Article:

Willett, P. (2015) Maximum Common Substructure-based Data Fusion in Similarity Searching. Journal of Chemical Information and Modeling. Published Online 20 January 2015. ISSN 1549-960X

https://doi.org/10.1021/ci5005702

Maximum Common Substructure-based Data Fusion in Similarity Searching

Edmund Duesbury,* John Holliday,* and Peter Willett*

Information School, University of Sheffield, 211 Portobello, Sheffield, S1 4DP, United Kingdom

E-mail: lip12ed@sheffield.ac.uk; j.d.holliday@sheffield.ac.uk; p.willett@sheffield.ac.uk

Abstract

Data fusion has been shown to work very well when applied to fingerprint-based similarity searching, yet little is known of its application to Maximum Common Substructure (MCS)-based similarity searching.

Two similarity search applications of the MCS will be focussed on here. Typically, the number of bonds in the MCS, as well as the bonds in the two molecules being compared, are used in a similarity coefficient. The power of this technique can be extended using data fusion, where the MCS similarities of a set of reference molecules against one database molecule are fused. This "group fusion" technique forms the first application of the MCS in this work. The other application is that of the chemical hyperstructure. The hyperstructure concept is an alternative form of data fusion, being a hypothetical molecule that is constructed from the overlap of a set of existing molecules.

This paper compares fingerprint group fusion (extended-connectivity fingerprints), MCS similarity group fusion, and hyperstructure similarity searching, and describes their relative merits and complementarity in virtual screening. It is concluded that

*To whom correspondence should be addressed

the hyperstructure approach as implemented here is less generally effective than conventional fingerprint approaches.

Introduction

Similarity-based approaches to virtual screening are very widely used in lead discovery programmes in the pharmaceutical and agrochemical industries[1,2]. In its simplest form, a known bioactive reference structure is matched against each of the structures in a chemical database to produce a ranking. The top-ranked molecules are those that are structurally most similar to the reference structure, using some quantitative measure of similarity, and are thus assumed to have the greatest likelihood of activity. Similarity searching is normally conducted using 2D fingerprints[3]. While these have been shown to provide both an effective and an efficient way of computing molecular similarity they are clearly a very simple type of structural representation and there has hence been much interest in alternative similarity measures based on 1D, 2D or 3D information of various kinds[4]. One such approach is based on the encoding of molecules in chemical databases as labelled graphs, so that similarity searching can be implemented using the maximum common subgraph (MCS), which is often referred to as the maximum common substructure in chemoinformatics. Algorithms which find the MCS align the graph representing the reference structure with the graphs representing each of the database structures, finding the database molecules that have the largest substructure(s) in common with the reference structure.

If not one but several reference structures are available then the results of searches for each of the individual structures can be combined using the methods of data fusion[5]. These yield consensus rankings that often exhibit a greater degree of clustering of actives at the top of the ranking than can be obtained using a single reference structure. Data fusion can also be used to combine the rankings resulting from the use of multiple similarity measures; however in this paper we focus on the use of multiple reference structures, an approach that has been called group fusion[5]. An alternative way of exploiting multiple reference structures in fingerprint-based similarity searching is to combine their individual fingerprints into a single consensus fingerprint[6,7]. This combines the representations of the reference structures, rather than the rankings resulting from their use as reference structures.

A concept parallel to the consensus fingerprint used by Shemetulskis et al.[7], is that of the chemical *hyperstructure* where, rather than fusing fingerprints into one, the actual chemical graphs themselves are fused into one chemical graph. In graph theory terms, the hyperstructure is a chemical abstraction of the "supergraph" concept[8]. The hyperstructure concept originates from multiple independent sources[9–11], though these will not be reviewed here.

Research on hyperstructures at Sheffield has stemmed from the work of Vladutz, and Gould[12], who proposed that hyperstructures be used to increase the efficiency of substructure searching. Brown et al.[13] utilised the maximum common substructure (MCS) in hyperstructure mapping and construction, both studies utilising genetic algorithms to find the MCS. Hyperstructures however, were found to be unsatisfactory from a virtual screening context when used in substructural analysis, being consistently outperformed by UNITY fingerprints in retrieving active compounds[14]. The reasons for the poor performance were unclear, though one proposed reason was that chemically non-meaningful artefacts (termed "ghost substructures") were present in the hyperstructures that resulted from the mappings of otherwise structurally different features between molecules. However, the random and non-deterministic nature of the genetic algorithms used may also have played a part in this.

In this paper we discuss the use of multiple reference structures for graph-based similarity searching, using both types of fusion: fusing the rankings resulting from searches that use the individual chemical graphs representing each of the reference structures; and fusing the individual chemical graphs into a single consensus graph, viz. a hyperstructure as discussed further below. Specifically, we report MCS-based similarity searches using both multiple reference structures and hyperstructures, and compare the results with those obtained using conventional fingerprint-based group fusion.

Materials and Methods

Hardware and Software

The hardware used in this study featured an Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz processor with 16 GB of DDR3 RAM clocked at 1333 MHz, running Kubuntu 13.10. The Konstanz Information Miner (KNIME) 2.9.2[15,16] running Java 1.6 was used for all experimental aspects in this study, and the Chemistry Development Kit 1.5.3 (CDK) was used for all chemoinformatics functionality unless otherwise noted[17]. 64-bit R 3.0.1[18] was used to calculate all univariate statistics reported, and the hyperstructure construction and search software was developed in Java, for use with KNIME.

