Bioorganic & Medicinal Chemistry 20 (2012) 5379–5387

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Hole filling and library optimization: Application to commercially available fragment libraries

Yuling An, Woody Sherman, Steven L. Dixon *

Schrödinger Inc., 120 West 45th Street, New York, NY 10036, United States

ARTICLE INFO

Article history:
Available online 24 March 2012

Keywords:
Cheminformatics
Chemical fingerprint
Compound library optimization
Hole filling

ABSTRACT

Compound libraries comprise an integral component of drug discovery in the pharmaceutical and biotechnology industries. While in-house libraries often contain millions of molecules, this number pales in comparison to the accessible space of drug-like molecules. Therefore, care must be taken when adding new compounds to an existing library in order to ensure that unexplored regions in the chemical space are filled efficiently while not needlessly increasing the library size. In this work, we present an automated method to fill holes in an existing library using compounds from an external source and apply it to commercially available fragment libraries. The method, called Canvas HF, uses distances computed from 2D chemical fingerprints and selects compounds that fill vacuous regions while not suffering from the problem of selecting only compounds at the edge of the chemical space. We show that the method is robust with respect to different databases and the number of requested compounds to retrieve. We also present an extension of the method where chemical properties can be considered simultaneously with the selection process to bias the compounds toward a desired property space without imposing hard property cutoffs. We compare the results of Canvas HF to those obtained with a standard sphere exclusion method and with random compound selection and find that Canvas HF performs favorably. Overall, the method presented here offers an efficient and effective hole-filling strategy to augment compound libraries with compounds from external sources. The method does not have any fit parameters and therefore it should be applicable in most hole-filling applications.

1. Introduction

Compound libraries, both real and virtual, are an essential component of the drug discovery process. Pharmaceutical companies use compound libraries in high throughput screening (HTS),[1] virtual screening,[2–4] fragment-based discovery,[5] de novo design,[6] combinatorial chemistry,[7] natural product design,[8] selectivity profiling,[9] and other applications. Compound libraries at pharmaceutical companies continue to grow in size, with many of them in the millions. However, the number of commercially available compounds is greater than the number of compounds at individual pharmaceutical companies, and the number of available compounds continues to increase rapidly, as evidenced by the growing size of the Chemical Abstract Service (CAS), which recently reported over 63 million substances.[10] Furthermore, millions of compounds are available from vendor collections, such as ChEMBL,[11] ChemExper,[12] and eMolecules.[13] The Zinc database has compiled a large set of commercially available compounds and the full set is now approaching 20 million.[14] With the growing number of compounds come additional complexities associated with storage, management, and maintenance. Therefore, a desirable strategy is to augment existing libraries with judiciously chosen new compounds so as to control the growth rate, while maximizing the incremental value of each compound.

The value attributable to a new compound cannot be determined prior to its use in drug discovery efforts and therefore one must decide in advance on the criteria for selecting new compounds. One approach is to focus on chemical diversity, with the expectation that spanning a broad chemical space will maximize the probability that useful compounds will be retrieved. The exact implementation of the diversity method is important, since it is not always desirable to find compounds that simply maximize the chemical distance from existing compounds. For example, an algorithm focused exclusively on maximizing the average distance to existing compounds may return only compounds with anomalous and undesirable properties.[15] Such an approach, typically referred to as edge design, results in compounds being selected around the edges of chemical space, as illustrated in Figure 1B. While the new compounds are indeed different from the existing ones, they do not satisfactorily cover certain important regions of chemical space.

To account for this, a more direct approach can be taken to 'fill holes' in the existing library (Fig. 1C). This approach assumes that there is information in the existing compounds with respect to

* Corresponding author. Tel.: +1 212 295 5800; fax: +1 212 295 5801.
E-mail address: steven.dixon@schrodinger.com (S.L. Dixon).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.03.037

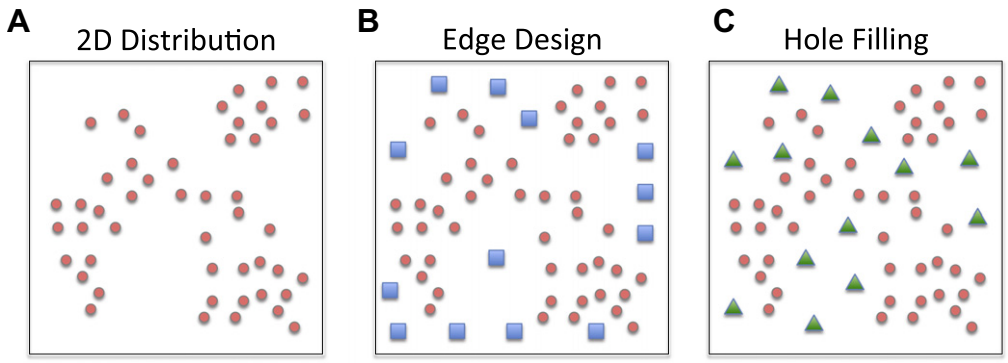

Figure 1. Examples of hole-filling strategies. (A) Illustration of an initial distribution of molecules in a reference compound library. The large dimensionality of the chemical space has been reduced to two dimensions in order to facilitate visualization. (B) New compounds added from an external library using an edge design approach (blue squares). (C) New compounds added from an external library using the Canvas HF hole-filling approach described in this work (green triangles).

desirability and therefore filling holes between existing compounds can yield a more suitable library selection. In addition, it is important to control the properties of new compounds during the selection process. While a pre-filter can be applied to ensure that all new structures conform to a predetermined set of rules, a preferable approach is to simultaneously optimize on diversity and properties. This ensures that interesting compounds are not lost simply because they fall slightly outside of the predefined rules. Furthermore, if multiple molecules can equally satisfy the hole-filling diversity objective, the ones with the more desirable properties should be retained.

In this work, we describe Canvas HF, a method for filling holes in chemical libraries according to the principles described above. The program, which is part of the Schrödinger Software Suite, combines a greedy ejection algorithm with simulated annealing to minimize nearest neighbor 2D fingerprint similarities among the structures being selected, and with respect to an existing library of compounds. The basic approach shares characteristics with other library design and optimization algorithms that have been published previously,[16–19] with the method described by Agrafiotis being the most similar by far. However, whereas both approaches utilize simulated annealing, the convergence rates and solutions will differ considerably due to the greedy component, which has been incorporated into Canvas HF. In addition, although previous studies have been performed to compare fingerprint methods for virtual screening applications,[20–23] we are not aware of a fingerprint comparison for hole-filling applications. Therefore, we compare the ability of different fingerprint methods to select compounds that contain new scaffolds.

