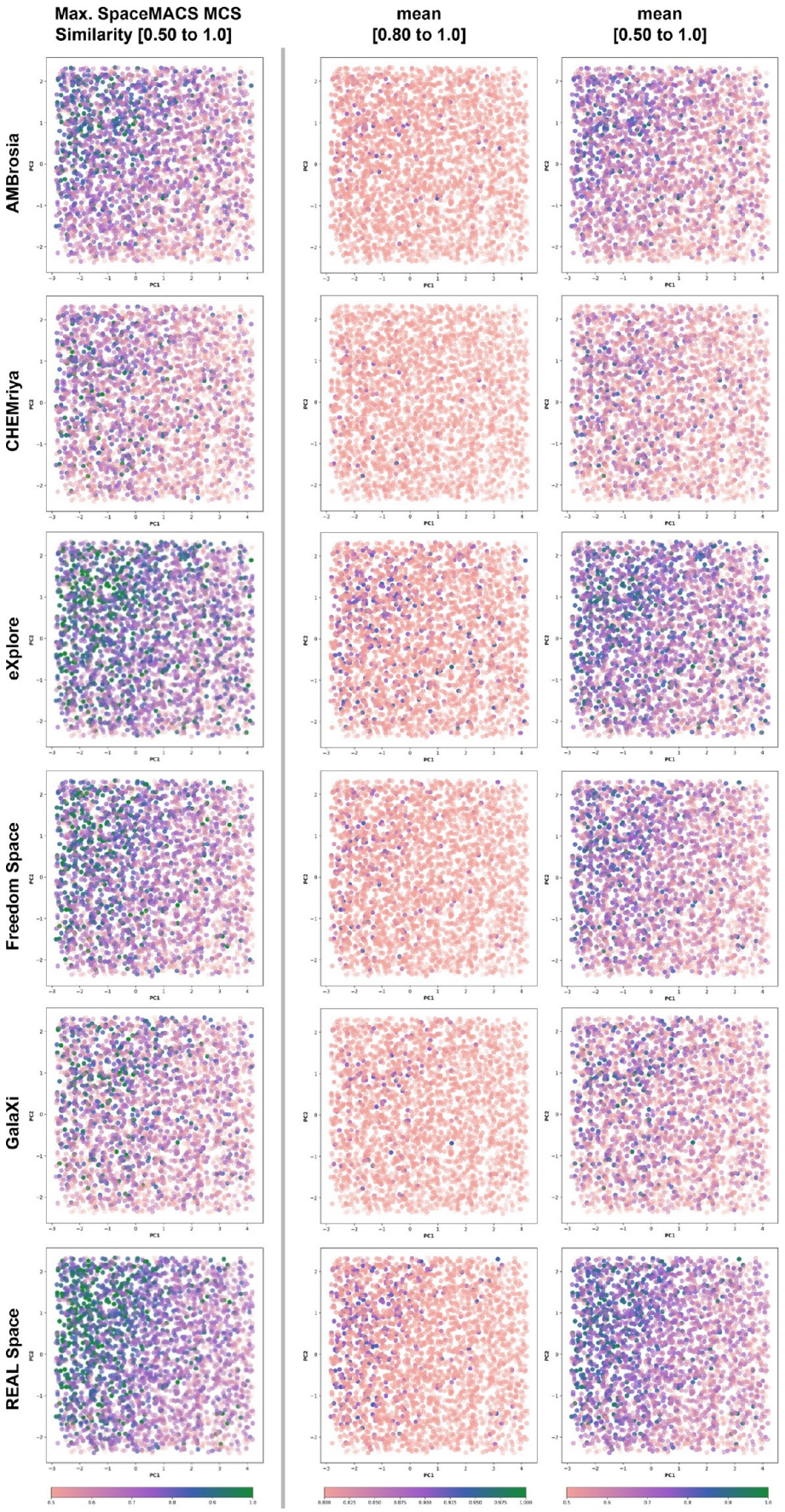


Figure 10. Overview of the SpaceMACS assessment of commercial Chemical Spaces. Point coordinates correspond to those of the associated Set *S* query compound. The SpaceMACS MCS score of the highest-ranking compound in the corresponding Chemical Spaces is shown on the left. For color

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

Figure 10. continued

coding, 0.50 was set as the lower cutoff (pink) and 1.0 as the upper cutoff (dark green). Values in between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided. To better assess the distribution of the mean values per query, two ranges were provided: 0.80 to 1.0 for similar results and 0.50 to 1.0 for more distantly related structures. In both cases, the lower cutoff is colored pink.

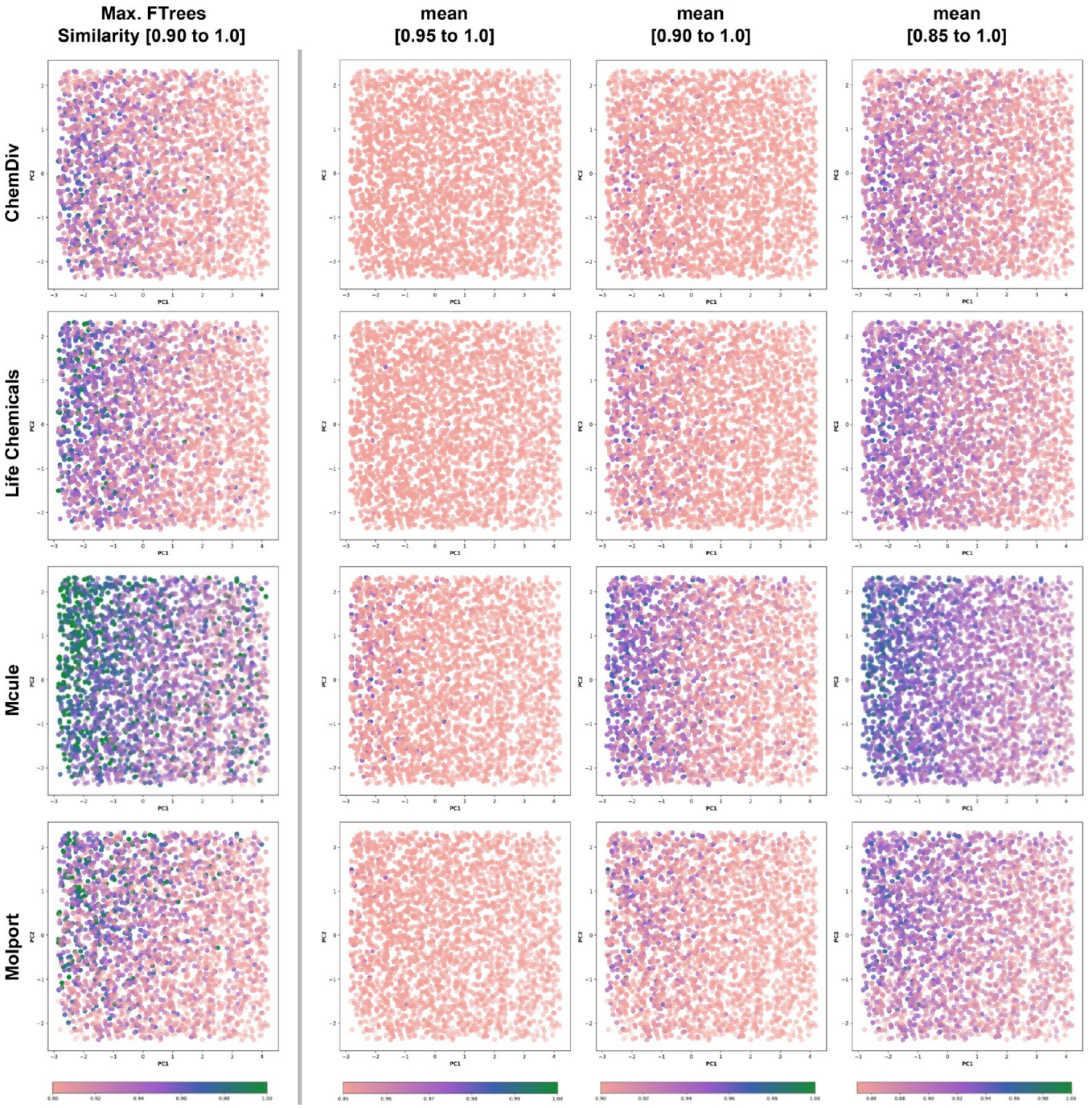


Figure 11. Overview of the FTrees assessment of commercial compound libraries. Point coordinates correspond to those of the associated Set *S* query compound. The FTrees similarity score of the highest-ranking compound in the corresponding vendor is shown on the left. For color coding, 0.90 was set as the lower cutoff (pink) and 1.0 as the upper cutoff (dark green). Values in between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided.

10. For the distribution among the quadrants, only minor fluctuations in the SpaceMACS results compared to the FTrees and SpaceLight results were observed.

Summarizing the observations made for all three search methods, the following statements can be made: All three search methods successfully mined relevant chemistry from the Chemical Spaces. The fundamental trends were consistent