Datasets

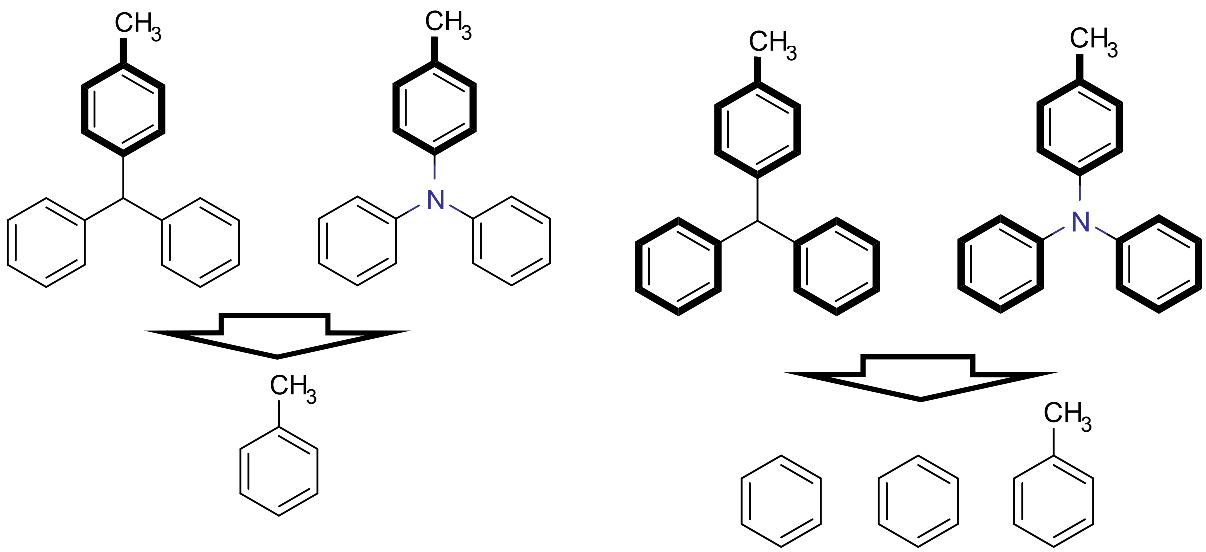
Three datasets have been used for the purposes of this study: MDDR; WOMBAT; and MUV. The MDDR and WOMBAT datasets have been used in much previous work, both at Sheffield and (in the case of MDDR) elsewhere. The MDDR dataset contains 102540 compounds, and WOMBAT contains 138049 compounds. With MDDR, 11 activity classes were used as active molecules, yielding 8184 unique active molecules. Molecules not belonging to the said activity class were treated as inactive. The same processes were also applied to 14 WOMBAT activity classes, yielding 8767 unique actives. The MUV dataset[19] involves compounds from 17 activity classes, with 30 actives in each activity class. MUV was included as an alternative benchmark, due to its design consideration of distance equivalence. The molecules in the MUV activity classes have been selected to be similar to their chosen decoy molecules, thus avoiding analogue bias which often characterises other datasets. Full details of these three datasets are provided by Hert et al.[6], Arif et al.[20] and Rohrer and Baumann[19] respectively.

For each activity class, 10 maximally diverse compounds were selected as a training set using the MaxMin algorithm as implemented in the KNIME version of RDKit, with a randomly assigned seed[21]. These 10 molecules would be subsequently removed from the dataset to remove self-similarity bias from the virtual screening statistics. RDKit standard Morgan fingerprints were used with a maximum radius of 3 (similar to the ECFP_6 fingerprint found in Pipeline Pilot[22], folded into 1024 bits) as the fingerprint descriptor in this study.

MCS Definition and Methodology

The MCS referred to in this work is the maximum common edge-induced substructure (MCES), as opposed to the maximum common induced subgraph, which yields smaller common mappings and is less chemically intuitive[23]. The MCES can be further abstracted into the *connected* (cMCES) and *disconnected* MCES (dMCES). The cMCES (Figure 1a) is the single MCES graph, where all the nodes in the subgraph are connected to at least one other node in the subgraph. The dMCES (Figure 1b) by contrast, sometimes known as the maximum overlap set (MOS), can contain multiple (separated) subgraphs, representing all the edges in common between the graphs being matched. In this study, the MCS will refer to the dMCES at all points, which we are using as it is better suited to comparing structurally dissimilar graphs (as shown in Figure 1, where bold-facing denotes common substructures).

The MCES was found using the `MaxCommonSubstructure` class in JChem 6.1.0, 2014, by ChemAxon[24,25]. This algorithm has been claimed by its authors to be the quickest inexact method for finding the MCES between two molecules, and also incorporates a number of



(a) (b)

Figure 1: MCS types between two molecules differing by a single central atom. 1a is the connected Maximum Common Substructure, whereas 1b is the disconnected Maximum Common Substructure

heuristics to adjust the mappings, making the alignments more chemically feasible[26].

Hyperstructure Construction and Application

The construction process used here is based on the process described by Brown et al.[27], an example being depicted in Figure 2. The process is as follows:

1. Select the largest molecule and remove it from the set of available molecules. This molecule is now the first hyperstructure.

2. Select the next most similar molecule, based on the number of bonds in common between the hyperstructure and this molecule.

3. Use the MCS procedure to overlap and then to append this molecule to the hyperstructure, and remove this molecule from the set. Bonds of different types may be overlapped, yielding degenerate bond types (depicted by dashed lines in the hyperstructures).