Canvas HF can be run effectively with default settings in a variety of situations, with the only real tradeoff being between the quality of results and the number of optimization cycles. This greatly simplifies the library selection process compared to methods that require manual adjustment of parameters, such as objective function weights, genetic crossover and mutation rates, and exclusion distance.

We apply this method to collections of fragment-like molecules from commercially available vendor databases and show that, without any user intervention, the new method is able to fill holes as effectively as the standard sphere exclusion approach, which involves multiple iterations and manual tuning of parameters. Furthermore, we present a metric for 'hole-filling efficiency' that combines the diversity of the compounds retrieved with the computational time for the algorithm to complete. The method presented here is shown to be highly efficient in this regard. We also show the value of optimizing properties during the selection process as opposed to pre-filtering the external library using hard parameter cutoffs. While the application of the method in this paper is restricted to fragment-like molecules, the method itself has no restrictions and therefore can be applied to any chemical library.

2. Methods

In the material that follows, $\boldsymbol{L}_1$ denotes a reference library to which diverse compounds will be added and $\boldsymbol{L}_2$ denotes a second library from which diverse compounds will be selected. The distance d_{ij} between compounds i and j is computed as $1 - sim_{ij}$, where sim_{ij} is the fingerprint-based similarity lying on the interval [0,1].

One of the most widely used methods for choosing diverse compounds is sphere exclusion.[24] Its popularity is likely attributable to the fact that it imposes a hard upper limit on the similarity between any two compounds and the algorithm scales only linearly with the size of $\boldsymbol{L}_2$. In standard sphere exclusion, a single seed compound is chosen at random from $\boldsymbol{L}_2$, and a single pass is made through the remaining members of $\boldsymbol{L}_2$, choosing only those compounds that lie outside a user-defined *exclusion distance* of all previously chosen compounds, as illustrated in Figure 2A. If a specific number of diverse compounds are sought and successfully found before making a full pass through $\boldsymbol{L}_2$, the selection process can be halted.

Standard sphere exclusion may be adapted to hole filling by replacing the randomly chosen seed compound with the members of $\boldsymbol{L}_1$. Thus a given compound from $\boldsymbol{L}_2$ will be selected only if it lies outside the user-defined distance of all previously chosen compounds as well as all the compounds in $\boldsymbol{L}_1$ (Fig. 2B). Note that the augmented collection that consists of $\boldsymbol{L}_1$ plus the diverse compounds chosen from $\boldsymbol{L}_2$ will contain pairs of compounds that lie within the exclusion distance of each other only if such compounds exist in $\boldsymbol{L}_1$.

Although sphere exclusion is among the fastest known ways of selecting diverse compounds, it suffers from a variety of complications that arise from assigning an appropriate exclusion distance. For example, if a specific number of compounds are sought and the exclusion distance is too large, the algorithm may fail to find the desired number of compounds in $\boldsymbol{L}_2$. Even more troubling is the fact that reducing the exclusion distance does not always lead to the selection of additional diverse compounds because of the deterministic nature of the algorithm. As illustrated in Figure 2C, compounds selected early in the process may trigger an unfortunate chain of events that exclude an unusually large fraction of the compounds in $\boldsymbol{L}_2$. Furthermore, while choosing a sufficiently small exclusion distance will ultimately afford the desired number of compounds, diversity and coverage will be suboptimal if the distance is too small to force a full pass through $\boldsymbol{L}_2$. Only trial and error can determine the largest such distance in any given situation.

Some of the anomalies associated with sphere exclusion can be alleviated by sorting the compounds in $\boldsymbol{L}_2$ by decreasing similarity

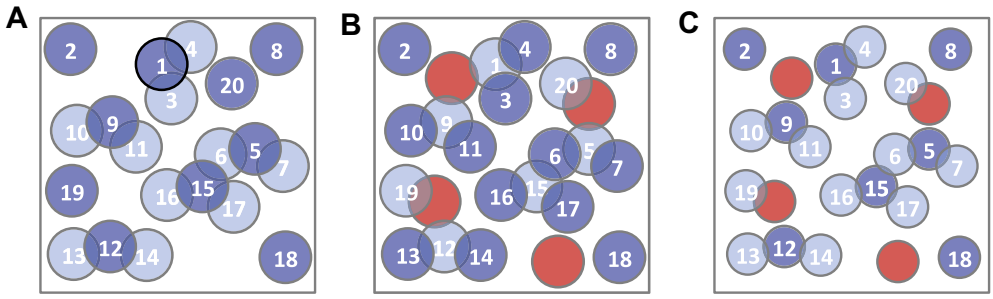

Figure 2. Illustration of the sphere exclusion method. Red circles represent compounds in a library $\boldsymbol{L}_1$ that contains holes, while blue circles represent a library $\boldsymbol{L}_2$ from which diverse compounds are being selected. Darker blue indicates that the compound was chosen by the sphere exclusion method, with the radius of each circle representing the exclusion distance. (A) In standard sphere exclusion, a single seed compound (1) is selected at random from $\boldsymbol{L}_2$, and the remaining $\boldsymbol{L}_2$ compounds are examined in order (2, 3, etc.), with a selection being made only if the compound lies outside the exclusion distance of all previously chosen compounds. (B) The method is adapted to hole filling by seeding the algorithm with the compounds in $\boldsymbol{L}_1$, followed by examination of all $\boldsymbol{L}_2$ compounds in order. (C) With the same collections $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$, use of a smaller exclusion distance may actually lead to fewer compounds being selected and sparser coverage.

to a small number of reference compounds and then examining the $\boldsymbol{L}_2$ compounds in sorted order. This technique, known as Directed Sphere Exclusion,[25] is more stable with respect to small changes in the exclusion distance, and it selects compounds that tend to be more uniformly spaced. However, it is not as readily adapted to hole filling as ordinary sphere exclusion, so it was not considered in this study.

