Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space

Alexios Koutsoukas,[†,∇] Shardul Paricharak,[†,§,∇] Warren R. J. D. Galloway,[‡] David R. Spring,[‡] Adriaan P. IJzerman,[§] Robert C. Glen,[†] David Marcus,[†] and Andreas Bender*[,†]

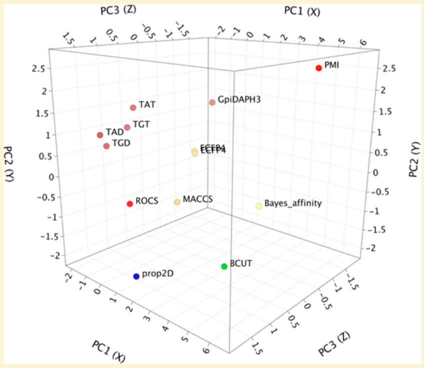
[†]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, United Kingdom

[‡]Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, United Kingdom

[§]Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands

Ⓢ *Supporting Information*

ABSTRACT: Chemical diversity is a widely applied approach to select structurally diverse subsets of molecules, often with the objective of maximizing the number of hits in biological screening. While many methods exist in the area, few systematic comparisons using current descriptors in particular with the objective of assessing diversity in *bioactivity space* have been published, and this shortage is what the current study is aiming to address. In this work, 13 widely used molecular descriptors were compared, including fingerprint-based descriptors (ECFP4, FCFP4, MACCS keys), pharmacophore-based descriptors (TAT, TAD, TGT, TGD, GpiDAPH3), shape-based descriptors (rapid overlay of chemical structures (ROCS) and principal moments of inertia (PMI)), a connectivity-matrix-based descriptor (BCUT), physicochemical-property-based descriptors (prop2D), and a more recently introduced molecular descriptor type (namely, “Bayes Affinity Fingerprints”). We assessed both the similar behavior of the descriptors in assessing the diversity of chemical libraries, and their ability to select compounds from libraries that are diverse in *bioactivity space*, which is a property of much practical relevance in screening library design. This is particularly evident, given that many future targets to be screened are not known in advance, but that the library should still maximize the likelihood of containing bioactive matter also for future screening campaigns. Overall, our results showed that descriptors based on atom topology (i.e., fingerprint-based descriptors and pharmacophore-based descriptors) correlate well in rank-ordering compounds, both within and between descriptor types. On the other hand, shape-based descriptors such as ROCS and PMI showed weak correlation with the other descriptors utilized in this study, demonstrating significantly different behavior. We then applied eight of the molecular descriptors compared in this study to sample a diverse subset of sample compounds (4%) from an initial population of 2587 compounds, covering the 25 largest human activity classes from ChEMBL and measured the coverage of activity classes by the subsets. Here, it was found that ”Bayes Affinity Fingerprints” achieved an average coverage of 92% of activity classes. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes were represented in the selected compounds respectively, followed by BCUT, prop2D, MACCS, and PMI (in order of decreasing performance). In addition, we were able to show that there is no visible correlation between compound diversity in PMI space and in bioactivity space, despite frequent utilization of PMI plots to this end. To summarize, in this work, we assessed which descriptors select compounds with high coverage of bioactivity space, and can hence be used for diverse compound selection for biological screening. In cases where multiple descriptors are to be used for diversity selection, this work describes which descriptors behave complementarily, and can hence be used jointly to focus on different aspects of diversity in chemical space.



■ INTRODUCTION

Computational methods play a pivotal role in modern drug discovery, extending from de novo computer-aided drug-design methods[1,2] to tasks such as storing and analyzing large combinatorial chemical libraries.[3] The size of chemical space is difficult to quantify, but there is no doubt that it is very large in nature and, according to one estimate, there are ~10^{63} organic small molecules that could be formed of up to 30 heavy atoms.[1,4] Considering the fact that likely only a very small fraction of that space is therapeutically relevant, it is of high importance to develop novel computational approaches that will allow efficient exploration and effective selection of molecules that could be tested for bioactivity against proteins of interest.[5]

Received: August 8, 2013
Published: December 2, 2013

dx.doi.org/10.1021/ci400469u | *J. Chem. Inf. Model.* 2014, 54, 230−242

One concept in this direction that has gained importance in recent decades, in particular with the advent of modern combinatorial chemistry and high-throughput screening (HTS), is chemical diversity. Chemical diversity, a concept complementary to (but not just the mirror image of) chemical similarity,[6,7] is routinely utilized in library design and compound selection to quantitatively evaluate the presence of distinct structural features (however defined) present in chemical libraries.[8,9] To this end numerous computational methods are currently available which allow for the selection of structural diverse subsets that maximize the chemical space, i.e., selecting a set of compounds with the maximum degree of structural variation, while retaining a manageable number of molecules to be screened or tested for novel activities.[10]

Chemical diversity is by no means a uniquely defined concept, and it has been argued that it could only be measured by relevant external criteria and thus cannot be inherently "objective".[11] Nevertheless, it has developed into a concept of high practical relevance in the field of cheminformatics and, in particular, the design of screening libraries, as it allows one to quantify the similarity (or dissimilarity) of two or more chemical libraries and rationally select chemically diverse compounds from a much larger population of molecules. This approach is particularly suitable when knowledge about the chemical matter active on a protein target is limited (in this case, more-focused compound selection would often be performed), or when the aim is to design a general screening set that can be applied to multiple protein targets.[8]

An important aspect when measuring chemical diversity is the choice of molecular descriptors, which is used as input for distance or similarity measures to quantify this measure. An ideal molecular descriptor might, for example, show good correlation with human perception of chemical diversity, so that it resembles the human mind in decision-making processes. However, this is, in practice, difficult to realize for at least two reasons: first, because it is difficult to say with certainty how humans actually assess chemical structures that are displayed in front of them,[12] and, second, because there is a remarkable inconsistency in assessment, both between chemists, and also when displaying the same structures to a chemist repeatedly.[13] Hence, resembling human perception might actually not be a desirable goal to pursue in the end, and quantitative measures related to the problem at hand might be more suitable to measure the performance of diversity assessment methods. As described in detail later, the quantitative "external" measure that we decided to pursue in the current work was bioactivity space coverage, given that in many cases diversity selection of compounds aims at assembling a library of small molecules with increased chances of identifying hits against both current and future targets.

Recent studies have demonstrated that the use of different descriptors could generate significantly different results when selecting subsets of diverse molecules,[14−16] underlining that the choice of molecular descriptor clearly affects (or biases) the perception of chemical diversity present in a library. Furthermore, as was shown by Fergus et al., the diversity of a library is dependent on the number of library members, and very small libraries could give counterintuitive estimates of diversity and should be treated with caution.[17] Different diversity assessment methods can yield vastly different results,[18] depending on what type of chemical libraries they are used to analyze, the size of the libraries, as well as the source of the molecules, which can be compounds stemming from combinatorial chemistry, or natural products.[19]

Currently, various methods are routinely being used to assess chemical diversity, including fingerprint-based,[16] shape-based,[20] and pharmacophore-based methods.[21] Fingerprint-based methods compare small molecules in terms of the presence or absence of a set of substructural or fingerprint features (derived from molecular graph representations), hereby taking into account atom connectivity, and are widely used in virtual screening.[22] Alternatively, shape-based methods encode molecular conformational information, which can be internal distances or external molecular properties, which are then applied to compare molecules based on those properties. Examples of such shape-based methods are ROCS descriptors, which compare molecules based on their molecular shapes, by assessing atom-centered overlapping Gaussians and calculating the maximal intersection of the volume between molecules.[23] Furthermore, pharmacophore-based methods compare molecular similarity in terms of the presence or absence of pharmacophoric features (which may, in turn, often be represented as fingerprints).[21]

Despite the wide usage of all the above-mentioned methods, each descriptor focuses only on one aspect of the chemical information available. For example, shape-based methods are scaffold-independent, whereas pharmacophoric descriptors focus on pharmacophoric points and do not take into account the entire molecular surface, and structural keys encode only the presence or absence of predefined substructural features but not the connectivity among them. An alternative approach to quantify molecular diversity that has been explored recently (and also in this work) is based on "in silico" bioactivity profiles, which maps chemical structural space into a predicted bioactivity profile against a large number of protein targets.[24] Given that diversity in bioactivity space is often the main aim of diversity selection projects, basing the decision of how diverse compounds are on their bioactivity profiles might well be a purely empirical, but still rather suitable, decision to make.