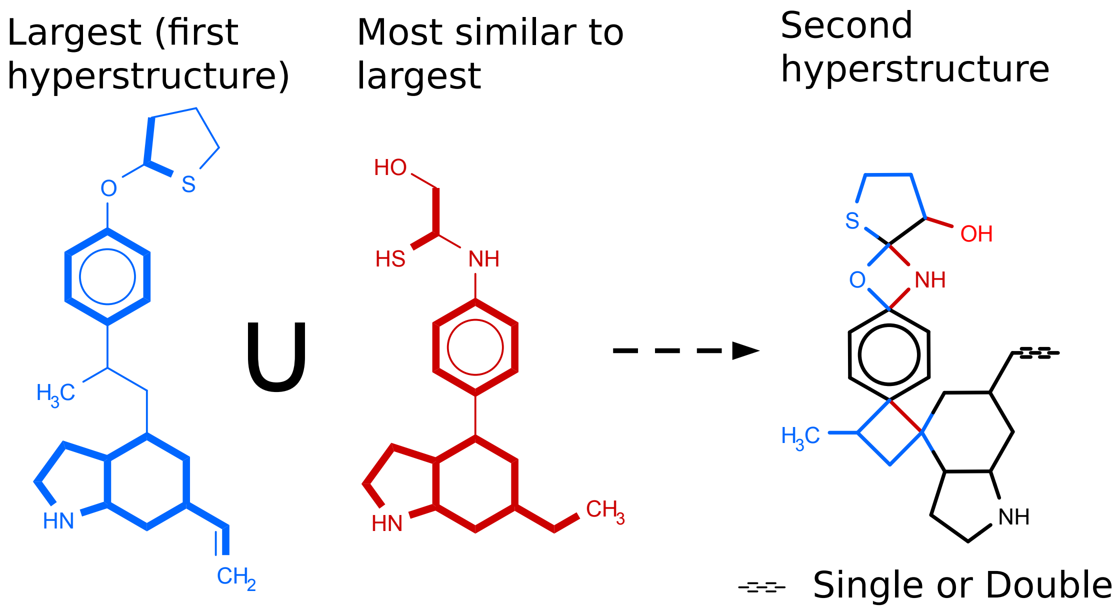


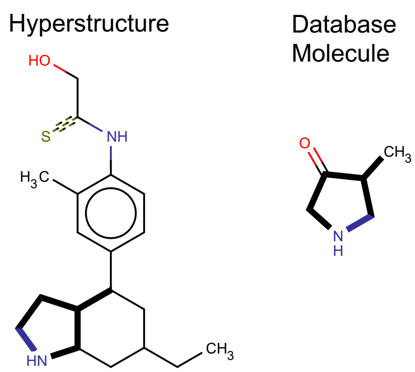
Figure 2: The hyperstructure construction process with two molecules, shown in blue and red - bold bonds indicate those in the MCS between the two molecules. In the resulting hyperstructure, black bonds and atoms indicate the MCS, while unique bonds and atoms are coloured based on the original molecules.

4. Repeat steps 2 and 3 until the set is empty.

The Tanimoto coefficient is the standard similarity coefficient for quantifying chemical structural similarity, and was used for the MCS and fingerprint scores described in the next section. However, the asymmetric Tversky coefficient was used for the hyperstructure similarity calculations. This is more suitable than the Tanimoto coefficient for similarity searching using hyperstructures, due to the size differences (notably in this work, the number of bonds), since the hyperstructure will be at least as large as the biggest molecule used to build the hyperstructure. We also expect substructure similarity to play a large role in similarity searching, since the hyperstructure collectively represents the scaffolds present in the input molecules. The Tversky coefficient can be biased towards substructure similarity. The Tversky coefficient is defined as

$$S_{Tv} = \frac{c}{\beta(a-c)+(1-\beta)(b-c)+c} \qquad \beta = 0 \rightarrow \frac{c}{b} \qquad \beta = 1 \rightarrow \frac{c}{a}$$

where c in this case represents the number of edges in the MCS, a is the number of bonds in the database molecule, and b the number of hyperstructure bonds. Higher values of β give a near substructure-like search, whereas lower values bias the results towards a superstructure-like search (as exemplified in Figure 3). One should note that exclusion of the terms β and $(1 - \beta)$ yields the Tanimoto coefficient. In internal studies, several values of β have been tested, and the value of 0.9 emerged as the most generally suitable for virtual screening recall - a value which has been found to also be beneficial in similarity searching using fingerprints[28,29], and has also shown potential in scaffold hopping with fingerprints[30].



$$a = 7 \tag{1}$$
$$b = 25 \tag{2}$$
$$c = 6 \tag{3}$$
$$S_{Tv} = 0.259 : \beta = 0.1 \tag{4}$$
$$S_{Tv} = 0.375 : \beta = 0.5 \tag{5}$$
$$S_{Tv} = 0.682 : \beta = 0.9 \tag{6}$$

Figure 3: An example of how the value of β influences similarity in the Tversky coefficient. Edges in the MCS are marked in bold

Similarity Search Methodology

The reference sets of active molecules mentioned were subjected to one of three search methods:

- Hyperstructure construction and searching, using the Tversky coefficient with a β of 0.9 as discussed above.

- MCS group fusion using the Tanimoto coefficient (with the MAX rule on similarity scores)
- Morgan fingerprint group fusion using the Tanimoto coefficient (with the MAX rule on similarity scores)

Data fusion was also implemented on the rankings obtained with fingerprints, MCS and hyperstructures. Given a set of similarity values (or ranks) S_1, S_2, ... S_n, the MAX fusion rule identifies the maximum similarity value S. The SUM fusion rule by contrast is the summation of S_1, S_2, ... S_n. SUM fusion of the ranks were used here, as this combination rule is more appropriate when fusing different similarity measures[5]. A summary of method names (including fusion types) are described in the Table 1.

Table 1: Descriptions of the abbreviations of search methods employed here.

Method	Description
HS	Hyperstructure search applied using Tversky similarity coefficient, with a β of 0.9
FP	Fingerprint Tanimoto similarity with the MAX fusion rule applied to the similarity scores
MCS	Tanimoto similarity (based on the bonds in the MCS and the two structures being compared), with the MAX fusion rule applied to the similarity scores
FP MCS	SUM fusion of FP and MCS ranks
FP HS	SUM fusion of FP and HS ranks
MCS HS	SUM fusion of MCS and HS ranks
FP MCS HS	SUM fusion of FP, MCS and HS ranks

Evaluation metrics

Two measures have been used to quantify the effectiveness of screening. The first is the Enrichment Factor (EF) for the top 1% of ranked compounds (shown as $EF_{1\%}$). EF is a simple statistic to interpret for determining recall. We appreciate however that the EF does not account for the relative rank of compounds. Two activity classes in MDDR also have a number of actives which exceed 1% of the proportion of the database (renin inhibitors and Substance P antagonists, having 1130 and 1246 active compounds respectively). We have therefore chosen to use the Boltzman-Enhanced ROC score (BEDROC) as an alternative evaluation score. BEDROC scales from 0.0 to 1.0, where a value closer to 1.0 indicates superior virtual screening performance[31].

A small problem with BEDROC is the need to set a tuning factor α, which determines how many of the top-ranked compounds contribute to the BEDROC score. A higher value of α means that a smaller percentage of the top-ranked compounds contributes to the majority of the BEDROC score. A value of 160.9 for α has been chosen in this study as it corresponds to $EF_{1\%}$, where approximately 80% of the BEDROC score is explained by the top 1% of compounds in the ranked list (refer to Table 2 of Truchon and Bayly[31]). It should be noted that whilst it has been shown that BEDROC and EF are often strongly correlated[32], BEDROC takes account of the ratio of actives to inactives, where EF does not. The same study notes that the two measures are uncorrelated when this ratio differs between activity classes.