A number of other techniques are routinely used for selecting diversity, including MAXMIN,[24] k-means clustering,[26] and hierarchical agglomerative clustering,[27] although each of these has its own disadvantages. The MAXMIN method begins with a single seed compound from $\boldsymbol{L}_2$, then proceeds with iterative selection of the compound in $\boldsymbol{L}_2$ that exhibits the largest minimum distance to any compound already selected, until the desired number of compounds has been chosen. MAXMIN has no parameters that can be varied to directly control the diversity, so an unfortunate choice of seed compound can result in suboptimal coverage. K-means clustering is specifically designed to provide near optimal coverage of a given space, but it can break down if the number of compounds requested exceeds the effective dimensionality of that space, with the result being empty clusters. Furthermore, there is no straightforward way of integrating hole filling into the k-means clustering algorithm, so it is difficult to obtain cluster centroids that are optimal for covering any particular chemical subspace. Hierarchical agglomerative clustering suffers from this same limitation, with the added drawback of near cubic scaling with respect to the size of $\boldsymbol{L}_2$, making it impractical when $\boldsymbol{L}_2$ contains more than a few thousand compounds.

To address the various limitations in the diversity selection methods described above, we have developed an alternative approach, Canvas HF, which incorporates both greedy and stochastic elements to choose diverse compounds, fill holes in an existing library, and optimize a set of pre-computed molecular properties. This method always returns the desired number of compounds, with the quality of the results being determined by the number of optimization cycles executed. Thus, a convenient tradeoff is achieved between the desired level of optimality and computational effort. The Canvas HF method is described in detail below.

Let n be the number of diverse compounds to be selected from $\boldsymbol{L}_2$; let $\boldsymbol{S}_k$ be the currently selected compounds at the beginning of optimization cycle k; let $NN(i)$ be the nearest neighbor of compound i, where $NN(i) \in \boldsymbol{S}_k$ or $NN(i) \in \boldsymbol{L}_1$; let $sim_{i,NN}$ be the fingerprint-based similarity between i and $NN(i)$; and let $\boldsymbol{F}$ be a set of m property filters of the form $P_{\min} \leqslant P \leqslant P_{\max}$, where P is a pre-computed molecular property and $P_{\min}$ and $P_{\max}$ are user-imposed bounds on that property. We define the similarity score SIM_k as the average nearest neighbor similarity for the compounds in $\boldsymbol{S}_k$, and the filter score $FILTER_k$ as the average fraction of property filters *failed* by the compounds in $\boldsymbol{S}_k$:

$$SIM_k = \frac{1}{n}\sum_{i\in\boldsymbol{S}_k} sim_{i,NN} \tag{1}$$

$$FILTER_k = \frac{1}{n}\sum_{i\in\boldsymbol{S}_k}\frac{1}{m}\sum_{f\in\boldsymbol{F}}[1-f(i)] \tag{2}$$

Here, $f(i) = 1$ or 0, depending on whether compound i passes or fails, respectively, property filter f. The total score is simply the sum of SIM_k and $FILTER_k$, a quantity that ranges from 0 to 2.

An initial subset $\boldsymbol{S}_0$ is arrived at by randomly choosing n compounds from $\boldsymbol{L}_2$. At each optimization cycle $k = 1,2,\ldots,N$, the objective is to adjust the membership of $\boldsymbol{S}_k$ so as to minimize the total score. A series of n replacements are attempted per cycle, where the candidate compound for removal, α, is always the member of $\boldsymbol{S}_k$ whose nearest neighbor similarity $sim_{\alpha,NN}$ is the largest. An alternate compound β is selected randomly from $\boldsymbol{L}_2$ and the exchange is made if *either* of the following conditions is satisfied:

1. $SIM_k^\beta < SIM_k \;\&\; FILTER_k^\beta \leqslant FILTER_k$.

2. $SIM_k^\beta \leqslant SIM_k \;\&\; FILTER_k^\beta < FILTER_k$.

The scores SIM_k^β and $FILTER_k^\beta$ are computed after temporarily removing compound α from $\boldsymbol{S}_k$ and replacing it with compound β. While the above formulae look highly similar, there is a difference in the relational operators ($<$ and $\leqslant$). This difference is needed to ensure that both the similarity and filter scores are non-increasing, thereby preventing replacements that reduce one score at the expense of the other.

If neither of the above conditions is met, the following temperature-related quantities are defined in preparation for a Monte Carlo test:

ε = user-defined Monte Carlo tolerance (default = 0.001)
$T_{\min} = -\varepsilon/\log(0.01)$.
$T_{\max} = -\varepsilon/\log(0.5)$.
$\Delta T = (T_{\max - T_{\min}})/(N-1)$.
$T_k = T_{\max} - k\Delta T$.

The Monte Carlo test is conducted as follows:

x = uniform random number on $[0, 1]$.
$\Delta SCORE_k^\beta = \max\{0, SIM_k^\beta - SIM_k\} + max\{0, FILTER_k^\beta - FILTER_k\}$.
$PTEST_k^\beta = \exp(-\Delta SCORE_k^\beta / T_k)$.

If $x < PTEST_k^{\beta}$, replace α with β.
If $x < PTEST_k^{\beta}$, replace α with β.

Observe that a score increase of ε is accepted with a probability of 50% in the first optimization cycle, whereas the probability of acceptance drops to 1% in the final cycle. Typically, after 5 optimization cycles the number of replacements per cycle is very small and the change in the total score from one cycle to the next is on the order of 0.001. Therefore, we set the default convergence criteria for the optimization to terminate when the total score changes by less than 0.001 over three consecutive cycles. This parameter can be adjusted by the user but has not been explored in this work.

In the remainder of this paper, we present the results of applying Canvas HF to the selection of compounds from commercially available fragment libraries. We compare the method to the hole-filling adaptation of sphere exclusion described previously. We also explore the behavior of Canvas HF with respect to fingerprint methods and the effect of using property biasing in the selection process.

3. Results

A collection of over 150,000 compounds was compiled from a number of commercially available fragments libraries ranging in size from dozens to tens of thousands of compounds. A subset of 50,000 compounds was randomly selected from the above collection to create the pool of compounds $\boldsymbol{L_2}$ from which all diverse selections were made. In order to test whether the method is robust to variations in $\boldsymbol{L}_1$, we used two independent sets of fragment compounds, Chembridge Fragment Library[28] and Key Organics Complete Fragments.[29] Each $\boldsymbol{L}_1$ library contains approximately 8000 compounds, and Figure 3 shows for these reference libraries the distribution of several commonly used physicochemical properties. The cheminformatics package Canvas (Schrödinger, Inc.) was used for all calculations presented here.