Common to all of the above descriptors, once chemical structures have been encoded in a computer-accessible way, selection algorithms come into play that, in principle, can be based on any of the descriptors mentioned above.[10] The objective of these methods is to select subsets of compounds with maximum structural diversity from an initial large pool of compounds, while retaining the overall diversity of the initial population of molecules. Because of the large size of libraries, which can be in the order of 10^6 or larger in HTS campaigns, these are mainly heuristic methods, since exhaustive enumeration of all possible subsets would be computationally unfeasible and can be categorized as[25] (i) maximum dissimilarity-based algorithms,[26] (ii) clustering,[27] (iii) partitioning of cell-based approaches,[28] and (iv) optimization methods.[29,30] All methods provide approaches to cherry pick diverse sets of compounds from large libraries; however, they still depend on molecular descriptors to compare compounds and, therefore, are affected by the shortcomings and the behavior of the descriptor applied.

In this study, we will focus on the descriptor aspect of chemical diversity selection, given that no comparative study using bioactivity coverage as an objective function has been published yet, according to the best of the authors' knowledge, despite its importance for selecting diverse screening subsets. To this end, the behavior of 13 widely employed descriptor types was assessed, which fall into four main categories, namely, fingerprints, shape-based methods, pharmacophoric methods, and two-dimensional (2D) properties. In addition, we utilized "Bayes Affinity Fingerprints" as descriptors,[31,24] that represent molecular structures based on their in silico bioactivity profiles.

Table 1. Molecular Descriptors Used in This Study and Software Implementation

descriptor type	descriptor name	implemented in	description
fingerprint-based	ECFP4	MOE v2011.10[42]	atom type, extended connectivity fingerprint, maximum distance = 4
	FCFP4	MOE v2011.10[42]	functional-class-based, extended connectivity fingerprint, maximum distance = 4
	MACCS	MOE v2011.10[42]	166 predefined MDL keys (public set)
connectivity-matrix-based	BCUT	MOE v2011.10[42]	atomic charges, polarizabilities, H-bond donor and acceptor abilities, and H-bonding modes of intermolecular interaction
shape-based	rapid overlay of chemical structures (ROCS), combo Tanimoto (shape and electrostatic score)	OpenEye v3.1.2[48]	shape-based molecular similarity method; molecules are described by smooth Gaussian function and pharmacophore points
	PMI	MOE v2011.10[42]	normalized principal moment-of-inertia ratios
pharmacophore-based	GpiDAPH3	MOE v2011.10[42]	graph-based 3-point pharmacophore, eight atom types computed from three atom properties (in pi system, donor, acceptor)
	TGD	MOE v2011.10[42]	typed graph distances, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
	TAD	MOE v2011.10[42]	typed atom distances, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
	TGT	MOE v2011.10[42]	typed graph triangles, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
	TAT	MOE v2011.10[42]	typed atom triangles, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
bioactivity-based	Bayes affinity fingerprints	in-house-developed in silico bioactivity prediction model[51]	bioactivity model based on multicategory Bayes classifier trained on data from ChEMBL v. 14
physicochemical-property-based	prop2D	MOE v2011.10[42]	physicochemical properties (such as molecular weight, atom counts, partial charges, hydrophobicity etc.)

These descriptors were applied for diversity selection across different chemical libraries varying both in size and diversity, and were assessed with respect to their similarities and differences in rank-ordering compounds in diversity selection procedures by employing the Spearman's rank correlation coefficient. In addition, coverage of bioactivity space was measured to assess descriptor performance, in addition to differences in behavior. Hence, the objective of this study is to assess correlation among widely employed chemical descriptors across a large set of libraries, in order to obtain a better understanding of the situations in which these descriptors correlate and when they do not, as well as to evaluate their ability to cover large numbers of bioactivity classes in the selected subsets. This is of relevance for selecting compound subsets for biological screening, in particular, in cases where either different target families or orphan targets will be screened.

■ MATERIALS AND METHODS

Molecular Datasets. The behavior of the different diversity assessment and selection approaches was assessed on diverse sets of small molecule libraries, namely, compounds generated via Diversity-Oriented Synthesis (DOS) approaches,[32] metabolites from HMBD,[33] DrugBank,[34] PubChem,[35] and ChEMBL.[36]

FN, CEH, HEB, DRS, and Da Libraries. All of the libraries from this category stem from Diversity-Oriented Synthesis (DOS) approaches, and they were used since they claim to contain large chemical diversity by their very nature.[37] The FN, CEH, and HEB datasets consist of 45 nonpeptidic macrocyclic compounds in total and are part of a larger library of such compounds generated using a DOS approach.[38] The compounds all contain macrocyclic rings, a structural motif which is argued to be of value for targeting the extended binding interface associated with protein−protein interactions.[39,40] The DRS library consists of 28 compounds, which were not based on any general scaffold type. The Da library consists of 27 small molecules generated using a branching DOS strategy.[41,42] All of these DOS libraries were synthesized with the intent of providing hits against diverse biological targets, and, indeed, the screening of the Da and DRS libraries has identified compounds with antibacterial activity already.[41,42]

HMDB: Human Metabolome Database (HMDB). The Human Metabolome Database[33] is a comprehensive resource of small endogenous molecule metabolites found in the human body. For this study, HMDB version 2.5 was stored locally in SDF format, containing 8535 compounds in total. Random selection was applied, followed by a filtering criterion of MW < 900 Da. In total, 981 molecules were selected for the current study.

DrugBank. DrugBank[34] constitutes a comprehensive resource for drugs and drug target information. For this study, DrugBank version 3 was used and stored locally in SDF format. Compounds with a molecular weight of 900 Da or less were randomly selected, leading to a total of 1036 drugs and druglike molecules considered in this study.

PubChem. PubChem[35] is a large open repository for small molecules and biological properties of small molecules for public access, hosted by the U.S. National Institutes of Health (NIH). For this study, the PubChem FTP service was accessed and 10 random subsets of compounds from the full database were selected and downloaded locally in SDF file format. Consecutive random selection steps were applied to select molecules. In addition, the molecular weight was set not to exceed 950 Da, leading to a total of 947 compounds selected and used in this study. No diversity selection algorithm was applied prior to analyzing the results, to avoid introducing descriptor bias.

ChEMBL. ChEMBL[36] is a database containing binding, functional, and ADMET information for a large number of druglike bioactive molecules maintained by EMBL-EBI. ChEMBL version 14 was utilized for this study, downloaded, and installed on a local MySQL server. The 50 most-populated human protein targets were selected (see File SI_2 in the Supporting Information), based on the number of compounds annotated with K_i, IC_{50}, EC_{50} and K_d values equal to or better than 1 μM. These data consisted of various types of targets including enzymes (proteases, lyases, reductases, hydrolases, and kinases) representing 48% of the classes, membrane receptors (GPCRs and non-GPCRs) representing 44% of the classes and transcription factors and transporters each representing 4% of the classes. The target classes contained 1573 compound associations on average, varying from 1014 to 2971 data points (see Table S1 and Figure S2 in the Supporting Information). Subsequently, 2587 compounds were randomly selected from the 25 largest classes. This resulted in classes containing 103 data points on average, with the smallest class containing 12 data points and the largest class containing 190 data points (see Table S2 in the Supporting Information). This dataset was utilized to compare the performance of how molecular descriptors sample diverse subsets of compounds and achieve protein target coverage.

TIMBAL. TIMBAL[43] is a database containing small molecules that modulate protein−protein interactions. All compounds with annotated K_i, IC_{50}, EC_{50}, or K_d values of 10 μM or better were selected, which resulted in a total of 1995 unique compounds across 34 target classes after standardization (as described in the following subsection, "Library Preparation").

Library Preparation. Molecules were standardized using ChemAxon's Standardizer with the options Remove salts (keep largest fragment), Neutralize, Remove Explicit Hydrogens, Aromatize, Mesomerize, and Tautomerize.[44] Following standardization, molecules were loaded to Molecular Operating Environment 2011.10[45] (MOE) and three-dimensional (3D) molecular conformations were calculated using MOE,[45] applying the Rebuild 3D option, while retaining existing chirality (default options). The standardization protocol utilized in ChemAxon's Standardizer is provided in File SI_3 in the Supporting Information.

Molecular Descriptors. Twelve (12) widely employed structural molecular descriptors and one descriptor based on predicted bioactivity spectra, namely, the "Bayes Affinity Fingerprints", were utilized in this study for the representation of molecules. Molecular descriptors are listed in Table 1 and briefly described in the following.

