JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

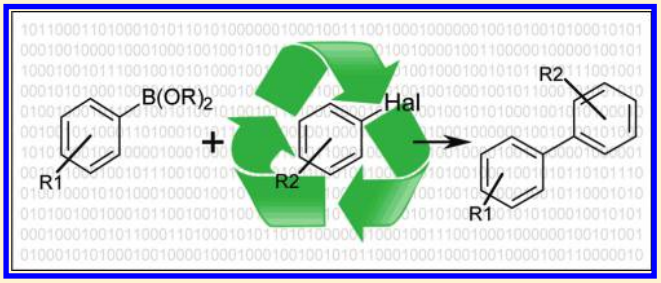
pubs.acs.org/jcim

Automated Recycling of Chemistry for Virtual Screening and Library Design

Mikko J. Vainio,[†] Thierry Kogej,*[,†] and Florian Raubacher[‡]

[†]Discovery Sciences Computational Sciences and [‡]R&I iMed, AstraZeneca R&D, Pepparedsleden 1, S-43183 Mölndal, Sweden

ABSTRACT: An early stage drug discovery project needs to identify a number of chemically diverse and attractive compounds. These hit compounds are typically found through high-throughput screening campaigns. The diversity of the chemical libraries used in screening is therefore important. In this study, we describe a virtual high-throughput screening system called Virtual Library. The system automatically "recycles" validated synthetic protocols and available starting materials to generate a large number of virtual compound libraries, and allows for fast searches in the generated libraries using a 2D fingerprint based screening method. Virtual Library links the returned virtual hit compounds back to experimental protocols to quickly assess the synthetic accessibility of the hits. The system can be used as an idea generator for library design to enrich the screening collection and to explore the structure−activity landscape around a specific active compound.



INTRODUCTION

Chemical library design is a central activity in the early phases of the drug discovery process. Library design is used in lead generation projects to produce series of analogues around hit and lead compounds in order to explore structure−activity relationships (SARs). A large number of current clinical candidates are based on hits found in high-throughput screening (HTS) experiments—at AstraZeneca, this is more than 50% of the candidate drugs since 2004.[1] In order to maintain a level of success in HTS, it is necessary to continuously enrich the HTS compound collection by design and synthesis of libraries of novel compounds that reflect the target portfolio. The HTS collection at AstraZeneca is expanded through our Compound Collection Enhancement program launched in 2003. The program has delivered more than 4000 chemical library designs and several hundred thousand compounds have been synthesized. These compounds currently represent a substantial part of the HTS screening collection and a wealth of chemistry knowledge has been accumulated through the design, evaluation, and synthesis of these compound libraries.[1]

In this work, we describe how the chemistry of the existing libraries is "recycled" in order to create a large number of virtual libraries[2] and how these are exploited. Software that enables a quick recall of ideas for chemical synthesis from such set of virtual libraries is likely to improve the efficiency of the hit identification and lead optimization phases of the drug discovery process. Indeed, other organizations have reported analogous infrastructure to what we describe here.[3,4]

The Virtual Library (VL) is a collection of virtual compound libraries based on validated synthetic protocols. The individual libraries are encoded using a list of reactants that is a superset of that used in the actual synthesis, which greatly increases the chemical space covered by the virtual compounds compared to the synthesized products. The VL can be screened in silico against one or more query compounds using several types of 2D fingerprints via a web application in a matter of minutes. Queries can be submitted programmatically, too, through a web service interface, which is used to embed the VL search functionality into in-house chemical library design applications.

The challenges in deploying virtual libraries as part of the drug discovery process include library definition and construction, and efficient screening. Library definition may be easy for a chemist but tricky to achieve algorithmically, because the process involves reasoning about feasible chemistry. The challenge with screening is the size of the search space—a virtual library may easily encode 10^{12} product molecules.[2,5] This problem of combinatorial explosion is commonly tackled by representing the product compounds as combinations of fragments and performing the search in multiple steps. The fragments are searched first and the combinations of the highest ranking fragments, that are the full product compounds, are screened in the last step. Examples of such divide-and-conquer approaches are the Ftrees-FS method,[6] Topomer Shape Similarity searching in ChemSpace virtual library database,[2,7] and Basis Product (BP) method.[8] Besides virtual screening, the divide-and-conquer approach has been used to generate molecular descriptors and fragment-based 2D fingerprints for combinatorial libraries from reaction- and precursor-based library descriptions without explicitly enumerating the product compounds.[5,9]

In this work, we make use of the BP approach to virtual screening that resembles the LEad-based Analog hoPping 2 (LEAP2) method used in Pfizer Global Virtual Library (PGVL).[4] While the overall structure of VL evolved to

Received: March 23, 2012
Published: June 3, 2012

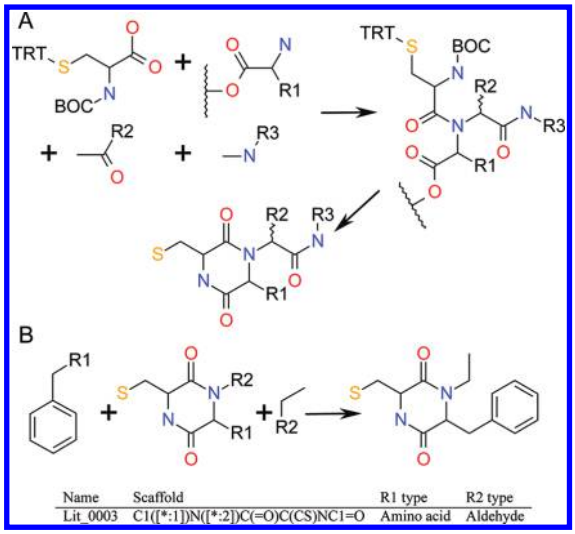
resemble LEAP2/PGVL, there are a number of differences: Compared to LEAP2/PGVL, VL uses a different similarity metric, binary fingerprint types, pruning of the fingerprint database during screening, avoids the computation of the fingerprints for enumerated hit compounds, and uses an automated process to generate virtual libraries, as described in Methods.

It is not unusual that HTS hits come from external chemistry sources from which the synthetic protocol is difficult (if not impossible) to retrieve. Consequently, the most common use case of VL at AstraZeneca is to find an alternative existing library synthetic protocol to speed up the structure−activity relationship exploration around a hit series coming from HTS within the lead generation phase.