Canvas HF (denoted as 'HF' in the figures and tables) and sphere exclusion (SE) were used to select diverse subsets of compounds from the pool to fill holes in each reference library. Based on internal validation studies, we have found that five cycles of hole filling provides a reasonable balance between the diversity of the selected compounds and the computational cost. Therefore, throughout the manuscript we use five cycles for Canvas HF, which is also the default in Canvas. In addition, we also ran as many cycles as needed until convergence, as defined by the total score changing by less than 0.001 in three consecutive cycles, which we refer to as HF(max). When property filtering was applied, we use the designations HF-props and HF-props(max). For property filtering, we used the following ranges: molecular weight $\leqslant 300$, $1 \leqslant A\log P \leqslant 3$, $40\ \text{Å}^2 \leqslant$ polar surface area $\leqslant 80\text{Å}^2$, $1 \leqslant$ hydrogen bond acceptors $\leqslant 3$, $1 \leqslant$ hydrogen bond donors $\leqslant 3$, and rotatable bonds $\leqslant 3$. The above ranges are similar to, but slightly more restrictive than, the common 'Rule of 3' for fragments.[30] We imposed the stricter criteria to ensure that approximately 10% of the 50,000-compound pool satisfied all filters. We chose four different exclusion distances for SE (0.5, 0.6, 0.7, and 0.8), using the Soergel distance measure, which is simply 1-Tanimoto.[31] The size of the selected subsets studied was 50, 500, 1000, 2500, 5000, and 10000. Random subsets of equal size were also selected from the same pool to provide a baseline comparison.

Canvas can be used to generate eight fingerprint types with various atom typing and scaling options. In previous work we explored all of these fingerprint methods and associated parameters for their ability to selectively retrieve known actives in virtual screening exercises.[17] In this work, we initially studied four of the fingerprint types that performed well in our previous study (dendritic, linear, molprint2D, and radial), with the objective being to determine which fingerprint method was best suited for the remainder of the study. While it is possible that some of the fingerprints that did not perform well for virtual screening could perform well in hole filling, the objective of this paper was not an exhaustive survey of the fingerprint methods. Using default settings for each fingerprint, subsets of 50 compounds were chosen from the pool of 50,000, using the Chembridge library as the reference.

The Canvas Scaffold Decomposition application was applied to each selected subset to generate an exhaustive list of Bemis-Murcko scaffolds.[32] We examined the number of unique scaffolds retrieved from the pool that did not exist in the Chembridge library and the associated computing time required (Table 1). As seen in Table 1, Molprint2D fingerprints generated the smallest number of new scaffolds, which is interesting given that this was the most effective method studied in our previous virtual screening work.[21] In addition, linear fingerprints required the longest time and did not yield a significant benefit in the number of new scaffolds (Table 2). Both dendritic and radial fingerprints performed best when considering the number of new scaffolds retrieved and the computational costs. It is important to note that retrieving the largest number of unique scaffolds may not be the only desired objective of a hole filling study and several other

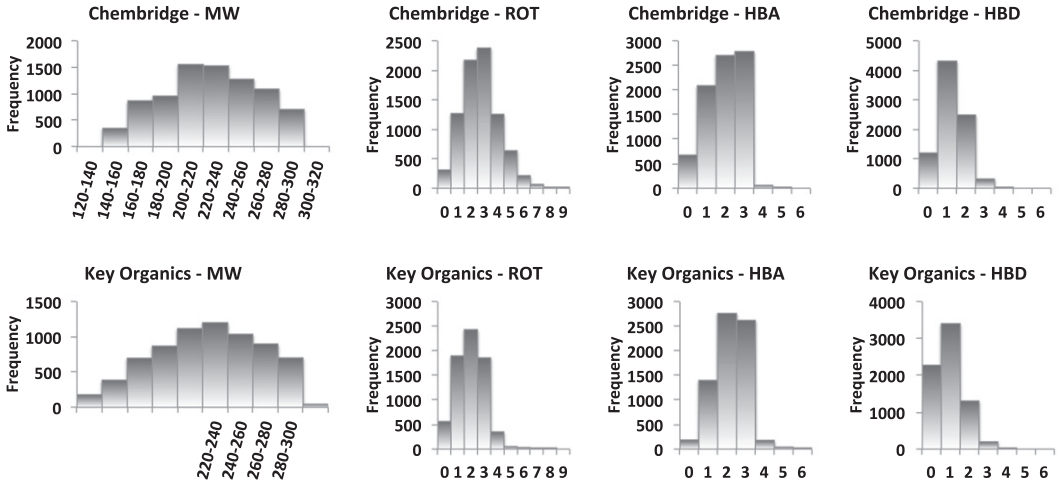

Figure 3. Property distributions for the two reference fragment libraries. MW = molecular weight; ROT = rotatable bonds; HBA = hydrogen bond acceptors; HBD = hydrogen bond donors.

Table 1
The number of molecules out of the 50 selected that contain new scaffolds[a]

Fingerprint	HF	HF (max)	HF-props	HF-props (max)	SE (0.5)	SE (0.6)	SE (0.7)	SE (0.8)	Total
Dendritic	40	45	41	38	38	38	37	38	315
Linear	42	47	37	41	45	45	45	45	347
Molprint2D	38	42	39	39	33	34	33	31	289
Radial	43	47	40	41	39	37	38	36	321

[a] HF = Canvas hole filling with default parameters (5 cycles); HF (max) = Canvas hole filling with maximum cycles; props = inclusion of properties in Canvas hole filling; SE (X) = sphere exclusion with X as the similarity cutoff; total = sum of all columns.

Table 2
CPU time in seconds required to select 50 diverse compounds[a]

	HF	HF (max)	HF-props	HF-props (max)	SE (0.5)	SE (0.6)	SE (0.7)	SE (0.8)
Dendritic	15	32	18	55	3198	2018	1105	378
Linear	17	51	21	58	5732	3988	2395	1098
Molprint2D	4	13	7	25	335	185	92	21
Radial	6	20	9	26	1546	1620	939	278

[a] HF = Canvas hole filling with default parameters (5 cycles); HF (max) = Canvas hole filling with maximum cycles; props = inclusion of properties in Canvas hole filling; SE (X) = sphere exclusion with X as the similarity cutoff.

approaches could be taken to assess the diversity of new compounds. Nonetheless, for the remainder of this work we use dendritic fingerprints, although the results from radial fingerprints are expected to be qualitatively similar at least for retrieving unique scaffolds.