(i) Fingerprint-based descriptors:

(1) *MACCS keys (MOE).*[46] 166 predefined substructural key sets of the public subset as implemented in MOE, which were originally designed for quicker database retrieval of compounds with certain predefined chemical functionalities.

(2) *ECFP4 and FCFP4 (MOE).*[47] Circular fingerprints as implemented in MOE, where E stands for atom type and F stands for functional class. Extended connectivity fingerprints are derived from variation of the Morgan algorithm,[48] and this descriptor type has been shown previously to capture much information relevant to the bioactivity of a compound.[49,50]

(ii) Pharmacophore-based descriptors:

(3) *GpiDAPH3 (MOE).*[51] Graph-based three-point pharmacophore employing any set of three possible atom types, namely, "in pi system", "donor", and "acceptor" atom.

(4) *TAD, TAT, TGD and TGT (MOE).* Typed atom distances (TAD), typed atom triangle (TAT), typed graph distances (TGD), and typed graph triangles (TGT). Six different atom types are possible: donor, acceptor, polar, anion, cation, and hydrophobic.

(iii) Shape-based descriptors:

(5) *ROCS (OpenEye).*[23,52,53] Molecular shapes are described by smooth Gaussian function and pharmacophoric points. "Combo Tanimoto" was used as a similarity function.

(6) *PMI (MOE).*[54] Three principal moments of inertia (PMI) derived from 3D structures as implemented in MOE.

(7) *BCUT (MOE).*[55] Four-dimensional (4D) BCUT_PEOE descriptors as implemented in MOE. BCUT descriptors are based on atomic charges, polarizabilities, H-bond donor, and acceptor abilities and H-bonding modes of intermolecular interaction.

(iv) 2D descriptors:

(8) *prop2D.* The first 10 principal components of all 2D physicochemical properties as implemented in MOE (v2011.10), containing properties such as molecular weight, atom counts, polar surface area, etc.

(v) Bayes Affinity Fingerprints:[56]

In silico predicted bioactivity spectra of small molecules generated using an in-house-developed bioactivity model based on the multicategory Naïve Bayesian classifier and bioactivity data extracted from ChEMBL. Compounds are initially described by circular molecular fingerprints and then are subjected to a target prediction model containing 134 450 bioactive compounds that cover 477 human protein targets. Protein targets are then ranked for each compound based on the likelihood of being active.

Distance Metrics. *Euclidean Distances.* Euclidean distances were employed as a distance function for descriptors that assume continuous values, namely, PMI, Bayes affinity fingerprints, ROCS, BCUT, and 2D physicochemical descriptors. Euclidean distances were calculated according to eq 1:[57]

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

where **p** and **q** are the descriptor vectors of two molecules containing ***n*** dimensions.

Hamming Distance. The Hamming distance was employed as a distance function for descriptors that take binary values (0 or 1), namely, circular, structural, and pharmacophoric fingerprint-based methods (ECFP4, FCFP4, MACCS, TGD, TGD, TAT, TAD, GpiDAPH3). The Hamming distance, which originates from information theory, is calculated for two equal n-length vectors as the number of positions at which the corresponding bits are different.[58] The use of the Hamming distance was preferred over the Tanimoto metric in this study, because the Tanimoto metric is a normalized metric taking values between the intervals [0,1], while the Hamming distance is not normalized.

For two vectors $\mathbf{p} = (p_1, p_2, ..., p_n)$ and $\mathbf{q} = (q_1, q_2, ..., q_n)$, the Hamming distance was calculated according to eq 2:

$$d(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\| = \sum_{i=1}^{n} |p_i - q_i| \qquad (2)$$

Spearman's Rank Correlation Coefficient. The Spearman's rank correlation coefficient (ρ) was employed as the correlation coefficient for assessing statistical dependence between rankings obtained among molecules as a result of applying different descriptors (each descriptor led to a unique ranking of molecules, depending on the properties encoded by the descriptor, which was then compared across descriptor pairs for each dataset used).[59] The Spearman's rank correlation coefficient can assume values in the interval [−1,1], where −1 or 1 indicates perfect positive or negative correlation, and values of zero indicate absence of any correlation between the two variables analyzed. The Spearman's rank correlation coefficient was calculated according to eq 3:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \qquad (3)$$

where x_i and y_i were the resulted rankings obtained from the distance matrices for each molecular descriptor used.

Principal Component Analysis (PCA). In order to obtain visualization of the arrangement of descriptors in multidimensional space, PCA was performed using the R statistical environment (version 2.15.2)[60] and the three first-principal components were visualized using Vortex.[61]

Assessment of Bioactivity Coverage. The performance of each compound diversity selection method was assessed based on the coverage of activity classes achieved by sampling a 1% diverse subset from the initial population of the compounds and by counting the number of activity classes being retrieved. The experiments were repeated three times, and the average number of protein targets presented in each sampled set was assessed. The descriptor-based diversity selection was performed utilizing MOE's function "Calculate Diverse Subset" with the option Output limit set to 100. In the case of Bayes Affinity Fingerprints, the approach as described by Nguyen et al.[24] was applied. Not all molecular descriptors were utilized in this step, because of high computational cost (e.g., calculating a similarity matrix among compounds in the PubChem dataset based on ROCS descriptors required approximately one week on an Intel Core 2 Duo desktop computer with 8 GB of RAM). Instead, eight representative descriptors were used, as shown in Table 2.

RESULTS AND DISCUSSION

Results obtained on the overall correlation of descriptors used in diversity assessment and averaged across all libraries are visualized in Figure 1. (The full numerical matrix and the matrices obtained for each library separately are provided in File SI_1 in the Supporting Information.) The PCA visualization performed on the matrix obtained from Spearman's rank correlation among molecular descriptors is presented in Figure 2, where the first three principal components explained 74% of the accumulative variance, as shown in Figure 3, whereas ~90% of the variance in descriptor space is explained by the first five principal components.

Table 2. Activity Classes Covered by Sampling a Diverse Subset of 4% (100 Compounds) from an Initial Set of 2587 Compounds Extracted from the 25 Largest ChEMBL Human Activity Classes[a]

		Number of Activity Classes Sampled, Out of 25 Classes			
descriptor	% activity classes sampled (subset size of 25 classes), averaged over three attempts (relative ranking)	1st attempt	2nd attempt	3rd attempt	average
Bayes Affinity Fingerprints ("Cut-off 30")	92% (1st)	22	24	23	23
ECFP4	91% (2nd)	22	23	23	22.7
GpiDAPH3	84% (3rd)	21	22	21	21
TGT	84% (4th)	20	21	22	21
random sampling	84% (5th)	21	21	19	21
BCUT	83% (6th)	21	20	21	20.7
prop2D	80% (7th)	21	21	19	20
MACCS	80% (8th)	19	20	20	20
PMI	75% (9th)	19	18	19	18.7

[a] Using Bayes Affinity Fingerprints, considering only predicted protein targets with a Bayes Score above 30, 92% of the initial activity classes were sampled, hence outperforming all other descriptors marginally. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes, respectively, were represented in the selected compounds. Random sampling retrieved ~84% of activity classes (averaged over three attempts), showing similar performance on this dataset with GpiDAPH3, TGT, and BCUT, while outperforming molecular descriptors such as prop2D (80%), MACCS (80%), and PMI (75%).

It can be seen in Figure 1 that the pharmacophore-based descriptors TAT, TAD, TGD, and TGT show a strong correlation with each other, with all Spearman's rank correlation coefficients observed being 0.74 (among TGT and TAT) or higher, indicating very similar behavior of those descriptors in diversity selection procedures. Accordingly, they were found to cluster together in Figure 2. The fifth pharmacophore-based descriptor utilized here, namely, GpiDAPH3, showed lower correlation with the previously mentioned descriptors, with Spearman's rank correlation coefficients observed with the descriptors TAD, TGD, and TGT of 0.46, 0.47, and 0.52, respectively, with the exception being TAT, which had a correlation coefficient with GpiDAPH3 of 0.65. This result is not surprising, considering that both of the latter descriptors capture pharmacophoric triangles (as opposed to TAD and TGD, which use pharmacophore points), with GpiDAPH3 taking into account three-point pharmacophore fingerprints calculated from molecular graphs, and TAT taking into account atom-typed triangles calculated from the 3D conformation of a molecule. In addition, the pharmacophore-based GpiDAPH3 descriptor showed intermediate correlations between 0.54 and 0.67 with the fingerprint-based descriptors ECFP4 and FCFP4 and MACCS structural keys. The rest of the descriptors showed correlations of 0.51 or lower with the pharmacophore-based descriptors, indicating low similarity in behavior.

