Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

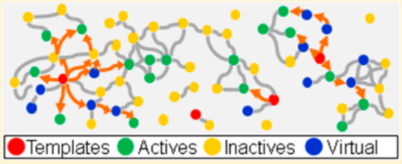
Hit Expansion Approaches Using Multiple Similarity Methods and Virtualized Query Structures

Andreas Bergner*,† and Serge P. Parel‡

BioFocus, Chesterford Research Park, Saffron Walden, Essex CB10 1XL, United Kingdom

Ⓢ *Supporting Information*

ABSTRACT: Ligand-based virtual screening and computational hit expansion methods undoubtedly facilitate the finding of novel active chemical entities, utilizing already existing knowledge of active compounds. It has been demonstrated that the parallel execution of complementary similarity search methods enhances the performance of such virtual screening campaigns. In this article, we examine the use of virtualized template (query, seed) structures as an extension to common search methods, such as fingerprint and pharmacophore graph-based similarity searches. We demonstrate that template virtualization by bioisosteric enumeration and other rule-based methods, in combination with standard similarity search techniques, represents a powerful approach for hit expansion following high-throughput screening campaigns. The reliability of the methods is demonstrated by four different test data sets representing different target classes and two hit finding case studies on the epigenetic targets G9a and LSD1.



1. INTRODUCTION

The search for new active hit compounds and novel scaffolds that are suitable for further development in hit-to-lead optimization campaigns is one of the key topics in early stage drug discovery. High-throughput screening (HTS) of large compound libraries has become a major paradigm for hit identification.[1,2] While HTS has not fulfilled the expectations of the initial hype, many of its shortcoming and pitfalls are now better understood,[3−5] thus preventing the hindrances of the early days. It is commonly believed that HTS remains a key technology for hit identification for the foreseeable future,[6] with the efficient HTS and uHTS technologies of today enabling the screening of huge libraries comprising more than one million compounds.[7]

It is still controversial as to how large, and how diverse or focused, the screening libraries should be, and how much computational compound selection approaches can contribute to hit identification, when compared with random HTS.[8−11] Hit finding is increasingly supported by smarter computational approaches,[12] and virtual screening (VS) can facilitate successful screening campaigns with very small compound decks, thus saving time and costs. It has already been demonstrated that VS has contributed to the discovery of compounds that ultimately reached the market.[8]

In reality, HTS and in silico screening approaches are allies that can often complement each other, and many efforts have been made to better align both approaches.[13] Rather than embracing one particular hit finding paradigm, in silico and screening components can be combined according to the project requirements, using the available target and active compound information. Where appropriate, the primary screen can be preceded by a VS campaign for defining a primary screening deck of suitable size and chemical composition. HTS data-mining techniques[14] enable post-screening analysis for detecting latent hit series and rescuing active scaffolds from low-activity data.[15−17] Bayesian learning methods allow knowledge about the chemical features that contribute to the activity to be distilled from such data.[18] An empirical post-HTS compound prioritization scheme that takes various physicochemical properties of the hits into account has been proposed by Oprea et al.[19] All of these techniques can be assembled in a linear sequential or iterative way, enabling the design of workflows that are tailored to each particular hit finding project.

A common approach for fast follow-up of hits obtained by the first round of HTS screening is hit expansion (HE). Validated hit structures and relevant structures obtained from data-mining procedures can be employed for defining small compound decks using in silico compound selection methods. The compound decks are then used in subsequent rounds of screening. Hit expansion can serve multiple purposes, depending on the outcome of the primary screen and the aim of a hit finding project. It can consolidate and expand knowledge about active regions in chemical space by scaffold hopping and by finding derivatives of active scaffolds, hence establishing or broadening SAR information. Alternative naming conventions include SAR expansion[20] and hit-directed nearest-neighbor searching.[21] Hit expansion is intended to strengthen the knowledge base for direct use in hit-to-lead campaigns.

Virtual screening (VS), or in silico screening,[12,22−27] encompasses a wide range of computational methods for selecting compounds that are predicted to be active, and can be considered as the mining of chemical space with the aim of distinguishing active and inactive molecules in a database.[28] VS approaches are usually grouped into structure-based VS (SBVS) and ligand-based VS (LBVS) methods.

Received: January 25, 2013

SBVS relies on the 3D structure of a protein with a suitable binding site of interest that can be used to adapt the different 3D conformational models (poses) of compounds that are generated by docking programs. These poses are subsequently analyzed in post-processing procedures in order to identify the compounds that are most likely to be active (see, for example, the recent reviews by Klebe[29] and Waszkowycz[30] for a comprehensive discussion).

In contrast, the starting point for LBVS are the structures of known active compounds, referred to as template (also query or seed) structures, which are used for screening large virtual compound databases. LBVS includes a large portfolio of machine learning and chemoinformatics methods.[31] Many of the chemoinformatics-based VS methods use the similar property principle, i.e., the hypothesis that similar compounds have similar properties.[32] The evidence for this principle has been asserted by Patterson et al. who wrote that "The very existence of medicinal chemistry teaches that similar molecules will tend to have similar biological properties. Were this not so, the major activity of medicinal chemists, lead optimization ..., would be futile!".[33] The concept of similarity perceived here is that of "local" similarity and concerns molecular changes, such as the exchange of R-groups. Another aspect of "local"' similarity is bioisosterism, i.e., the exchange of a chemical moiety or a molecular fragment that does not perturb the biological activity of a molecule.[34] Various approaches for deriving bioisosteric relationships in silico and compiling them in dictionaries have been published.[35−39] In contrast to "local" similarity, the concept of "global"' (or "holistic") similarity[40] is only of significance with respect to the composition of entire molecules. An example is the use of fingerprints (FPs) for similarity searching because FPs encode the composition of entire molecules in terms of certain features, and all FP components contribute to the overall similarity value. These considerations show that molecular similarity is not an absolute and unambiguous category. Rather, it is a concept that is only meaningful in the context of how one looks at molecules and which molecular aspects are important. This is also true for the way in which similarity is measured by means of some metric. The question as to whether or not two molecules contain the same substructure is expressed by a binary match/no match metric. For most other similarity methods, a metric is defined that allows the quantification of a normalized level of similarity between 0 and 1. Conversely, dissimilarity can be used as a metric for comparing compounds.