There are many ways to compare the diversity of compounds. In this study, we focus on filling the chemical space, as described by the substructures in the 2D fingerprints, which is not covered by the reference library $\boldsymbol{L}_1$. To assess the diversity of the selected compounds, we used the dendritic fingerprint nearest neighbor Tanimoto similarity sim_{NN}. Furthermore, the following efficiency formula was devised as a metric to quantify the performance of each method, by combining the diversity of the compounds and the computational cost:

$$\text{Efficiency} = \frac{1 - (SIM_{N,N,cross} + SIM_{N,N,self})/2}{1 - sim_{PW}} \cdot \frac{1}{\log(\text{CPU})} \qquad (3)$$

Here, $sim_{NN,cross}$ is the average nearest neighbor similarity for the compounds selected from $\boldsymbol{L}_2$, with nearest neighbors coming from the reference library $\boldsymbol{L}_1$, and $sim_{NN,self}$ is the average nearest neighbor similarity for the same compounds, with nearest neighbors coming from within the selected subset itself. The normalizing term sim_{PW} is the average pairwise similarity within the 50,000-compound pool $\boldsymbol{L}_2$. When a method did not produce the desired number of compounds, for example, SE with a large exclusion distance, an efficiency of zero was assigned. Values of this metric for each method are shown in Figure 4. Comparisons of the individual components of efficiency are shown in Figure 5. It should be noted that hole-filling efficiency is not explicitly part of the Canvas HF algorithm and was devised after the results were attained as a way to simultaneously assess the quality of the results and the computational efficiency.

Overall, Canvas hole filling with five optimization cycles (HF) yields the highest efficiency. Using small exclusion distances, for example, 0.5 or 0.6, and relatively small subsets, SE requires significantly longer computing time than Canvas HF, because SE must compare each compound in the pool of 50,000 to each compound in the reference library. Thus a large number of similarities must be computed in order to choose a small number of compounds. By contrast, Canvas HF merely selects compounds at random from the pool, rather than making a full pass through it. Thus while SE diversity is comparable to that of Canvas HF when selecting smaller subsets at shorter exclusion distances, its larger computational time leads to lower efficiency. On the other hand, SE at higher exclusion distances, such as 0.7 or 0.8, completes quickly but does not always find the desired number of compounds, resulting once again in a lower efficiency than the other methods. This tradeoff between speed and number of compounds makes it difficult to

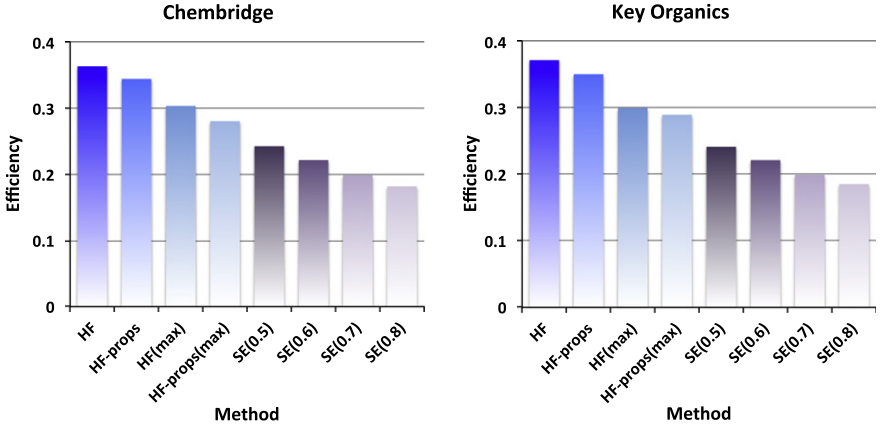

Figure 4. Efficiency of hole filling strategies. Results are averaged over all 6 selection sizes (50, 500, 1000, 2500, 5000, and 10000) for each of Chembridge (left) and Key Organics (right) fragment libraries. Canvas HF with various settings are shown in shades of blue. Sphere exclusion (SE) with various exclusion distances shown in shades of purple. Refer to Eq. 3 for the definition of efficiency.

