JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

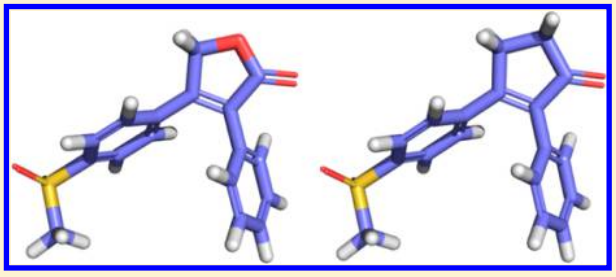
pubs.acs.org/jcim

Scaffold Hopping by Fragment Replacement

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

Discovery Sciences Chemistry Innovation Centre, AstraZeneca R&D, Pepparedsleden 1, 43186 Mölndal, Sweden

ABSTRACT: This work describes a data driven method for scaffold hopping by fragment replacement. A search database of scaffolds is created by cutting bonds of existing compounds in a combinatorial fashion. Three-dimensional structures of the scaffolds are then generated and made searchable based on the relative orientation of the broken bonds using an auxiliary index file. The retrieved scaffolds are ranked using volume overlap and electrostatic similarity scores. A similar approach has been used before in the program CAVEAT and others. The present work introduces a novel indexing scheme for the attachment vector geometry, which allows for fast searching. A scaffold shape descriptor is defined, which allows for queries with a single attachment vector (R-groups) and improves the shape similarity between the query and the suggested replacement fragments. The program, called Scaffold Hopping, is shown to retrieve relevant bioisosteric replacement scaffolds for a set of example queries in a reasonable time frame, making the program suitable to be used in drug design work.



INTRODUCTION

Scaffold hopping (or lead hopping) is one of the most common tasks in rational drug design. In scaffold hopping, a central part (scaffold) of a molecule is replaced by another chemical moiety while the biological activity of interest is retained. This replacement aims to solve biological, chemical, or intellectual property issues attributed to the scaffold of a lead series in a drug discovery project. More specifically, scaffold hopping is often used to improve the ADMET (absorption, distribution, metabolism, excretion, and toxicity) profile such as solubility, toxicity, metabolic liabilities, to improve potency or selectivity toward the target protein, or to create a compound that is analogous to a patented compound but not covered by the patent.[1−6]

Scaffold hopping has a long history as a concept, and a wide range of computational approaches that are capable of scaffold hopping exists, including 3D shape based similarity search, fingerprint based similarity search, pharmacophore matching, and fragment replacement techniques.[1−6] The latter is, in the context of drug design, equivalent to bioisosteric replacement, where the prefix "bio" is used to denote equipotency[4−8] and overlaps with the concept of a matched pair,[9] that is defined as a pair of compounds that differ only by a single-site substructure change. Taken together, what is referred to as a scaffold hop in this study is a bioisosteric fragment replacement operation that creates a matched pair where activity is retained.

Sun et al.[1] proposed a categorization of scaffold hops according to the degree of change relative to the query molecule. According to their scheme, a heterocycle replacement, such as replacement of a pyrrole with a thiazole, is a 1° hop. An example of a 1° hop is shown in Figure 1, where the now-withdrawn drug rofecoxib was used as query from which another active compound was created by the replacement of a central 5-membered ring. A ring-opening or closure is a 2° hop.

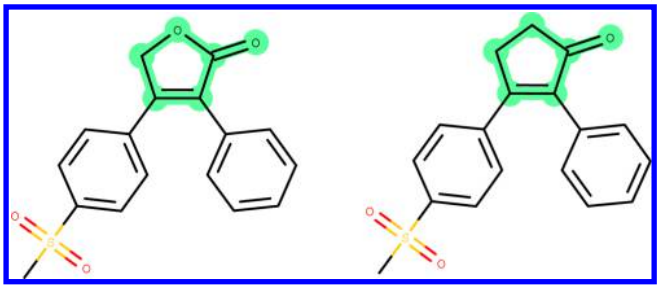


Figure 1. Example of a 1° scaffold hop. A cyclooxygenase-2 (COX-2) inhibitor rofecoxib, shown on the left, was used as a template structure. The program described in this study was used to replace the highlighted part of rofecoxib with other fragments. The compound shown on the right, which was among the 100 top-ranked virtual hit compounds, has a reported IC_{50} value of 11 nM toward COX-2.[53] See text for details.