The fingerprint-based descriptors ECFP4 and FCFP4 showed a Spearman's rank correlation coefficient of 0.89 with each other, indicating stronger correlation with each other than the pharmacophore-derived descriptors, as is also clearly visible in

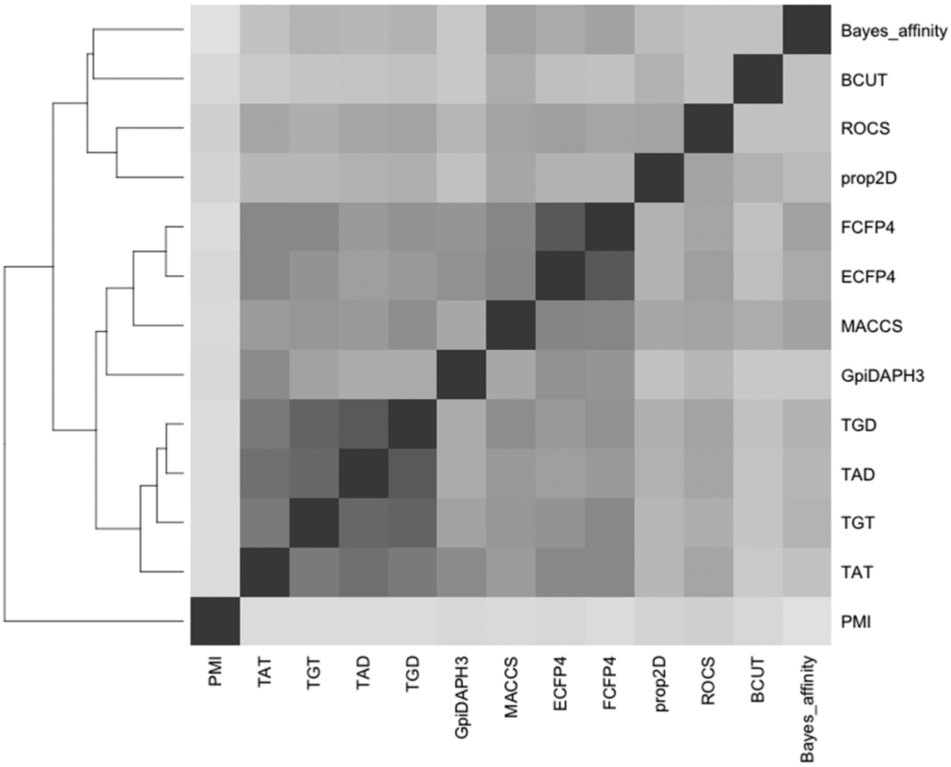


Figure 1. Spearman's rank correlation coefficients obtained from descriptors used in this study averaged over all included libraries. Darker colors indicate higher Spearman's rank correlation, whereas lighter colors indicate lower correlation. The descriptor PMI demonstrated the least correlation with any other descriptor, indicating that the behavior of this type of descriptor is significantly different from any others included here. Overall, it can be seen that pharmacophore-based descriptors such as TAT, TGT, TAD, and TGD show some correlation with fingerprint-based descriptors (ECFP4, FCFP4, and MACCS, with MACCS keys also showing some correlation with other descriptor types).

Figure 2. This high correlation can be explained by the fact that both ECFP4 and FCFP4 are both derived from radial atom connectivity, and their similar behavior has been observed before in the context of similarity assessment.[62] In addition, MACCS structural keys have been shown to be correlated with ECFP4 and FCFP4 with a Spearman's rank correlation coefficient of 0.69 and 0.68, respectively, which can be explained by the local nature of both descriptor types. Rather surprisingly, the descriptors ECFP4 and GpiDAPH3 demonstrated a Spearman's rank correlation of 0.62, indicating that these two types of fingerprints demonstrate relatively similar behavior, which is more difficult to rationalize, given that the GpiDAPH3 descriptor takes into account graph-based pharmacophoric representations, as opposed to 2D structural features or atom type or counts. On the other hand, this is still lower than the correlation between TAT and GpiDAPH3 descriptors, so the relative ordering of descriptor pairs remains consistent with our initial expectations. The results discussed here indicate that molecular descriptors derived from atom topology or graph-based pharmacophoric representations tend to behave rather similarly overall, as shown in detail in Figures 1 and 2.

When now visiting descriptors of very different nature, we can see that overall PMI shows Spearman's rank correlations of 0.22 or lower with other type of descriptors, indicating very different behavior, whereas for ROCS, the Spearman's rank correlation coefficient with all other descriptors was also 0.55 or lower. ROCS and PMI also demonstrated significantly different behavior from each other, as measured by the Spearman's rank correlation coefficient, which was found to be only 0.22 (see Figures 1 and 2), even though they both capture molecular shapes. This can be attributed to the fact that while PMI project molecular shapes (i.e., how similar molecules are to archetypal shapes such as spheres, disks, and rods) and are size-independent, ROCS considers not only molecular shape by comparing molecular shapes by overlapping Gaussian volumes, but also chemical similarity/atom types, thus being size-dependent. Similar differences exist for other descriptor types, where, e.g., 2D physicochemical descriptors, such as atom counts, molecular weight, and polar surface area, tend to increase as the size of the molecule increases and, hence, are often inherently size-dependent, while molecular fingerprints encode only the presence or absence of chemical substructures, and accordingly pay less attention to size.

As a nonstructural descriptor considered here, Bayes Affinity Fingerprints describe molecules by their predicted bioactivity spectra, and they showed low correlation with the rest of the descriptors utilized in this study: Spearman's rank correlations with all other descriptors did not exceed 0.53 in the case of FCFP4 and MACCS keys, and it was as low as 0.28 with GpiDAPH3 pharmacophores.

In order to further illustrate the extent to which different descriptors assess different aspects of chemical diversity, an example of comparison between two molecules in the Da library is given in Figure 4, where the ranks are shown for each descriptor employed (lower ranks mean higher similarity, whereas higher ranks mean lower similarity). It can be observed that, overall, there are significant differences among ranking positions obtained (ranks range from 3 to 27, given the library size of 27). MACCS keys and the pharmacophore-based descriptors TAT, TGT, TAD, and TGD showed very similar

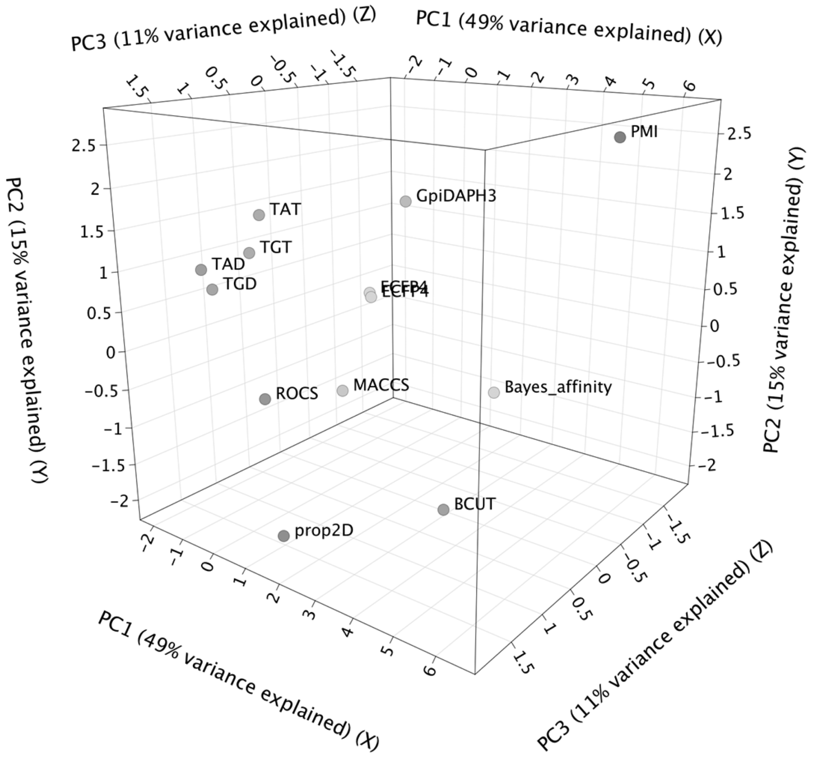
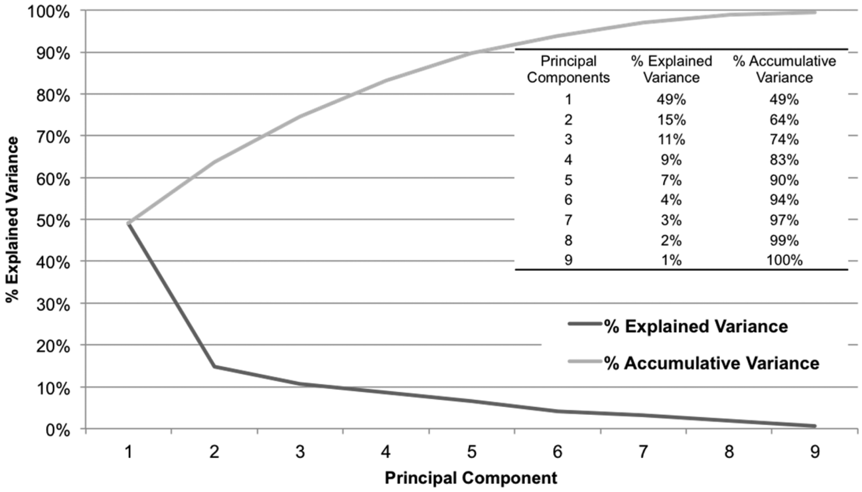