Both the concepts of "local" and "global" similarity outlined above have been used in VS approaches, although the "global" approaches are far more prominent. In spite of their known limitations, fingerprint-based similarity searches remain the most commonly applied LBVS methods,[40] along with 3D pharmacophore search methods.[41] Recent advances in FP-based searches include turbo similarity searching,[42,43] a powerful method that uses the nearest neighbors of the template molecules, as determined in a first similarity search, as the templates for a second similarity search. A study using bioisosteric equivalence, i.e., a "local" similarity concept, in context with VS has been published Birchall et al.[44] This study is based on reduced graphs (RG) that are matched using clique detection. The reduction scheme, i.e., the rules for transforming groups of atoms into objects that embody the nodes of the RGs, already represents bioisosteric equivalence to some extent. Furthermore, bioisosteric information has been explicitly incorporated into the RG nodes by allowing the nodes to represent known bioisosteric fragments. Probably the most well-known variants of RGs are feature trees[45] and pharmacophore graphs.[46] Feature trees and pharmacophore graphs are reduced graphs, which use a reduction scheme that condenses atom groups into the nodes of a tree, representing molecular features such as donors, acceptors, aromatic moieties, etc. These trees can be considered as generic 2D pharmacophores. Graph-theoretical methods enable rapid comparison of template (query) structure trees with trees calculated from compounds in large compound databases.[46] These methods somehow embrace both "local" and "global" views on similarity because the reduction of atom groups to nodes representing pharmacophoric features is local, while the comparison and scoring of entire trees is a global concept. More recently, shape-based search methods[47−49] and high-dimensional similarity searches[40] have been established as powerful tools for LBVS that could be used to complement the approaches described above. However, these have not yet been evaluated in the context of the concepts presented here, and are therefore out of the scope of this article.

There is consensus that the success of VS is hard to predict and that it is difficult to gauge a priori which VS method is best suited to a particular project situation.[22] In general, it is believed that the parallel use of various VS approaches promises a higher rate of success.[50] This is not surprising given that different molecular descriptors describe different aspects of molecular structure, and hence, different methods probe chemical space in different ways.[51] Consequently, various methods and metrics have been used in parallel for LBVS. Several 2D and 3D descriptors for HTS follow-up have been used for a hit-directed nearest-neighbor searching approach.[21] Holliday et al. have applied 25 different similarity search methods simultaneously, followed by aggregation and analysis of the results.[52] The aim of this study was to systematically assess the assumption that multiple searches truly increase the likelihood of finding more active compounds, which was found to be correct. Aggregation rules are required to assess and rank the combined results; this procedure is referred to as data fusion. The article describes the application of two data fusion methods, group fusion and similarity fusion, and provides justification for the use of data fusion in VS. Other, more advanced data fusion approaches for aggregating multi-search results have been developed, including theoretically derived fusion rules[53,54] and a probabilistic framework that applies belief theory.[55] Data fusion approaches have also recently been reviewed by Willet.[56]

In this paper, we present an approach designed to facilitate fast hit expansion: LBVS using the structures of validated hits from a primary screening campaign as template (query, seed) structures. We have developed a fully integrated hit expansion search tool, referred to as the Hit Expansion Toolbox (HE Toolbox), which combines various similarity metrics and search methods, and performs the searches in parallel and in multiple consecutive steps. The searches not only use the template structures but also virtualized molecules that are constructed from the templates using rules embodying various concepts of chemical similarity. Given the objective of establishing an approach that allows very fast turnaround of compound selection for HE, all methods incorporated in the HE Toolbox protocol have been selected such that the methods are generally applicable, highly automated and fast, without the need for further validation or customization. We have therefore excluded 3D-based methods such as SBVS and pharmacophore searches.

Table 1. Data Sets Were Extracted from the ChEMBL Database, with Information Corresponding to the Chemical Diversity of the Different Data Sets

target class	target name	Uniprot-ID	number of compounds	number of BM scaffolds	SSE	MACCS[a]	FCFP6[a]	ECFC4[a]
kinase	vascular endothelial growth factor receptor 2 (VEGFR2)	P35968	3,281	1,328	0.90	607	423	1,586
GPCR	melanin-concentrating hormone receptor 1 (MCHR1)	Q99705	1,284	648	0.90	249	174	625
nuclear receptor	estrogen receptor β (ERβ)	Q92731	834	525	0.95	370	367	579
protease	renin	P00797	2,154	838	0.87	186	115	593

[a] Clustering performed in Pipeline Pilot[60] by applying the maximum dissimilarity method for cluster center selection and using up to two rounds of cluster recentering. Distance cutoff values were 0.2, 0.5, and 0.75, respectively (MACCS, FCFP6, ECFC4), corresponding to Tanimoto similarities of 0.8, 0.5, and 0.25, respectively. The resulting number of clusters is given.

Simple searches incorporated in the HE Toolbox include FP-based similarity searches using different FPs. The HE Toolbox also includes an adaptation of parallel turbo similarity searches using different FPs. The pharmacophore graphs approach, described above, has also been incorporated. A simple data fusion scheme is presented that allows the results from multiple searches to be condensed into hit lists that define the follow-up hit expansion screening deck.

Following a detailed description of the methods, the utility of the HE Toolbox is demonstrated by an analysis of the retrieval of true active compounds based on data sets that were compiled using the ChEMBL database. Four targets representing four different target classes have been chosen for this assessment: protein kinase VEGFR2, GPCR MCHR1, nuclear hormone receptor ERβ, and aspartyl protease renin. Furthermore, two hit-finding case studies with the epigenetic targets G9a and LSD1 are briefly described.

2. MATERIALS AND METHODS

2.1. Data Sets. A local copy of the ChEMBL database[57] (version 13) was used for selecting and classifying all potency data according to their target family (using the "L2" Field). Only entries where the relationship operator between the potency field (STANDARD_VALUE = "KI") and the corresponding value was either "equal" or "less than" were retained. For this analysis, the four most common target families were kept: protein kinases, GPCRs, nuclear receptors, and proteases. Then, from each target family set, the target with the largest number of entries was retained, resulting in the data sets listed in Table 1.

Various measures for assessing the chemical diversity of the four data sets were calculated. These include the number of clusters obtained from FP-based clustering with the FPs used for similarity searches; the same similarity thresholds were used as for the similarity searches, see Table 2. Also, the number of different Bemis-Murcko scaffolds[58] was determined. Furthermore, the Scaled Shannon Entropy (SSE), a diversity measure that takes the frequency distribution of different scaffolds into account,[59] was evaluated for each data set. BM scaffolds were used for representing the cyclic systems. SSE values range between 0 and 1, with higher values indicating higher scaffold diversity.