METHODS

Encoding of Virtual Libraries. In the library encoding scheme used in this study, a library is defined by a unique name, a scaffold structure with attachment points for substituents and lists of types of reactants (amine, alcohol, etc.) to substitute into these positions. The structures are stored as SMILES (Simplified Molecular Input Line Entry System)[10] strings with dummy atoms marking the attachment points. An example of the encoding format is shown in Scheme 1.

Scheme 1. Example of the Reaction Encoding Scheme[a]



[a]The encoded reaction is that for library ACL0938 from ref 12 used as the literature example 3 in this study. The original synthesis route A (Scheme 2 in the work of Szardenings et al.[12]) is a two-step solid phase reaction with a cycle formation in the second step. There are three R-groups in the original scheme, of which two were used in ACL0938. The depiction B shows the encoded reaction abstraction. The scaffold, shown in the middle, has two attachment points to which either amino acids (R1) or aldehydes (R2) can be substituted. The final encoded reaction, as stored in the VL database, is the line of text at the bottom.

The library encoding scheme is a reaction abstraction that does not explicitly include all the reaction steps and does not relate to a specific type of reaction. The absence of this information reduces the flexibility of the encoding format with regard to the replacement of the scaffold by other chemically compatible structures. Moreover, the synthesis of the scaffold and its availability is not captured. However, the format allows for efficient enumeration of the library using off-the-shelf cheminformatics toolkits. The current infrastructure links the names of the encoded libraries to a record in a database of synthetic protocols, from which a chemist can extract the necessary information to reproduce the synthesis.

The lists of reactants of different types were generated using a set of Python scripts written on top of cheminformatics toolkits from OpenEye Scientific Software, as follows. SMARTS (SMiles ARbitrary Target Specification)[11] language patterns were used to match reactive substructures of available reactants. The matched atoms were then replaced by a dummy atom and the result, a building block or "product-fragment"[2] SMILES string, stored to be later used as a substituent in a scaffold SMILES of a library. Building blocks created via a match to a given SMARTS pattern are assumed to be interchangeable within a reaction, that is, of the same reactant type. A single structure can give rise to several building blocks depending on the reaction type and the way the scaffold is defined. For example, Figure 1 shows an aniline that produced two building blocks, one to be used in amide formations and another for

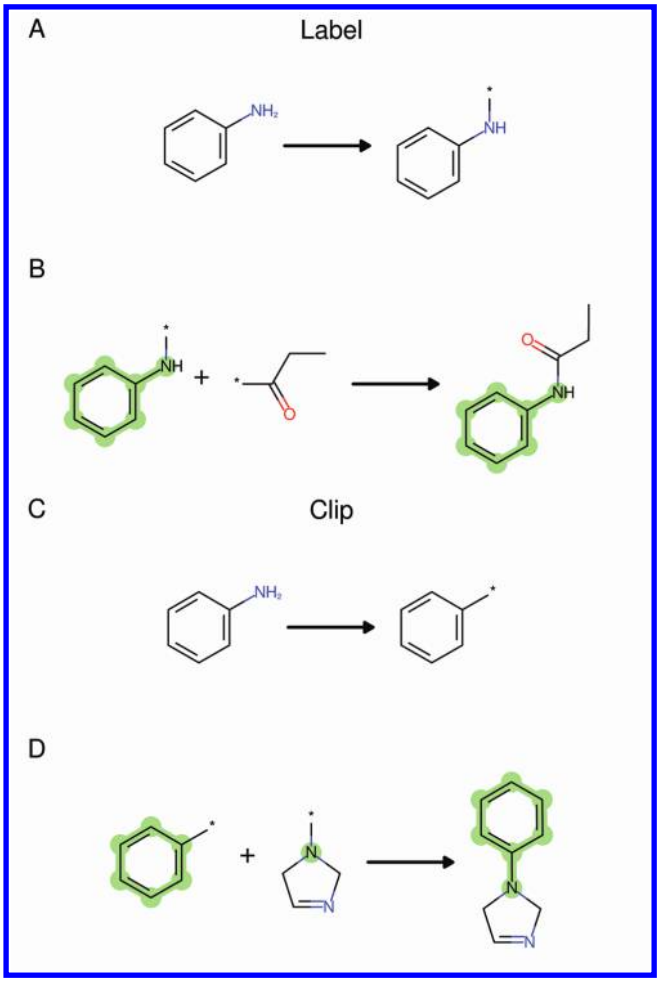


Figure 1. Preparation of a reactant structure for the reaction encoding scheme used in this study. The aniline can be processed in two different ways: (A) Label the aniline nitrogen as attachment point by adding a bond to a dummy atom. (B) Example reaction using the labeled reactant. (C) Replace the aniline nitrogen with a dummy atom. The clipped reactant structure can be also used to form ring closures. (D) Cycle formation is captured in the scaffold structure, which contains the nitrogen that was clipped away from the reactant structure. The atoms originating from the aniline are highlighted.