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

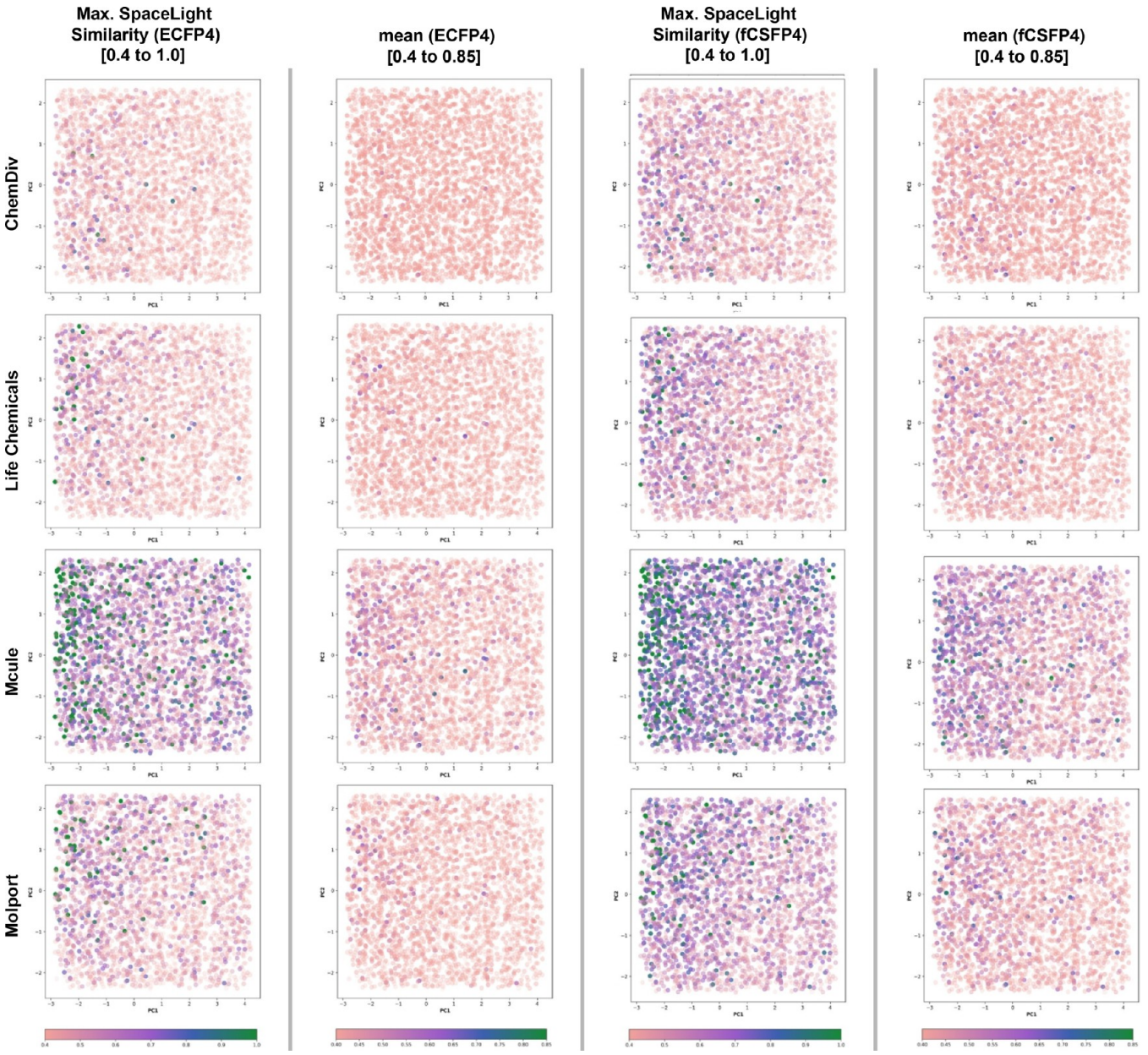


Figure 12. Overview of ECFP4 SpaceLight assessment of the commercial compound libraries. Point coordinates correspond to those of the associated Set *S* query compound. The molecular fingerprint similarity of the highest-ranking compound in the corresponding library is shown on the left. For color coding, 0.40 was set as the lower cutoff (pink) and 1.0 as the upper cutoff (dark green). Values in between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided; 0.40 was selected as the lower cutoff (pink) and 0.85 as the upper cutoff (dark green).

across all three methods, with Q1 providing the best coverage and Q3 the second-best coverage of the result molecules with a similarity of 1.0. Consequently, relatively simpler and lipophilic molecules were predictably the most frequently found in the Chemical Spaces. Due to the stereochemically agnostic scoring of FTrees, the percentage was further increased. Significantly lower coverage of the chemical landscape was observed for Q2 and particularly Q4, which may be associated with the increased complexity of the compounds and the corresponding multistep synthesis. This could also be linked to missing reactions in the definition of the Chemical Spaces or the absence of appropriate building blocks.

Commercial Libraries. To put the results into a broader context, conventional enumerated commercial compound libraries were also analyzed for their chemical diversity in a manner analogous to that of the aforementioned methods. The undeniable advantage of these enumerated catalogs is that their data format (typically SMILES or SDF) is natively supported by most computational tools, making them accessible for a broader client base. The combinatorial nature of Chemical Spaces, on the other hand, demands the usage of dedicated algorithms that were developed to efficiently operate in their architecture.

For the comparative analysis, four libraries from different compound providers were selected: "Representative Diversity Library" by ChemDiv (1.5×10^5 compounds), "HTS Compound Collection" by Life Chemicals (5.7×10^5), "Drug-Like Compounds Library" by Molport (1.8×10^6), and "Mcule Full" by Mcule (5.9×10^6). We decided to investigate

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

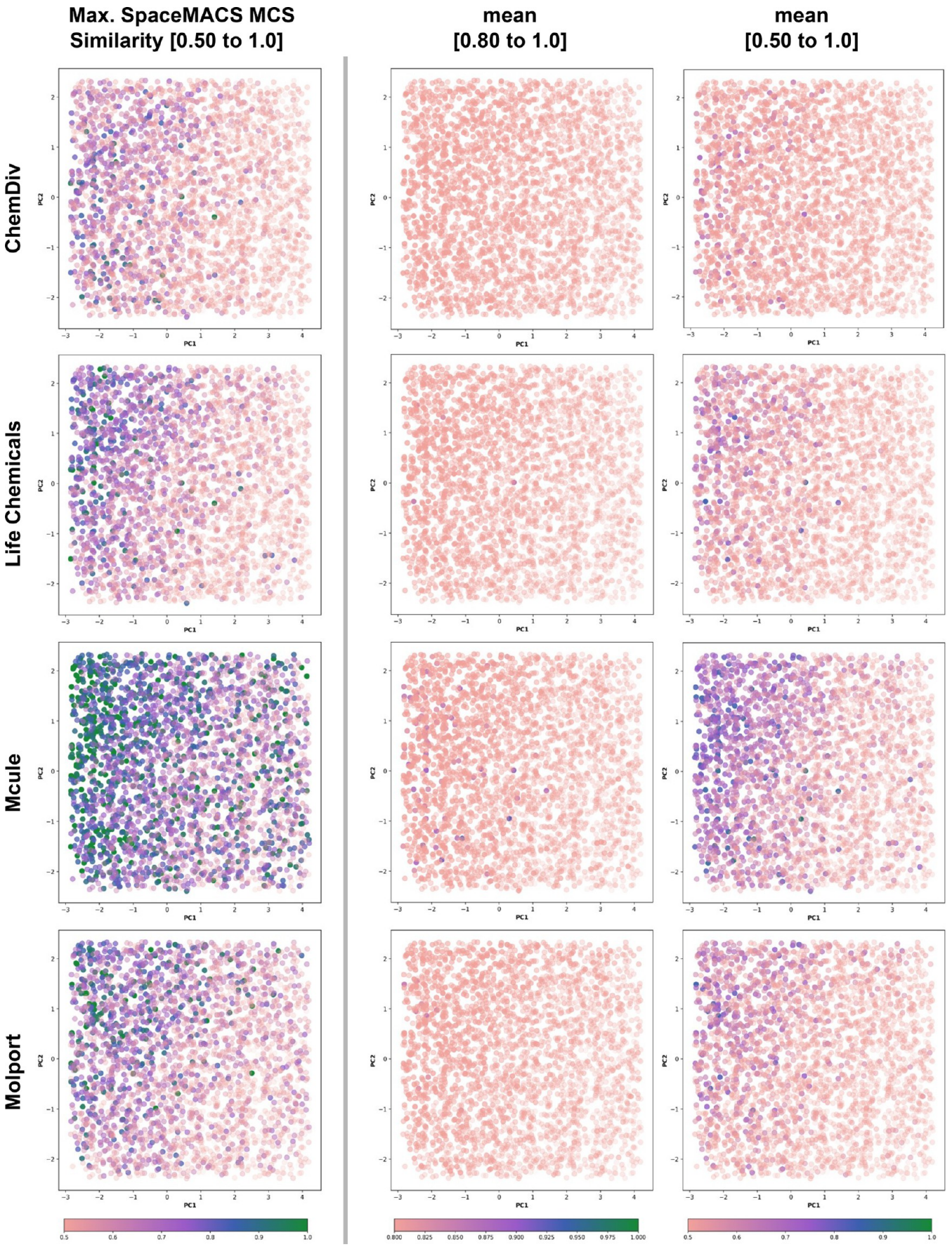


Figure 13. Overview of the SpaceMACS assessment of commercial compound libraries. Point coordinates correspond to those of the associated Set *S* query compound. The SpaceMACS MCS score of the highest-ranking compound in the corresponding library is shown on the left. For color coding,

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

Figure 13. continued

0.50 was set as the lower cutoff (pink) and 1.0 as the upper cutoff (dark green). Values in between are color-coded using a gradient. On the right, an overview of the means for the results of the corresponding query is provided. To better assess the distribution of the mean values per query, two ranges were provided: 0.80 to 1.0 for similar results and 0.50 to 1.0 for more distantly related structures. In both cases, the lower cutoff is colored pink.

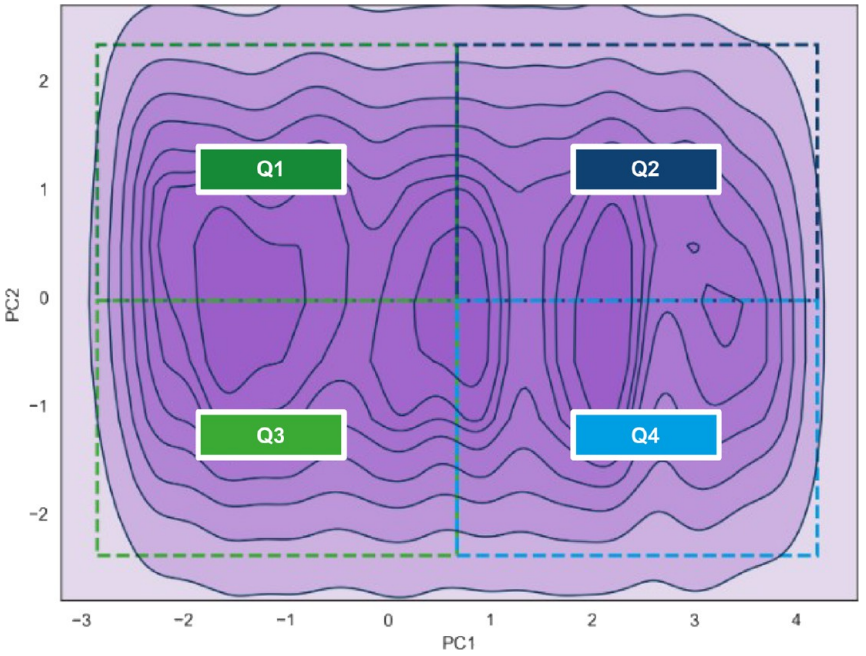


Figure 14. Density plot for the PCA of Set *S*. Ten density levels are displayed with the more intense hues in the highly dense areas. The previously discussed quadrants are represented as dashed areas: Q1 in dark green (upper left corner), Q2 in dark blue (upper right corner), Q3 in light green (bottom left corner), and Q4 in light blue (bottom right corner).

enumerated sets from vendors that do not possess a Chemical Space to evaluate the results in a highly distinguishable manner. In doing so, we ensured that libraries of different sizes were selected to identify any potential effect of the number of molecules on the diversity of the compounds. All libraries were accessed in December 2024 and analyzed with the aforementioned three algorithms (FTrees, SpaceLight, and SpaceMACS). Using Set *S* as the query, 100 results were retrieved for each entry.