A ring closure may improve binding free energy due to reduced loss of entropy upon binding by locking a compound to its bioactive conformation. A ring-opening may improve solubility due to increased degrees of freedom. Transition from a peptide to a peptidomimetic structure is a 3° hop. Shape or topology based virtual screening may return hits that do not share (nontrivial) common substructures (substituents) with the query. These whole-compound transitions, which are not matched pairs, are categorized as 4° hops.

The program CAVEAT[10] was a seminal work in the area of scaffold hopping by fragment replacement. Other programs using variations of the same methodology as CAVEAT, or aiming at the same end point of replacing a defined scaffold in a 3D structure, are Brood,[11] sparkV10,[12] Core Hopping,[13]

Received: February 11, 2013
Published: July 4, 2013

dx.doi.org/10.1021/ci4001019 | J. Chem. Inf. Model. 2013, 53, 1825−1835

Scaffold Replacement,[14] SHOP,[15] ParaFrag,[16] and ReCore.[17] The input to these programs is a 3D structure of an active compound or just a scaffold part of it, with a few marked bonds to be used as attachment vectors that connect the scaffold to substituents (R-groups). All these programs perform a search in predefined databases of fragments and return matching 3D structures that can be used to replace the original scaffold. The suggested new scaffolds are then organized according to their size, predicted physicochemical properties, pharmacophoric features, common structural framework, or other features that can help the user to view the results and access the potential of the suggested structural replacement.

Against the categorization proposed by Sun et al.,[1] fragment replacement methods can find 1°, 2°, and 3° scaffold hops but are intrinsically incapable of 4° hops (in a single run) because of the need to preserve at least one substituent.

This study describes a method for scaffold hopping by fragment replacement. The implementation, simply titled Scaffold Hopping, builds upon and extends the methodology of CAVEAT and uses shape based virtual screening technology to score hits. At first approximation, it differs from CAVEAT in the way the 3D configuration of the attachment vectors of fragments is indexed, in the cascade of filters applied to candidate fragments during screening, and on the final scoring of hit scaffolds. A qualitative assessment of the performance of the method is made based on its ability to retrieve known bioisosteres of a set of example queries.

The Scaffold Hopping program is made available to the computational chemistry community of AstraZeneca as a web application. A web service interface is provided, too, in order to allow integration with other chemical structure design software packages.

MATERIALS AND METHODS

Reference Chemical Structures. Reference structures were obtained from three sources: GOSTAR,[18] eMolecules,[19] and the AstraZeneca internal compound registry. The structures were extracted from the proprietary Chemistry Connect database, which collates chemical information from a number of public and proprietary sources and stores molecular structures standardized according to AstraZeneca in-house chemistry business rules.[20]

Preparation of a Fragment Database. A database of 3D fragment structures to be searched was derived as follows.

The fragment generation process cuts a set of acyclic bonds of an input structure according to user-defined rules and enumerates the resulting fragments in a combinatorial manner. A hard-coded restriction of the algorithm prevents cyclic bonds from being cut. The user-defined rules are SMARTS[21] substructure search definitions that match the types of bonds to be cut. Two very general bond types were used here, namely "[D2,D3,D4,D5,D6]-[D2,D3,D4,D5,D6]" that targets single bonds between nonterminal atoms, and "[!#1]-a" that targets single bonds between a non-hydrogen and an aromatic atom. The aim of this general bond cutting scheme is to maximize the diversity of the generated fragments and therefore increase the chance of finding interesting scaffold replacements. This might have a detrimental effect on the synthetic feasibility of the resulting fragments compared to the usage of some other bond cutting scheme based on retrosynthetic analysis, such as the RECAP[22] rules. However, synthetic feasibility is explicitly addressed in later steps of the method as described below.