Enrichment involves retrieving as many active compounds as possible, but it is often just as important in a virtual screening context, to find a few structurally dissimilar compounds rather than a large set of close analogues. A method that identifies multiple different "cores" or "scaffolds" is said to be proficient at scaffold-hopping. Scaffolds in this work are represented by Bemis-Murcko frameworks with bond and atom labels removed[33], but with R-groups kept when attached to linker atoms (as is the method for the RDKit definition of Murcko frameworks in KNIME). Two statistics will be presented in this study to assess the ability of a method to obtain unique scaffolds. The first is the "First-Found" scaffold enrichment factor, using the top 1% of the ranked list of molecules (represented here as $EF_{FF1\%}$). "First-Found" refers to the rank of the top-ranked molecule belonging to a scaffold, and ignoring all subsequent molecules belonging to the scaffold. Although this measure of obtaining diversity has been criticised for several statistical flaws[34], we use it here as we are only interested in whether an active scaffold is identified or not in a ranked list. We are not interested in how the compound ranks are distributed for the given active scaffold. This represents a common goal in a virtual screening project if one is seeking new scaffolds. The other measure is the mean number of active molecules per active scaffold in the top 1% ranking (represented here as $F_{1\%}$). A lower value for $F_{1\%}$ indicates that a search method retrieves on average less actives per scaffold, and is therefore less biased towards finding analogues for a small number of scaffolds

To get an idea of whether the search techniques retrieved different sets of molecules, we calculated a number of physicochemical descriptors. These were calculated for the active compounds in the top 1% of the ranked database resulting from each similarity search method. The descriptors used were the counts of the number of rotatable bonds, heavy atoms, heteroatoms and rings; the length of the largest acyclic chain, and the fragment complexity. This last descriptor was the CDK implementation of the work described by Nilakantan et al.[35]:

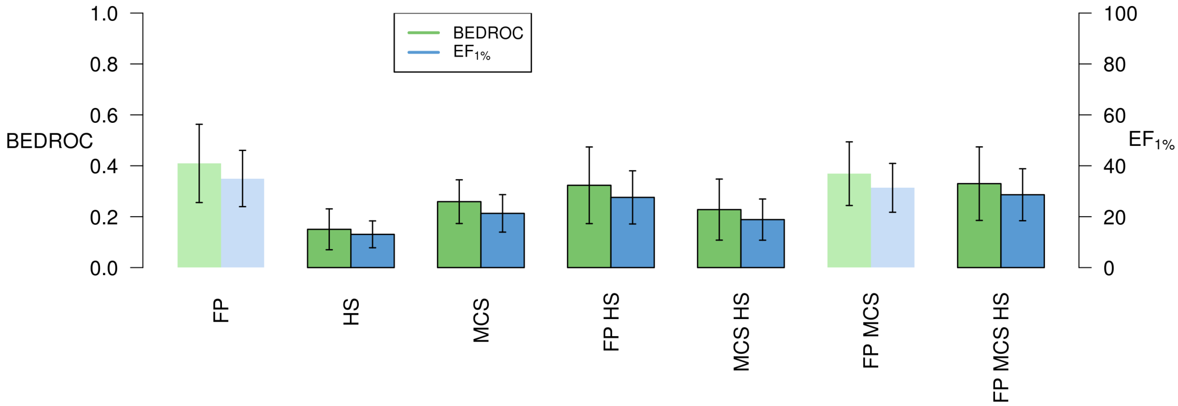
$$C = |B^2 - A^2 + A| + H/100$$

where A is the number of non-hydrogen atoms, B is the number of bonds, C is the fragment complexity and H is the number of heteroatoms.

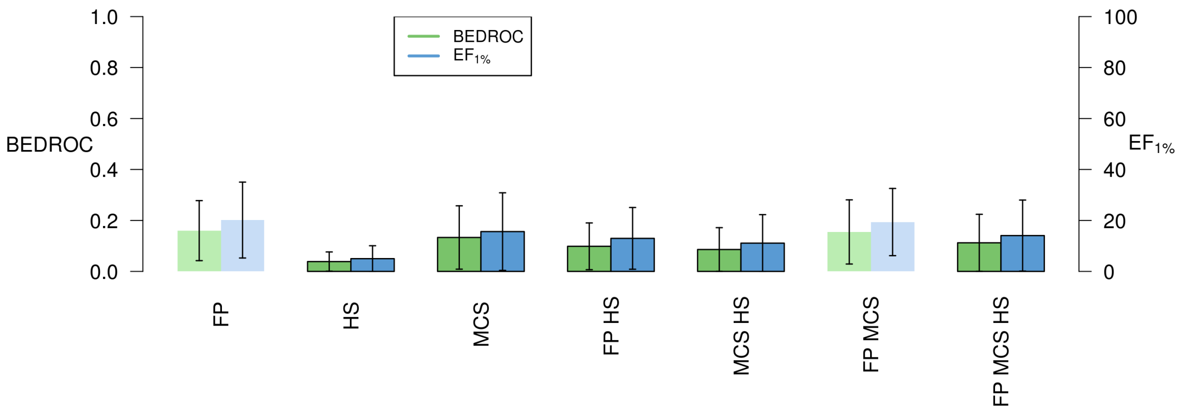
To assess the overlap between the ranked lists resulting from two different search methods, we calculated the number of actives common to the top 1% of the two ranked lists.

Results and Discussion

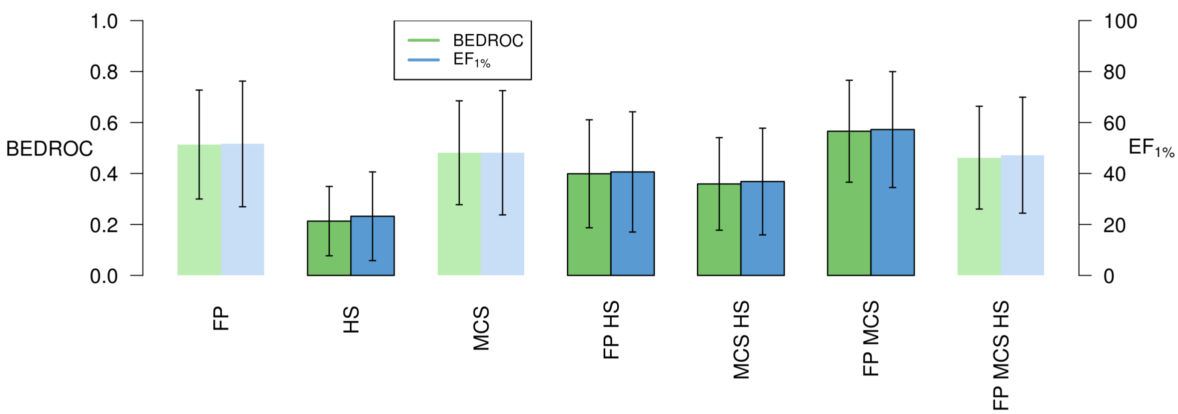
Figure 4 summarises recall performances for the similarity search methods tested, and Figure 5 summarises scaffold-retrieval, averaged over the different activity classes in each dataset. The immediate conclusion from the statistics tables is, on comparing the mean values of BEDROC and $EF_{1\%}$, that FP is significantly superior to HS, with MCS lying