The FTrees insights are summarized in Figure 11. The highest similarity mean score, based on the 100 results for each processed query molecule, was calculated for Mcule followed by those for Molport, Life Chemicals, and finally ChemDiv. The lowest SD for the top 100 averages was observed for Molport followed by Life Chemicals, Mcule, and ChemDiv.

Similar to the investigated Chemical Spaces, the commercial libraries also show that a majority of closely rated hit compounds are located in Q1.

The SpaceLight search with ECFP4 delivered the same ranking in terms of compounds with a similarity of 1.0: Mcule on top followed by Molport, Life Chemicals, and ChemDiv last. The SpaceLight results are summarized in Figure 12. Applying fCSFP4 in SpaceLight led to similar results. Furthermore, the majority of the top results with a similarity of 1.0 for both ECFP4 and fCSFP4 were again located in Q1 for all of the investigated sources.

The SpaceMACS screening runs are summarized in Figure 13.

To compare the results of combinatorial Chemical Spaces and enumerated library searches, several design aspects need to be addressed.

We take into account that the results discussed here do not fully represent the complete coverage capacity of the chemical space in terms of the availability of individual compounds by each vendor, as only individual libraries were selected. A set tailored for chemical diversity can deliberately not include all in-house compounds, with the aim of maintaining a manageable size that can be efficiently screened using common programs. Furthermore, it should be emphasized that the size difference between the libraries and the Chemical Spaces influences the availability of closely related molecules and consequently also the associated values such as mean and SD. Given a difference of 7 orders of magnitude between the smallest library (ChemDiv with 1.5×10^5 compounds) and the largest investigated Chemical Space (eXplore with 5.0×10^{12} compounds), it is undeniably likely that the latter will contain significantly more related substances to a query compound than a library that is 10 million times smaller. Another point is the fact that libraries can be deliberately enriched with relevant molecules (those reported as bioactive) to increase the relevance of the set for research purposes. The observations below regarding SD differences suggest, at least, that for particularly closely related rank 1 results, there is a much more pronounced decline in similarity scores for the enumerated libraries compared to the Chemical Spaces.

In contrast, Chemical Spaces can be indirectly enriched with bioactive substances only by incorporating appropriate building blocks and reactions. The coverage of the bioactive landscape is merely a consequence of the combinatorial explosion of possible compounds. However, in the case of SpaceMACS, it should be noted that if a building block itself already matches a query, then

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

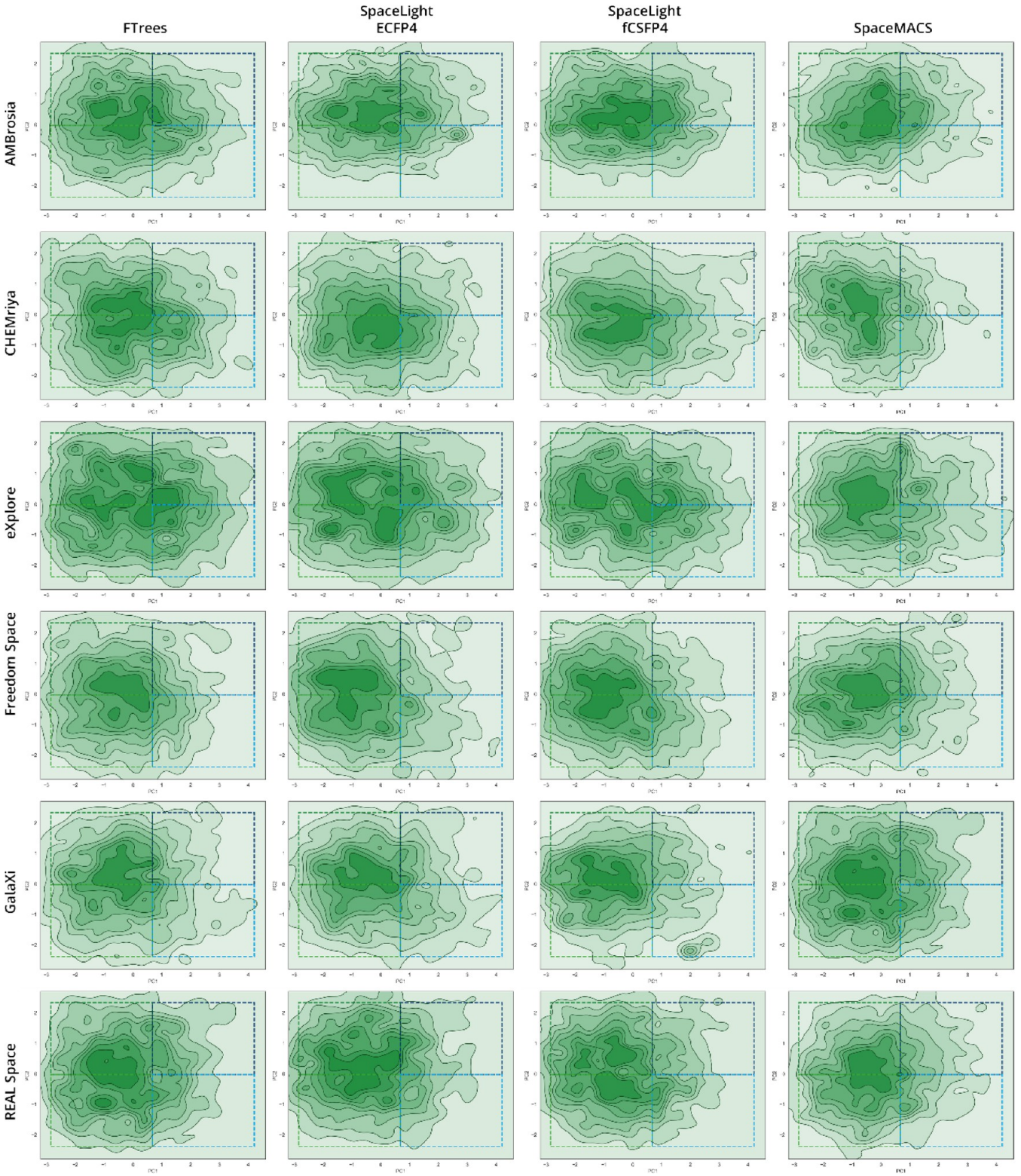


Figure 15. Overview of the chemical landscape distribution for rank 1 results of investigated commercial Chemical Spaces and the applied search algorithms as density plots. For each distribution, a 10-level representation was selected. Corresponding quadrants of Set *S* are depicted in overlay: Q1 in dark green (upper left corner), Q2 in dark blue (upper right corner), Q3 in light green (bottom left corner), and Q4 in light blue (bottom right corner).

it will appear as a result. This means that larger molecules similar to a product can theoretically be introduced into Chemical Spaces as needed. In practice, however, this is only applied to a limited extent, as size filters exclude such molecules to keep the size of the products within a relevant range (e.g., drug-like).

In the context of this study, collection-focused decisions can lead to underrepresentation of certain compound classes. For

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

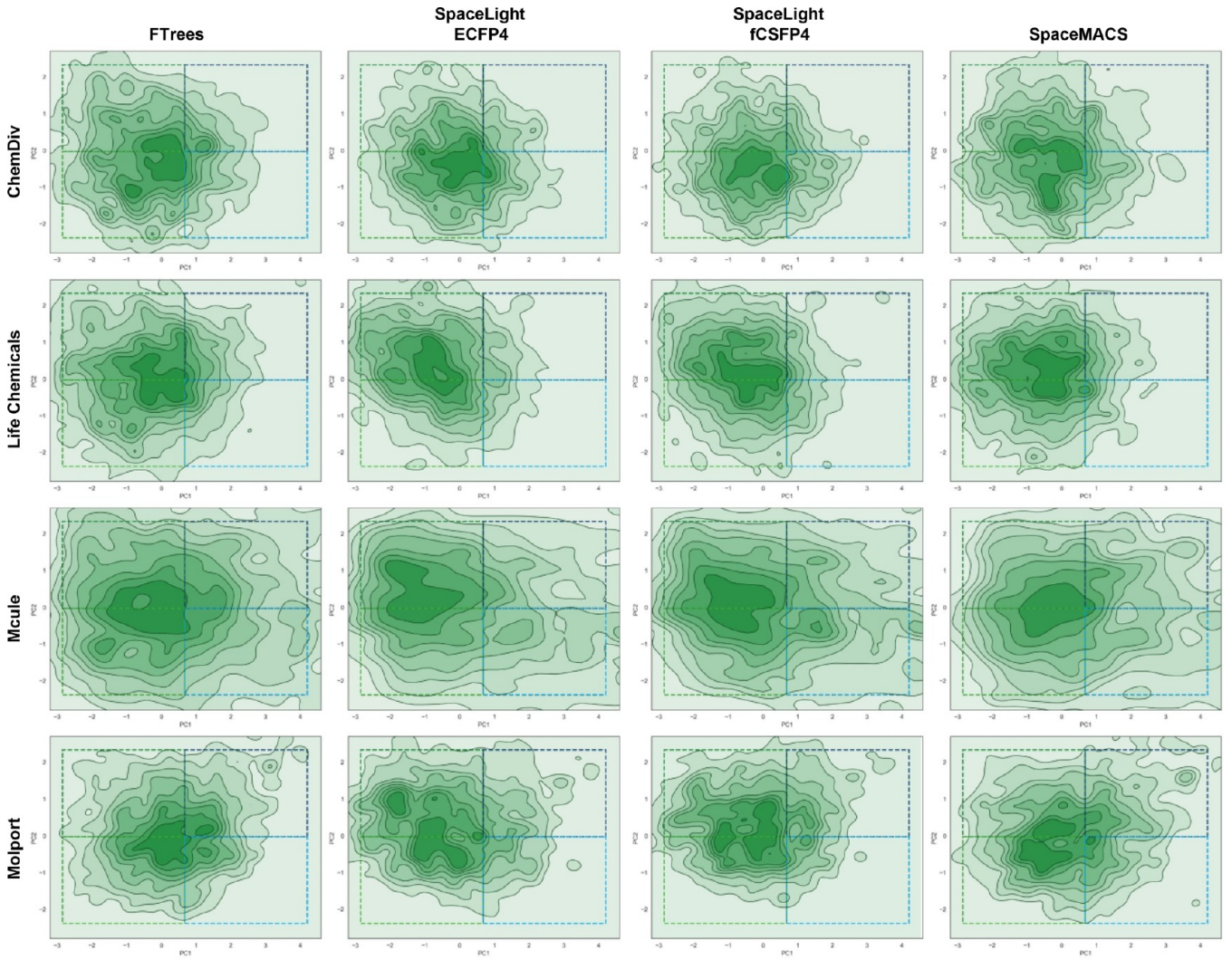


Figure 16. Overview of the chemical landscape distribution for rank 1 results of investigated commercial libraries and the applied search algorithms as density plots. For each distribution, a 10-level representation was selected. Corresponding quadrants of Set *S* are depicted in overlay: Q1 in dark green (upper left corner), Q2 in dark blue (upper right corner), Q3 in light green (bottom left corner) and Q4 in light blue (bottom right corner).

example, the deliberate choice to offer a drug-like library may result in the exclusion of bRo5 compounds or natural-product-like classes (e.g., nucleotides and aminoglycosides), which is reflected in the chemical coverage of the results. In Chemical Spaces, this aspect is controlled by setting or omitting size filters on the building blocks, thereby determining the molecular weights of the resulting products. Consequently, filtering based on other parameters such as substructure, number of H-bond donors/acceptors, and others influences the chemical space coverage.