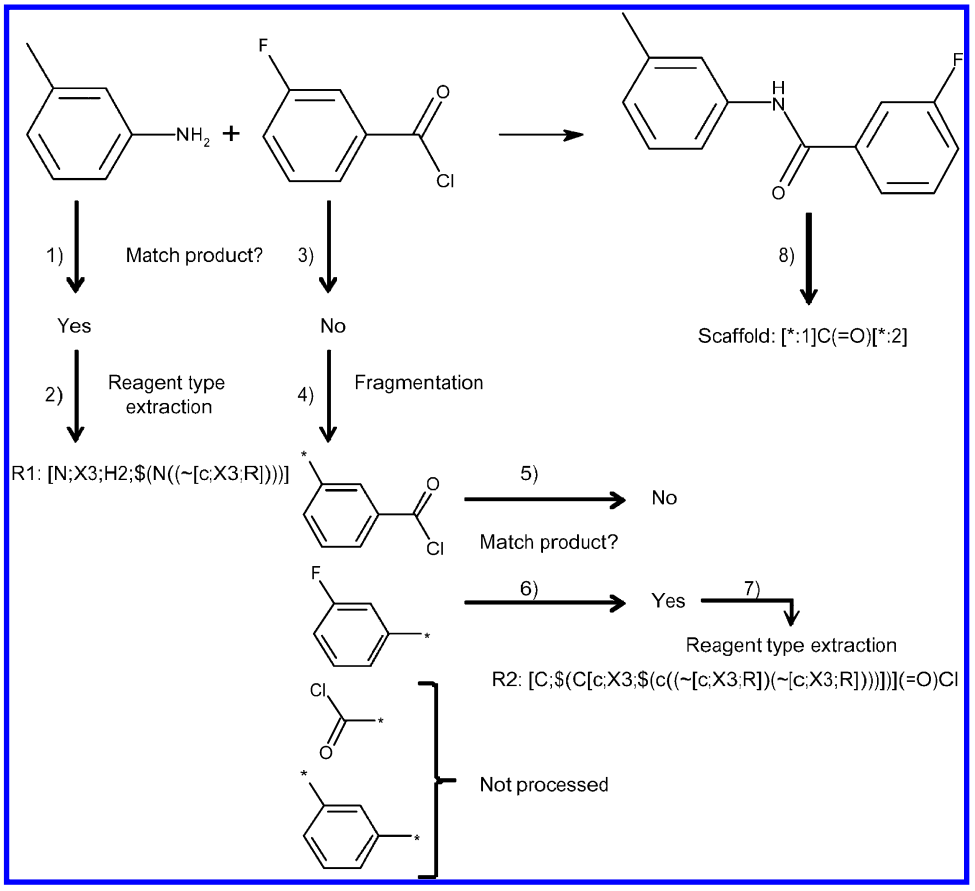


Figure 2. Example of the workflow for automatic extraction of a virtual library encoding from a reaction stored in the electronic laboratory notebook database.

cycle formations where the nitrogen is included in the scaffold structure. The application of preprocessing rules to reactants allows for the encoding of ring closures and other "complex" chemistry despite the minimalistic nature of the abstracted reactions.

In order to increase the coverage of the chemical space by the VL, assumptions were made on the types of the reactants to be considered in a number of encoded libraries. For example, some amide formation reactions were originally designed to combine a set of acid chlorides and primary amines. In such cases, we assumed that the list of reactants can be extended by carboxylic acids and sulfonamides for the acidic part and secondary and tertiary amines for the amine part without dramatically reducing the synthetic accessibility of the virtual product compounds. Whether this assumption holds for a particular combination of reactants and reaction conditions (a hit compound) is left to the chemist to assess. The same hit may be encoded by other libraries, too, and indeed VL frequently returns alternative synthetic protocols for the same hit compound.

Product compounds are enumerated by replacement of the dummy atoms in a scaffold SMILES string by building block SMILES strings. Because the replacements are made using standard character array manipulations available in every programming language, enumeration is extremely fast and no specialized software packages are needed—these were the reasons to prefer the current library encoding format over the more common SMIRKS[11] format.

A software tool with a graphical user interface was implemented to simplify the manual reaction encoding process. The tool has an embedded chemical drawing widget for sketching the scaffold structure and a predefined list of reactant types (corresponding to the 100 SMARTS patterns mentioned above) that can be assigned to each attachment point on the scaffold.

Currently, more than 4000 unique library synthetic protocols have been manually encoded. The protocols have been limited to those which have yielded compounds into the AstraZeneca corporate collection. For the purposes of this publication, experiments were made on ten publicly available reaction schemes (vide infra).

Automatic Derivation of Libraries from Electronic Laboratory Notebook Records. The reaction records from the electronic laboratory notebook (ELN) database were extracted using the program HazELNut from NextMove Software, Ltd.[13] After filtering out purifications and other nonreaction records, approximately 400 000 reaction records remained to be processed. In our setting, the output file format was reaction SMILES.[11] The conversion of one reaction to the virtual library encoding format described above is made as follows.

In the current implementation of the automatic reaction encoding process, only reactions with exactly one product structure are used. An example of such reaction, and the encoding process, is shown in Figure 2.

All reactants are sequentially matched against the product structure using an exact substructure matching algorithm. The matching considers only non-hydrogen atoms. If the reactant is found to be a substructure of the product, the matching atoms are removed from the product and the next reactant is matched

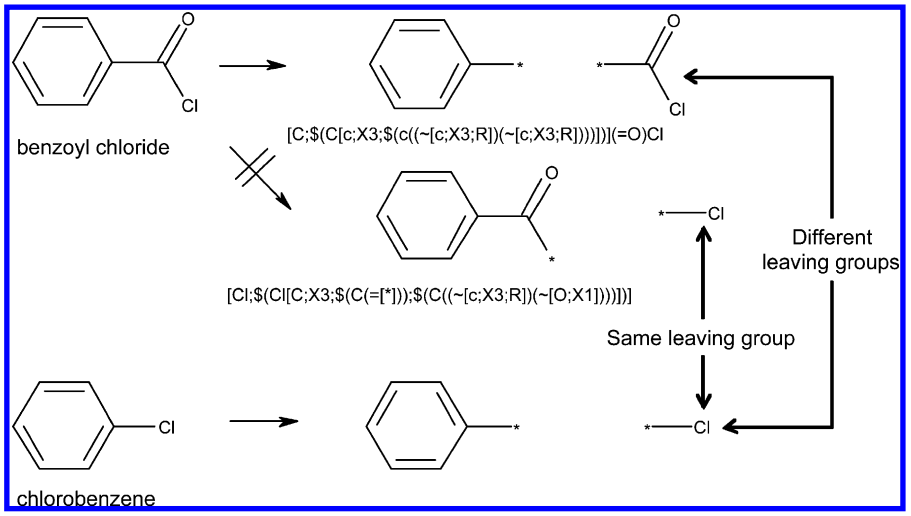


Figure 3. Example of the reactant fragmentation procedure applied to benzoyl chloride (top) and chlorobenzene (bottom). The functional groups acid chloride and aryl chloride are detected by the bond breaking rules, defined as SMARTS patterns. The matched bond is cleaved to produce a product-fragment form of the reactant and a leaving group.

against the remainder of the product. This procedure is repeated until all the reactants have been processed. At this point, the remaining atoms in the product will be used as the "scaffold" of the reaction (step 8 in Figure 2). The first reactant (3-methylaniline) in Figure 2 matches the product (step 1) and is progressed to reactant type SMARTS pattern generation (step 2).

In most cases, the reactant structures do not exactly match the product due to the presence of a leaving group in the reactant. In order to identify and remove the leaving group, reactants that are not substructures of the product are subjected to a fragmentation procedure and the fragments are matched one at a time against the product. The processing of the second reactant (3-fluorobenzoyl chloride) in Figure 2 follows this path (steps 3 and 4).

As a reactant can be fragmented in many different ways, the fragments are processed in the order of increasing number of attachment points (primary criterion) and decreasing size (secondary criterion), i.e., the largest fragment with one attachment point is chosen first. In step 5 in Figure 2, the largest fragment does not match the product, but the second largest one does (step 6), which is subject to reactant type generation (step 7) and the matching procedure for that reactant is terminated.