From each set, a maximally diverse selection of 10 compounds was performed using FCFP4 fingerprints. For all targets, these templates corresponded to either 10 different BM scaffolds (VEGFR2, MCHR1) or nine different BM scaffolds plus one acyclic molecule (ERβ, renin). These compounds were used as the templates (seeds, queries) for the HE Toolbox analysis. In most of our HE projects, the template number ranged from 30 to 100. However, we have been performing HE projects with template numbers from one up to several hundreds. The relatively small number of 10 templates was chosen for assessing the effectiveness of the combined HE methods with only a few starting points.

Table 2. Tanimoto Similarity Threshold Values for Different Fingerprints Applied in This Study[a]

Similarity search from original templates	
MACCS	0.80
FCFP6	0.50
ECFC4	0.75
Similarity search after bioisosteric replacement	
MACCS	0.85
FCFP6	0.55
ECFC4	0.80
Similarity search after ring expansion/contraction	
MACCS	0.80
FCFP6	0.50
ECFC4	0.75
Turbosim search from original templates	
MACCS	0.9
FCFP6	0.6
Turbosim search after bioisosteric replacement	
MACCS	0.95
FCFP6	0.65
Turbosim search after ring expansion/contraction	
MACCS	0.95
FCFP6	0.65

[a] Different thresholds were set depending on the template virtualization methods.

As a negative set of putatively inactive compounds, 500,000 compounds from the entire Screening Compounds Directory[61] comprising a total of 9.7 million compounds were randomly selected using a Pipeline Pilot protocol.[60] The number of Bemis-Murcko scaffolds in this negative set was 182,775. Since the inactivity of these compounds was not confirmed experimentally, a certain proportion of the compounds could be, unknowingly, active. Scior et al[28] have discussed the pitfalls associated with such situations in VS. However, in the context of our study, each compound fished from the negative set that is in fact active would decrease the number of false negatives and hence increase the number of correctly found true positives. Given a hypothetical hit rate of 0.1% in HTS, this could correspond to 500 compounds. However, given the arbitrary mix of active compounds from the ChEMBL database with a large diverse screening deck, this should only be

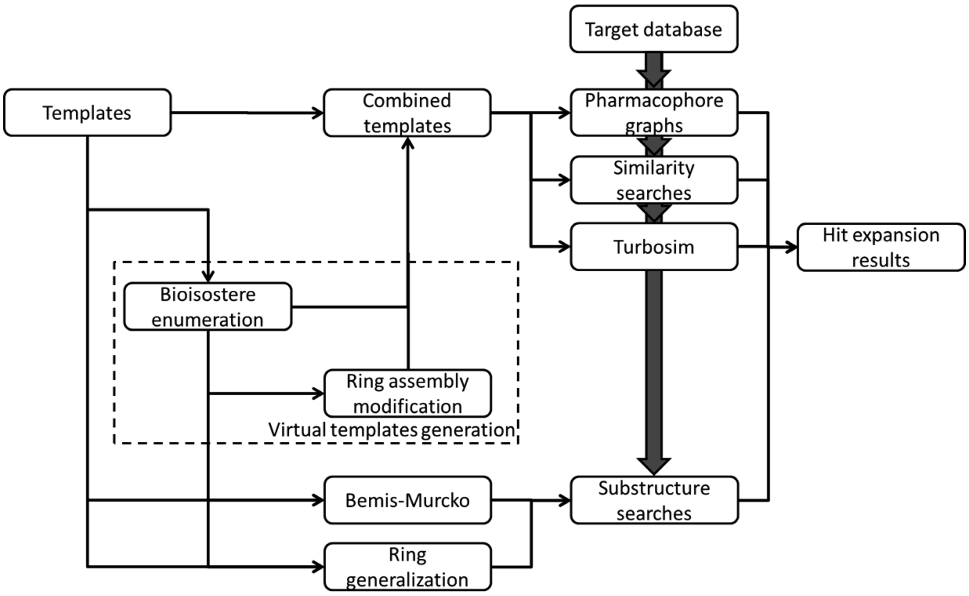


Figure 1. Schematic representation of the workflow for template virtualization and different search methods, as implemented in the HE Toolbox. Initial templates and virtual templates are combined before being subjected to various similarity searches that are executed in parallel. Similarity searches and Turbosim searches are executed with different fingerprint sets. Results are aggregated using data fusion techniques in a separate protocol. See text for further details.

considered as a hypothetical benchmark in order to get a rough idea of the impact of this effect.

2.2. Hit Expansion Toolbox. The HE Toolbox is a modular implementation of the template virtualization and multi-method similarity search approach that is described in this paper. The combination and parallel execution of all the methods have been implemented in Pipeline Pilot;[60] the workflow is illustrated in Figure 1. The HE Toolbox comprises the following modular components.

Substructure Searches Using Bemis-Murcko Scaffolds. Bemis-Murcko (BM) ring assemblies[58] were generated from the input templates using the *Generate Fragments* component of Pipeline Pilot. Exocyclic and linker double bonds were not retained, according to default settings. All BM assemblies were subsequently used as substructure search queries.

Similarity Searches. Fingerprint-based similarity searches were performed using MDL MACCS keys,[62] FCFP6,[63] and ECFC4 fingerprints. The Tanimoto coefficient was used as the similarity metric, as it has been demonstrated to be the metric of choice for LBVS.[43] Fragment-occurrence fingerprints such as ECFC4 have been included in the workflow based on a recent study that demonstrates higher effectiveness of this fingerprint type in similarity-based searches, when compared with standard binary fingerprints.[64] Defining appropriate thresholds for similarity searches is nontrivial, and relies on a chemically meaningful notion of significant similarity and the database studied.[65] A simple empirical approach has been successfully applied in many of our HE projects. Similarity thresholds were chosen such that the numbers of compounds retrieved with each fingerprint search was balanced, i.e., in the same order of magnitude. The threshold values applied in this study are listed in Table 2; these correspond to the default settings commonly used in our projects. It is noteworthy that the selected threshold values for the MDL MACCS keys and FCFP6 fingerprints virtually coincide with the threshold values used in a study by Muchmore et al. for rescuing 50% of all true actives (0.82 for MDL MACCS keys and 0.45 for FCFP6 fingerprints).[55]

Turbo Similarity. The turbo similarity algorithm (Turbosim) was implemented as described in the literature.[42,43] After the first similarity search, up to 50 of the top ranking compounds with a similarity value to the template >0.3 were kept, as recommended.[42] After the second search round, based on the results from the first round (the new templates), the MAX fusion rule[66] was applied, i.e., the larger similarity value from both the first and the second search step was used as the MAX score. All compounds with a final MAX similarity score above a predefined threshold (as given in Table 2) were kept. Searches were performed using FCFP6 fingerprints and MDL MACCS Keys.