Figure 2. Similarity of molecular descriptors in perceiving the diversity of chemical libraries visualized in PCA space. The Spearman's rank correlations between the descriptors were subjected to principal component analysis (PCA): ~74% of the accumulative variance is captured in the first three principal components shown here; therefore, descriptors located close to each other show a stronger correlation based on the Spearman's rank-correlation coefficient. Overall, it can be observed that two clusters of descriptors emerged that show high correlation within their groups: first, the fingerprint-based descriptors ECFP4 and FCFP4, and, second, the pharmacophore-based descriptors TAT, TGT, TAD, and TGD. In addition, MACCS keys and GpiDAPH3 descriptors did not show any significant correlation with other type of descriptors, nor did the descriptors BCUT, PMI, and prop2D.



Principal Components	% Explained Variance	% Accumulative Variance
1	49%	49%
2	15%	64%
3	11%	74%
4	9%	83%
5	7%	90%
6	4%	94%
7	3%	97%
8	2%	99%
9	1%	100%

Figure 3. Scree plot of the PCA of the 13-dimensional descriptor space (compared to the 37-dimensional descriptor space in an earlier study by Bender et al.[62]). The first three principal components capture 74% of the total variance, while five principal components are required to capture 90% of the total variance. In an earlier study by Bender et al.,[62] the first three components captured only 50% of the total variance, while 10 principal components were required to capture 80% of this measure. It cannot be definitely concluded whether more variance is captured in fewer dimensions, because of more similar behavior of descriptors or simply their lower number; however, it can be seen that a small number of dimensions already is sufficient to capture similarities and dissimilarities of the molecular descriptors when it comes to molecular diversity assessment, as employed in this study. When compared to the study by Bender et al.,[62] the increase in the percentage of accumulative explained variance for the first five principal components is very similar (41% here, compared to 38%), indicating that the first part of the curve is similar in both cases, albeit with a different percentage of variance, explained by the first principal component (49% here, compared to only 24% in the study by Bender et al.[62]).

results (they all ranked the molecules within the library at similar positions), and the fingerprint-based descriptors ECFP4 and FCFP4 also showed very similar results, compared among themselves (they both assigned a moderate similarity ranking).

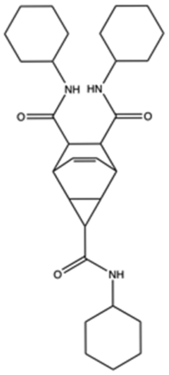
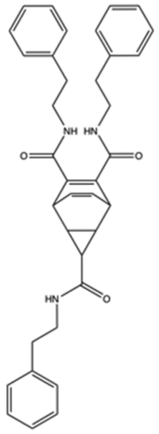
Query molecule	Target molecule	Descriptor	Ranking (%)
		2D properties	93
		BCUT	89
		Bayes affinity fingerprints	41
		ECFP4	26
		FCFP4	26
		MACCS	11
		GpiDAPH3	100
		PMI	26
		ROCS	100
		TAD	19
		TAT	11
		TGD	11
		TGT	11

Figure 4. Two molecules from the Da library that showed significant differences in ranking positions obtained from different descriptors used. These molecules were selected from the Da library, which was shown to contain multiple molecules with antibacterial properties.[41,42] Pharmacophore-based and atom environments fingerprint descriptors showed very similar results, most likely because they encode only the presence or absence of chemical features, but do not take into account the size and/or molecular surface of the molecules, with the exception of GpiDAPH3. On the other hand, shape-based and 2D descriptors perceive the two molecules as significantly different, most likely because they take into consideration the entire molecular shape, hereby perceiving larger molecules as different from smaller ones (even if they consist of similar substructures). Hence, two chemical compounds subjectively perceived to be chemically similar by medicinal chemists could be considered as very similar, moderately similar, or even very different according to the descriptors used to assess chemical diversity.

However, the descriptors GpiDAPH3, ROCS, prop2D all perceived the molecules considered to be highly *dissimilar*. One possible explanation for the results observed above is that the descriptor MACCS was originally designed for indexing molecules and contains a small key set comprised of only 166 bits, which is the public subset (corresponding to predefined chemical functionalities), and it therefore appeared to be unable to distinguish well between relatively more similar molecules (much the same as the pharmacophore-based descriptors TAT, TGT, TAD, and TGD). However, the fingerprint-based descriptors ECFP4 and FCFP4 do take into account fingerprint features that are present in each molecule (independent of any predefined keyset), and they are hence able to also identify more subtle dissimilarities between those overall similar molecules. On the other hand, the descriptors prop2D and ROCS take into account substructural feature counts, molecular shape, and atom connectivity, and hence aggregate differences among the compounds, resulting in a perceived significance, with respect to physicochemical properties and shape of the structures. Hence, our findings illustrate that two molecules subjectively perceived to be chemically similar by visual inspection could be considered as highly similar, moderately similar, or even extremely different, according to the descriptors used to assess chemical diversity.

A comparison of descriptors on multiple datasets revealed that, although the Spearman's rank correlations between descriptors for larger datasets (such as PubChem, see Figure 5) resemble the correlations between descriptors for the compounds averaged across all libraries, analyses for particular datasets do not always show the same trend. It can be seen that the descriptors correlate better for the DRS dataset (see Figure 6), where the average correlation for the DRS dataset across all descriptors was 0.53, whereas, for the PubChem dataset, this was only 0.31 (see Figure 5). The Bayes Affinity Fingerprints show more similarity to fingerprint-based descriptors (which could be explained by the fact that this descriptor type was generated by a model trained on fingerprint-based descriptors) and shape-based descriptors, whereas PMI shows very low correlation with all other descriptors. Some descriptors show much higher correlations with other descriptors in the DRS dataset than either the PubChem or the overall dataset. For example, BCUT correlates very poorly with other descriptors (with an average correlation of 0.19) for the PubChem dataset, whereas for the DRS dataset, it shows a higher correlation of 0.49. Similarly, ROCS correlates poorly with other descriptors for the PubChem dataset (0.28), whereas for the DRS dataset, the average correlation was twice as high (0.57). These findings illustrate that the behavior of descriptors is highly dependent on the dataset analyzed and therefore, size and chemical composition of datasets should be taken into account when interpreting chemical diversity.