The fragmentation is made by cutting bonds matched by a set of SMARTS patterns, analogous to the bond breaking rules in retrosynthetic analysis.[14] For example in Figure 3, an aromatic chloride (chlorobenzene) is cut at the Cl−Ar bond, while the acid chloride function of benzoyl chloride is cut at the bond connecting the carbonyl to the remaining substituent, not at the bond connecting the carbonyl to the chlorine atom (the route show in the middle). The latter route would yield the same leaving group (chlorine atom) for the two types of reactants, which would make it more difficult in a later step to generate reactant type SMARTS patterns that would differentiate between these two reactants that are used in different types of reactions. We experimented with a maximum common substructure based approach to match reactants to the product, too, but preferred the fragmentation/exact match based approach because of the ability to explicitly define the bond breaking rules.

The reactant type is a SMARTS pattern generated from the reactant by analyzing the atoms within a distance of two bonds from the attachment point atom, that is, the fragment atom vicinal to the leaving group. For reactants found to be exact substructures of the product, the leaving group is a hydrogen atom. Several atomic properties, such as aromaticity, connectivity, and ring membership, are used in the SMARTS pattern. Examples of reactant type SMARTS patterns are shown in Figures 2 and 3.

In order to avoid any dependency on the order of processing the reactants, all permutations of the reactants up to the first four reactants (24 permutations) are considered. For a reaction with five components (which are rare), the fifth reactant is always used last. Duplicate reaction encodings generated in the permutation process are filtered out and the encoding with the smallest scaffold is retained—a smaller scaffold indicates a better matching of reactants to the product.

Finally, the perceived reactant type SMARTS patterns are applied on the reactants of original reaction in order to obtain the corresponding product-fragments, which are then combined with the perceived scaffold to form the product structure. If the product structure formed by this round-trip test does not match that of the original reaction, the reaction encoding is discarded. Upon successful test, the reactant type SMARTS patterns are added to the list of patterns used to preprocess the available reactants to form building blocks.

Several ELN reactions may lead to the same virtual library encoding. To disambiguate the link between an encoding and the original reaction, the identifier for the reaction with the highest yield is used to annotate the encoding.

The automated library generation procedure on the 400 000 reaction records resulted in approximately 20 000 library encodings. In total, VL contains more than 24 000 libraries that encode approximately 10^{15} product molecules. The exact numbers vary between the regular updates of VL depending on changes in the list of available reactants and new synthetic protocols in the ELN but the number of products does not grow wildly, because only a few protocols with new type of chemistry are introduced at each update.

Fingerprints. For a general introduction to chemical fingerprints the reader is referred to Chapter 6 of the Daylight

Theory Manual.[11] The different types of fingerprints used in this study, listed in Table 1, are 2048-bit binary hashed

Table 1. Chemical Fingerprints Used in This Study[a]

name	type	ref
Foyfi	path atomic property, Daylight-like	Blomberg et al.[15]
Alfi	path pharmacophoric feature	Blomberg et al.[15]
ECfi	circular, ECFP-like	Rogers et al.[16]

[a]All fingerprints were calculated using in-house software.[15] See text for details.

fingerprints generated using in-house software. The procedure to generate a Foyfi fingerprint, described in detail in Blomberg et al.,[15] enumerates all possible paths in the input molecular graph up to a predefined number of bonds (here seven). For each path, an integer number is calculated by recursively hashing the atomic number of the atoms and the types of connecting bonds encountered in the path. The resulting large integer is iteratively divided by the fingerprint length, and on each iteration, the remainder is used to set the corresponding bit. Usually several bits are set for a given path, and consequently, there is no direct correspondence between a specific bit and an atom or substructure.

The Alfi fingerprint type encodes pharmacophoric features instead of atomic numbers as Foyfi does. The features are detected using user-defined SMARTS patterns. Atoms that are not matched by a SMARTS pattern are assigned a default integer label for hashing. Bond types are treated as for the Foyfi fingerprint.[15] The SMARTS patterns used in this work can label the atoms as hydrogen-bond donors or acceptors, acid or basic centers, or aliphatic or aromatic. Under this scheme, for example, phenol and aniline have very similar fingerprints because both consist of an aromatic ring and a donor/acceptor. In order to produce nonidentical fingerprints in such cases, the last nine bits in the Alfi fingerprint are used to indicate the presence of the elements carbon, nitrogen, oxygen, sulfur, fluorine, chlorine, bromine, iodine, and other elements, respectively.

The ECfi fingerprint is an adaptation of the extended-connectivity fingerprint.[16] The procedure to calculate an ECfi fingerprint is analogous to that for Foyfi fingerprint but instead of paths, shells of connected atoms are expanded from each atom in the input molecular graph up to a predefined number of bonds, here three, and the atoms at each level of expansion are then used for hashing and setting 1-bits in the fingerprint.

Basis Product Approach to Screening Virtual Libraries. The screening method implemented in this study, listed in Scheme 2, is based on the method described by Zhou et al.,[8] in which subsets of compounds from a virtual combinatorial library are enumerated substituentwise by varying the substituent for one attachment point on the scaffold and using small capping groups for the other attachment points (here, the building block with the shortest SMILES string). The enumeration produces lists of valid molecular structures, called Basis Products (BPs), for which binary fingerprints are calculated and stored in a database. The database is then screened one library at a time in a two-step procedure: First, each list of BPs is screened separately and the substituents corresponding to the top N highest scoring compounds recorded. Second, the fingerprints for the top N BPs are bitwise OR'ed in all combinations in order to produce an approximation of the fingerprint of the full product molecule

Scheme 2. VL Screening Algorithm[a]

```
Require: 2D structure of a query compound
Compute binary fingerprint for the query compound
Initialize an empty master hitlist
for each library in VL do
    for each attachment point for the scaffold of the library do
        Initialize an empty hitlist for the attachment point
        for each BP fingerprint for the attachment point do
            Calculate score for the fingerprint
            Update hitlist for the attachment point
        end for
    end for
    Initialize an empty hitlist for the library
    for each OR'ed combination of the highest-scoring BP fingerprints do
        Calculate score for the OR'ed fingerprint
        Update library hitlist with the OR'ed fingerprint (product molecule)
    end for
    Merge the library hitlist to the master hitlist
end for
Calculate library scores
Output the master hitlist
```

[a]The implementation is multithreaded on the level of the first for-loop, i.e., libraries are screened in parallel.