Template Virtualization and Multi-Step Searches. The concept of turbo similarity searching mentioned above, i.e., the idea of using first nearest neighbor search results as templates for searching second nearest neighbors, inspired us to incorporate multiple step searches in the system. Furthermore, the multi-step searches are not restricted to the search templates as the starting points. Instead, the template (query, seed) structures are also transformed into virtual templates that are subsequently used for executing the similarity searches described above, in addition to searching with just the templates. The transformation rules for template virtualization are aimed at mimicking the viewpoint of "local" similarity that a medicinal chemist would apply in lead optimization. From this perspective, templates and virtual templates are similar, although the transformation can bridge a gap in chemical space. Virtual templates hence represent interesting additional seeds for exploring the chemical space.

One virtualization rule is the simple enumeration of the bioisostere templates using a dictionary. Another set of rules that has been encoded by SMIRKS strings[67] involves the expansion and contraction of ring assemblies, leaving the R-groups unchanged. Also, generalized rings that are used for substructure pattern searches, a special case of bioisosterism, are used in the system.

By combining all these methods, both "global" and "local" metrics are incorporated into the VS search workflow. The different searches, together with real and virtualized template molecules, generate a complex network of search paths that sample different representations of chemical space, in order to detect regions that contain further active molecules. The following three virtualization concepts have been implemented in the HE Toolbox:

- *Bioisosteric enumeration*: In this study, the bioisosteric transformation rules published by Wagener and Lommerse[68] were applied to the templates, resulting in an enumeration of virtual templates. The *Enumerate Bioisosters* component of Pipeline Pilot was used. All compounds generated by either one single transformation or all possible transformations were used as additional virtual search templates.
- *Ring generalization ("Genring")*: An approach that works inversely to substructure searches with Bemis-Murcko scaffolds, i.e., the removal of all side groups, is ring generalization. Here, the side groups of the template structures are kept, and all atoms in the ring systems are transformed into unspecified atoms. These substructure patterns are then used as queries. For this, SMARTS[69] strings were generated from the input templates by converting the type of all ring atoms and bonds to atom type or bond type "Any". This approach can be considered as a special case of bioisosteric replacement, and fundamentally mimics the exchange of the core scaffold while retaining the topology of the ring system. This is a common scaffold-hopping technique that medicinal chemists apply in lead optimization projects, see a recent publication on B-Raf kinase inhibitors as an example.[70]
- *Ring assembly modification*: Another classical scaffold-hopping approach is to modify the size of rings in fused ring systems. For instance, a typical bioisosteric transformation used for scaffold hopping is to replace an indole ring system with a quinoline ring system, i.e., to expand a [6.5] to [6.6] ring system. Conversely, the conversion of a quinoxaline into a benzimidazole ring system represents a [6.6] to [6.5] ring contraction, see Figure 2. These types of transformations were encoded as SMIRKS-based ring assembly modification rules.[67] For ring expansions, carbon and nitrogen atom insertions were allowed, whereas for ring contractions, only the removal of carbon and nitrogen ring atoms without exocyclic substituents was permitted. Template structures were processed with these rules for generating new virtual templates. In contrast to the generalized rings which were used for substructure searches, the modified ring assemblies were used for true similarity searches.

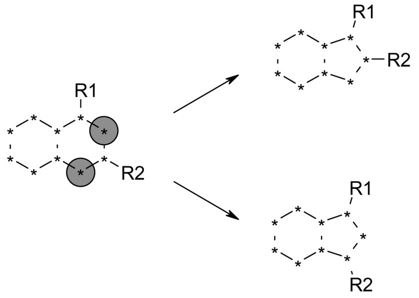


Figure 2. Example of [6.6] to [6.5] ring contractions. Contraction, i.e., the deletion of ring atoms, is only applied to unsubstituted positions.

Pharmacophore Graphs. DiscNgine's Chemistry Collection[46] includes pharmacophore graph searching methods that are similar to the concept of feature trees,[45] and also incorporates newer extensions such as graph reduction methods.[44] Graph-theoretical methods enable the rapid comparison of different pharmacophore trees and for similarity searches to be performed against a vast compound database. The default settings of the Pipeline Pilot implementation for this package were applied against the template structures. All hits with a "best score" ≥ 0.75 (DiscNgine default scoring function) were kept.

2.3. Post-Processing and Data Fusion.

In order to aggregate the results from all the different searches, a simple data fusion approach was applied similar to the group and similarity fusion scheme introduced by Holliday et al.[52] However, the search methods used in the HE Toolbox represent different metrics, namely, binary metrics (match/no match) for substructure searches with BM-scaffolds and generalized ring substructure searches and real number metrics for all other score-associated methods. Hence, a data fusion scheme based on just the ranking of real numbers was not possible. Furthermore, the data fusion scheme presented here aims to control the impact of different template structures on the results, and allows the balance between different templates to be governed, thus avoiding over-representation of certain templates. As a result, all matches for non-score associated methods and all hits found according to globally set thresholds for score-associated methods were collected, merged, and kept for further analysis. For searches with virtualized templates, it is possible that the virtually enumerated structures generated from one template produce multiple hits. For example, a subset of the bioisostere enumeration produced from one template can be similar to a given structure, while the inverse set is not similar (given a search threshold). In such cases, only the similarity value produced by the most similar virtual search structure is kept, referring back to the original template. Similarly, multiple matches of virtual template ensembles from substructure searches are aggregated to one match per template. Currently, there is no tracking of individual virtual templates but rather tracking to initial templates and methods.

The prototypic data fusion workflow consists of, first, the initial removal of unspecific search results obtained from substructure searches and, second, the definition of the three subsets (set 1−3) described below; this method has also been applied to the test sets. Depending on the results and requirements, specific adaptations can easily be incorporated, such as template prioritization schemes.