(a) MDDR

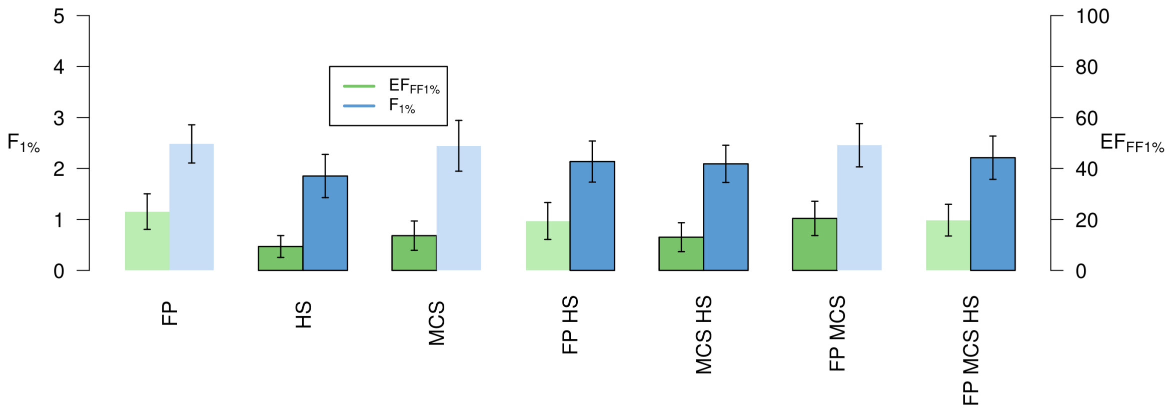


(b) MUV

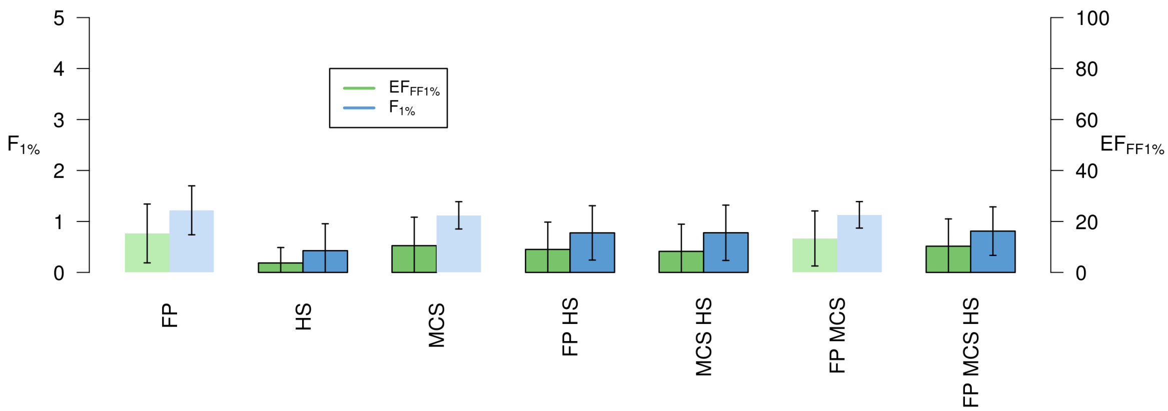


(c) WOMBAT

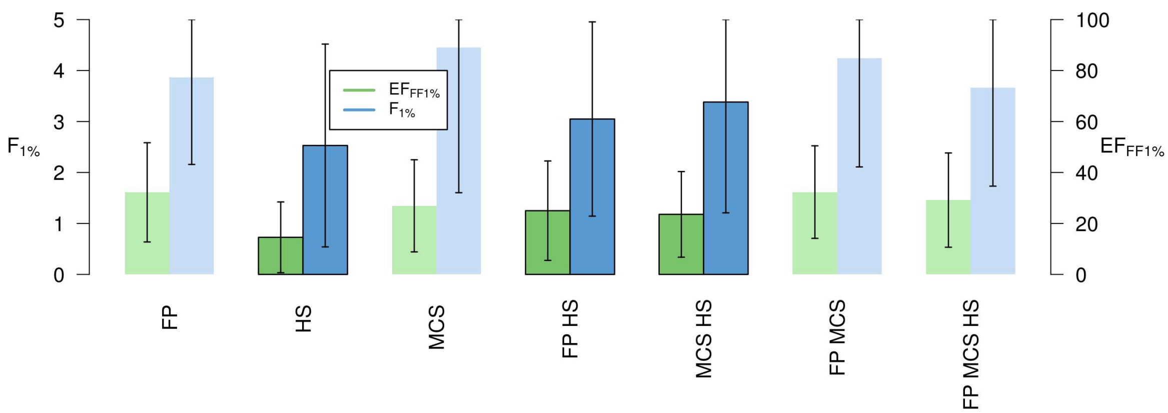
Figure 4: Bar charts showing mean recall statistics for the datasets. Dark-coloured bars indicate that the method has a mean significantly different from that of FP ($p \leq 0.05$), as determined by a paired 2-tailed Wilcoxon signed-rank test. Error bars represent one standard deviation above and below the mean.



(a) MDDR



(b) MUV



(c) WOMBAT

Figure 5: Bar charts showing mean scaffold-retrieving statistics for the datasets. Dark-coloured bars indicate that the method has a mean significantly different from that of FP ($p \leq 0.05$), as determined by a paired 2-tailed Wilcoxon signed-rank test. Error bars represent one standard deviation above and below the mean.

between the two. This suggests that fingerprints are better suited to virtual screening, the p-values for the Wilcoxon signed-rank tests being significant. In all three datasets the fingerprints significantly (and consistently) outperformed hyperstructures in virtual screening ability. The observation that fingerprints generally match or outperform MCS-based methods is consistent with the related work of Raymond and Willett[36], though in this work the fingerprints used are of a better standard in terms of virtual screening recall than those used by Raymond and Willett[36]. A further conclusion from the figures is, as has been noted by previous studies, that the MUV dataset presents a much harder challenge for ligand-based virtual screening methods such as those presented here, since the performance is inferior compared to the other two datasets.