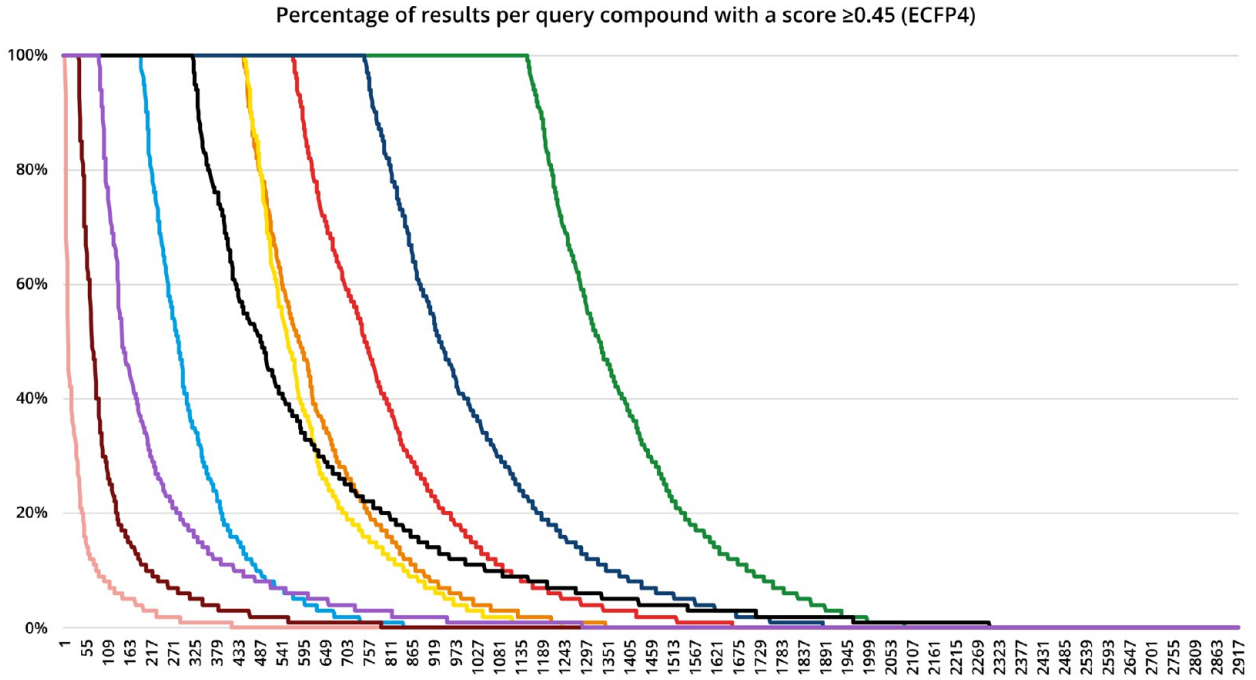
Assessment of Retrieved Results in regard to Chemical Space Coverage. One of our major interests was to employ the generated benchmark Set *S* to investigate the chemical space coverage by commercial compound sources. Considering the collected observations, the search results of the Chemical Spaces and enumerated libraries were then further analyzed. In particular, the question of the coverage of the chemical landscape defined by Set *S* is of interest.

To address this, the corresponding PC1 and PC2 values of the mined molecules were calculated along with their computed parameters (atomic polarizability, cLogP, number of H-bond donors and acceptors, stereocenters and rotatable bonds, PSA, and shape index) and the coefficients from the PCA to define their position in chemical space. Subsequently, the following density plots were created: one for Set *S*, as well as density plots for all sources and search methods to represent the results with rank 1 and the top 10 ranking results, to allow a visual evaluation of the Chemical Spaces, libraries, and the algorithms used.

As expected, the density plot of Set *S* (see Figure 14) shows a homogeneous distribution, which corresponds to the nature of the random selection of up to 30 possible compounds from the 10 × 10 segment matrix described above. Dense clusters (highest density level) can be found particularly in three central areas: in the intersection of Q1 and Q3, around the central point where all four quadrants meet, and in the middle between Q2 and Q4.

The Chemical Spaces consistently exhibit high compound densities proximal to the centroid of the PC matrix, with a drift toward Q1 and Q3. Molecules in this area have a molecular weight (MW) between approximately 250 and 500 Da, a clogP between −1 and 5, hydrogen bond acceptors between 0 and 9, and donors between 0 and 3, classifying them as typically drug-like. Consequently, the Chemical Spaces particularly cover the

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124



	ChemDiv	Life Chemicals	Molport	CHEMriya	Mcule	GalaXi	AMBrosia	Freedom Space	REAL Space	eXplore
# queries with 100% of results ≥0.45 (ECFP4)	4 (0.01%)	36 (1.2%)	88 (3.0%)	192 (6.6%)	320 (11.0%)	449 (15.4%)	448 (15.4%)	571 (19.6%)	746 (25.6%)	1,152 (39.5%)
# queries with ≥50% of results ≥0.45 (ECFP4)	10 (0.03%)	70 (2.4%)	146 (5.0%)	287 (9.8%)	492 (16.9%)	557 (19.1%)	588 (20.2%)	753 (25.8%)	938 (32.2%)	1,334 (45.7%)

Figure 17. Mean similarity score development within the top 100 retrieved results for queries with an SD ≥ 0.05 (SpaceLight) or SD ≥ 0.03 (SpaceMACS). Results for the Chemical Spaces are shown on the left, while results for the enumerated libraries are on the right.

area that is commonly explored for hit discovery and extension via analogs.

This trend is evident both for the rank 1 results (Figure 15) and the top 10 ranking compounds (see Figure S4). While the overall coverage of Q1 and Q3 is good across all examined Chemical Spaces, significantly weaker coverage is observed for Q2 and Q4. In particular, the edges of both quadrants are underpopulated, leading to the following conclusions about the retrieved compounds: Chemical Spaces provide relatively less good results for complex, hydrophilic compounds (e.g., nucleotides/nucleoside analogs, an increased proportion of charged groups such as amines, carboxylic acids, or guanidines). The same applies to compounds with a natural-product-like character (e.g., sp^3-rich carbon systems, opioid-related structures). Possible causes for this behavior include the availability of functionalizable building blocks for natural-product-like compounds; the absence of reactions to create the involved molecular scaffolds; the increased reactivity of more hydrophilic building blocks, which can lead to their exclusion in reaction definitions; as well as the synthetic complexity of these compounds, which may not be achievable through a one- or two-step synthesis (e.g., nucleotides[67]). The less complex compounds, which can be straightforwardly assembled from two to four building blocks, are more accessible and therefore constitute the majority of the retrieved results.

None of the examined Chemical Spaces exhibited extremely divergent behavior in terms of coverage based on physicochemical properties of the retrieved compounds. Proportionally, eXplore showed the greatest coverage of Q2 and Q4 among all Spaces. Additionally, the density clustering of compounds in eXplore is closer to the centroid of the matrix compared to all other Spaces, which consequently leads to an increase in the average molecular weight as well as the number of hydrogen bond donors and acceptors. Regarding drug-like results, Freedom and REAL Space consistently maintained a strong focus on them for all three applied methods. In the case of AMBrosia and GalaXi, the zone with the highest population density projects the least into the Q3 region, indicating slightly lower coverage of complex compounds.

As for the methods, all three algorithms produced similar density distributions across the different Spaces. In detail, FTrees occasionally exhibits the formation of clustering islands in the case of AMBRosia, CHEMriya, and eXplore due to delineations, which may indicate homogeneity in the physicochemical properties of the compounds. This behavior could be driven by the building blocks used and their assembly into specific molecular scaffolds.

Compared to FTrees, the ECFP4 and fCSFP4 SpaceLight results indicate no significant change in chemical landscape coverage of the Chemical Spaces. Again, clustering was observed for eXplore. Interestingly, REAL Space also showed clustering points in the fCSFP4 search that were not observed in the other methods. Such behavior may, in individual cases, be related to the captured features of the used fingerprints and the coupling

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

reactions of the chemical spaces, leading to structurally different assembled result molecules.

Lastly, a trend observed for SpaceMACS was the increase in density in the drug-like region for all the examined Spaces. A possible reason for this could be that the MCS matching predominantly favors more common scaffolds, which are also found in drug-like compounds.

In summary, the largest proportion of results retrieved by the three methods was in the range of classical drug-like structures. The evaluation of the top 10 ranking results showed no significant changes in the coverage landscapes. The addition of further compounds led to an increased enrichment in the drug-like region accompanied by a decline in isolated clusters (see Figure S5).

For comparison, the results from commercial compound libraries were also analyzed for their distribution (Figure 16). Interestingly, significant differences were observed compared to combinatorial Chemical Spaces. The libraries of ChemDiv, Life Chemicals, and Molport are noticeably more compact, translating into the reduced presence of borderline compounds. All four examined libraries maintained the highest density in the drug-like area.

A unique characteristic of Molport's method was the presence of molecules in Q2. Mcule, on the other hand, exhibited a broader distribution compared with the other three libraries and Chemical Spaces. Interestingly, it projected more into Q2 and Q4, indicating greater diversity in the physicochemical properties of the results. As already observed with the Chemical Spaces, expanding to the top 10 ranking results centers the density in the area of drug-like structures for all commercial libraries (see Figure S5).

The density plots suggest that the commercial libraries also may serve as sources for drug-like compounds, with Mcule, as the largest investigated library, showing the widest coverage beyond conventional structures among the examined sets.

Analysis of the Scoring Behavior within the Retrieved Results. To gain further insights into the performance of the Chemical Spaces and libraries, the scoring within the top 100 retrieved results was analyzed in more detail. For SpaceLight, query results from the respective Chemical Spaces, and consequently the libraries, with an SD $\geq$ 0.05 (approximately 2-fold of the average SDs) were uniformly examined. The top-ranking results, as well as the scores for the remaining results, were considered and grouped into rank clusters of 20. The corresponding mean score was then calculated for each cluster (see Tables S6 and S7). The results are visualized in Figure 17.