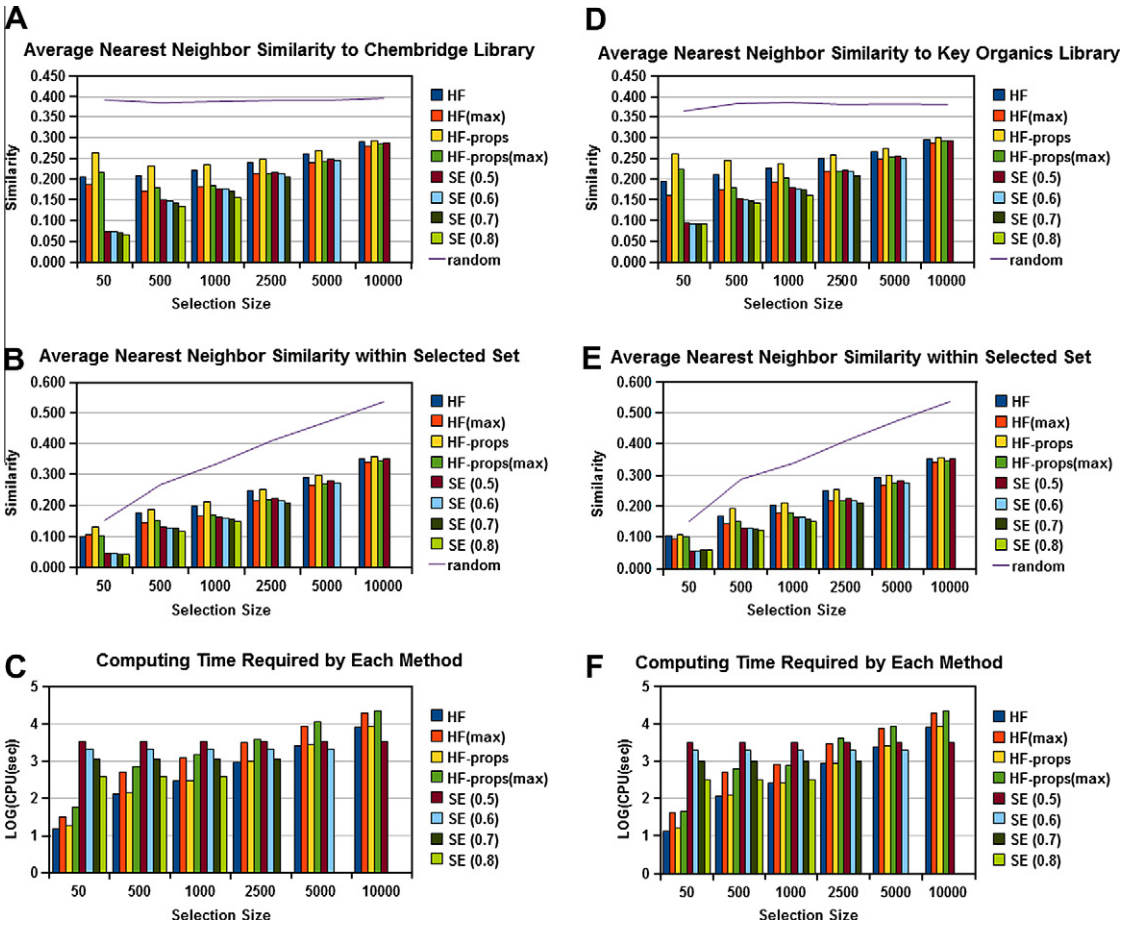

Figure 5. Assessment of the ability of each method to select diverse compounds. (A) Average nearest neighbor similarity between the selected subsets and the Chembridge reference library at different selection sizes. (B) Average nearest neighbor similarity within selected subsets for the Chembridge library. (C) CPU time required at different subset sizes to fill holes in Chembridge library. (D) Average nearest neighbor similarity between the selected subsets and Key Organics reference library. (E) Average nearest neighbor similarity within selected subsets for Key Organics library. (F) CPU time required at different subset sizes to fill holes in Key Organics library.

determine the optimal exclusion distance for a given subset size. For example, an exclusion distance of 0.7 failed to produce the desired number of diverse compounds at selection sizes of 5000 and 10,000 for both reference libraries. In general, the optimal distance depends on the reference library $\boldsymbol{L}_1$ the external pool $\boldsymbol{L}_2$, the fingerprint method, the fingerprint parameters, and the metric used to compute distances.

To account for any inherent diversity in the pool of compounds $\boldsymbol{L}_2$, it is useful to compare the results of diversity-based selection to that of pure random selection. As shown in Figure 6, compounds chosen by Canvas HF and SE show greater dissimilarity among themselves and with respect to the reference library than equal numbers of compounds chosen at random. To ensure statistical significance of the average similarities being compared, pair-wise *T*-values (based on the difference in average similarity divided by combined standard deviation of any given two sets) were calculated between all sets at each selection size. The lowest *T*-value (4.01) at the smallest size (50) corresponds to a confidence of 99.9%, with the rest having even higher confidence (see Supplementary data). The differences from random are less pronounced for larger subsets because chemical space necessarily becomes more crowded as additional compounds are selected. Thus diversity-based methods are not as advantageous when sampling larger and larger fractions of a given collection.

Observe that for random selections, the average nearest neighbor self-similarity increases with subset size, but the average nearest neighbor similarity to the reference library does not. A simple analogy is that of adding more and more swimmers to a pool with a fixed number of lifeguards. There is less and less space between swimmers as more swimmers enter the pool, but the average distance from a swimmer to the nearest lifeguard does not change.

To visually illustrate the concept of hole filling with a concrete example, a fingerprint-based Kohonen map (self-organizing map; SOM)[33,34] was trained on each reference library $\boldsymbol{L}_1$ plus the compounds selected by Canvas HF (Fig. 7B and D), then applied to the reference library itself (Fig. 7A and C). Individual SOM cells were colored based on population using a white (0 compounds) to red ($\geqslant$10 compounds) heat map. As shown in Figure 7, nearly all of the empty cells in the reference library become occupied after adding compounds selected by Canvas HF.

It is worth noting that the process of mapping high dimensional fingerprint space to a two-dimensional grid involves a fairly drastic loss of information. For example, we found that compounds mapped to the same cell do not always exhibit higher similarities than compounds mapped to different cells, which implies that some cells cover larger regions of chemical space than others. Furthermore, close proximity of cells does not necessarily imply higher similarity, as illustrated by the shading of cell borders in Figure 7, where darker borders indicate a larger distance between adjacent cells. Thus while Kohonen maps are convenient for illustration purposes, they provide only a qualitative picture of the hole-filling process.

In addition to achieving efficient hole filling, Canvas HF provides a means for biasing selections toward a desired set of property

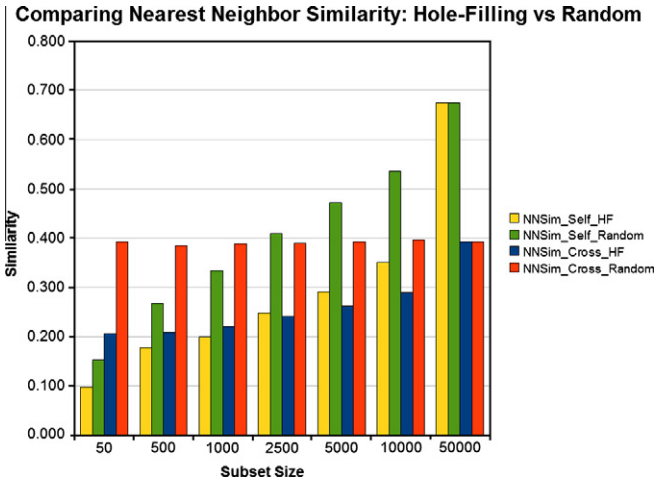

Figure 6. Comparison of average nearest neighbor similarities within selected subsets with respect to the Chembridge library. Yellow: average nearest neighbor similarity within the subset selected by Canvas HF. Green: average nearest neighbor similarity within randomly selected subsets. Blue: average nearest neighbor similarity between subset selected by HF and Chembridge library. Red: average nearest neighbor similarity between randomly selected compounds and the Chembridge library.

ranges, without requiring every compound to satisfy all property filters. To illustrate the advantage of this approach, Canvas HF with properties (HF-props), was compared to SE run on a subset of $\boldsymbol{L}_2$ that satisfied all of the following property filters: molecular weight $\leqslant 300$, $1 \leqslant A\log P \leqslant 3$, $40\ \text{Å}^2 \leqslant$ polar surface area $\leqslant 80\ \text{Å}^2$, $1 \leqslant$ hydrogen bond acceptors $\leqslant 3$, $1 \leqslant$ hydrogen bond donors $\leqslant 3$, and rotatable bonds $\leqslant 3$. A total of 5900 compounds from the original pool of 50,000 remained after applying these filters.