Structures with more than 15 breakable bonds were not processed due to the exponentially increasing runtime of the combinatorial fragment enumeration algorithm. Furthermore, fragments with more than 3 rotatable bonds or more than 20 non-hydrogen atoms were discarded in order to limit the size of the search space. These limitations should have a marginal impact on the perceived quality of the retrieved scaffolds, because according to the authors' experience, chemists usually prefer simple cyclic scaffolds over flexible acyclic linkers.

A cut bond is marked as an attachment vector, a bond between a *base* atom (**b**) that is part of the scaffold and a *tip* atom (**t**) that is not part of the scaffold and marks the other end of the vector. A hydrogen atom with a nonphysical isotope value of four is used as a tip atom in order to perform conformational analysis on the scaffold structures.

The AstraZeneca corporate collection, the GOSTAR, and the eMolecules databases were used as sources of input compounds for the fragment database generation, which gave a total of ∼10 M unique structures after removing duplicates. These compounds were either made in-house, reported in the literature, or are available for purchase; this implies some degree of synthetic feasibility for the fragments as they originate from validated chemistry. The fragment database stores connections between generated fragments and their parent structures. This enables the method to return useful additional information such as the frequency of occurrence for each fragment, and link out to Chemistry Connect[20] for the chemical and biological profiles of the parent compounds. More than half (∼8 M) of the initially created ∼14 M fragments occurred only in a single parent structure. These so-called singleton fragments were discarded for two reasons. First, the risk of including wrongly drawn, chemically unfeasible fragments (e.g., pentavalent carbon atoms) in the database is mitigated. The probability of occurrence of identical unfeasible fragments in several input structures is low; therefore, a wrongly drawn structure can be expected to produce a singleton fragment. The risk of encountering wrongly drawn fragments will become non-negligible if input compounds are sourced via automated structure extraction methods, for example, text mining in patent databases.[23] Second, fragments present in a chemical series are assumed to be more synthetically accessible than singleton fragments; therefore, the removal of the singletons should leave us with more tractable chemistry.

As the scaffold replacement search is performed in 3D, the 2D fragment structures (in canonical SMILES[24] format) were converted to 3D structure models using CORINA version 3.60.[25] The defined stereocenters were preserved and stereoisomers were enumerated when undefined stereocenters were present, hydrogen atoms were added to the models, and alternative ring conformations were generated (options "-d wh,rc,de=20,stergen,preserve"). The 3D structures were then subject to a conformational analysis using OMEGA[26] version 2.4.6, with an energy window of 5 kcal/mol, no ring conformer search, and a maximum of 255 output conformers (options "-includeInput -enumRing false -fromCT false -ewindow 5.0 -maxconfs 255 -rms 0.1"). Properties of the resulting conformer database are listed in Table 1.

3D Index Creation. In order to allow fast searching, the generated conformer database was indexed according to the relative geometry of the attachment vectors. The procedure described here does not apply to fragments with only one attachment vector. For those, a shape descriptor index was generated instead (see below).

Table 1. Statistics of the Database of Fragments Used in This Study[a]

attachment vectors	fragments	conformers	index records
1	1 511 279	182 709 238	N/A
2	2 609 177	339 483 359	1 620 755 066
3	1 849 112	239 726 085	614 455 053
4	737 119	93 682 242	93 682 242
total	6 706 687	855 600 924	2 328 892 361

[a]See text for details.

In CAVEAT, the search phase was made efficient by the use of an index file to the 3D structures of alternative scaffolds.[10] The indexing scheme was based on the relative spatial orientation of the attachment vectors. For a pair of attachment vectors, defined by atoms ($\mathbf{b}_1$–$\mathbf{t}_1$) and ($\mathbf{b}_2$–$\mathbf{t}_2$), the distance d ($|\mathbf{b}_1$–$\mathbf{b}_2|$), the dihedral angle δ ($\mathbf{t}_1$–$\mathbf{b}_1$–$\mathbf{b}_2$–$\mathbf{t}_2$), and the angles α_1 ($\mathbf{t}_1$–$\mathbf{b}_1$–$\mathbf{b}_2$) and α_2 ($\mathbf{t}_2$–$\mathbf{b}_2$–$\mathbf{b}_1$) were canonicalized and binned into user-defined intervals. Figure 2 illustrates these measures.