Consistently, the score for rank 1 and the cluster means exceeded the average values for the respective Space, with eXplore, Freedom Space, and REAL Space continuing to deliver the best results. As previously described, it became evident that rank 1 results were followed by the highest drops in scores for eXplore, Freedom Space, and REAL Space in the case of ECFP4. For the fCSFP4 results, the trend shifted: Fingerprint scores dropped for AMBrosia the most between rank 1 and ranks 1 to 20, followed by GalaXi, Freedom Space, CHEMriya, REAL Space, and eXplore. In contrast to ECFP4, which showed double-digit percentage decreases in the mean of ranks 1 to 20 compared to the mean of the rank 1 compounds (19.1 to 22.0%), only single-digit percentage decreases were observed in the case of fCSFP4 (2.4 to 9.9%). This may imply that fCSFP4 is particularly capable of delivering more similar compounds to the query in the top ranks than ECFP4 when applied to combinatorial Chemical Spaces. Another possible explanation is that fCSFP4's distribution behaves fundamentally differently from ECFP4's due to its feature capture, which has already been reported for other fingerprint methods.[64] It should be noted that a similar SD assessment for FTrees is only partially applicable and comparable as FTrees provides a narrower range of similarity scores. Furthermore, FTrees can assign identical scores to stereoisomers, meaning that multiple results may have a score of 1.0, which, in turn, reduces the SD. Therefore, the proportionality of the SD to the quality of the results must be considered in the context of each algorithm.

While the Chemical Spaces showed a range of 87 to 331 queries with an SD $\geq$ 0.05 for ECFP4, the libraries ranged from 77 to 725. For fCSFP4, the ranges were 96 to 281 (Chemical Spaces) and 173 to 785 (libraries), and for SpaceMACS, they were 113 to 296 (Chemical Spaces) and 1,111 to 1,800 (libraries). Approximately 10% of the queries yielded results with a broad similarity distribution for the Chemical Spaces, and this proportion increased up to over 50% when applying the same parameters to the libraries. The 10% observation for Chemical Spaces can be connected to the utilization of the 0.05 SD cutoff, which is approximately 2 times larger than the average SD of the Spaces. This effect was particularly pronounced in the Mcule set, which exhibited the greatest fluctuations in the SD across all three searches.

Furthermore, analyzing the similarity scores within the top 100 ranks per query can provide valuable insights into the set's ability to include relevant chemistry for a search query. In direct comparison, the similarity of the top-ranking compound to the next 20 ranks decreases much more rapidly for the libraries than for the Chemical Spaces. It is worth noting that the means of the rank 1 results with an increased SD were above the mean of all rank 1 results for the respective data set, suggesting that well-scoring structures were found in these cases, whereas for those with a lower SD value, generally lower-scoring compounds were retrieved. It can therefore be generalized that when good results are found, they inevitably bring along additional results with lower scores, which are reflected in an increased SD.

In the case of the SpaceMACS results, the lowest SD values were obtained for GalaXi followed by eXplore, CHEMriya and AMBrosia, and Freedom Space and REAL Space. Since we were interested in how those values can be coupled to the quality of the results, we performed an analogous assessment to the SpaceLight results mentioned above. For SpaceMACS, an SD cutoff of $\geq$ 0.03 was selected (approximately 2-fold of the average SDs) for the results per query with high fluctuations in their mean similarity score (see Table S6 of the Supporting Information). The mean of MCS similarities of results per query with SD $\geq$ 0.03 was higher than the mean of the results for the whole respective Chemical Space, which equates to better results for those queries. Since the ranges of score decreases per cluster transition are quite similar across all Chemical Spaces, no general conclusion about the behavior and parameters of the results can be drawn. Given the fact that many identical molecules (those with an MCS similarity of 1.0) as well as those with fairly good scores are found among those with an SD > 0.03, it suggests the possibility that the higher average scores of a Chemical Space for individual results contribute to fluctuations, while the majority of lower scores pull the SD values down. An additional interplay with a larger number of structurally similar analogs, influenced by the size of the Chemical Space, adds further complexity to these relationships.

This insight is important because it provides information about the quality of the retrieved results for the respective data

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

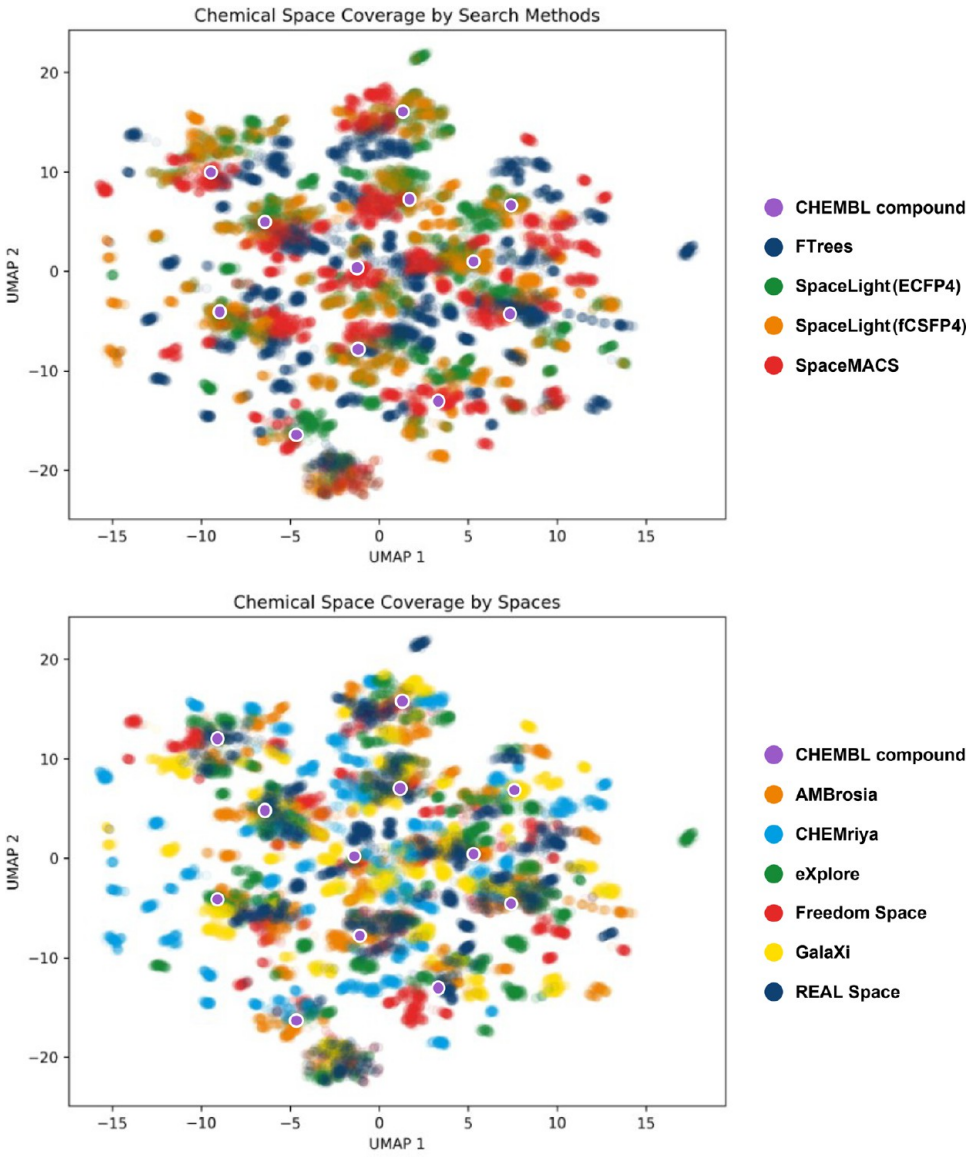


Figure 18. UMAP analysis of Chemical Spaces with the respective coloring by the source and applied search method. The used ChEMBL query compounds (**9**−**20**; for structures, see Figure 4) are highlighted in purple.

sets. In comparison, the combinatorial Chemical Spaces performed better across all three searches (SpaceLight ECFP4 and fCSFP4, as well as SpaceMACS) than the enumerated libraries. The decline in mean similarities across the rank clusters is more pronounced for the libraries than for the Chemical Spaces, meaning that the best-performing library manages to keep up with only the worst-performing Chemical Space.

To further look into the capacities to provide related structures to a query compound, we performed another analysis of the compound sources. Using the ECFP4 SpaceLight results, we examined how many of the 100 retrieved results for a query compound of Set *S* have a score ≥ 0.45. The premise is that an ECFP4 score of ≥ 0.45 is still considered to be of a relevant similarity. It was observed that Chemical Spaces were able to provide at least 50 structurally related analogs for up to 46% of the queries in the best case (eXplore), whereas the best-performing library (Mcule) achieved this in only 17% of the cases (2.7-fold difference) (see Figure 17). The performance ranking of the sources remained the same with regard to the ability to deliver at least 100 relevant results for a query compound. Mcule was the only library able to outperform CHEMriya, but it still fell short of the capacities of the other Chemical Spaces.

The percentage decline in similarity from one rank cluster to the next converges to similar levels. Combined with the previous observation that Chemical Spaces perform better on average, this suggests that Chemical Spaces provide more similar compounds that also exhibit higher similarity scores than the investigated libraries. This leads to the presence of more analogs in Chemical Spaces alongside the top-ranking result, expanding the molecular portfolio for hit exploration, lead optimization, scaffold hops, and SAR studies.