Table 3 compares the number of new scaffolds found for subsets of 50 and 500 diverse compounds. A scaffold is counted as new if it is unique and it does not exist in the reference library. Observe that it is possible for a compound to contribute multiple scaffolds, as evidenced by some of the approaches finding more than 50 new scaffolds when only 50 compounds were chosen. From Table 3, it is evident that Canvas HF finds significantly more new scaffolds than SE at any cutoff, with up to a twofold improvement in some cases.

As expected, many of the new scaffolds found by HF-props came from compounds that were eliminated by the pre-filter (about 90% of the compounds from the original pool were filtered out), and there is little overlap between the compounds and the new scaffolds found by HF-props and SE. Since HF-props does not require every compound to satisfy all property filters, only a portion of compounds selected by HF-props exist in the property-filtered pool. For example, with Chembridge as the reference library, at a selection size of 50, only 13 compounds chosen by HF-props satisfied all property filtering criteria. These 13 compounds contain 13 unique scaffolds (although not each compound has a single new scaffold; some have more than one and some have none). Figure 8 shows the 10 molecules selected by HF-props that satisfy all property filtering criteria and also contain at least one new scaffold, with new scaffolds highlighted in yellow. For comparison, the nearest neighbors in the SE combined set are also shown.

4. Conclusions and future directions

In this work we presented Canvas HF, a simple and versatile method for selecting diverse compounds, filling holes in an existing library, and optimizing a set of molecular properties. Canvas HF rapidly selects compounds with high diversity in both fingerprint space and scaffold space, and performs robustly across

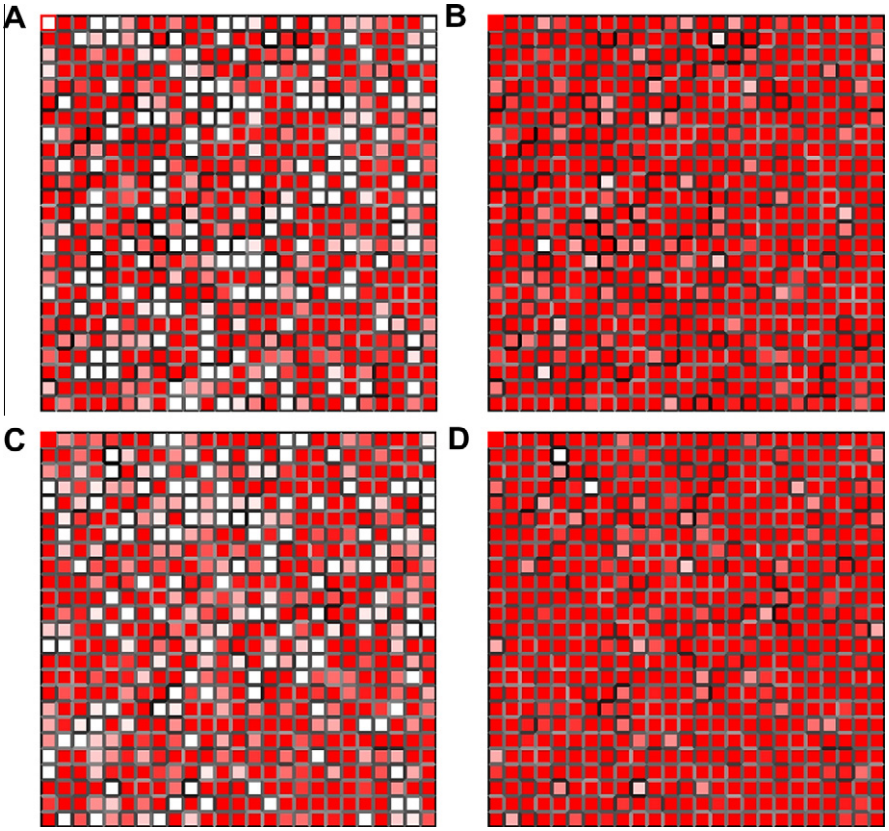

Figure 7. Illustration of Canvas hole filling using 25×25 Kohonen maps. Cells are colored by population, with white for empty cells, and red for cells containing 10 or more compounds. (A) Chembridge library; (B) Chembridge library combined with 10,000 compounds selected by Canvas HF; (C) Key Organics library; (D) Key Organics library with 10,000 compounds selected by Canvas HF.

Table 3
Comparison of the number of new scaffolds found by Canvas HF with property filters selecting from a pool of 50,000 compounds, and with SE selecting from a property-filtered pool of 5900 compounds

	HF-props	Pre-filter SE (0.5)	Pre-filter SE (0.6)	Pre-filter SE (0.7)	Pre-filter SE (0.8)
Chembridge 50 compds	83	53	52	43	46
Key Organics 50 compds	85	50	51	48	40
Chembridge 500 compds	831	425	430	436	440
Key Organics 500 compds	885	452	429	443	467