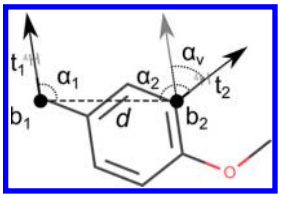


Figure 2. Schematic representation of the measures used to describe the relative geometry of two attachment vectors defined by atoms ($\mathbf{b}_1$–$\mathbf{t}_1$) and ($\mathbf{b}_2$–$\mathbf{t}_2$). Distance d ($|\mathbf{b}_1$–$\mathbf{b}_2|$) and angles α_1 ($\mathbf{t}_1$–$\mathbf{b}_1$–$\mathbf{b}_2$), α_2 ($\mathbf{t}_2$–$\mathbf{b}_2$–$\mathbf{b}_1$), and α_V ($\mathbf{t}_1$–$\mathbf{b}_1$, $\mathbf{t}_2$–$\mathbf{b}_2$) are binned into predefined intervals and the bin numbers are used to construct an integer hash value for fast searching of compatible attachment vector geometries. The dihedral angle δ, formed by the normals of planes ($\mathbf{t}_1$–$\mathbf{b}_1$–$\mathbf{b}_2$) and ($\mathbf{b}_1$–$\mathbf{b}_2$–$\mathbf{t}_2$), was used in CAVEAT[10] but is not used in this work. See text for details.

The index file was sorted according to the bin numbers for fast searching. Tolerance in the query parameters was introduced by searching all the bins that fall within the limits of the tolerance. All records (structures) that contained at least one matching vector pair for each query vector pair were collected and subsequently checked for simultaneous fulfillment of the constraints of the query.[10]

The indexing scheme implemented in the current method borrows ideas from CAVEAT[10] and from ref 27, where, for three-connected scaffolds, the array of the three distances d between the attachment vector base atoms and the corresponding angles α_v between the vectors was canonicalized by sorting in increasing order according to the distance values. It should be emphasized that the angles α_v are not dihedral angles as in CAVEAT. It was also noted that the distances and angles are discrete rather than continuous due to the limited size of the scaffold structure and the nature of chemical bonding, where the bond lengths and angles assume preferred values according to the hybridization state and atomic number of the involved atoms.[27] The approach in the present study builds on this notion; here, given a pair of attachment vectors the distance d is binned to 32 intervals between 1.0 and 37.0 Å, where the interval width grows progressively from 0.5 to 2.0 Å. The angle α_v is binned into eight intervals of a width of 22.72° (the last bin is narrower). In the binary numeral system, the angle bin number (values 0–7) can be represented using three bits and the distance bin number (values 0–31) using five bits. The relative geometry of a pair of attachment vectors can be therefore compactly represented by combining the bin numbers into a single eight-bit byte using bitwise left-shift and OR operations. Figure 3 illustrates the packing of decimal numbers in a binary byte.

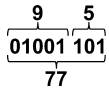


Figure 3. Illustration of the packing of distance and angle interval bin numbers into an eight-bit byte. The angle bin number 5 in the decimal number system is represented by three bits (101) in the binary number system, and the distance angle bin number 9 by the remaining five bits (01001), where the bits are read from right to left. The value of the whole eight-bit byte is 77 in the decimal number system. Four bytes, sorted in decreasing order according to their decimal value, constitute a 32-bit hash number. See text for details.

For scaffolds with two attachment vectors (one pair), a more accurate representation of the geometry is obtained by the construction of an additional byte as follows. The angle values α_1 and α_2 (see Figure 2) are ordered so that $\alpha_1 \leq \alpha_2$ and binned as described above. A one-byte binary representation is then derived from the bin numbers so that the bin number for α_1 is represented by the three lowest bits and the bin number for α_2 by bits 4–6.