Furthermore, the lower percentage decrease within the fCSFP4 search compared with the ECFP4 search indicates that this fingerprint variant performs more robustly in Chemical Spaces. While both fingerprint variants exhibited comparable performance in the enumerated sets, even across the respective libraries—similar to SpaceMACS on the same sets—the performance of fCSFP4 in Chemical Spaces was on par with that of SpaceMACS. While all methods can be applied to

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

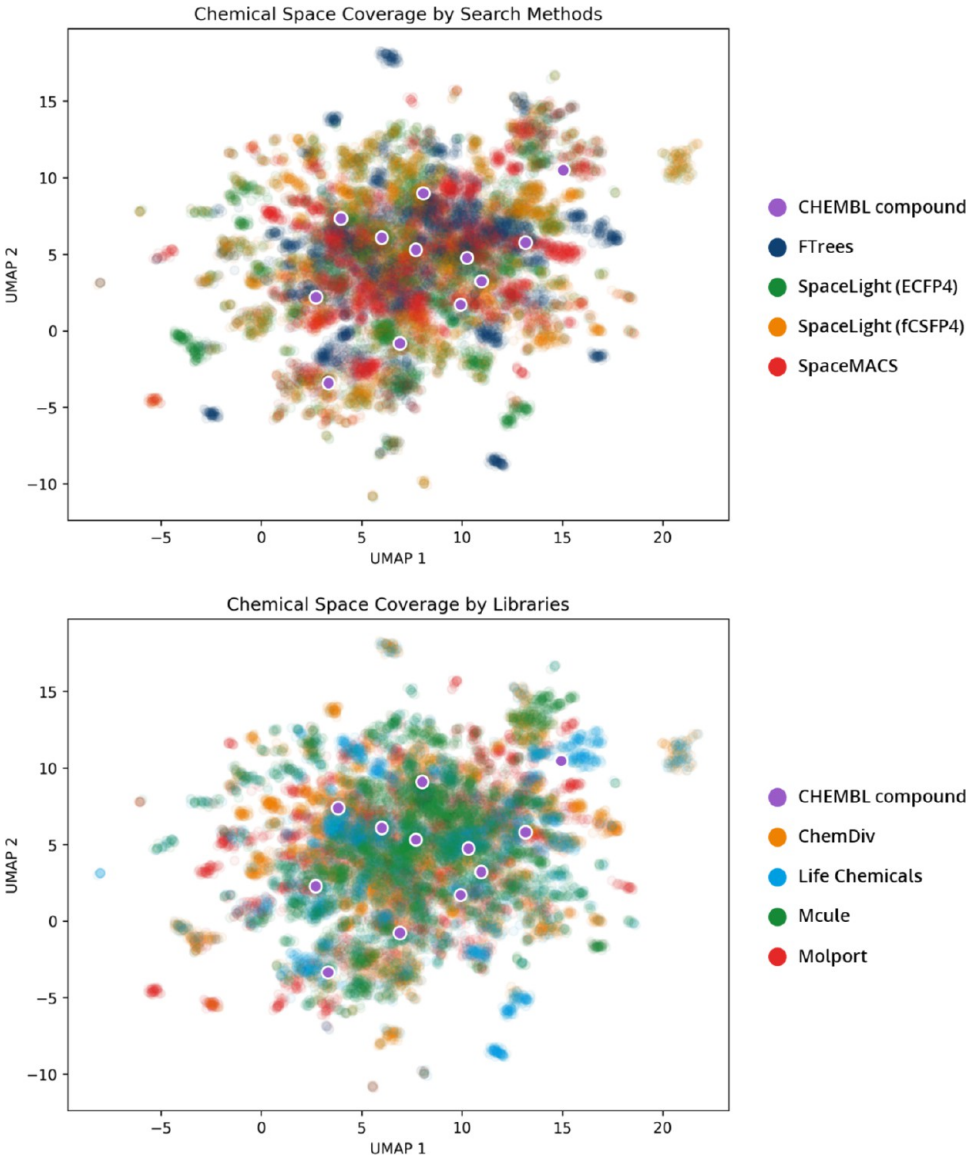


Figure 19. UMAP analysis of commercial libraries with the respective coloring by source and applied search method. The used ChEMBL query compounds (**9**−**20**; for structures, see Figure 4) are highlighted in purple.

enumerated data sets, the fCSFP4 variant and SpaceMACS were specifically developed for use on combinatorial Chemical Spaces which is not the case for ECFP4.[13,14] Thus, these observations are consistent with previous studies and the corresponding expectations.

For the final comparative assessment of the chemical diversity of the retrieved molecules, a uniform manifold approximation and projection (UMAP) analysis was conducted. UMAP, like PCA, is widely applied to reduce high-dimensional molecular descriptor data into a visual representation to facilitate the interpretation of molecular relationships within a compound set. An example of typical UMAP mapping of a query compound and results of the three search methods used in the study are shown in Figure S6 of the Supporting Information.

For the UMAP visualization, the 12 compounds from Figure 4 of Set *S* and their retrieved results from the Chemical Spaces and libraries were selected. For the generation of the 2D UMAP analysis, the ECFP4 fingerprint was used by default as a similarity score for determining the relationships of the molecules with the following parameters: *n_neighbors* = 50, *min_dist* = 0.8, *random_state* = 42. The corresponding 2D representations by method and compound source are summarized in Figures 18 and 19.

From the representations of the Chemical Spaces, the following insights can be drawn: For each query compound, a distinct cluster of results is formed that is visually separated from the others. Within the clusters, the applied search methods can largely be distinguished from one another.

As expected, the FTrees results were the farthest from the coordinates of the query compound because the search algorithm operates independently of the connectivity of heavy atoms within the query molecule and rather captures the pharmacophore features of its increments. By using ECFP4 for mapping, it is placed in context with the query compound through an orthogonal method, leading to a shift of the results. In contrast, the methods SpaceLight and SpaceMACS, which depend on the connectivity of heavy atoms, show a significantly closer proximity to the query compound.

Regarding the differences within the Chemical Spaces, CHEMriya and GalaXi particularly exhibit the most outliers

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

beyond the dense clusters. In contrast, the retrieved results from eXplore and REAL Space are predominantly located close to the coordinates of the query compounds. The still-possible visual distinction by source within the clusters themselves also leads to the conclusion that each Chemical Space carries a high proportion of unique scaffolds and chemistry, which aligns with previous studies.[15,21,22]

Subsequently, an analogous analysis was also conducted for the commercial enumerated libraries. The visualization is depicted in Figure 19.

In direct comparison to the visualization of the Chemical Spaces, it is noticeable that no distinct clusters are formed around or in the periphery of the query compounds, making the results appear significantly more homogeneous. Again, a trend emerges where FTrees exhibit the greatest distance to the query compounds, with the fingerprint-based results from SpaceLight also frequently deviating. The previously observed uniqueness of the Chemical Space results within a cluster cannot be observed in the case of the libraries. The sets seem to provide much more unified results where a precise assignment to the query compound or a source is not clearly possible. The increased absence of Mcule and Molport outside the cluster assemblies provides an indication that the associated results are closest to those of the query compounds. Most of the outliers therefore emerged from the ChemDiv and Life Chemicals sets, which in turn reflect the calculated means for the libraries and applied methods.

In summary, UMAP analysis showed that each of the applied search methods was able to enrich individual chemical diversity. Additionally, the different Chemical Spaces provided unique molecular scaffolds with only minimal overlap between the various sources. However, structural distinction was absent in the case of the libraries: The results suggest that even among the top 100 ranking compounds, it becomes difficult with the applied methods to distinguish which query a result belongs to. Considering which similarity score range can still be deemed "acceptable" to classify a result as similar to a query compound, especially when lower ranks display significant similarity to another, structurally unrelated query compound, this issue emphasizes the capacity limits of the commercial libraries or too-small molecule sets in general.

Computation Time. To process the searches using Set *S* as queries (2,917 entries), computation times for the Chemical Spaces ranged from 11 to 1386 min. Notably, REAL Space showed particularly long computation times. In the case of the enumerated libraries, computation times ranged from 9 to 37,470 min. The results and the description of the hardware used can be found in Figure S7 of the Supporting Information.

When the number of compounds contained in a source is compared relative to the time required for screening, the following performance ratios can be established for the tools used: FTrees was able to search Chemical Spaces 3000 to 1.5×10^7 times faster than the libraries, and SpaceLight achieved 2500 to 2.8×10^6 times faster performance in the case of ECFP4 and 2200 to 1.9×10^6 times faster for fCSFP. SpaceMACS demonstrated the highest efficiency, being 48,000 to 1.8×10^8 times faster.

However, it must be explicitly emphasized that the tools used were designed for combinatorial Chemical Spaces and therefore operate optimally within them. While they are capable of processing standard SD files, they do so less efficiently. In this sense, the resulting bias should be acknowledged.

CONCLUSIONS

Following is a summary of the efforts of this study: With the vision of creating a highly relevant and versatile molecular set, the ChEMBL database was searched for bioactive substances. The raw set, consisting of approximately 11 million entries, was systematically filtered to retain compounds commonly used in modern drug discovery screenings. To exclude isolated activity events outside of a compound series, molecules with fewer than five members sharing a Bemis−Murcko scaffold were removed. This process resulted in Set *L* ("large-sized," 379,169 compounds), followed by additional downsizing steps leading to Set *M* ("medium-sized," 25,234 compounds) and Set *S* ("small-sized," 2,917 compounds). Set *M* contains only the smallest representatives of each Bemis−Murcko scaffold group from Set *L*. Set *S* was derived from a PCA analysis of Set *M*, where outliers with extreme physicochemical and topological properties were excluded. Random molecules were then selected from the resulting matrix to ensure homogeneous coverage of chemical space and a comfortable size for computationally slim assessments.

In the context of the applicability in drug discovery campaigns, each set can be further filtered based on project-specific needs, such as selecting compounds with drug-like properties, fragments, or bRo5 characteristics. Furthermore, the generated benchmark sets are suitable for a wide range of computational tasks and applications in the context of both ligand- and structure-based drug design. The three sets, each 1 order of magnitude larger than the previous one, can benefit not only extensive but also moderate computations. Additionally, they are well-suited as seeds for machine learning or artificial intelligence applications, serving as starting points for the de novo design of compounds with improved physicochemical or pharmacological properties and iterations of molecular scaffolds.

Subsequently, Set *S* was used as queries to evaluate commercial sources for compounds, namely, combinatorial Chemical Spaces and enumerated vendor libraries, in terms of their ability to provide similarity-based chemistry. In terms of similarity means, eXplore and REAL Space consistently performed best among the Chemical Spaces. Among the libraries, Mcule's "Full" library achieved the highest scores. A trend was observed between library size and similarity scores (both fingerprint- and MCS-based), where an increase in size was associated with higher scores. Taking into account the standard ECFP4 fingerprint, it was particularly the larger sets that were able to deliver structurally related molecules for a given drug-like query (see Figure S8 of the Supporting Information). However, it cannot be ruled out that the prior rational design of the library contributes to this effect.

While both source architectures were able to cover the chemical space around structures that comply with drug-like physicochemical properties, the Chemical Spaces provided more compounds whose average similarity was higher than that of the libraries. In direct comparison, the best-performing library, Mcule "Full", was at the level of the Chemical Spaces with the lowest mean similarity scores and chemical landscape coverage. Especially considering that libraries can conveniently be enriched with reported bioactive compounds, it is worth highlighting that Chemical Spaces, solely through the use of building blocks and reactions, can generate more and structurally similar compounds related to a query. This is particularly relevant when a compound is not identically present in the collection, making alternative scaffolds the only option.

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

Furthermore, this scenario extends to more complex queries that require multiple intricate synthesis steps, compounds that depend on expensive or proprietary building blocks, and structures that have not yet been commercially registered or included in catalogs.