(the bitwise OR operator assigns a 1-bit at position i in the result fingerprint if either of the OR'ed fingerprints has a 1-bit at that position), and a score is calculated for the approximated fingerprint. Finally, the highest-scoring product molecules are enumerated and reported.

Similarity Index Score. For scoring fingerprints during screening, we used the Tanimoto index that quantifies molecular similarity. The Tanimoto index between molecules A and B, described by binary fingerprints A and B, is calculated as

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

where $|A \cap B|$ is the number of mutual 1-bits in the fingerprints (population count on the result of the bitwise operation A AND B) and $|A \cup B|$ is the number of 1-bits in the union of the fingerprints (bitwise A OR B). For binary fingerprints the range of the Tanimoto index is from 0 to 1. As can be seen from the right-hand side of eq 1, the Tanimoto index can be computed from the population counts of the individual fingerprints and that of their union. The individual counts can be precomputed and stored together with the fingerprints in a database, thus only the mutual 1-bits need to be computed while screening the database against a query fingerprint. The population count can be computed very efficiently, especially on modern CPUs that implement the POPCNT instruction in hardware.[17]

Recently, Swamidass and Baldi[18] introduced formulas to calculate upper and lower bounds on the number of 1-bits that the target fingerprint may have in order to achieve a given threshold of Tanimoto similarity index.[18] Our fingerprint screening code supports the use of a threshold value for both the Tanimoto index and the number of compounds to collect on a hit list and recalculates the bounds whenever a compound is added on a hitlist that has reached the maximum size. Within the BP approach, the bounds do not hold exactly; therefore, only the higher limit ("too many 1-bits") is applied on the products of the OR'ed BP fingerprints.

Library Score. The purpose of the VL is to recycle synthetic protocols and generate ideas for new chemical libraries. A search in the VL, however, returns a set of hit compounds scored according to one of the methods described

above. Therefore, a library scoring scheme was devised in order to aid the selection of interesting candidate libraries for further analysis. The score for a given library L is the average of the score values over all the hit compounds H originating from the library

$$S_{\text{lib}} = \frac{1}{m} \sum_{H \in L} S(H) \tag{2}$$

where $S(H)$ is the score for the hit compound and m is the maximum limit for the number of hits collected per library during screening.

The library score closely reflects the average similarity of the hits from the library but discredits libraries that produce only few hits and therefore may not be the most interesting ones for the chemist to follow up. The library scores are computed in a postprocessing step after the similarity screening has completed, as listed in Scheme 2.

According to our observations, similarity values calculated using different fingerprint types for the same set of structures generally do not have the same distribution, for example, ECfi tends to yield lower values that Foyfi. Therefore, S_{lib} is computed separately for each query structure and fingerprint type.

Data Sets. Data sets were prepared for two experiments. The first experiment assessed the accuracy of the BP screening method against a set of reactions from literature in comparison with fingerprint screening made on exhaustively enumerated libraries. In the second experiment, VL was queried using a set of compounds extracted from literature in order to exemplify the degree to which VL covers current medicinal chemistry space.

Reactions from Literature. Ten diverse libraries with two attachment points on the scaffold were selected from literature; see Table 2. For each of these libraries, 2 sets of 2000 reactants were randomly selected from the Available Chemicals Directory[19] and internal reactant sets. No physicochemical property filters were applied during the selection of reactants, because we aimed to assess the accuracy of the BP method for retrieving hits irrespective of any drug-likeness criteria. Exhaustive enumeration gave rise to libraries of 4 million (= 2000 × 2000) compounds—still small enough in order to be exhaustively enumerated while they cover a large chemical space around the respective scaffold. In addition to the exhaustively enumerated libraries, the reactions were encoded in the BP format as described above, which gave rise to 4000 (= 2000 + 2000) BPs per library, and 1000 libraries of size of 2500 compounds (50 × 50 reactants) were made by random sampling of the sets of 2000 reactants selected earlier for each attachment point of each library. The different encodings of the 10 libraries thus cover (subsets of) the same chemical space.

Table 2. Reactions from Literature[a]

library	reaction
1	three-component Strecker reaction for the synthesis of α-aminonitriles[20]
2	two-step solid-phase "catch and release" strategy for the synthesis of isoxazoles and 3,4,5-trisubstituted pyrazoles[21]
3	solid-phase Ugi route for the synthesis of diketopiperazines[12]
4	three-component aza-Diels–Alder reaction for the synthesis of tetrahydroquinolines[22]
5	three-component condensation for the synthesis of 3-aminoimidazo[1,2- *a*]pyridines and -pyrazines[23]
6	Ugi three-component condensation, scheme 2 from Zhang et al.[24] was used in this study.
7	Ugi four-component condensation[25]
8	condensation between a hydrazine and an aldehyde to yield an acyl hydrazone, Scheme 4 from Kim et al.[26] used in this study
9	five-step scheme for the synthesis of 2-substituted 4,5-dihydroxypyrimidine-6-carboxamides[27]
10	a textbook amide formation example

[a]These reactions were used to create a set of virtual libraries on which the accuracy of the implemented BP screening method was evaluated.

From each exhaustively enumerated library, 100 query compounds were randomly extracted, which gave a set of 1000 query compounds in total. Similarity searches targeting the exhaustive libraries using these query compounds will always return some hit compounds that are similar to the query, including the query compound itself. Thus, by using a subset of the target libraries as query structures, we avoid the risk of working within a low similarity range and can use the exhaustively enumerated libraries as a positive control for benchmarking the approximative representations of the libraries.

For each target library, the average of the 50 highest Tanimoto index values (most similar compounds) per query was returned.

Compounds from Literature. A set of 626 molecular structures that appeared in issue 13 of the *Journal of Medicinal Chemistry* (Vol. 54, 2011)[31] were extracted from the Gostar database[28] and clustered using an in-house implementation of the Taylor–Butina[29,30] algorithm based on Foyfi fingerprints with similarity threshold set to 0.7. The clustering resulted in subsets (70 singletons) from which the cluster centroids were extracted. This set of representative structures was then submitted as queries to VL search with a maximum limit of 100 hits per query per fingerprint type and a Tanimoto similarity threshold of 0.4.

RESULTS AND DISCUSSION

Basis Product versus Full Compound Searches. The accuracy of the approximation of the fingerprint for the full product molecule depends on the capping groups used for creating the BPs. The capping groups are usually small; therefore, they have fewer features that are encoded as 1-bits in the binary fingerprint than a "real" substituent would have. This does not introduce error if and only if the features present in the capping groups are a subset of features present in all of the "real" substituents. However, this is not the case in general and some of the features present in the capping group may not be present in a "real" substituent, thus the approximated fingerprint may have some 1-bits that the fingerprint for the full product molecule would not have.