- Removal of unspecific search results: For relatively small template molecules, the corresponding Bemis-Murcko scaffolds can be very small (e.g., benzene, pyridine); hence, the subsequent substructure searches can produce large numbers of hits. The same holds true for substructure searches with generalized rings. Such cases represent situations in which the substructure definition is not specific enough and the BM scaffold is not a meaningful scaffold approximation; hence, the results of such searches are omitted. An omission threshold T_{omit}

Table 3. Number of True Active and False Positive Compounds Retrieved as Obtained at Different Stages of the Search Procedure for All Four Target Class Data Sets[a]

		initial retrieval	after removing unspecific queries	data fusion total (%) [set1/set2/set3]
ERβ				
	actives	442 (53.6)	430 (52.2)	322 (39.1) [209/58/55]
	inactives	11,512 (2.3)	10,150 (2.0)	2,550 (0.5) [224/76/2,250]
	retrieval rate (%)	3.7	4.1	11.2
VEGFR2				
	actives	1,550 (47.4)	1,470 (44.9)	1,093 (33.4) [557/270/266]
	inactives	57,334 (11.5)	36,514 (7.3)	4,147 (0.8) [931/484/1,448]
	retrieval rate (%)	2.6	3.9	20.9
MCHR1				
	actives	938 (73.6)	938 (73.6)	802 (63.0) [629/58/112]
	inactives	37,238 (7.5)	37,130 (7.4)	6,883 (1.4) [1,811/1,318/3,754]
	retrieval rate (%)	2.5	2.5	10.4
Renin				
	actives	1,987 (92.7)	1,840 (85.8)	1,624 (75.8) [1,189/229/206]
	inactives	140,268 (28.1)	22,185 (4.4)	4,302 (0.9) [672/645/2,985]
	retrieval rate (%)	1.4	7.7	27.4

[a] The percentage of compounds retrieved, in relation to the total number of compounds in the respective data sets, is given in parentheses. For the final results, the number of compounds retrieved by the individual data fusion methods (method consensus set (set 1), template consensus set (set 2), balanced template representation set (set 3)) are also indicated in square brackets.

can be set independently for each search type. T_{omit} specifies the number of hits (per method and per template) above which the search results are completely ignored. The default value for T_{omit} for both Bemis-Murcko scaffold and Genring substructure searches was 100.

- Following the removal of unspecific search results, the first set (set 1) is defined. This "method consensus set" comprises all compounds that have been independently found by at least a given number N of search methods. This follows the similarity fusion concept by Willet;[43] however, only the number of hits per compound and method is counted. In the present study, N was set to three.
- The second set (set 2), referred to as the "template consensus set", comprises all remaining compounds that were found independently by two or more templates. This represents a simplified implementation of the group fusion concept.
- For all remaining compounds, i.e., those that were not included in either the method or the template consensus set, the following procedure is applied on a per method and per template basis in order to generate a third set called the "balanced template representation set" (set 3):
    - For all score-associated searches, the N top-ranking compounds were kept. Different thresholds defining this number can be set for all individual search methods (DiscNgine, DiscNgine-bioiso, FP, FP-bioiso, FP-Turbo with FP = MACCS, FCFP6 resp. ECFC4). The default value N for all of these parameters is 50.
    - For searches without scores, i.e., substructure searches with binary match/no-match results, a maximally diverse subset of up to T_{div} compounds was kept per method and template. Given that, as described above, unspecific search results were completely omitted depending on the threshold T_{omit}, the two thresholds T_{div} and T_{omit} govern the selection of hits from the substructure searches on a per method and per template basis. The values of T_{div} and T_{omit} can be set independently for each search type. If the number of hits is smaller than or equal to T_{div}, all hits are kept. If the number of hits is between T_{div} and T_{omit}, a maximal diverse subset of T_{div} hits is selected and kept. If the number of hits is larger than T_{omit}, all hits are omitted. Default values for T_{omit} and T_{div} for both Bemis-Murcko scaffold and Genring substructure searches are 100 and 50, respectively.

3. RESULTS AND DISCUSSION

3.1. Analysis of ChEMBL Data Sets. The HE Toolbox, i.e., the protocol for template virtualization and execution of all similarity search methods described above in parallel, was used for assessing the retrieval of true active compounds from each of the four ChEMBL data sets. In all cases, 10 selected compounds were used as the template structures for searching against the corresponding data set (following omission of the template structures). Each set of templates was also used as input against a diverse subset of the SCD database comprising 500,000 compounds in order to obtain an estimate of the potential false positive hit rate. The numbers of compounds retrieved initially, after removing unspecific searches, and after the full data fusion process are given in Table 3. In all cases, the proportion of true active compounds retrieved from the ChEMBL data sets was significantly higher than the proportion of (presumably) false positive compounds retrieved from the SCD data set. For example, 442 out of 834 true active ERβ (53.6%) could be rescued with the HE Toolbox, while only 2.3% of the SCD compounds were found at the same time. For renin, these proportions were 92.7% and 28.1% for true positives and false positives, respectively. In this case, unspecific substructure searches representing small molecules such as benzene or pyridine, resulting from BM-scaffold/Genring substructure searches, produced a very high number of hits. With the removal of unspecific queries, the true and false positive retrieval rates obtained were 85.8% and 4.4%, and full data fusion figures were 75.8% and 0.9%, respectively. In all cases, the proportion of initially retrieved true actives was significantly higher than the proportion of false positives (inactives). Removal of unspecific queries and subsequent data fusion dramatically improved the retrieval rate of true actives. For example, the method consensus set (set 1) for ERβ comprised 209 active (true positives) compounds and only 224 inactive compounds (false positives); a similar positive prediction value could be achieved for all four data sets. For the template consensus set (set 2), the situation was overall comparable; however, for MCHR1, 58 versus as much as 1,318 true and false positives were retrieved, respectively. As expected, this ratio was significantly smaller for the balanced template representation set (set 3).

Furthermore, as can be seen from analyzing the combined results in Tables 3 and 4, the proportion of Bemis-Murcko scaffolds in the final sets is generally very similar to the proportion of hits obtained. This shows that the high hit rates obtained are not solely due to the retrieval of single large scaffold families belonging to the same chemotype, but that chemical diversity is retained throughout the compound selection and data fusion processes.