For scaffolds with three or four attachment vectors, there are three or six pairs, respectively, for which a (d, α_v) byte representation of the relative geometry is derived. In these cases, the bytes are sorted in descending order.

The four largest of the ordered bytes are combined into a 32-bit unsigned integer *hash* (padding null bytes are added for two and three attachment vector cases). The hash thus describes the relative geometry of the attachment vectors in the scaffold structure in a canonical manner. The second byte used for the two vector cases increases the discriminative power of the hash. Because only internal coordinates of the system are used to derive the value, it is invariant under translation and rotation. In contrast, it does not unambiguously define the geometry; for three vector pairs, an additional fixed internal coordinate would be needed in order to make the description unambiguous, and for six vector pairs only the four largest bytes are used to derive the hash value.

A query scaffold with two attachment vectors can match a pair of vectors of a database scaffold with three or four attachment vectors. Any nonmatching vectors are simply ignored in downstream processing, with the tip atom left as a hydrogen atom (without the isotope four flag). In order to be able to search for subset matches, all combinations of two and three vector pair hashes are generated for scaffolds with three or four attachment vectors.

The hash is used as a search key in a binary file that stores a record for each scaffold conformation. The number of records in the database is listed in Table 1. The records are sorted according to the hash value and can be therefore searched in $O(\log_2 n)$ time with respect to the number of records in the file n. Because disk access is orders of magnitude slower than access to memory, a subset of the file offsets where a particular hash can be found is stored in the file, too, and read into memory at program start for quickly narrowing down the region of the file needed to scan during the search phase. Moreover, records are stored in separate files according to the number of vector pairs

dx.doi.org/10.1021/ci4001019 | J. Chem. Inf. Model. 2013, 53, 1825–1835

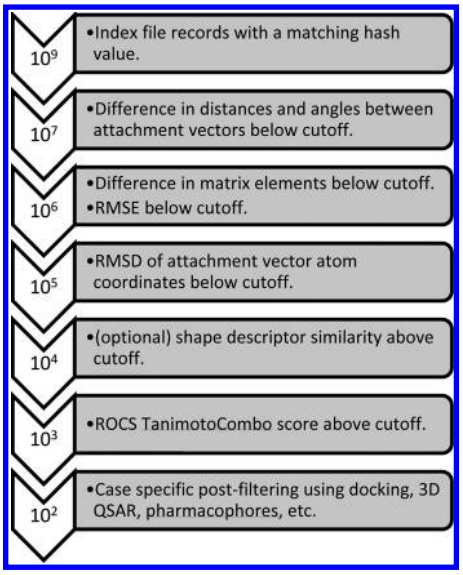
in the hash so that a query will be compared only to records with the same number of vectors as in the query.

3D Index Searching. Hash values are generated for a query scaffold using the same procedure as for the search database scaffolds described above. However, no subset combinations of attachment vectors are used for generating query hash values, because all attachment vectors of a query need to be matched simultaneously.

Some tolerance in matching the distances and angles is desirable in the search procedure. For a query scaffold, several hashes are generated by offsetting the distance and angle bin numbers by −1, 0, or 1 in all possible combinations in each byte (honoring the bin number value limits, of course), resorting the bytes and forming the 32-bit integer hash as described above. This procedure gives rise to at most 9 (3^2) hashes per byte. As the number of bytes (vector pairs) increases, the number of resulting combinations of bin numbers and corresponding hashes grows exponentially, which is the reason that only four out of the six possible pairs for a four attachment point scaffold is used in this study: disk-based seeking of the index file is relatively slow, therefore searching for $9^4 = 6561$ instead of $9^6 = 531\,441$ query hashes was deemed as a reasonable trade-off.