Another source of approximation error are features that extend over the scaffold and a substituent (or several substituents if the scaffold is small) in the full product molecule—these features may not be encoded in the approximated fingerprint, thus the approximated fingerprint may not have some of the 1-bits that the fingerprint for the full product molecule would have.

The approximation error can be quantified by comparison of the BP approximated similarity index value S_{BP} with the value calculated for the fully enumerated library compound S_{full} against a query molecule. Another way of quantifying the error is to calculate the similarity index between the approximated and exact fingerprints for the library compounds. Results of such comparisons on the set of example libraries

from the literature are listed in Table 3; the correlation coefficients r calculated between the Tanimoto similarity values

Table 3. Correlation between the BP Approximated and Fully Enumerated Compound Tanimoto Similarity Values (S_{BP} and S_{full}, Respectively) for Compounds (10 Libraries against 1000 Query Fingerprints, Each Query Returns 50 Hit Compounds)[a]

	r (S_{BP}, S_{full})	$\overline{S_{self}}$	Stdev (S_{self})
AlFi	0.88	0.80	0.10
FoyFi	0.98	0.93	0.05
ECFi	0.99	0.82	0.05

[a]The self similarity S_{self} is the Tanimoto similarity index between the bitwise OR'ed BP fingerprints and the full compound fingerprint.

are moderately high but the BP approach is obviously a fuzzy method. The selection of capping groups used in the BP approach can be optimized against the measures listed in Table 3; however, no such optimization is made in the current implementation of the method but the building block with the shortest SMILES string was used in order to minimize the number of 1-bits set by the capping group.

The BP approximation error was further quantified for 100 randomly selected query compounds from the literature example library 2. The query compounds were used to search the exhaustively enumerated libraries, the BP approximated libraries (for which the approximated and the exact similarity values between the hit and the query fingerprints were returned), and the 1000 randomly selected subsets of the exhaustively enumerated libraries. The distributions of the average similarity for the 50 highest-scoring compounds $\overline{S}$ is depicted in Figure 4. The left-hand side of Figure 4 shows the distributions of $\overline{S}$ for searches targeting library 2. Searches in the exhaustively enumerated library show the highest $\overline{S}$ as expected. The BP approach yields slightly lower similarities than the exhaustive searches and the exact similarities ("BP_full") but on average higher than searching the random samples of the libraries ("Random"). The trend is observed for all fingerprint types.

For the searches using queries from library 2, the BP method should return the query structure with very high rank. However, this is not always the case as seen from Table 4 that lists the number of times a query was returned within the set of 50 highest-scoring compounds. This observation prompted us to introduce the library score.

ECfi shows higher query retrieval rates in Table 4 than the other fingerprint types. A putative explanation for this observation is the way the molecular graph is traversed while encoding the fingerprints: Paths in the molecular graph, which are encoded up to length of seven bonds in Alfi and Foyfi, can span over the attachment bond between the library scaffold and substituents and are therefore disrupted in the BP approach more frequently than the radial atom environments encoded up to radius of three bonds in ECfi. Moreover, Alfi abstracts atom types with pharmacophoric features that are a more general feature than the atomic numbers used in Foyfi and ECfi, which may lead to identical Alfi fingerprints for structures that Foyfi or ECfi would not consider equal. Whether the generalization is desirable or not depends on the use case, for example the hits from VL can be submitted to a pharmacophore matching analysis, therefore it is desirable to have Alfi fingerprints included in VL.

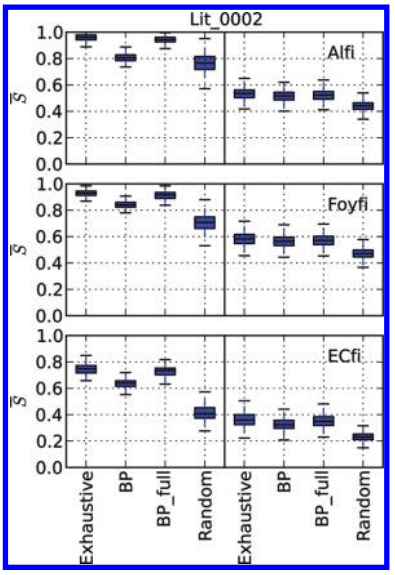


Figure 4. Distributions of average similarity index values $\overline{S}$ for the top 50 hits for 100 randomly selected query compounds from the literature example library 2. The columns plot data for searches targeting the exhaustively enumerated libraries ("Exhaustive"), BP approach ("BP"), BP approach with full enumeration of the hit compounds ("BP_full"), and 1000 randomly selected samples of the exhaustively enumerated libraries ("Random"). The left half of the plot shows the distributions for searches targeting library 2, from which the query structures were selected. The right half of the plot shows the average results for searches targeting the other libraries. Each box shows the upper and lower quartiles of the distribution and the whiskers extend to the extreme data points within 1.5 quartiles from the median value. See text for discussion.

Table 4. Number of Times a Query Structure Was Retrieved in Searching the Literature Example Libraries Using the BP Approach[a]

library	Alfi	Foyfi	ECfi
1	57	79	99
2	84	87	100
3	58	83	100
4	56	82	100
5	79	95	100
6	83	81	100
7	40	77	100
8	70	81	100
9	95	88	100
10	79	77	100

[a]For each library, 100 query structures were randomly selected. See text for details.

No experiments were made to compare the speed of the BP method with exhaustive search times for two reasons. First, such timing experiments are very dependent on hardware and the implementation of the population count algorithm[17] and may not generalize. Assuming that the fingerprints are preloaded into memory (as in VL), an exhaustive search using Tanimoto similarity threshold based heuristics[18] and hardware instruction based population count may well be faster than the BP method. However, when the search space grows to 10^{10} compounds and beyond, the storage requirements for exhaustive search grow to terabyte level, which is impractical. That is the second and compelling reason for us to omit timing

dx.doi.org/10.1021/ci300157m | J. Chem. Inf. Model. 2012, 52, 1777−1786

experiments and just note that the BP approach is fast enough for practical purposes; a typical run time of VL for a single query structure is 2 min on the current hardware.

Library Retrieval. The ability of the BP approach to retrieve the correct library for a query structure was tested on the sample of 10 libraries from literature. The average of Tanimoto similarity index values over the top 50 hits for all query structures from a library are plotted in Figure 5, where a row corresponds to the library where the queries were selected from and columns correspond to the library the hits originated from. The diagonal in each subplot is lighter than other areas, which indicates that hits from the source library of the queries were scored higher than hits from other libraries. The signal is clear, especially for ECfi fingerprint. On the basis of these results, we consider the BP approach to be able to detect the correct (source) library for the query compounds.