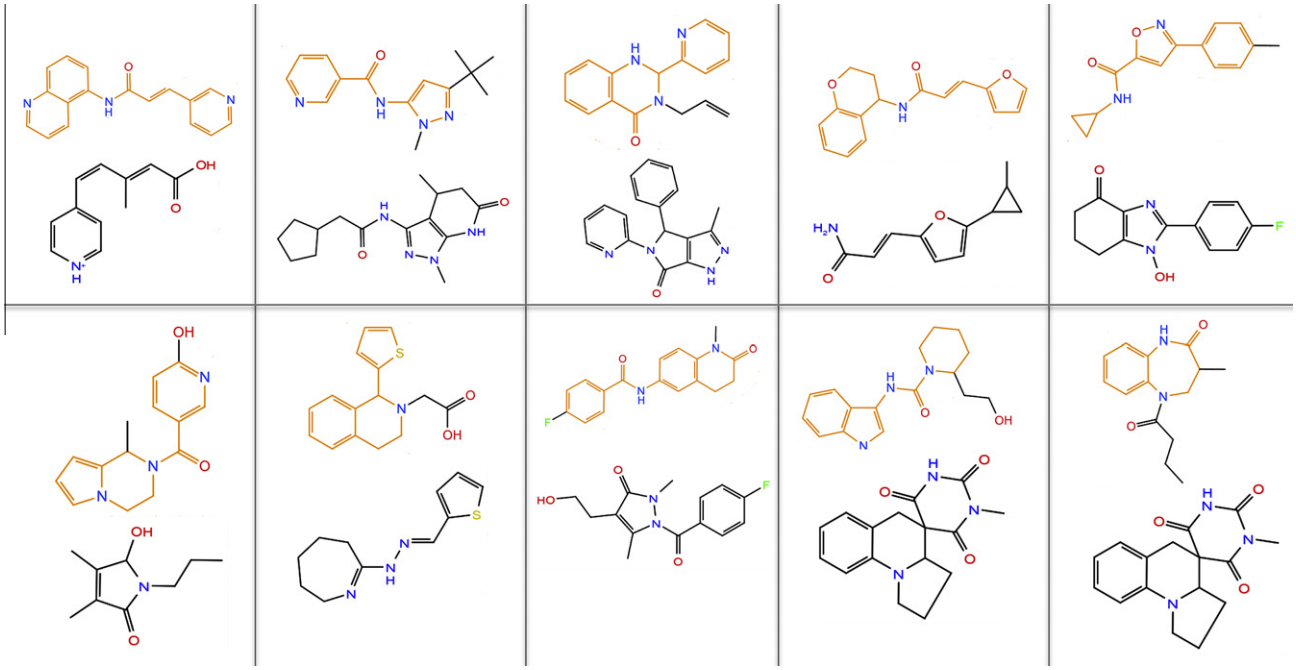

Figure 8. New scaffolds found by Canvas HF-props (upper image in each pair) from compounds that are also present in the property-filtered pool. The closest molecule found by SE from the property-filtered pool is shown for reference (lower image in each pair). The unique scaffold for each molecule is highlighted with yellow bonds, which sometimes represents the entire molecule. Note that the two bottom-right structures from SE are the same, signifying that this molecule found by SE is the closest to two different molecules found by Canvas HF-props.

different fragment libraries using different subset selection sizes. Canvas HF consistently outperformed sphere exclusion when considering the diversity of the retrieved compounds and the computational cost. In addition, Canvas HF with default parameters performed well in all tests, which illustrates that trial-and-error approaches to parameter tuning are unnecessary.

While the results presented here are encouraging, more work is needed to compare Canvas HF with other methods and across other data sets. For example, all work done here was with fragment libraries. Although we have no evidence that the conclusions will not hold up for drug-like libraries, it is necessary for us to confirm that in future work. Furthermore, the performance of Canvas HF needs to be explored on larger libraries. For example, typical pharmaceutical libraries have millions of compounds and commercially available compound databases also contain millions of molecules. A Canvas HF calculation to select one thousand compounds from a one million compound external library to fill holes in a one million compound in-house library takes approximately 5 h on a single processor and uses a maximum of 1.5 GB of memory, implying that jobs of this size can run on a standard desktop computer. However, pushing the method into the tens of millions of compounds could become prohibitively expensive. We are currently exploring ways to extend the method to handle libraries of this size. Finally, numerous other methods that improve upon sphere exclusion have been published (see Introduction for references) and we were not able to compare with those methods. We hope to see future publications from research groups that have access to the other methods and can compare with Canvas HF.

We believe Canvas HF offers a fast, robust, and effective approach for the important tasks of hole filling and library optimization in pharmaceutical research. With the expanding number of available compounds and the complexities of maintaining huge corporate compound collections, it is important to have a method that can intelligently select compounds to complement in-house libraries. Canvas HF offers a solution to this problem. Future development work will be done to extend the capabilities of the method and future studies will show the applicability of the method to other problems.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmc.2012.03.037.

References and notes

1. Broach, J. R.; Thorner, J. *Nature* **1996**, *384*, 14.
2. Bajorath, J. *Nat. Rev. Drug Disc.* **2002**, *1*, 882.
3. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Disc.* **2004**, *3*, 935.
4. Sastry, M.; Dixon, S.; Sherman, W. *J. Chem. Inf. Model.* **2011**, *51*, 2455.
5. Murray, C. W.; Rees, D. C. *Nat. Chem.* **2009**, *1*, 187.
6. Loving, K.; Alberts, I.; Sherman, W. *Curr. Top. Med. Chem.* **2010**, *10*, 14.

7. Terrett, N. K.; Gardner, M.; Gordon, D. W.; Kobylecki, R. J.; Steele, J. *Tetrahedron* **1995**, *51*, 135.
8. Breinbauer, R.; Vetter, I. R.; Waldmann, H. *Angew. Chem., Int. Ed.* **2002**, *41*, 2878.
9. Huggins, D. J.; Sherman, W.; Tidor, B. *J. Med. Chem.* **2012**, *55*, 1424.
10. Chemical Abstracts Service. http://www.cas.org/cgi-bin/cas/regreport.pl.
11. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. *Nucleic Acids Res.* **2011**, 1.
12. ChemExper. http://www.chemexper.com.
13. eMolecules. http://www.emolecules.com/.
14. Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177.
15. Dixon, S. L.; Villar, H. O. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192.
16. Agrafiotis, D. *IBM J. Res. Dev.* **2001**, *45*, 545.
17. Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375.
18. Martin, E. J.; Critchlow, R. E. *J. Comb. Chem.* **1999**, *1*, 32.
19. Clark, R. D. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181.
20. Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. *J. Mol. Graph. Model.* **2010**, *29*, 157.
21. Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. *J. Chem. Inf. Model.* **2010**, *50*, 771.
22. Tovar, A.; Eckert, H.; Bajorath, J. *ChemMedChem* **2007**, *2*, 208.
23. Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. *J. Chem. Inf. Model.* **2009**, *49*, 108.
24. Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. *J. Mol. Graph. Model.* **1997**, *15*, 372.
25. Gobbi, A.; Lee, M.-L. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 317.
26. Lloyd, S. *Inform. Theory IEEE Trans.* **1982**, *28*, 129.
27. Murtagh, F. In *Compstat Lectures*. Physica-Verlag: Vienna, 1985; Vol. 4.
28. Chembridge Fragment Library. http://www.chembridge.com/screening_libraries/fragment_library/.
29. Key Organics Complete Fragments. http://www.keyorganics.co.uk/Solutions/BIONET/Fragment-Libraries/.
30. Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. *Nat. Rev. Drug Disc.* **2004**, *3*, 660.
31. Fechner, U.; Schneider, G. *ChemBioChem* **2004**, *5*, 538.
32. Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887.
33. Kohonen, T. In *Self-Organizing Maps. Springer Series in Information Sciences*; Springer: Heidelberg, 1997; Vol. 30,
34. Kohonen, T.; Somervuo, P. *Neurocomputing* **1998**, *21*, 19.