The three search methods used—FTrees, SpaceLight, and SpaceMACS—were each able to extract individually diverse chemistry based on different interpretations of similarity. This can be leveraged in a project-specific manner to generate tailored libraries. Given the vast volume of Chemical Spaces, this opens up opportunities to independently create customized enumerated compound libraries, comparable to the studied sets, on the scale of millions. Given the size of the investigated sets, the search algorithms performed more efficiently on the combinatorial Chemical Spaces based on the required computation time per compound.

Our analysis suggests that there is still significant development potential for bRo5 compounds in both Chemical Spaces and commercial libraries. As expected, more complex compounds were less well represented in commercial sources compared to classic drug-like compounds. In the case of Chemical Spaces, this also covers compounds with hydrophilic groups that cannot be made in one or two steps due to the potential reactivity of the functionalities in the associated building blocks that may lead to unwanted byproducts during the synthesis. This suggests that additional reactions are needed to capture this uncharted area, a problem well-known in medicinal chemistry,[54,68,69] or that an additional layer of synthesis processing must be introduced, including the possible use of protecting groups, which could ultimately lead to the desired product. Furthermore, our analysis suggests that the coverage of more lipophilic substances can be improved by expanding the portfolio of functionalizable natural product-like and sp^3-rich building blocks.

The highlighted potential extends beyond ligand-based applications: access to more and more relevant structures is equally crucial in a structure-focused context to identify the best possible candidates for follow-up. Several virtual screening campaigns involving combinatorial Chemical Spaces have already demonstrated the superiority of larger hunting grounds.[70−72]

These findings will contribute to the development of a holistic understanding of the chemical space. Furthermore, they will aid in identifying gaps in the molecular class landscape that require improvement and in enhancing their accessibility.

■ ASSOCIATED CONTENT

Data Availability Statement

All data from the ChEMBL database were retrieved in December 2024. All Chemical Spaces and commercial vendor libraries were publicly accessed in December 2024. Their input data are not publicly available. In the case of the eXplore Space, however, a collection of the reactions used, including their SMARTS representations, is freely available for download (https://www.biosolveit.de/infiniSee/cookbook). The input data for the REAL Space can be licensed from BioSolveIT. A license is required to use FTrees, SpaceLight, and SpaceMACS.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00719.

Raw score data in table form, additional quadrant and score distribution assessment, and computational times and setup (PDF)

Benchmark_Set_L [379k]: the large-sized generated molecule collection of bioactive compounds (CSV)

Benchmark_Set_M_[25k]: the medium-sized generated molecule collection of bioactive compounds (CSV)

Benchmark_Set_S_[3k]: the small-sized generated molecule collection of bioactive compounds (CSV)

■ AUTHOR INFORMATION

Corresponding Author

Alexander Neumann − *BioSolveIT GmbH, Sankt Augustin 53757, Germany;* orcid.org/0000-0002-1446-4389; Phone: +49-2241-2525-566; Email: alexander.neumann@biosolveit.de; Fax: +49-2241-2525-525

Author

Raphael Klein − *BioSolveIT GmbH, Sankt Augustin 53757, Germany;* orcid.org/0000-0002-5087-8730

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00719

Author Contributions

A.N.: Conceptualization of the study design, generation of Set *S* and Set *S′*, analysis of the raw search result data, and visualization. R.K.: Collection of the ChEMBL raw data, filtering and processing to create Set *L* and *M*, initiation of the Chemical Space and commercial library search runs, confirmation of identical molecules, and calculation of target distribution. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank BioSolveIT for providing the licenses to the Chemical Space exploration tools. The authors have no conflicts of interest to declare.

■ ABBREVIATIONS

AI,artificial intelligence; DEL,DNA-encoded library; CSFP,connected subgraph fingerprint; ECFP,extended-connectivity fingerprint; HERG,human Ether-à-go-go-Related Gene; HTS,high-throughput screening; Kir,inward-rectifier potassium channel; Kv,voltage-gated potassium channels; MCS,maximum common substructure; ML,machine learning; MW,molecular weight; PAH,polycyclic aromatic hydrocarbons; PC,principal component; PCA,principal component analysis; PSA,polar surface area; QX,quadrant number X; SAR,structure−activity relationship; SD,standard deviation; Set L,benchmark set L (denoted for "large-sized"); Set M,benchmark set M (denoted for "medium-sized"); Set S,benchmark set S (denoted for "small-sized"); Sim = 1,similarity score of 1.0; UMAP,uniform manifold approximation and projection

■ REFERENCES

(1) Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B. Evolution of Commercially Available Compounds for HTS. *Drug Discovery Today* **2019**, *24* (2), 390−402.

(2) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (9), 2021−2034.

(3) Tingle, B. I.; Tang, K. G.; Castanon, M.; Gutierrez, J. J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y. S.; Irwin, J. J. ZINC-22—A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J. Chem. Inf. Model.* **2023**, *63* (4), 1166−1176.
(4) *EnamineStore*. https://new.enaminestore.com/ (accessed 2025−03−29).
(5) Gloriam, D. E. Bigger Is Better in Virtual Drug Screens. *Nature* **2019**, *566* (7743), 193−194.
(6) Lyu, J.; Irwin, J. J.; Shoichet, B. K. Modeling the Expansion of Virtual Screening Libraries. *Nat. Chem. Biol.* **2023**, *19* (6), 712−718.
(7) Cherkasov, A. The 'Big Bang' of the Chemical Universe. *Nat. Chem. Biol.* **2023**, *19* (6), 667−668.
(8) Korn, M.; Ehrt, C.; Ruggiu, F.; Gastreich, M.; Rarey, M. Navigating Large Chemical Spaces in Early-Phase Drug Discovery. *Curr. Opin. Struct. Biol.* **2023**, *80*, No. 102578.
(9) Klingler, F.-M.; Gastreich, M.; Grygorenko, O.; Savych, O.; Borysko, P.; Griniukova, A.; Gubina, K.; Lemmen, C.; Moroz, Y. SAR by Space: Enriching Hit Sets from the Chemical Space. *Molecules* **2019**, *24* (17), No. 3096.
(10) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discovery Today* **2019**, *24* (5), 1148−1156.
(11) Protopopov, M. V.; Tararina, V. V.; Bonachera, F.; Dzyuba, I. M.; Kapeliukha, A.; Hlotov, S.; Chuk, O.; Marcou, G.; Klimchuk, O.; Horvath, D.; Yeghyan, E.; Savych, O.; Tarkhanova, O. O.; Varnek, A.; Moroz, Y. S. The Freedom Space − a New Set of Commercially Available Molecules for Hit Discovery. *Mol. Inf.* **2024**, *43*, No. e202400114.
(12) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput. Aided. Mol. Des.* **1998**, *12* (5), 471−490.
(13) Bellmann, L.; Penner, P.; Rarey, M. Topological Similarity Search in Large Combinatorial Fragment Spaces. *J. Chem. Inf. Model.* **2021**, *61* (1), 238−251.
(14) Schmidt, R.; Klein, R.; Rarey, M. Maximum Common Substructure Searching in Combinatorial Make-on-Demand Compound Spaces. *J. Chem. Inf. Model.* **2022**, *62* (9), 2133−2150.
(15) Perebyinis, M.; Rognan, D. Overlap of On-demand Ultra-large Combinatorial Spaces with On-the-shelf Drug-like Libraries. *Mol. Inf.* **2023**, *42* (1), No. 2200163.
(16) Pikalyova, R.; Akhmetshin, T.; Horvath, D.; Varnek, A. CoLiNN: A Tool for Fast Chemical Space Visualization of Combinatorial Libraries Without Enumeration. *Mol. Inf.* **2025**, *44*, No. e202400263.
(17) Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. Meta-GTM: Visualization and Analysis of the Chemical Library Space. *J. Chem. Inf. Model.* **2023**, *63* (17), 5571−5582.
(18) Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Horvath, D.; Gavrilenko, K. S.; Marcou, G.; Moroz, Y. S.; Oksiuta, O.; Varnek, A. A Close-up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry. *J. Chem. Inf. Model.* **2022**, *62* (9), 2171−2185.
(19) Pikalyova, R.; Zabolotna, Y.; Volochnyuk, D. M.; Horvath, D.; Marcou, G.; Varnek, A. Exploration of the Chemical Space of DNA-encoded Libraries. *Mol. Inf.* **2022**, *41* (6), No. 2100289.
(20) Bellmann, L.; Klein, R.; Rarey, M. Calculating and Optimizing Physicochemical Property Distributions of Large Combinatorial Fragment Spaces. *J. Chem. Inf. Model.* **2022**, *62* (11), 2800−2810.
(21) Bellmann, L.; Penner, P.; Gastreich, M.; Rarey, M. Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. *J. Chem. Inf. Model.* **2022**, *62* (3), 553−566.
(22) Neumann, A.; Marrison, L.; Klein, R. Relevance of the Trillion-Sized Chemical Space "EXplore" as a Source for Drug Discovery. *ACS Med. Chem. Lett.* **2023**, *14* (4), 466−472.
(23) Revillo Imbernon, J.; Jacquemard, C.; Bret, G.; Marcou, G.; Kellenberger, E. Comprehensive Analysis of Commercial Fragment Libraries. *RSC Med. Chem.* **2022**, *13* (3), 300−310.
(24) Orlov, A. A.; Akhmetshin, T. N.; Horvath, D.; Marcou, G.; Varnek, A. From High Dimensions to Human Insight: Exploring Dimensionality Reduction for Chemical Space Visualization. *Mol. Inf.* **2025**, *44*, No. e202400265.
(25) Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T. Comparative Analyses of Structural Features and Scaffold Diversity for Purchasable Compound Libraries. *J. Cheminform.* **2017**, *9* (1), 25.
(26) Zhu, H. Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu. Rev. Pharmacol. Toxicol.* **2020**, *60* (1), 573−589.
(27) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminform* **2017**, *9* (1), 45.
(28) Keshavarzi Arshadi, A.; Salem, M.; Firouzbakht, A.; Yuan, J. S. MolData, a Molecular Benchmark for Disease and Target Based Machine Learning. *J. Cheminform.* **2022**, *14* (1), 10.
(29) Tian, T.; Li, S.; Zhang, Z.; Chen, L.; Zou, Z.; Zhao, D.; Zeng, J. Benchmarking Compound Activity Prediction for Real-World Drug Discovery Applications. *Commun. Chem.* **2024**, *7* (1), 127.
(30) Isigkeit, L.; Chaikuad, A.; Merk, D. A Consensus Compound/Bioactivity Dataset for Data-Driven Drug Design and Chemogenomics. *Molecules* **2022**, *27* (8), No. 2513.
(31) Zhang, J.-Y.; Wang, Y.-T.; Sun, L.; Wang, S.-Q.; Chen, Z.-S. Synthesis and Clinical Application of New Drugs Approved by FDA in 2022. *Mol. Biomed.* **2023**, *4* (1), 26.
(32) Tran-Nguyen, V.-K.; Rognan, D. Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement. *Int. J. Mol. Sci.* **2020**, *21* (12), No. 4380.
(33) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. *J. Comput. Aided. Mol. Des.* **2015**, *29* (9), 885−896.
(34) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, *52* (D1), D1180−D1192.
(35) Prudent, R.; Lemoine, H.; Walsh, J.; Roche, D. Affinity Selection Mass Spectrometry Speeding Drug Discovery. *Drug Discovery Today* **2023**, *28* (11), No. 103760.
(36) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162* (6), 1239−1249.
(37) Holenz, J.; Stoy, P. Advances in Lead Generation. *Bioorg. Med. Chem. Lett.* **2019**, *29* (4), 517−524.
(38) Shultz, M. D. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. *J. Med. Chem.* **2019**, *62* (4), 1701−1714.
(39) Roskoski, R. Rule of Five Violations among the FDA-Approved Small Molecule Protein Kinase Inhibitors. *Pharmacol. Res.* **2023**, *191*, No. 106774.
(40) Doak, B. C.; Kihlberg, J. Drug Discovery beyond the Rule of 5 - Opportunities and Challenges. *Expert Opin. Drug Discovery* **2017**, *12* (2), 115−119.
(41) Egbert, M.; Whitty, A.; Keserű, G. M.; Vajda, S. Why Some Targets Benefit from beyond Rule of Five Drugs. *J. Med. Chem.* **2019**, *62* (22), 10005−10025.
(42) DeGoey, D. A.; Chen, H.-J.; Cox, P. B.; Wendt, M. D. Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection. *J. Med. Chem.* **2018**, *61* (7), 2636−2651.
(43) O'Reilly, M.; Cleasby, A.; Davies, T. G.; Hall, R. J.; Ludlow, R. F.; Murray, C. W.; Tisi, D.; Jhoti, H. Crystallographic Screening Using Ultra-Low-Molecular-Weight Ligands to Guide Drug Design. *Drug Discovery Today* **2019**, *24* (5), 1081−1086.
(44) Doak, B. C.; Norton, R. S.; Scanlon, M. J. The Ways and Means of Fragment-Based Drug Design. *Pharmacol. Ther.* **2016**, *167*, 28−37.
(45) Yu, X.; Sun, D. Macrocyclic Drugs and Synthetic Methodologies toward Macrocycles. *Molecules* **2013**, *18* (6), 6230−6268.