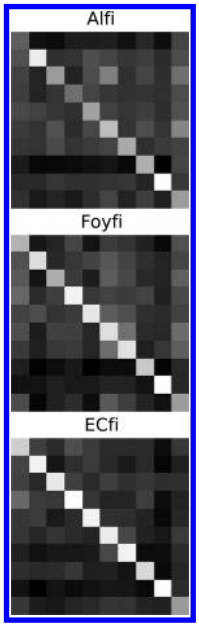


Figure 5. A heatmap of average Tanimoto similarity index values $\overline{S}_{BP}$ over the top 50 hits found by BP approach for 100 query compounds originating from each literature example library screened against all the libraries. Libraries are plotted from 1 to 10 starting from top and left. Diagonal elements represent searches targeting the same library that the query compounds originated from. The values of $\overline{S}_{BP}$ are encoded in shades of gray: zero is shown in black, and one, in white. The subplots correspond to different fingerprint types. See text for details.

The library score values averaged over the 100 query compounds from the literature example libraries are plotted in Figure 6. As was seen for the average Tanimoto similarity values in Figure 5, the library score is on average higher for the library from which the query structure originated from. This result is of course trivial in the current test setting but serves as a sanity check for the library score. Table 5 lists the number of times the source library of a query compound obtained the highest library score. The values for Alfi and Foyfi fingerprints are higher compared to the frequencies of query retrieval in Table 4. The library score selects the correct library in the majority of cases, with library 1 detected the least number of times when using Alfi fingerprint. Apparently, the pharmacophoric abstraction of atom types used in the generation of Alfi fingerprints is misleading in this case: The attachment atom of the capping group for one attachment point is labeled as a hydrogen bond acceptor but for the majority of the reactants for that attachment point the corresponding atom is labeled hydrophobic. The attachment atom is included in many paths in the molecular graph and the discrepancy in atom type labels between the BP and the query compounds leads to different sets of 1-bits when the paths are hashed to give the Alfi fingerprints.

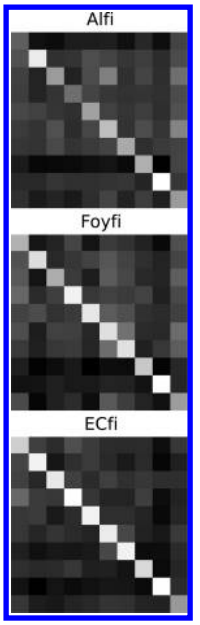


Figure 6. A heatmap of average library score values $\overline{S}_{lib}$ over the 100 query compounds originating from each literature example library screened against all the libraries. Libraries are plotted from 1 to 10 starting from top and left. Diagonal elements represent searches targeting the same library that the query compounds originated from. The values of $\overline{S}_{lib}$ are encoded in shades of gray: zero is shown in black, and one, in white. The subplots correspond to different fingerprint types. See text for details.

Table 5. Number of Times the Source Library of the Query Structure Obtained the Highest Library Score in Searching the Literature Example Libraries Using the BP Screening Method[a]

library	Alfi	Foyfi	ECfi
1	57	99	100
2	100	100	100
3	82	99	100
4	88	100	100
5	99	100	100
6	90	100	100
7	98	100	100
8	100	100	100
9	100	100	100
10	99	96	100

[a]For each library, 100 query structures were randomly selected. See text for details.

Compounds from Literature. The purpose of this exercise is to give some flavor to the reader of the extent to which the VL covers the current medicinal chemistry space. A diverse set of query structures was extracted from *J. Med. Chem.* vol. 54 issue 13,[31] as described in Methods. The distribution of the nearest neighbor similarity values of the virtual hits found in the VL is shown in Figure 7. There were four, five, and five

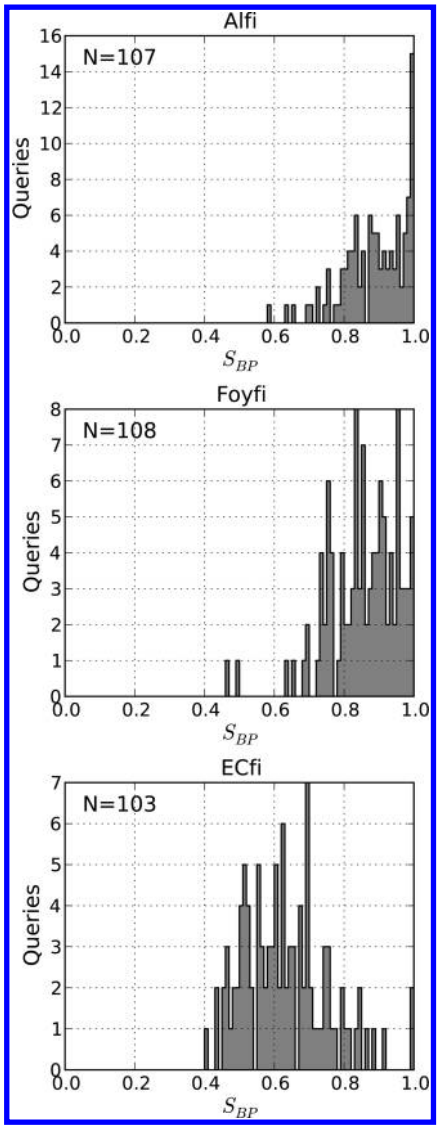


Figure 7. Histograms of the nearest neighbor Tanimoto similarity of query structures extracted from *J. Med. Chem.* vol. 54 issue 13.[31] *N* is the number of nearest neighbors returned. See text for details.

exact matches found by the Alfi, Foyfi, and ECfi fingerprints, respectively, of which were unique exact matches. Moreover, the right-hand tails of the distributions for Alfi and Foyfi lean toward Tanimoto index equal to 1.0 indicating that very similar virtual compounds to the literature queries were found. Because the VL web application automatically creates links to a synthesis scheme and reactants, a synthesis plan can be suggested for these virtual hits.

Some of the query compounds did not have a nearest neighbor within the applied similarity cutoff. These "missed" query compounds would make good starting points for traditional library design in order to enrich the HTS screening collection—assuming the missed compounds were reported active against a biological target of interest, are not available for acquisition, and do not already exist in the compound collection.