In order to correlate our findings with previous related studies, we compared our findings with the main results in the study previously reported by Bender et al.,[15] where PCA of the molecular descriptor space was performed, with respect to the

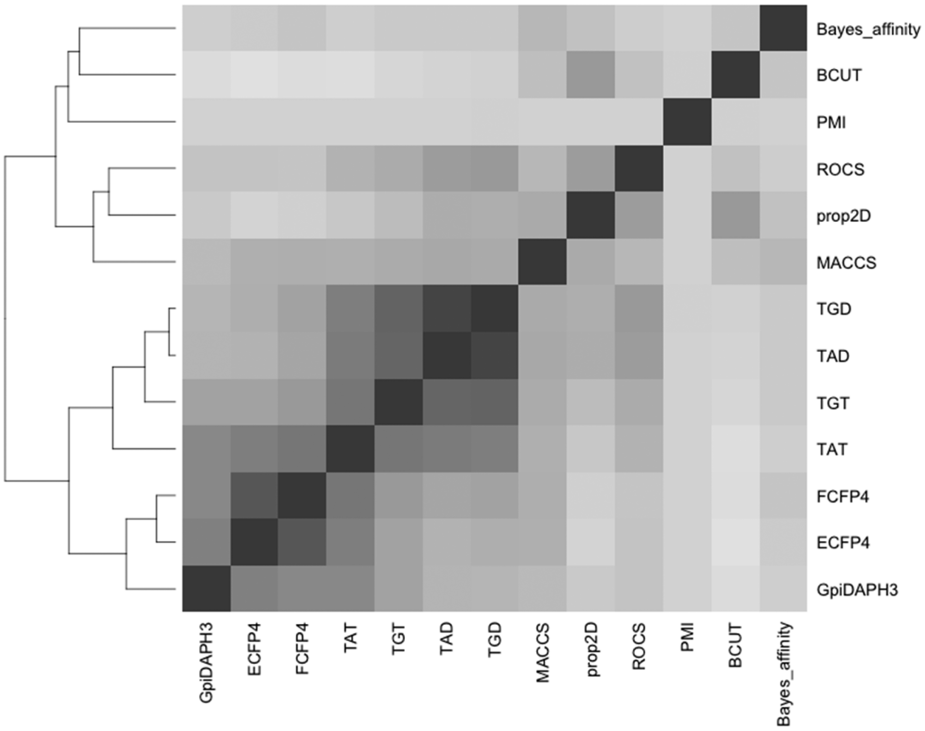


Figure 5. Spearman's rank correlation coefficient based on the overall averaged results over the PubChem library. Darker colors show higher correlation, whereas lighter colors show lower correlation among molecular descriptors. Overall, it can be seen that the Spearman's rank correlation pattern for the PubChem library is very similar to the overall Spearman's rank correlation pattern averaged over all libraries used in this study (recall Figure 1), with fingerprint-based and pharmacophore-based descriptors showing the most correlation among each other and within their respective group.

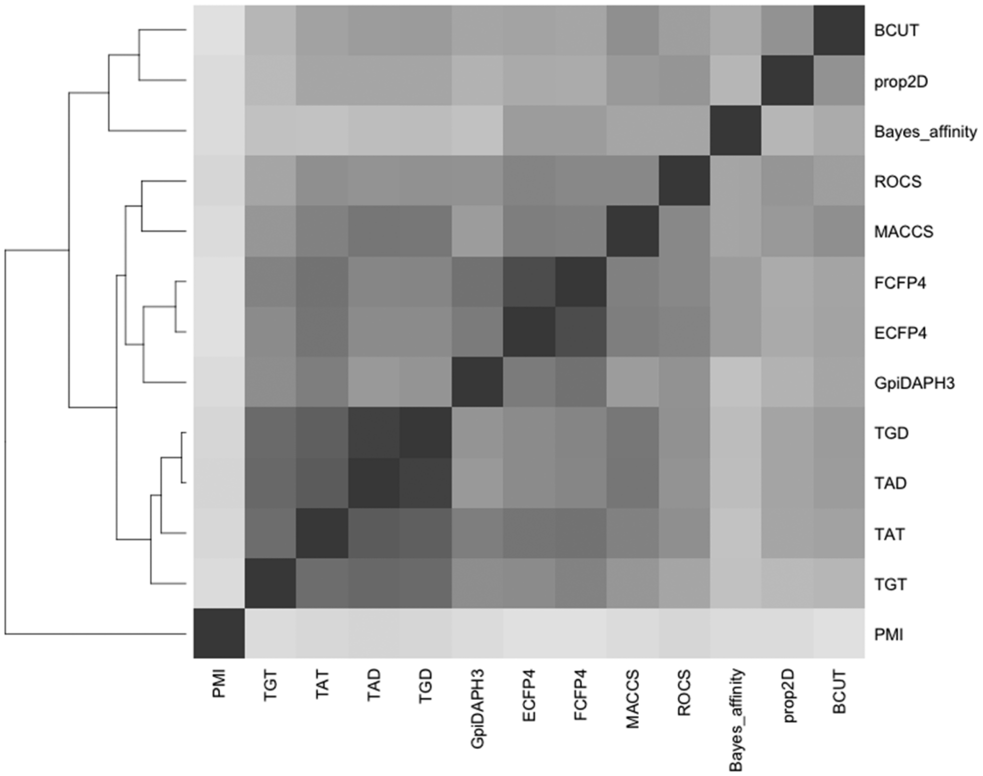


Figure 6. Spearman's rank correlation coefficient based on the overall averaged results over the DRS library. Darker colors show higher correlation, whereas lighter colors show lower correlation among descriptors. It can be observed that the descriptors correlate better for the DRS dataset than for the much larger PubChem dataset (see Figure 5): the average correlation for the DRS dataset across all descriptors was 0.53, whereas, for the PubChem dataset, this was only 0.31.

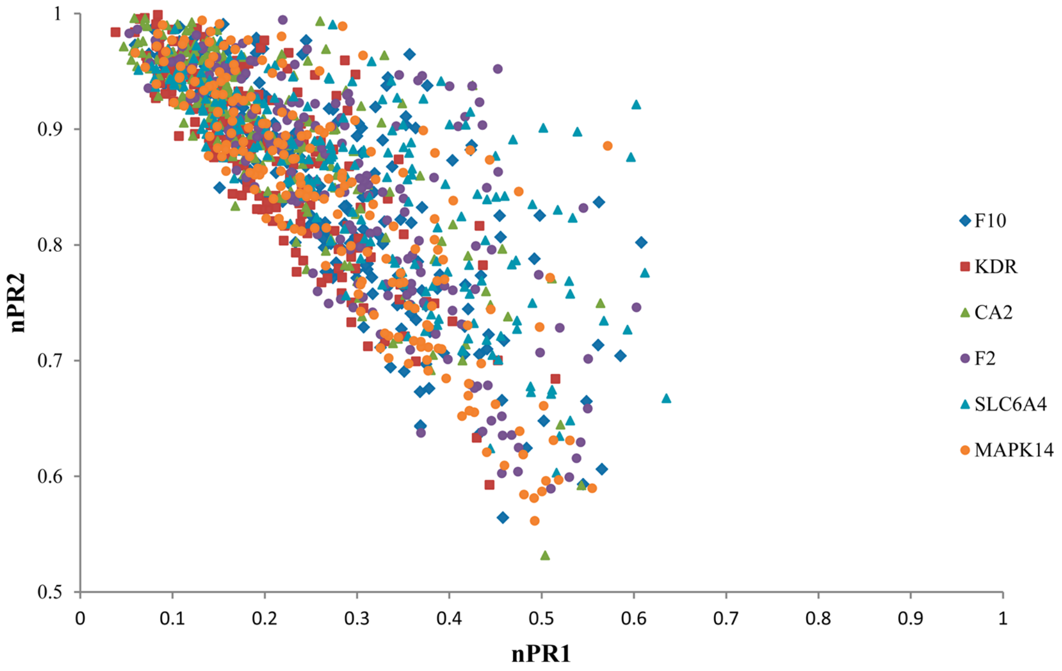


Figure 7. Normalized PMI ratios (nPR1 and nPR2) plot of 6 out of 50 ChEMBL activity classes utilized; namely, the coagulation factor X (F10), vascular endothelial growth factor receptor 2 (KDR), carbonic anhydrase 2 (CA2), prothrombin (F2), sodium-dependent serotonin transporter (SLC6A4), and mitogen-activated protein kinase 14 (MAPK14). It can be seen that molecular shape diversity of chemical libraries, measured by PMI, does *not* correlate with or indicate diversity of coverage in bioactivity space, as compounds binding to different protein families occupy similar space, and, in return, compounds from different areas of PMI space are bioactive against the same protein. Hence, the authors would recommend that, although PMI analyses give an insight into the shape properties of a chemical library, they might not be the most suitable tool to assess compound diversity in bioactivity space.

similarity of molecules, with the objective being to understand which descriptors contain orthogonal information and which descriptors are correlated with each other. Many agreements and disagreements were observed regarding the correlation among the descriptors used between our PCA (Figure 2) and the PCA conducted by Bender et al.[15] First, fingerprint-based descriptors such as ECFP and FCFP cluster together in both PCA plots, as do the 3D pharmacophore-based descriptors TAT and TGT. However, the pharmacophore-based descriptor GpiDAPH3 is positioned away from other pharmacophore-based descriptors, but closer to the fingerprint-based descriptors in the current study, whereas, in the previous analysis by Bender et al.,[15] it has been located near other pharmacophore-based descriptors. In addition, MACCS keys (referred to as MDL in the study by Bender et al.[15]) are correlated with fingerprint-based descriptors in our study; however, this is not the case in the study by Bender et al.,[15] where MACCS keys are more correlated to pharmacophore-based descriptors instead. Differences observed in this study compared to the previous study reported can be attributed to different objectives, because, in this study, the objective was to evaluate the correlation of molecular descriptors based on rankings obtained from calculated distances among compounds present in chemical libraries for diversity assessment, instead comparing the performance of molecular descriptors in retrieving active compounds for virtual screening assessment.