The data fusion techniques show improved BEDROC scores over the results for HS and MCS, though none of them are superior to FP alone. There exists only one technique which FP doesn't consistently outperform, this being FP MCS, though the time requirements (see next section) to calculate the MCS in this technique makes FP preferable.

FP interestingly outperforms all the other techniques in terms of $EF_{FF1\%}$ as for BEDROC, implying that fingerprint group fusion is good for scaffold hopping. Effective scaffold hopping has been observed with single reference structures[37,38], thus it is unsurprising that group fusion of fingerprints also yields a favourable scaffold hopping potential. Although HS generally retrieves a lower number of unique active scaffolds, it can be seen that HS almost always obtains a significantly lower $F_{1\%}$ than the FP and MCS methods. From this, it can be inferred that the top-ranked active molecules retrieved by hyperstructures have less analogues per scaffold than those retrieved with fingerprints (in relation to the number of actives retrieved). MCS by comparison has no significant difference in scaffold retrieval compared to fingerprints, and also has an inferior virtual screening performance to fingerprints. The data fusion techniques, from their BEDROC and $EF_{1\%}$ values, show compromises in diversity retrieval. Of note, FP HS and MCS HS for all three datasets have significantly lower $F_{1\%}$ values than FP.

Table 2: Time statistics on the constructed hyperstructures for each class in the MDDR dataset. HS_C is the hyperstructure construction time. FP_S, HS_S and MCS_S are times for fingerprint, hyperstructure and MCS searching respectively. All times in this table are in milliseconds.

targetID	FP_S	HS_C	HS_S	MCS_S
6233	$1.5 \cdot 10^4$	$1.5 \cdot 10^3$	$9.1 \cdot 10^5$	$1.8 \cdot 10^6$
6235	$5.9 \cdot 10^4$	$1.0 \cdot 10^3$	$1.4 \cdot 10^6$	$2.3 \cdot 10^6$
6245	$1.4 \cdot 10^4$	$6.3 \cdot 10^2$	$1.1 \cdot 10^6$	$2.0 \cdot 10^6$
7701	$1.5 \cdot 10^4$	$9.0 \cdot 10^2$	$1.5 \cdot 10^6$	$2.1 \cdot 10^6$
31420	$3.4 \cdot 10^4$	$1.6 \cdot 10^3$	$2.4 \cdot 10^6$	$4.9 \cdot 10^6$
31432	$1.9 \cdot 10^4$	$1.6 \cdot 10^3$	$2.1 \cdot 10^6$	$3.4 \cdot 10^6$
37110	$1.7 \cdot 10^4$	$1.0 \cdot 10^3$	$1.9 \cdot 10^6$	$2.9 \cdot 10^6$
42731	$4.4 \cdot 10^4$	$2.3 \cdot 10^3$	$3.2 \cdot 10^6$	$4.0 \cdot 10^6$
71523	$2.4 \cdot 10^4$	$2.3 \cdot 10^3$	$3.0 \cdot 10^6$	$4.0 \cdot 10^6$
78331	$2.0 \cdot 10^4$	$1.1 \cdot 10^3$	$1.3 \cdot 10^6$	$2.2 \cdot 10^6$
78374	$2.1 \cdot 10^4$	$1.0 \cdot 10^3$	$2.0 \cdot 10^6$	$2.6 \cdot 10^6$

Table 2 shows the performance times of the hyperstructure and MCS searches. The typical search time requirement for hyperstructures has a mean of 82.6 times greater than than fingerprint searches, but the hyperstructure searches are consistently faster than MCS. The fraction of time required for hyperstructure searches compared to MCS is between 0.49 and 0.8 with a mean of 0.638, depending on the dataset and activity class.

Table 3: Mean values of physicochemical properties of the actives of the top 1% ranked lists for given methods. Values with the prefix "p_" reflect the p-values from paired 2-tailed Wilcoxon signed rank tests, tested against FP.

Method	Complexity	Largest Chain	Rotatable Bonds	Heavy Atoms	Heteroatoms	Rings
FP	3570	15.022	11.511	32.253	7.714	3.874
MCS	3146	13.188	10.554	29.995	6.793	3.655
HS	3969	16.469	12.486	33.913	7.632	3.873
p_MCS	0.002	0.005	0.054	0.005	0.005	0.175
p_HS	0.010	0.042	0.083	0.032	1.000	0.765

The physicochemical properties shown in Table 3 highlight some statistically significant differences in the molecules between hyperstructure, MCS and fingerprint searches. Of immediate note is that hyperstructures retrieve significantly larger (more atoms) active molecules than fingerprints. In addition to being larger, the molecules are more complex

Table 4: Complementarity of search methods. Each cell gives the proportion of identified active compounds in common with the two methods, divided by the number of actives identified by the method in the column.

Method	MCS	FP	HS
MCS	1.000	0.312	0.352
FP	0.609	1.000	0.583
HS	0.293	0.248	1.000

(a) Renin-Angiotensin inhibitors

Method	MCS	FP	HS
MCS	1.000	0.329	0.105
FP	0.392	1.000	0.105
HS	0.016	0.013	1.000

(b) Cyclooxygenase Inhibitors

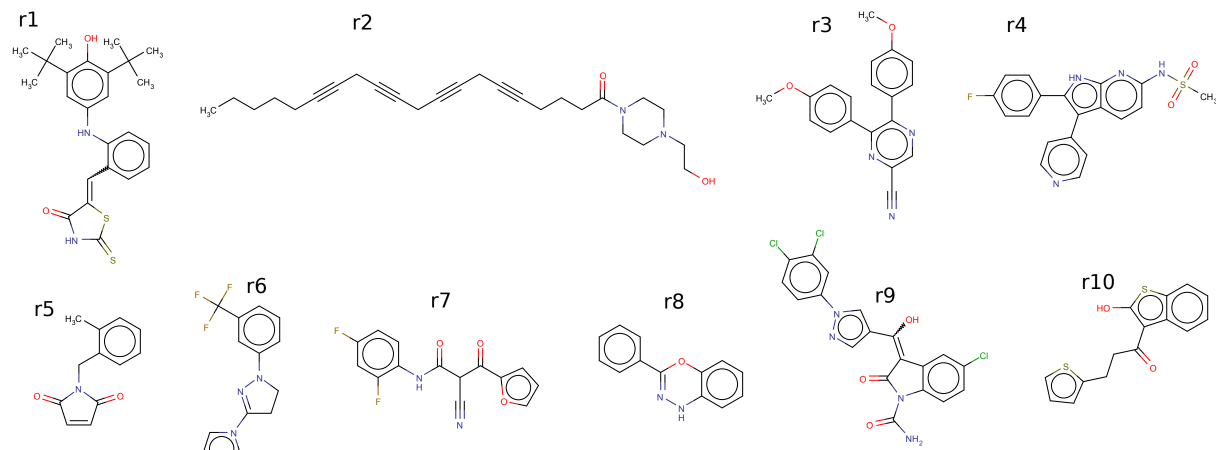
Method	MCS	FP	HS
MCS	1.000 $\pm$ 0.000	0.445 $\pm$ 0.122	0.153 $\pm$ 0.085
FP	0.268 $\pm$ 0.074	1.000 $\pm$ 0.000	0.123 $\pm$ 0.074
HS	0.230 $\pm$ 0.087	0.317 $\pm$ 0.163	1.000 $\pm$ 0.000

(c) Mean values across all classes $\pm$ one standard deviation.