The search procedure, illustrated in Scheme 1, is a cascade of filters where each step discards a large fraction of the index records that passed the previous filter. A record in the index file contains the information needed in each step so that no additional disk file seeks are needed until the very last step(s): the hash value, coordinates of the attachment vector atoms, angle(s) between attachment vectors, and distance(s) between the base atoms, offset to the structure file, a unique identifier for the structure record, and a sparse matrix of distances between all combinations of base and tip atoms excluding distances between the base and tip atoms in the same attachment vector. The sparse matrix is stored in a canonical form as a sorted array of distance values, where the distances between base atoms have a higher priority and occur first in the array, followed by distances between base and tip atoms. Distances between tip atoms are the last elements. Within each priority class, the distance values are sorted in ascending order.

Scheme 1. Scaffold Hopping Search Procedure[a]



[a]A search in the Scaffold Hopping database uses several consecutive filtering steps to funnel down the number of index records. A typical number of records that passes each step is indicated on the left. The information needed on the scaffold geometry in each step is stored in the binary index record file until the optional shape descriptor similarity step, thereafter the 3D structure of the scaffold needs to be fetched from a structure file. The last step is performed by the user outside of the Scaffold Hopping application. See text for details.

During a search, a database record T, that has the same hash as the query record Q, is discarded if any of the following comparisons with Q fails (see Scheme 1): (a) find a permutation of attachment vectors that satisfies the base atom distance and vector angle difference cutoffs (with default values of 1.0 Å and 30°, respectively), (b) check that the relative differences between the sparse distance matrix elements d_i to the larger element are below a cutoff value c_d (default 15%),

$$\frac{|d_i^Q - d_i^T|}{\max(d_i^Q, d_i^T)} < c_d \;\forall\; i$$

and that the root mean squared error (RMSE) thereof is below a cutoff value c_{RMSE} (default 0.6 Å)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} |d_i^Q - d_i^T|^2} < c_{\mathrm{RMSE}}$$

(c) the RMSD of the coordinates of the attachment vector atoms in an optimal least-squares matching,[28] obtained by explicitly trying out all possible permutations of query and target attachment vectors, must be below a cutoff value c_{RMSD} (default 0.3 Å).

If T passes these pruning steps, it is considered for inclusion in a *shortlist* of hit scaffolds. The shortlist is allowed to contain 10 times the maximum number of hits requested by the user. The excess hits are discarded later in a pruning step (see below). Only one record per unique scaffold identifier is included in the shortlist in order to maintain diversity of hit scaffolds. A record is included if it is better than the current worst record in the list based either on the RMSD value or on shape descriptor similarity with the query. The shape descriptor is calculated on-the-fly. For that, the molecular structure needs to be fetched from the structure file according to the offset stored in the index record, which makes the shape descriptor based screening more time-consuming than that based on the comparison of RMSD values. The calculation of the shape descriptor and similarity values between them is described in the following section.

Scaffold Shape Descriptor (SSD). The Ultrafast Shape Recognition (USR) descriptor[29] and its derivatives[30−35] describe a molecular shape by 3 or 4 central moments of the distribution (mean, variance, skewness, and kurtosis) of interatomic distances measured from 4 reference atoms (or points in space), which results in a vector of 12 floating point numbers (16 if the fourth moment is included[30]). The inverse Manhattan distance is frequently used as the similarity measure of USR descriptors[29,31,33] and is defined as

$$S_{AB} = \left(1 + \frac{1}{N}\sum_{i}^{N} |A_i - B_i|\right)^{-1}$$

where N is the number of elements in descriptor vectors A and B.

dx.doi.org/10.1021/ci4001019 | J. Chem. Inf. Model. 2013, 53, 1825−1835

The calculation of USR descriptor requires only the computation of $4N_{atoms}$ distances and the four moments, and similarity comparison of short descriptor vectors is obviously very efficient. The USR approach is therefore well suited for the screening of a large number of conformers in the scaffold database (see Table 1).