Overall, the distributions of similarity values for the nearest neighbors in Figure 7 peak at levels that most practitioners would consider "fairly similar".

CONCLUSIONS

The Basis Product approach combined with different 2D fingerprints and scoring schemes provides a versatile tool for the exploration of the synthesizable chemical space around a seed compound. Despite the fuzzy nature of the search procedure, it frequently returns relevant virtual hits, nearly parallel to those returned by exhaustive screens on the same chemical space. Because the virtual hits are annotated with a validated synthetic protocol, the chemistry turnaround time is expected to be short. The virtual libraries are regenerated from the ELN on a regular basis; therefore, the protocol for each library presents the latest highest yield scheme for that type of chemistry. Inclusion of reactions from external sources, such as commercial reaction databases and patents, will further increase the diversity of chemistry accessible via VL in the future.

The main objective of the method developed here is to prioritize the virtual libraries, rather than individual compounds, for rapid exploration of structure−activity relationship. In the light of the results obtained in this study, and the initial experiences from the deployment of VL in production, we conclude the objective is well met.

AUTHOR INFORMATION

Corresponding Author

*E-mail: thierry.kogej@astrazeneca.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. David A. Cosgrove for help with the code for computing the fingerprints, Dr. Roger Sayle for developing HazELNut, Dr. Sorel Muresan for suggestions on the manuscript, and Dr. Christian Tyrchan for implementing the software for manual library encoding. Petra Båth, Anna Said, Linda Nilsson, and numerous chemists at AstraZeneca are acknowledged for the manual encoding of virtual libraries.

REFERENCES

(1) Brickmann, K. Compound collection enhancement (CCE) increasing quality and size of the AZ compound collection. *Presented at the 4th International Conference on Compound Libraries*, Düsseldorf, Germany, October 6−8, 2008.

(2) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010−1023.

(3) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49*, 270−279.

(4) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. LEAP into the Pfizer Global Virtual Library (PGVL) Space: Creation of Readily Synthesizable Design Ideas Automatically. In *Chemical Library Design*; Zhou, J. Z., Ed.; Methods in Molecular Biology; Humana Press: New York, 2011; Vol. *685*, pp 253−276.

(5) Downs, G. M.; Barnard, J. M. Techniques for Generating Descriptive Fingerprints in Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 59−61.

(6) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497−520.

(7) Andrews, K. M.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.

(8) Zhou, J.; Shi, S.; Na, J.; Peng, Z.; Thacher, T. Combinatorial library-based design with Basis Products. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 725−736.
(9) Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R. D. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J. Mol. Graph. Modell.* **2000**, *18*, 452−463.
(10) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(11) Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/index.html (accessed December 1, 2011).
(12) Szardenings, A. K.; Antonenko, V.; Campbell, D. A.; DeFrancisco, N.; Ida, S.; Shi, L.; Sharkov, N.; Tien, D.; Wang, Y.; Navre, M. Identification of Highly Selective Inhibitors of Collagenase-1 from Combinatorial Libraries of Diketopiperazines. *J. Med. Chem.* **1999**, *42*, 1348−1357.
(13) NextMove Software Ltd, Cambridge, England; http://www.nextmovesoftware.com/ (accessed February 14, 2012).
(14) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP − Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(15) Blomberg, N.; Cosgrove, D.; Kenny, P.; Kolmodin, K. Design of compound libraries for fragment screening. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 513−525.
(16) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(17) Haque, I. S.; Pande, V. S.; Walters, W. P. Anatomy of High-Performance 2D Similarity Calculations. *J. Chem. Inf. Model.* **2011**, *51*, 2345−2351.
(18) Swamidass, S. J.; Baldi, P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. *J. Chem. Inf. Model.* **2007**, *47*, 302−317.
(19) Accelrys, Inc.: San Diego, CA, USA; http://accelrys.com/ (accessed March 5, 2012).
(20) Jarusiewicz, J.; Choe, Y.; Yoo, K. S.; Park, C. P.; Jung, K. W. Efficient Three-Component Strecker Reaction of Aldehydes/Ketones via NHC-Amidate Palladium(II) Complex Catalysis. *J. Org. Chem.* **2009**, *74*, 2873−2876.
(21) Ma, W.; Peterson, B.; Kelson, A.; Laborde, E. Efficient Synthesis of Trisubstituted Pyrazoles and Isoxazoles Using a Traceless "Catch and Release" Solid-Phase Strategy. *J. Comb. Chem.* **2009**, *11*, 697−703.
(22) Legros, J.; Crousse, B.; Ourévitch, M.; Bonnet-Delpon, D. Facile Synthesis of Tetrahydroquinolines and Julolidines through Multi-component Reaction. *Synlett* **2006**, *2006*, 1902−1899.
(23) Shaabani, A.; Soleimani, E.; Maleki, A. One−Pot Three−Component Synthesis of 3-Aminoimidazo[1,2-*a*]pyridines and -pyrazines in the Presence of Silica Sulfuric Acid. *Monatsh. Chem.* **2007**, *138*, 73−76.
(24) Zhang, J.; Jacobson, A.; Rusche, J. R.; Herlihy, W. Unique Structures Generated by Ugi 3CC Reactions Using Bifunctional Starting Materials Containing Aldehyde and Carboxylic Acid. *J. Org. Chem.* **1999**, *64*, 1074−1076.
(25) Keating, T. A.; Armstrong, R. W. Postcondensation Modifications of Ugi Four-Component Condensation Products: 1-Isocyanocyclohexene as a Convertible Isocyanide. Mechanism of Conversion, Synthesis of Diverse Structures, and Demonstration of Resin Capture. *J. Am. Chem. Soc.* **1996**, *118*, 2574−2583.
(26) Kim, Y.; Koh, M.; Kim, D.-K.; Choi, H.-S.; Park, S. B. Efficient Discovery of Selective Small Molecule Agonists of Estrogen-Related Receptor γ using Combinatorial Approach. *J. Comb. Chem.* **2009**, *11*, 928−937.
(27) Boyd, V. A.; Mason, J.; Hanumesh, P.; Price, J.; Russell, C. J.; Webb, T. R. 2-Substituted-4,5-Dihydroxypyrimidine-6-Carboxamide Antiviral Targeted Libraries. *J. Comb. Chem.* **2009**, *11*, 1100−1104.
(28) *GOSTAR Online Structure-Activity Relationship Database*; GVK Biosciences Private Limited: Hyderabad, India; http://gostardb.com/gostar/ (accessed February 15, 2012).
(29) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(30) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.
(31) *J. Med. Chem.* **2011**, *54*, 4283−4936