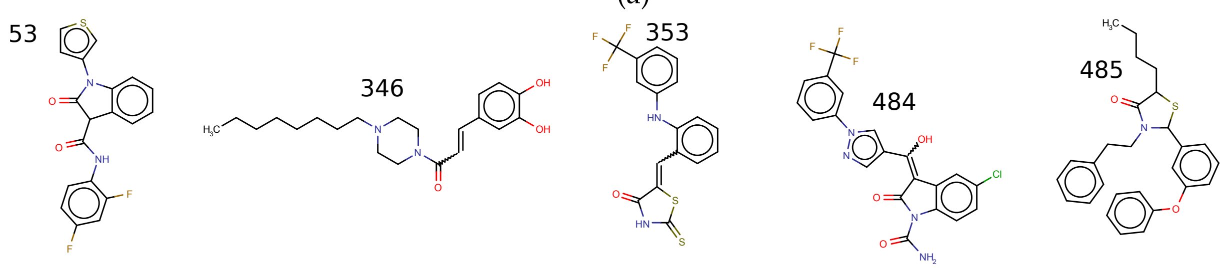
and possess larger acyclic chains, but have no significant difference in heteroatom and ring count. This implies that the molecules are less "feature-rich" and more chain-rich. By contrast, the MCS-retrieved active molecules are smaller and less chain-rich, though also possessing significantly less heteroatoms.

In the MDDR dataset, the Renin-Angiotensin class is the most intrinsically similar class, whereas cyclooxygenase inhibitors are the most diverse. This is on the basis of the mean pairwise similarities between all pairs of active molecules for said activity classes using Unity 2D fingerprints and the Tanimoto coefficient[6]. Table 4 shows a strong lack of overlap in the retrieved actives of the techniques presented, both for the two activity classes as well as with the mean values. Of note is that FP and MCS have a greater overlap with each other than with the hyperstructure searches. These observations are more pronounced with the cyclooxygenase inhibitors than with renin, though even with the former there is still little overlap.

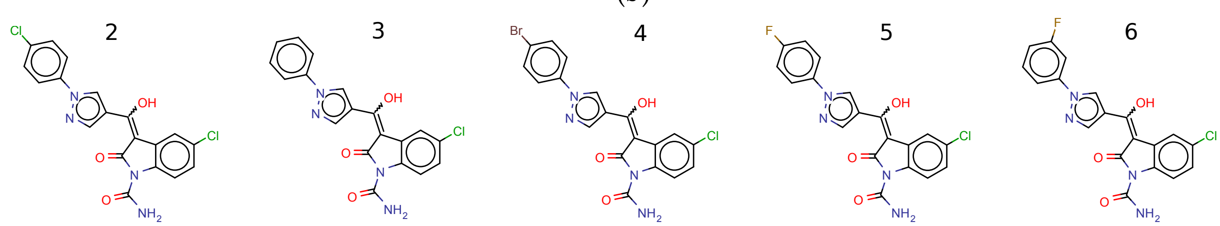
One of the potential attractions of the HS approach is that it may provide a way of identifying novel scaffolds. The top five ranked actives involved in the MDDR COX inhibitors search are shown in Figure 6b, with the reference structures from which the



(a)



(b)



(c)

Figure 6: Compounds involved in the MDDR COX inhibitors search. 6a shows the molecules used to construct the hyperstructure (numbered arbitrarily). 6b shows the top five-ranked active compounds retrieved by the hyperstructure, with their ranks displayed. 6c shows the top five-ranked active compounds retrieved by FP, with their ranks displayed.

Table 5: Similarities of molecules in Figure 6. The similarity reported uses the Tanimoto coefficient with the MAX fusion rule to the reference molecules, the reference molecule beinq quoted. The fingerprint used is as with the FP method.

Molecule	Similarity	Ref
56	0.260	r7
346	0.228	r2
353	0.570	r1
484	0.685	r9
485	0.126	r2

(a) Retrieved actives with HS (Figure 6b)

Molecule	Similarity	Ref
2	0.769	r9
3	0.756	r9
4	0.732	r9
5	0.732	r9
6	0.718	r9

(b) Retrieved actives with FP (Figure 6c)

hyperstructure was constructed being in Figure 6a. Comparisons of these two will reveal that the search has identified three scaffolds not present in the reference structures (actives 56, 346 and 485). Of note, the top five compounds retrieved by FP (Figure 6c) all share the same scaffold (with r9) and generally differ from each other by just one atom. The similarities are also evident from Table 5, where the FP similarities for the FP-retrieved actives are much higher than those retrieved by HS. This ability to prioritise analogues over non-analogues is a major reason for why the fingerprint-MAX rule outperforms HS (and it is unsurprising this works so well given that 10 molecules with different scaffolds are used as the reference set).

Conclusions

The results of this investigation showed that both hyperstructures and MCS fusion, at least for the datasets used, are inferior to a conventional fingerprinting method for performing virtual screening in terms of enrichment, and scaffold retrieval. MCS group fusion is significantly slower than hyperstructure-based similarity searching, albeit with a slight gain in virtual screening performance. Although hyperstructures retrieve fewer scaffolds than fingerprints, they retrieve a better spread of compounds across scaffolds compared to all the other techniques tested, implying that they are less likely to find analogues than

MCS and fingerprint group fusion techniques. MCS and, in particular, hyperstructure searches, have low overlaps with fingerprints in the active compounds retrieved. The physicochemical properties of the actives retrieved often differ between the three techniques as well: compared to fingerprints, hyperstructures tend to retrieve larger molecules with greater chains and fewer heteroatoms, whereas the opposite is observed with MCS searches. Data fusion of the techniques used in this study generally yields compromises in virtual screening performance. All fusion techniques here outperformed MCS and hyperstructure-based searches alone, but failed to outperform fingerprint searching.