In order to assess not only similarities and dissimilarities in the behavior of different descriptors when applied to diversity selection but also their *performance, with respect to a relevant measure*, eight out the 13 descriptors were used for diversity selection and their performance was assessed by the coverage of protein targets in bioactivity space. (Some methods could not be used due to computational demands, given that the full compound similarity matrix needed to be computed for diversity selection using the methods employed here.) The performance of each method was assessed based on the number of activity classes covered by sampling a diverse subset of 4% (100 compounds) from an initial set of 2587 randomly selected compounds covering the 25 largest ChEMBL activity classes of human protein targets (target classes shown in Table S2 in the Supporting Information). These 25 activity classes contain 103 data points on average and vary in size from 12 data points to 190 data points, indicating that this is an unbalanced dataset.

Results of this analysis are shown in Table 2. Bayes Affinity Fingerprints, considering only predicted protein targets with a Bayes score above 30 ("Cut-off 30"), sampled an average of 92% of bioactivity classes, hence outperforming all other descriptors marginally. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes, respectively, were represented in the selected compounds. Random sampling retrieved ~84% of activity classes (averaged over three attempts), showing similar performance on this dataset to the descriptors GpiDAPH3, TGT, and BCUT, while outperforming molecular descriptors such as prop2D (80%), MACCS (80%), and PMI (75%). Despite the seemingly high performance of random sampling, it should be noted that random selection would only yield the best results (compared to any other method based on molecular descriptors) in the case where all classes are of equal size by picking up the most diverse subsets in bioactivity space. This could lead one to the misconception that random selection is a better option than the currently available methods used for diversity selection. However, in more realistic situations, such as here, some biological targets are more promiscuous than others, and hence they can accommodate a more diverse set of compounds in their binding sites than others (therefore leading to "larger activity classes" against those more promiscuous proteins). In such cases, random selection would struggle to pick compounds from small classes and, thus, does not seem to be the most suitable approach to be applied for diversity selection. Diversity selection methods based on molecular descriptors appear to be less successful in retrieving bioactive compounds against a broad range of protein targets, since no prior knowledge of which bits in the fingerprint matter (and lead to bioactivity differences) is considered by these methods. Instead, Bayes Affinity Fingerprints, which are trained on active compounds covering a large part of chemogenomic space, subselect compounds by taking into account known bioactive chemistry (and modifications leading to bioactivity changes), and thus appear to be able to achieve better protein target coverage. Our results are consistent with our previously reported study by Nguyen et al.,[24] where diversity selection based on bioactivity spectra fingerprints outperformed commonly employed circular fingerprint-based methods by up to 10%, when sampling bioactive compounds.

Finally, we attempted to assess whether PMI plots, which have frequently been used to assess the diversity of, e.g., DOS libraries, can be used to this end, when also paying attention to bioactivity coverage. This analysis is presented in Figure 7, as a PMI ratios plot (nPR1 and nPR2) for 6 out of 50 ChEMBL activity classes. It can be seen that molecular shape diversity of chemical libraries, as measured by PMI, does not correlate or indicate a diversity of coverage in bioactivity space, because compounds binding to different protein families occupy similar space, and compounds from different areas of PMI space are bioactive against the same protein. However, one could argue that the targets from ChEMBL represent a biased part of biological space (the space where it has been easy to identify bioactive molecules with traditional medicinal chemistry efforts). Therefore, a similar analysis on 1995 compounds with annotated K_i, IC_{50}, EC_{50}, or K_d values of 10 μM or better from the TIMBAL database was performed,[43] which involves a database of compounds with protein−protein inhibitory properties (see Figure S1 in the Supporting Information). It also can be seen that, in this case, molecular shape diversity does not correlate or indicate diversity of coverage in bioactivity space, because the compounds have a similar distribution to that in Figure 7. Hence, in the authors' opinion, PMI analyses, while giving insight into the shape properties of a chemical library, should rather not be used to assess diversity in bioactivity space. One might find this assumption tempting to make; however, Figure 7 illustrates the rather small correlation between diversity in PMI space and diversity in bioactivity space; moreover, from this illustration, it is apparent that diversity in one space has little predictive value for diversity in the other. In the opinion of the authors, diversity in bioactivity space—even if only computationally established—would be a practically more relevant measure for assessing the biologically relevant diversity of small molecule libraries.

■ CONCLUSION

This study aimed at assessing both the correlations between molecular descriptors in the context of diversity assessment, as well as their performance, with respect to covering bioactivity space. It was found that descriptors derived from atom topology (i.e., pharmacophore-based descriptors, such as TAT, TAD, TGD, and TGT) and fingerprint-based descriptors (such as ECFP4 and FCFP4) generally showed strong correlation within each group (all with Spearman's rank correlations of 0.74 or higher) and between both groups (all with Spearman's rank correlations of 0.59 or higher). On the other hand, shape-based descriptors such as rapid overlay of chemical structures (ROCS) and principal moments of inertia (PMI) demonstrated behaviors that were significantly different from each other, as measured by the Spearman's rank correlation coefficient, which was found to be only 0.22. Moreover, it was observed that descriptors correlate differently, depending on the dataset used. For example, the average correlation of descriptors for the DRS dataset, encompassing 28 dissimilar compounds (with multiple scaffolds being present in the dataset), is 0.53, whereas for the much larger and more diverse (on an absolute scale) PubChem dataset, the average correlation reached only 0.31. Hence, overall our results indicate that molecular descriptors differ in the way they assess chemical diversity, depending on the diversity and size of the datasets used, and selecting the appropriate descriptor is a nontrivial task, which must take into account both of these aspects. Moreover, the shape-based descriptor PMI showed no correlation with any other type of descriptor utilized in this study, demonstrating very different behavior from all other descriptors employed here.

Given that diversity in bioactivity space is often the primary objective of compound diversity selection procedures, we furthermore benchmarked the descriptors employed here, with respect to covering 25 bioactivity classes upon selecting 4% of 2587 compounds from a subset of ChEMBL. Here, it was found that the Bayes Affinity Fingerprints showed the best performance by covering 92% of bioactivity classes, hence outperforming all other descriptors marginally. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes, respectively, were represented in the selected compounds, followed by BCUT, prop2D, MACCS, and PMI (in order of decreasing performance). Random sampling retrieved ∼84% of activity classes (averaged over three attempts), showing similar performance of this dataset to the descriptors GpiDAPH3, TGT, and BCUT, while outperforming molecular descriptors such as prop2D (80%), MACCS (80%), and PMI (75%). In addition, we were able to show that there is no visible correlation between compound diversity in PMI space and in bioactivity space, despite frequent utilization of PMI plots to this end.

Overall, this study assessed which descriptors are able to select compounds with high coverage of bioactivity space, and which can hence be used for diverse compound selection for biological screening. It also gives guidelines as to which descriptors behave rather collinear and which descriptors behave more orthogonal in diversity selection tasks. We propose that a combination of complementary descriptors such as Bayes Affinity Fingerprints, PMI, and prop2D be used to computationally select a diverse set of compounds for screening purposes. Such a computational "filter" might also add value to current endeavors of assembling screening libraries against diverse targets, such as the European Lead Factory,[63] as well as library enhancement initiatives that are taking place continuously in pharmaceutical companies.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Matrices of Spearman's rank correlations as well as heatmap visualizations for each library (file SI_1.zip). Dataset covering 50 activity classes from ChEMBL is provided with UniProt identifier as activity class identifiers (file SI_2). Protocol utilized for molecule standardization for ChemAxon's Standardizer (file SI_3.xml). Principal moments of inertia (PMI) plot of the chemical library TIMBAL (Figure S1). Distribution of the 50 activity ChEMBL classes used in the study (Figure S2). Summary of the 50 largest activity classes used in the study (Table S1). Unbalanced dataset assembled for comparison of molecular descriptors in sampling diverse datasets in bioactivity space (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +44 (1223) 762 983. E-mail: ab454@cam.ac.uk.