While the exact details vary depending on the variant of the USR descriptor, they all use the discontinuous minimum and maximum functions to select the reference atoms; therefore, a small conformational change can lead to an abrupt change in the descriptor values, and the similarity thereof, when a different atom is chosen as a reference point. This discontinuity probably has minor practical impact given that USR has demonstrated good performance in prospective virtual screening campaigns.[36,37] However, in the context of scaffolds, the use of discontinuous functions can be avoided because the attachment vector provides an unambiguous frame of reference. In this study, two reference points and a reference *direction* are used instead of four points in order to define a scaffold shape descriptor (SSD) as follows.

One descriptor, an array of 16 elements, is generated for each attachment vector in the scaffold. The first four elements of the descriptor are the four central moments (scaled to linearity by taking the cube root of the third moment and fourth root of the fourth moment as described in ref 38) of the distribution of distances from the base atom to non-hydrogen atoms in the scaffold. Elements 5−8 of the descriptor are the moments of distribution of the angles (in radians) formed by the direction of the attachment vector (from the tip to the base atom), and the directions from the base atom to the non-hydrogen atoms in the scaffold. A second reference point, which is not necessarily coincident with any atomic coordinates, is placed at a distance of 8 Å from the base atom in the direction of the attachment vector (that is, away from the tip atom). The scaled central moments of the distribution of the distances and angles measured from that point are the remaining elements of the descriptor.

The distance of 8 Å was chosen based on the average dimensions of the fragments so that the distances and angles would have a reasonable distribution; if the second reference point was placed very close to the first reference point, the distributions measured from the two reference points would be very similar and therefore the second reference point would provide little additional information in the descriptor. If the second reference point was placed far from the first point, the measured distances would be large, angles would be small, and the information content again reduced.

SSD cannot distinguish between stereoisomers because no vector cross product, or other operation that is equivariant under translation and rotation but not under reflection,[34] is used in the derivation of the reference points; however the similarity between SSDs is a smooth function of the geometry of the scaffolds since the frame of reference is fixed by the attachment vector and no abrupt changes of reference points occur.

In previous USR approaches, scaling factors were introduced to the similarity calculation where quantities of different units were used in the derivation of the elements of the descriptor vector.[32,33] Here, the interatomic distances are in angstroms, and the angles are in radians; therefore, scaling factors could be used, too. However, no scaling scheme was implemented in this work, because the numerical values for both quantities are of the same magnitude in this context. The scaffold structures are rather small, and the distances are expected to range from 1.5 to 20 Å for the majority of structures, whereas the angle values have the domain $[0, \pi]$. The effect of distance on the similarity measure will be somewhat emphasized.

SSDs were generated for scaffolds with one attachment vector and stored into a binary file together with a link to the corresponding 3D structure record. That file is then screened and a shortlist of hit scaffolds collected based on the inverse Manhattan distance similarity. For scaffolds with two to four attachment vectors, SSDs are generated on-the-fly if needed when collecting the shortlist of hit scaffolds.

For benchmarking, a program for the calculation and comparison of USR descriptors with four scaled central moments (an array of 16 elements) was implemented, too.

Alignment and Scoring. The shortlist of hit structures from the 3D shape index or SSD descriptor search phase are input to the program ROCS (rapid overlay of chemical structures)[39] that superimposes structures by maximization of the overlap of volume and chemical features with respect to rigid-body rotation and translation starting from four initial alignments. A custom weighting scheme for the chemical features is employed that forces the attachment vector atoms to be aligned. In this scheme, a higher weight value of 15.0 is assigned to the superimposition of attachment vector atoms than to other chemical features that use a weight of 1.0. The higher weight value was chosen to be approximately equal to the average number of atoms in the scaffold database fragments.