The results above demonstrate clearly that fingerprints out-perform hyperstructure searches in terms of numbers of retrieved actives. One obvious reason for this behaviour is fingerprints' ability to identify large numbers of close analogues to an entire reference structure, something that is much more difficult for a hyperstructure that has been constructed from multiple individual molecules, especially when these are structurally disparate (as is the case here). It should also be noted that the baseline of comparison is a type of fingerprint that is known to be extremely effective for virtual screening. Finally, the MCS algorithm used in this work is inexact, and also generates only a single MCS, even if several different (and potentially more chemically feasible) ones are possible. These factors influence the performance of the hyperstructure concept, both in hyperstructure construction and similarity searching. It would thus be worth investigating potential performance changes using alternative MCS algorithms.

Acknowledgement

The authors thank Peter Englert for information on the features of the JChem MCS algorithm, John May for usage advice on the Chemistry Development Kit, and the University of Sheffield for the Faculty Scholarship funding and hardware used in this PhD project.

Conflict of Interest

The authors declare no competing financial interest.

References

(1) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204.

(2) Willett, P. Similarity Methods in Chemoinformatics. *Annu. Rev. Inform. Sci* **2009**, *43*, 1–117.

(3) Willett, P. Similarity-based Virtual Screening Using 2D Fingerprints. *Drug Discov.Today* **2006**, *11*, 1046–1053.

(4) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; Wiley-VCH: Weinheim, 2009.

(5) Willett, P. Combination of Similarity Rankings Using Data Fusion. *J. Chem. Inf. Model.* **2013**, *53*, 1–10.

(6) 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*, 3256.

(7) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.

(8) Bunke, H.; Jiang, X.; Kandel, A. On the Minimum Common Supergraph of Two Graphs. *Computing* **2000**, *65*, 13–25.

(9) Dubois, J. E.; Laurent, D.; Aranda, A. Méthode de Perturbation D'Environnements Limites Concentriques Ordonnés (PELCO). *J. Chim. Phys.* **1973**, *70*, 1608–1615.

(10) Menon, G. K.; Cammarata, A. Pattern Recognition II: Investigation of Structure—activity Relationships. *J. Pharm. Sci.* **1977**, *66*, 304–314.

(11) Simon, Z.; Badilescu, I.; Racovitan, T. Mapping of Dihydrofolate-reductase Receptor Site by Correlation with Minimal Topological (steric) Differences. *J. Theor. Biol.* **1977**, *66*, 485–495.

(12) Vladutz, G.; Gould, S. R. *Chemical structures: the international language of chemistry*; Springer Verlag: Berlin, 1988; Vol. 1; pp 371–384.

(13) Brown, R. D.; Downs, G. M.; Willett, P.; Cook, A. P. F. Hyperstructure Model for Chemical Structure Handling: Generation and Atom-by-Atom Searching of Hyperstructures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 522–531.

(14) Brown, N. Generation and Application of Activity-weighted Chemical Hyperstructures. PhD Thesis, University of Sheffield, Sheffield, 2002.

(15) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, P. D. H., Schmidt-Thieme, P. D. L., Decker, P. D. R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer Berlin Heidelberg, 2008; pp 319–326.

(16) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME - the Konstanz Information Miner: Version 2.0 and beyond. *SIGKDD Explor.* **2009**, *11*, 26–31.

(17) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

(18) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(19) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.

(20) Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and Use of Fragment-occurrence Data in Similarity-based Virtual Screening. *J. Comput. Aided Mol. Des.* **2009**, *23*, 655–668.

(21) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets Using Property-Based and Fragment-Based Molecular Descriptions. *QSAR* **2002**, *21*, 598–604.

(22) Rogers, D.; Hahn, M. Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(23) Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput. Aided Mol. Des.* **2002**, *16*, 521–533.

(24) Kovács, P.; Englert, P. MaxCommonSubstructure (JChem API Documentation (c) 1998-2013 ChemAxon Ltd.). 2013; `http://www.chemaxon.com/jchem/doc/dev/java/api/com/chemaxon/search/mcs/MaxCommonSubstructure.html`.

(25) Kovács, P.; Englert, P. Making the Most of Approximate Maximum Common Substructure Search. 2014; `http://www.slideshare.net/penglert/mcs-poster`.

(26) Englert, P.; Kovács, P. Making the Most of Approximate Maximum Common Substructure Search. *J. Cheminform.* **2014**, *6*, P29.

(27) Brown, R. D.; Downs, G. M.; Jones, G.; Willett, P. Hyperstructure Model for Chemical Structure Handling: Techniques for Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 47–53.

(28) Wang, Y.; Eckert, H.; Bajorath, J. Apparent Asymmetry in Fingerprint Similarity Searching Is a Direct Consequence of Differences in Bit Densities and Molecular Size. *ChemMedChem* **2007**, *2*, 1037–1042.

(29) Horvath, D.; Marcou, G.; Varnek, A. Do Not Hesitate to Use Tversky—and Other Hints for Successful Active Analogue Searches with Feature Count Descriptors. *J. Chem. Inf. Model.* **2013**,

(30) Senger, S. Using Tversky Similarity Searches for Core Hopping: Finding the Needles in the Haystack. *J. Chem. Inf. Model.* **2009**, *49*, 1514–1524.

(31) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(32) Riniker, S.; Landrum, G. A. Open-source Platform to Benchmark Fingerprints for Ligand-based Virtual Screening. *J. Cheminform.* **2013**, *5*, 26.

(33) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(34) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154–1162.

(35) Nilakantan, R.; Nunn, D. S.; Greenblatt, L.; Walker, G.; Haraki, K.; Mobilio, D. A Family of Ring System-Based Structural Fragments for Use in Structure-Activity Studies: Database Mining and Recursive Partitioning. *J. Chem. Inf. Model.* **2006**, *46*, 1069–1077.

(36) Raymond, J. W.; Willett, P. Effectiveness of Graph-based and Fingerprint-based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput. Aided Mol. Des.* **2002**, *16*, 59–71.

(37) Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D Fingerprints for Scaffold Hopping. *Future. Med. Chem.* **2011**, *3*, 405–414.

(38) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

Graphical TOC Entry