Author Contributions

$^\nabla$Contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.K., D.M., A.B., and R.C.G. thank Unilever for funding. S.P. and A.P.I.J. thank The Netherlands Organisation for Scientific Research (NWO, Grant No. NWO-017.009.065) and the Prins Bernhard Cultuurfonds for funding. D.R.S. and W.R.J.D.G. were funded by grants from the European Research Council, Engineering and Physical Sciences Research Council, Biotechnology and Biological Sciences Research Council, Medical Research Council, Frances and Augustus Newman Foundation, and Wellcome Trust. The authors thank Dotmatics for providing the Vortex software for visualizations in this manuscript.

REFERENCES

(1) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16* (1), 3−50.

(2) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813−1818.

(3) Huggins, D. J.; Venkitaraman, A. R.; Spring, D. R. Rational methods for the selection of diverse screening compounds. *ACS Chem. Biol.* **2011**, *6* (3), 208−217.

(4) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432* (7019), 824−828.

(5) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432* (7019), 855−861.

(6) Maggiora, G. M.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(7) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204−3218.

(8) Perez, J. J. Managing molecular diversity. *Chem. Soc. Rev.* **2005**, *34* (2), 143−152.

(9) Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J.; Selzer, P.; Glick, M. Biodiversity of small molecules—A new perspective in screening set selection. *Drug Discovery Today* **2013**, *18* (13−14), 674−680.

(10) Willett, P. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol.* **1999**, *6* (3−4), 447−457.

(11) Roth, H. J. There is no such thing as "diversity"! *Curr. Opin. Chem. Biol.* **2005**, *9* (3), 293−295.

(12) Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery. *PloS One* **2012**, *7* (11), e48476.

(13) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* **2004**, *47* (20), 4891−4896.

(14) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39* (16), 3049−3059.

(15) Bender, A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discovery* **2010**, *5* (12), 1141−1151.

(16) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29* (2), 157−170.

(17) Fergus, S.; Bender, A.; Spring, D. R. Assessment of structural diversity in combinatorial synthesis. *Curr. Opin. Chem. Biol.* **2005**, *9* (3), 304−309.

(18) Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14* (3), 325−330.

(19) Clemons, P. A.; Wilson, J. A.; Dancik, V.; Muller, S.; Carrinski, H. A.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (17), 6817−6822.

(20) Naylor, E.; Arredouani, A.; Vasudevan, S. R.; Lewis, A. M.; Parkesh, R.; Mizote, A.; Rosen, D.; Thomas, J. M.; Izumi, M.; Ganesan, A.; Galione, A.; Churchill, G. C. Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol.* **2009**, *5* (4), 220−226.

(21) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (3), 569−574.

(22) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23−24), 1046−1053.

(23) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17* (14), 1653−1666.

(24) Nguyen, H. P.; Koutsoukas, A.; Mohd Fauzi, F.; Drakakis, G.; Maciejewski, M.; Glen, R. C.; Bender, A. Diversity selection of compounds based on "Protein Affinity Fingerprints" improves sampling of bioactive chemical space. *Chem. Biol. Drug. Des.* **2013**, *82* (3), 252−266.

(25) Gillet, V. J. Diversity selection algorithms. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (4), 580−589.

(26) Lajiness, M. S. Molecular similarity-based methods for selecting compounds for screening. In *Computational Chemical Graph Theory*; Nova Science Publishers, Inc.: Commack, NY, USA, 1990; pp 299−316.

(27) Lipkowitz, K. B.; Boyd, D. B. Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*, Vol. *18*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2003.

(28) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity measures for rational set selection and analysis of combinatorial libraries: The Diverse Property-Derived (DPD) approach. *J. Chem. Inf. Comput. Sci.* **1997**, 37 (3), 599−614.

(29) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2* (1−2), 64−74.

(30) Waldman, M.; Li, H.; Hassan, M. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graphics Modell.* **2000**, *18* (4−5), 412−426, 533−536.

(31) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept? *J. Chem. Inf. Model.* **2006**, *46* (6), 2445−2456.

(32) Galloway, W. R.; Isidro-Llobet, A.; Spring, D. R. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun.* **2010**, *1*, 80.

(33) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; NeveuV.; Greiner, R.; Scalbert, A.HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801−D807 (Database Issue).

(34) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.***2008**, *36*, D901−D906 (Database Issue).

(35) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633 (Web Server Issue).

(36) Gaulton, A.; Bellis, L. J.; BentoA. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107 (Database Issue).

(37) Spandl, R. J.; Bender, A.; Spring, D. R. Diversity-oriented synthesis; a spectrum of approaches and results. *Org. Biomol. Chem.* **2008**, *6* (7), 1149−1158.

(38) Beckmann, H. S. G.; Nie, F.; Hagerman, C. E.; Johansson, H.; Tan, Y. S.; Wilcke, D.; Spring, D. R. A New Strategy for the Diversity-Oriented Synthesis of Macrocyclic Scaffolds using Multi-Dimensional Coupling. *Nat. Chem.* **2013**, *5*, 861−867.

(39) Mullard, A. Protein-protein interaction inhibitors get into the groove. *Nat. Rev. Drug Discovery* **2012**, *11* (3), 173−175.

(40) Wolfson, W. Grabbing for the ring: macrocycles tweak the conventions of drug making. *Chem. Biol.* **2012**, *19* (11), 1356−1357.

(41) Wyatt, E. E.; aaFergus, S.; Galloway, W. R.; Bender, A.; Fox, D. J.; Plowright, A. T.; Jessiman, A. S.; Welch, M.; Spring, D. R. Skeletal diversity construction via a branching synthetic strategy. *Chem. Commun. (Cambridge, U. K.)* **2006**, *31*, 3296−3298.

(42) Wyatt, E. E.; Galloway, W. R.; Thomas, G. L.; Welch, M.; Loiseleur, O.; Plowright, A. T.; Spring, D. R. Identification of an anti-MRSA dihydrofolate reductase inhibitor from a diversity-oriented synthesis. *Chem. Commun. (Cambridge, U. K.)* **2008**, *40*, 4962−4964.

(43) Higueruelo, A. P.; Schreyer, A.; Bickerton, G. R.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.* **2009**, *74* (5), 457−467.

(44) *ChemAxon Standardizer,* version 5.12; ChemAxon, Ltd: Budapest, Hungary, 2012.

(45) *Molecular Operating Enviroment (MOE)*, version 2011.10; Chemical Computing Group, Inc: Montreal, Canada, 2012.

(46) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.

(47) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(48) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107−113.

(49) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2* (22), 3256−3266.

(50) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708−1718.

(51) Williams, C. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Diversity* **2006**, *10* (3), 311−332.

(52) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48* (5), 1489−1495.

(53) *OEChem vROCS*, version 3.1.2; OpenEye Scientific Software: Santa Fe, NM, USA, 2011.

(54) Sauer, W. H.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 987−1003.

(55) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 28−35.

(56) Koutsoukas, K.; Lowe, R.; KalantarMotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, J. B. O.; Glen, R. C.; Bender, A. In silico target predictions: defining a benchmarking dataset and comparison of performance of the multiclass Naïve Bayes and Parzen−Rosenblatt Window. *J. Chem. Inf. Model.* **2013**, *53* (8), 1957−1966.

(57) Deza, M. M. *Encyclopedia of Distances*, Second Edition; Springer: New York, 2012.

(58) Hamming, R. W. Error detecting and error correcting codes. *Bell System Tech. J.* **1950**, *29*, 147−160.

(59) Myers, J. L.; Well, A.; Lorch, R. F. *Research Design and Statistical Analysis*, Third Edition; Routledge: New York, 2010; 809 pp.

(60) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010.

(61) *Dotmatics Vortex*, version 2013.03.20719, Dotmatics: The Old Monastery, Windhill, Bishops Stortford, Herts, U.K., 2013.

(62) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108−119.

(63) Mullard, A. European lead factory opens for business. *Nat. Rev. Drug Discovery* **2013**, *12* (3), 173−175.