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124

(46) Zalessky, I.; Wootton, J. M.; Tam, J. K. F.; Spurling, D. E.; Glover-Humphreys, W. C.; Donald, J. R.; Orukotan, W. E.; Duff, L. C.; Knapper, B. J.; Whitwood, A. C.; Tanner, T. F. N.; Miah, A. H.; Lynam, J. M.; Unsworth, W. P. A Modular Strategy for the Synthesis of Macrocycles and Medium-Sized Rings via Cyclization/Ring Expansion Cascade Reactions. *J. Am. Chem. Soc.* **2024**, *146* (8), 5702−5711.
(47) Zhu, Z.; Shaginian, A.; Grady, L. C.; O'Keeffe, T.; Shi, X. E.; Davie, C. P.; Simpson, G. L.; Messer, J. A.; Evindar, G.; Bream, R. N.; Thansandote, P. P.; Prentice, N. R.; Mason, A. M.; Pal, S. Design and Application of a DNA-Encoded Macrocyclic Peptide Library. *ACS Chem. Biol.* **2018**, *13* (1), 53−59.
(48) Chai, J.; Arico-Muendel, C. C.; Ding, Y.; Pollastri, M. P.; Scott, S.; Mantell, M. A.; Yao, G. Synthesis of a DNA-Encoded Macrocyclic Library Utilizing Intramolecular Benzimidazole Formation. *Bioconjugate Chem.* **2023**, *34* (6), 988−993.
(49) Pognan, F.; Beilmann, M.; Boonen, H. C. M.; Czich, A.; Dear, G.; Hewitt, P.; Mow, T.; Oinonen, T.; Roth, A.; Steger-Hartmann, T.; Valentin, J.-P.; Van Goethem, F.; Weaver, R. J.; Newham, P. The Evolving Role of Investigative Toxicology in the Pharmaceutical Industry. *Nat. Rev. Drug Discovery* **2023**, *22* (4), 317−335.
(50) Ye, L.; Ngan, D. K.; Xu, T.; Liu, Z.; Zhao, J.; Sakamuru, S.; Zhang, L.; Zhao, T.; Xia, M.; Simeonov, A.; Huang, R. Prediction of Drug-Induced Liver Injury and Cardiotoxicity Using Chemical Structure and in Vitro Assay Data. *Toxicol. Appl. Pharmacol.* **2022**, *454*, No. 116250.
(51) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460−473.
(52) Han, Z.; Shen, Z.; Pei, J.; You, Q.; Zhang, Q.; Wang, L. Transformation of Peptides to Small Molecules in Medicinal Chemistry: Challenges and Opportunities. *Acta Pharm. Sin. B* **2024**, *14* (10), 4243−4265.
(53) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54* (10), 3451−3479.
(54) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59* (10), 4443−4458.
(55) Ertl, P.; Altmann, E.; McKenna, J. M. The Most Common Functional Groups in Bioactive Molecules and How Their Popularity Has Evolved over Time. *J. Med. Chem.* **2020**, *63* (15), 8408−8418.
(56) Mueller, L. G.; Slusher, B. S.; Tsukamoto, T. Empirical Analysis of Drug Targets for Nervous System Disorders. *ACS Chem. Neurosci.* **2024**, *15* (3), 394−399.
(57) Vasaikar, S.; Bhatia, P.; Bhatia, P.; Chu Yaiw, K. Complementary Approaches to Existing Target Based Drug Discovery for Identifying Novel Drug Targets. *Biomedicines* **2016**, *4* (4), No. 27.
(58) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discovery* **2017**, *16* (1), 19−34.
(59) Schmidt, R.; Krull, F.; Heinzke, A. L.; Rarey, M. Disconnected Maximum Common Substructures under Constraints. *J. Chem. Inf. Model.* **2020**, *61*, 167.
(60) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(61) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for in Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093−3098.
(62) Venkatraman, V.; Gaiser, J.; Demekas, D.; Roy, A.; Xiong, R.; Wheeler, T. J. Do Molecular Fingerprints Identify Diverse Active Drugs in Large-Scale Virtual Screening?(No). *Pharmaceuticals* **2022**, *17*, 992.
(63) Skinnider, M. A.; Dejong, C. A.; Franczak, B. C.; McNicholas, P. D.; Magarvey, N. A. Comparative Analysis of Chemical Similarity Methods for Modular Natural Products with a Hypothetical Structure Enumeration Algorithm. *J. Cheminform.* **2017**, *9* (1), 46.
(64) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. *Expert Opin. Drug Discovery* **2016**, *11* (2), 137−148.
(65) Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Orhan, I. E.; Banach, M.; Rollinger, J. M.; Barreca, D.; Weckwerth, W.; Bauer, R.; Bayer, E. A.; Majeed, M.; Bishayee, A.; Bochkov, V.; Bonn, G. K.; Braidy, N.; Bucar, F.; Cifuentes, A.; D'Onofrio, G.; Bodkin, M.; Diederich, M.; Dinkova-Kostova, A. T.; Efferth, T.; El Bairi, K.; Arkells, N.; Fan, T.-P.; Fiebich, B. L.; Freissmuth, M.; Georgiev, M. I.; Gibbons, S.; Godfrey, K. M.; Gruber, C. W.; Heer, J.; Huber, L. A.; Ibanez, E.; Kijjoa, A.; Kiss, A. K.; Lu, A.; Macias, F. A.; Miller, M. J. S.; Mocan, A.; Müller, R.; Nicoletti, F.; Perry, G.; Pittalà, V.; Rastrelli, L.; Ristow, M.; Russo, G. L.; Silva, A. S.; Schuster, D.; Sheridan, H.; Skalicka-Woźniak, K.; Skaltsounis, L.; Sobarzo-Sánchez, E.; Bredt, D. S.; Stuppner, H.; Sureda, A.; Tzvetkov, N. T.; Vacca, R. A.; Aggarwal, B. B.; Battino, M.; Giampieri, F.; Wink, M.; Wolfender, J.-L.; Xiao, J.; Yeung, A. W. K.; Lizard, G.; Popp, M. A.; Heinrich, M.; Berindan-Neagoe, I.; Stadler, M.; Daglia, M.; Verpoorte, R.; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. *Nat. Rev. Drug Discovery* **2021**, *20* (3), 200−216.
(66) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminform.* **2020**, *12* (1), 43.
(67) Roy, B.; Depaix, A.; Périgaud, C.; Peyrottes, S. Recent Trends in Nucleotide Synthesis. *Chem. Rev.* **2016**, *116* (14), 7854−7897.
(68) Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M. Expanding the Medicinal Chemistry Synthetic Toolbox. *Nat. Rev. Drug Discovery* **2018**, *17* (10), 709−727.
(69) Dombrowski, A. W.; Aguirre, A. L.; Shrestha, A.; Sarris, K. A.; Wang, Y. The Chosen Few: Parallel Library Reaction Methodologies for Drug Discovery. *J. Org. Chem.* **2022**, *87* (4), 1880−1897.
(70) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; Zvonok, N.; Jain, M. K.; Savych, O.; Radchenko, D. S.; Nikas, S. P.; Petasis, N. A.; Moroz, Y. S.; Roth, B. L.; Makriyannis, A.; Katritch, V. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature* **2022**, *601* (7893), 452−459.
(71) Beroza, P.; Crawford, J. J.; Ganichkin, O.; Gendelev, L.; Harris, S. F.; Klein, R.; Miu, A.; Steinbacher, S.; Klingler, F.-M.; Lemmen, C. Chemical Space Docking Enables Large-Scale Structure-Based Virtual Screening to Discover ROCK1 Kinase Inhibitors. *Nat. Commun.* **2022**, *13* (1), 6447.
(72) Müller, J.; Klein, R.; Tarkhanova, O.; Gryniukova, A.; Borysko, P.; Merkl, S.; Ruf, M.; Neumann, A.; Gastreich, M.; Moroz, Y. S.; Klebe, G.; Glinca, S. Magnet for the Needle in Haystack: "Crystal Structure First" Fragment Hits Unlock Active Chemical Matter Using Targeted Exploration of Vast Chemical Spaces. *J. Med. Chem.* **2022**, *65* (23), 15663−15678.

https://doi.org/10.1021/acs.jcim.5c00719
J. Chem. Inf. Model. 2025, 65, 9097−9124