Table 4. Number of Bemis-Murcko Scaffolds Retrieved at Different Stages of the Searching Procedure for All Four Data Sets[b]

		total	initial retrieval	data fusion[b]
ERβ				
	actives	525	227 (43.2)	179 (34.1) [124/42/37]
	inactives	18,2775	5,938 (3.2)	1,491 (0.8) [35/60/1,368]
VEGFR2				
	actives	1,328	604 (45.5)	454 (34.2) [398/139/146]
	inactives	18,2775	20,317 (11.1)	2,087 (1.1) [398/484/1,448]
MCHR1				
	actives	648	482 (74.4)	415 (64.0) [309/53/88]
	inactives	18,2775	18,439 (10.1)	3,963 (21.7) [994/826/2,562]
renin				
	actives	838	796 (95.0)	636 (75.9) [450/135/117]
	inactives	182,775	74,673 (40.9)	2,669 (14.6) [384/476/1,992]

[a] Since a scaffold can be retrieved by more than one method, the sum of the scaffold count for individual methods is larger than the total number of unique scaffolds. Percentages as compared with the complete data set are given in parentheses. [b] For the final results (complete data fusion), the number of BM scaffolds retrieved by the individual data fusion methods (method consensus set (set 1), template consensus set (set 2), balanced template representation set (set 3)) are also given in square brackets.

The combination of bioisosteric enumeration and other virtualization techniques of course generates large numbers of virtual molecules, many of which are chemically, and particularly in the context of drug discovery, meaningless. However, the system will only find a hit molecule if the last step of a search path finds a real molecule present in the database; hence, virtualization steps leading away from real molecules do not produce any hits. Clearly, not all bioisosteric and other transformations comply with the stringent criteria applied to bioisosteric replacement in LO, such as 3D shape similarity, matching number of H-bonds, etc. This holds true even for "LO-compatible" bioisosteric transformations. As argued by Sheridan,[50] a carboxylate may bind to two target structures A and B. This does not, however, imply that the bioisosteric tetrazole derivative binds as well to both targets. This caveat has been accepted knowingly, given that the aim of the approach was to quickly select compounds for a fast turnaround hit expansion campaign, and that the likelihood of finding active compounds using bioisosteric replacement should be significantly higher than the selection of random compounds or compound with low fingerprint similarity.

The power of this sequential approach involving the generation of virtual templates is illustrated in Figure 3. This example is taken from the VEGFR2 data set. Bioisosteric enumeration of (A) led to 69 virtual bioisosteric compounds, four of which being template structures that found true hits in subsequent Turbosim searches. The bioisosteric transformation "Amide to Oxadiazole" leading to the true active molecules with the highest MAX similarity converts the keto-sulfonamide template structure (A) into the virtual oxadiazole template (B). Subsequently, an oxazole (C) similar to (B) is found in the first step of a Turbosim (FCFP6) search. In the second Turbosim (FCFP6) search step, the similar oxazole (D) with different decorator groups attached to one of the phenyl rings is found. While the similarity and the existence of common pharmacophore features in the molecule pairs A−B, B−C, and C−D is evident, A and D appear to be quite different. The Tanimoto similarity between A and D (FCFP6) is only 9%.

3.2. Case Studies for Epigenetic Targets.

The utility of the approaches presented here has also been demonstrated in two hit finding projects for epigenetic targets. In a case study by Ahrens et al.[71] for finding inhibitors for the protein lysine methyl transferase G9a, a predecessor of the HE Toolbox was used for a LBVS-based selection. Two known inhibitors that bind to the substrate binding site, and a cofactor analogue that binds to the cofactor binding site, were used as the template structures. A small screening deck of 1,242 compounds was selected from the entire BioFocus compound collection comprising approximately 900,000 compounds using all the search methods described in this article. (In addition, a subset of compounds was selected using SBVS in this case study.) However, simpler data fusion rules based on graded post-processing were applied. A combined hit rate of 5.4% percent was achieved in this study (3.4% for pharmacophore graphs, and 6.9% for the other methods of the HE Toolbox).

A second case study by Beyer et al.[22,72,73] was carried out to identify inhibitors of the histone demethylase LSD1. The primary screening deck consisted of two subsets each comprising 656 compounds. The first subset was determined by SBVS, i.e., docking in the FAD cofactor binding site. The BioFocus compound collection comprising 870,000 compounds at the time of the study were used. The second subset consisted of fragments from the BioFocus fragment library. A total of 33 experimentally validated hits for which dose response curves could be fitted were subsequently used as templates for hit expansion using the HE Toolbox approach as described in this paper. The hit expansion led, finally, to 27 further validated hits with the highest activity being 0.2 μM. It was interesting that hit expansion hits derived from fragment deck templates exhibited much better lipophilic effiency[74] than those derived from the SBVS deck. This raises the question as to whether or not a more direct path from fragments to drug-like molecules by means of hit expansion could help to advance

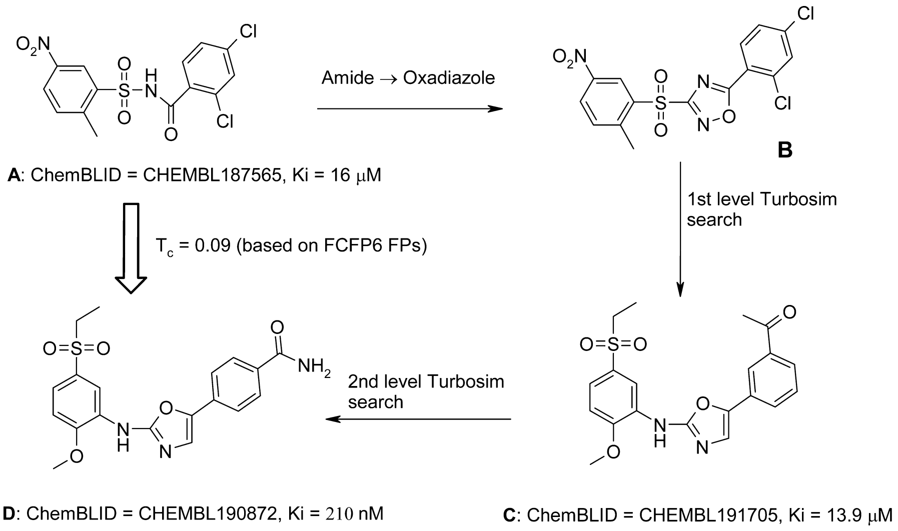

Amide → Oxadiazole
A: ChemBLID = CHEMBL187565, Ki = 16 μM
B
Tc = 0.09 (based on FCFP6 FPs)
1st level Turbosim search
2nd level Turbosim search
D: ChemBLID = CHEMBL190872, Ki = 210 nM
C: ChemBLID = CHEMBL191705, Ki = 13.9 μM

Figure 3. Example of structures involved in a search pathway leading to an active structure (D) that would not be found with simple fingerprint or pharmacophore graph-based similarity searches.

early drug discovery, when compared to mainstream fragment-based drug discovery.[75−77]

3.3. Outlook. There are many ways in which the search methods reported here could be extended: Generic similarity transformations as applied in LO could easily be added by extending the SMIRKS-based transformation rules. Such rules could be established depending on the target classes and the typically preferred synthesis paths applied in the respective medicinal chemistry laboratories. Shape-based searches could be incorporated into the workflows as a further complementary LBVS method, and these searches could then easily be combined with the template virtualization methods. Initial investigations indicate that shape-based methods are likely to retrieve additional true active compounds from test data sets (unpublished results). Additional filtering steps using data shaving[78] or naïve Bayesian classification[79] methods could be applied for deselecting inactive or inappropriate compounds. Data fusion could be enhanced along the lines of the references provided earlier or by using Pareto methods.[80] Furthermore, the determination of similarity thresholds could be advanced by analysis of compound sets retrieved with the various search methods, along the lines of a recent publication by Medina-Franco.[65]

4. CONCLUSIONS

This paper describes a chemoinformatics-driven hit expansion approach that applies various similarity search methods (fingerprint based, Turbosim, pharmacophore trees) in parallel. The approach allows additional active compounds and novel scaffolds in a library to be found quickly, starting from validated HTS hits used as the template (seed or query) structures. In addition to using only the template structures present in the screening libraries, the virtualization of template molecules generates additional starting points for probing the active chemical space. The virtualization rules applied aim to mimic some of the concepts that medicinal chemists use in LO, in particular, the concept of bioisosteric replacement is incorporated in various ways.

Our study confirms the notion that the parallel execution of multiple VS methods, in combination with data fusion, enhances the performance of VS. The extension of VS workflows using template virtualization generates similarity search paths in chemical space that are not, as far as we can see, accessible using conventional methods. Clearly, the enumeration of virtualized template sets by applying all the different transformation methods generates large numbers of individual search paths, leading to uncontrollably large number of hits. However, this can be controlled by using consensus search paths through chemical space that lead to the same hits. Such control is intrinsically implemented by the data fusion rules that appear to impose the restrictions that are necessary for reducing the search space, and hence for selecting meaningful HE hits. Our approach has been validated by four test sets using different target classes. Furthermore, we have successfully applied this approach or variations of this approach in a number of hit finding projects, including the two case studies on epigenetics targets briefly described. This also demonstrates the suitability of the approach for LBVS. In summary, we conclude that the knowledge about active chemical space can be extended by combining different concepts of chemical similarity, including bioisosteric replacement and related methods applied in LO. Template virtualization can facilitate the finding of active areas in chemical space that cannot be accessed easily by conventional similarity search methods.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Detailed retrieval matrices (method × template) for each target and each stage of the selection and fusion process. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: andreas.bergner@boehringer-ingelheim.com.

Present Addresses

†Andreas Bergner, Boehringer-IngelheimRCV GmbH & Co KG, Dr. Boehringer Gasse 5-11, 1120 Vienna, Austria.
‡Serge P. Parel: Exquiron Biotech AG, Kägenstrasse 17, 4153 Reinach, Switzerland.

Author Contributions

The manuscript was written through contributions of both authors. Both authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Kate Hilyard and Chris G. Newton for valuable feedback on the manuscript. We thank Werner Stürmer for his constant support. We also thank Karen J. Bergner for careful revision of the manuscript and for help with the English.

■ ABBREVIATIONS

BM, Bemis-Murcko (scaffold); FP, fingerprint; HE, hit expansion; HTS, high-throughput screening; LBVS, ligand-based virtual screening; LO, lead optimization; RG, reduced graph; SBVS, structure-based virtual screening; SSE, scaled Shannon entropy; VS, virtual screening

■ REFERENCES

(1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.

(2) Mayr, L. M.; Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580−588.

(3) Bender, A.; Bojanic, D.; Davies, J. W.; Crisman, T. J.; Mikhailov, D.; Scheiber, J.; Jenkins, J. L.; Deng, Z.; Hill, W. A.; Popov, M.; Jacoby, E.; Glick, M. Which aspects of HTS are empirically correlated with downstream success? *Curr. Opin. Drug Discovery Devel.* **2008**, *11*, 327−337.

(4) Gribbon, P.; Sewing, A. High-throughput drug discovery: What can we expect from HTS? *Drug Discovery Today* **2005**, *10*, 17−22.

(5) Shoichet, B. K. Screening in a spirit haunted world. *Drug Discovery Today* **2006**, *11*, 607−615.

(6) Snowden, M.; Green, D. V. The impact of diversity-based, high-throughput screening on drug discovery: "Chance favours the prepared mind". *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 553−558.

(7) Wölcke, J.; Ullmann, D. Miniaturized HTS technologies – uHTS. *Drug Discovery Today* **2001**, *6*, 637−646.

(8) Clark, D. E. What has virtual screening ever done for drug discovery? *Exp. Opin. Drug Discovery* **2008**, *3*, 841−851.

(9) Koppen, H. Virtual screening: What does it give us? *Curr. Opin. Drug Discovery Dev.* **2009**, *12*, 397−407.

(10) Lipkin, M. J.; Stevens, A. P.; Livingstone, D. J.; Harris, C. J. How large does a compound screening collection need to be? *Comb. Chem. High Throughput Screening* **2008**, *11*, 482−493.

(11) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273−276.

(12) Langer, T.; Hoffmann, R.; Bryant, S.; Lesur, B. Hit finding: towards 'smarter' approaches. *Curr. Opin. Pharmacol.* **2009**, *9*, 589−593.

(13) Davies, J. W.; Glick, M.; Jenkins, J. L. Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr. Opin. Chem. Biol.* **2006**, *10*, 343−351.

(14) Harper, G.; Pickett, S. D. Methods for mining HTS data. *Drug Discovery Today* **2006**, *11*, 694−699.

(15) Posner, B. A.; Xi, H.; Mills, J. E. Enhanced HTS hit selection via a local hit rate analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2202−2210.

(16) Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J. H.; Raman, P.; Ertl, P.; Schuffenhauer, A. Compound set enrichment: A novel approach to analysis of primary HTS data. *J. Chem. Inf. Model.* **2010**, *50*, 2067−2078.

(17) Varin, T.; Didiot, M. C.; Parker, C. N.; Schuffenhauer, A. Latent hit series hidden in high-throughput screening data. *J. Med. Chem.* **2012**, *55*, 1161−1170.

(18) Hassan, M.; Brown, R.; Varma-O'Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol. Divers.* **2006**, *10*, 283−299.

(19) Oprea, T. I.; Bologa, C. G.; Edwards, B. S.; Prossnitz, E. R.; Sklar, L. A. Post-high-throughput screening analysis: an empirical compound prioritization scheme. *J. Biomol. Screening* **2005**, *10*, 419−426.

(20) Baringhaus, K. H.; Hessler, G. Fast similarity searching and screening hit analysis. *Drug Discovery Today: Technol* **2004**, *1*, 197−202.

(21) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. *J. Med. Chem.* **2005**, *48*, 240−248.

(22) Glick, M.; Jacoby, E. The role of computational methods in the identification of bioactive compounds. *Curr. Opin. Chem. Biol.* **2011**, *15*, 540−546.

(23) McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494−502.

(24) Muegge, I.; Oloff, S. Advances in virtual screening. *Drug Discovery Today: Technol* **2006**, *3*, 405−411.

(25) Rester, U. From virtuality to reality - Virtual screening in lead discovery and lead optimization: A medicinal chemistry perspective. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 559−568.

(26) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.

(27) Schnecke, V.; Bostrom, J. Computational chemistry-driven decision making in lead generation. *Drug Discovery Today* **2006**, *11*, 43−50.

(28) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing pitfalls in virtual screening: A critical review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.

(29) Klebe, G. Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−594.

(30) Waszkowycz, B. Towards improving compound selection in structure-based virtual screening. *Drug Discovery Today* **2008**, *13*, 219−226.

(31) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, *16*, 372−376.

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

(33) 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*, 3049−3059.

(34) Wermuth, C. G. Molecular Variations Based on Isosteric Replacements. In *The Practice of Medicinal Chemistry*, 2nd ed.; Wermuth, C. G., Ed.; Academic Press: London, 2003; pp 189−214.

(35) Ertl, P. In silico identification of bioisosteric functional groups. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 281−288.

(36) Jakobi, A. J.; Mauser, H.; Clark, T. ParaFrag—An approach for surface-based similarity comparison of molecular fragments. *J. Mol. Model.* **2008**, *14*, 547−558.

(37) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103−108.

(38) Wagener, M.; Lommerse, J. P. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.

(39) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. SwissBioisostere: A database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41*, D1137−D1143.

(40) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225−233.

(41) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539−558.

(42) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *J. Med. Chem.* **2005**, *48*, 7049−7054.

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

(44) Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1330−1346.

(45) Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des* **1998**, *12*, 471−490.

(46) *DiscNgine*, version 1.0.4. DiscNgine: Romainville, France, 2012.

(47) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular overlay based on shape and electrostatic potential. *J. Chem. Inf. Model.* **2009**, *49*, 492−502.

(48) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.

(49) *ROCS*, version 3.1.2. Open Eye Scientific Software: Santa Fe, NM, 2013.

(50) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(51) 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*, 108−119.

(52) Holliday, J. D.; Kanoulas, E.; Malim, N.; Willett, P. Multiple search methods for similarity-based virtual screening: Analysis of search overlap and precision. *J. Cheminf.* **2011**, *3*, 29 DOI: 10.1186/1758-2946-3-29.

(53) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: Similarity and group fusion. *J. Chem. Inf. Model.* **2006**, *46*, 2206−2219.

(54) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: theoretical model. *J. Chem. Inf. Model.* **2006**, *46*, 2193−2205.

(55) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941−948.

(56) Willett, P. Combination of similarity rankings using data fusion. *J. Chem. Inf. Model.* **2013**, *53*, 1−10.

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

(58) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(59) Medina-Franco, J. L.; Martinez-Mayorga, K.; Bender, A.; Scior, T. Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR Comb Sci* **2009**, *28*, 1551−1560.

(60) *Pipeline Pilot*, version 8.5.0.200. Accelrys: San Diego, CA, 2012.

(61) *Screening Compounds Directory*, version 2011.4, Accelrys: San Diego, CA, 2011.

(62) 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*, 1273−1280.

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

(64) Arif, S.; Holliday, J.; Willett, P. Analysis and use of fragment-occurrence data in similarity-based virtual screening. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 655−668.

(65) Medina-Franco, J. L. Scanning structure-activity relationships with structure-activity similarity and related maps: from consensus activity cliffs to selectivity switches. *J. Chem. Inf. Model.* **2012**, *52*, 2485−2493.

(66) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(67) *SMIRKS*. http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed December 10, 2012).

(68) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.

(69) *SMARTS*. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed December 10, 2012).

(70) Gopalsamy, A.; Shi, M.; Hu, Y.; Lee, F.; Feldberg, L.; Frommer, E.; Kim, S.; Collins, K.; Wojciechowicz, D.; Mallon, R. B-Raf kinase inhibitors: Hit enrichment through scaffold hopping. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2431−2434.

(71) Ahrens, T.; Bergner, A.; Sheppard, D.; Hafenbradl, D. Efficient hit-finding approaches for histone methyltransferases: the key parameters. *J. Biomol. Screen.* **2012**, *17*, 85−98.

(72) Beyer, K.; Rye, P.; Fasler, S.; Hafenbradl, D.; Bergner, A. Identification and Characterization of Inhibitors of the Histone Demethylase LSD1. Presented at Miptec, Basel, Switzerland, September 24−29, 2012.

(73) Bergner, A.; Allen, V.; Beyer, K. Making Epigenetic Target Screening Smart: Computational Compound Selection and Hit Expansion Approaches for Identifying Inhibitors of LSD1. Presented at Miptec, Basel, Switzerland, September 24−29, 2012.

(74) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.

(75) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51*, 3661−3680.

(76) Hajduk, P. J. Fragment-based drug design: How big is too big? *J. Med. Chem.* **2006**, *49*, 6972−6976.

(77) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660−672.

(78) Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. Data shaving: A focused screening approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 470−479.

(79) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682−686.

(80) Svensson, F.; Karlén, A.; Sköld, C. Virtual screening data fusion using both structure- and ligand-based methods. *J. Chem. Inf. Model.* **2011**, *52*, 225−232.