The shortlist of hits may contain up to 10 times the number of scaffolds requested by the user. ROCS TanimotoCombo score is used to prune the final list of hits, which are then assigned shape and electrostatic similarity scores using EON.[40]

Virtual Hit Compounds. The virtual hit compounds are generated by replacing the query scaffold by the hit scaffold while maintaining the original scaffold substitution in the query molecule unchanged. Each new virtual compound geometry is optimized using the MMFF94s force field[41] with the Sheffield solvation model[42] as implemented in the Szybki toolkit.[43] A flat-bottom harmonic constraint,

$$V = k_c(\max(r - k_d, 0))^2$$

where the bottom width k_d is 0.15 Å and force constant k_c equals 1.0, is placed on non-hydrogen atoms in order to allow relaxation of bond lengths, angles, and torsions, but not to diverge too much from the shape of the query parent compound unless the virtual compound is very strained. The virtual hits are then realigned on the query compound using ROCS and scored using EON analogous to hit scaffolds. Any strained and subsequently diverged structures can be expected to obtain a low similarity score.

Estimation of Synthetic Feasibility. The synthetic feasibility of virtually generated compounds is always questionable. Although rarely encountered fragments were discarded from the search space (see above) and any hit scaffolds should be chemically feasible as such, new bonds are created when the substituents of a query molecule and a hit scaffold are merged into a virtual hit compound and these new chemical environments may be unfeasible. A synthetic accessibility scoring method was implemented according to the work of Ertl and Schuffenhauer.[44] The method is based on substructure frequency counts derived from a large database of existing compounds and complexity penalty terms. The method is general and fast and can be used to rank virtual hit compounds relative to each other.[45] The substructures used in

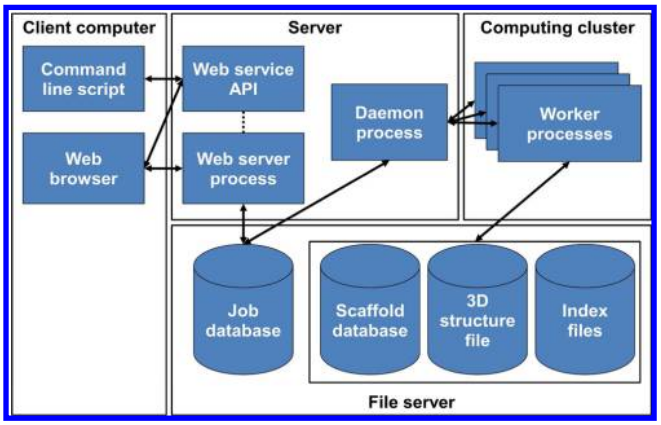
this study were Extended Connectivity[46] fingerprint-type radial substructures with a radius of two bonds, commonly known as ECFP_4 atom identifiers. The reference substructures were collected from the AstraZeneca corporate collection. Different compound collections, including eMolecules and GOSTAR, and combinations thereof, were experimented with but the resulting synthetic accessibility scores of a small set of test compounds were found to be almost unaffected as long as the reference compound set was sufficiently large, around 1 000 000 structures.

The synthetic accessibility score can be used to complement the similarity scores from ROCS and EON to rank the virtual hit compounds. Other, physicochemical structure descriptors are optionally calculated for the hit compounds in order to facilitate the selection of interesting candidates for synthesis efforts.

Implementation Notes. The web application and service interface were implemented in Python and the back-end fragmenting, indexing, and screening applications were written in C++, using the OEChem toolkit.[47]

The overall view of the application architecture is depicted in Scheme 2. A modular architecture was chosen for the scaffold hopping application implemented in this study for extendibility and variability, because the requirements for the similarity comparison between a query and a hit scaffold (or virtual hit compound) vary depending on the existing knowledge of the target protein and existing ligands. For example, an X-ray crystal structure or a homology model may be available for docking and scoring the virtual hits, or a set of known active ligands may have been used to derive a pharmacophore model. Here, the screening phase only considers the attachment vector geometry and optionally shape in order to accommodate many possible use cases and provide candidate hits for further scoring. The hits are ranked using ROCS and scored using EON as a sensible default option. Other means of ranking can be applied as a postfilter.

Scheme 2. Modular Architecture of the Scaffold Hopping Web Application[a]



[a]The web application has a modular architecture, and it is deployed on server machines specifically configured for each type of service: A web server machine handles the requests from client programs and stores incoming queries to a job database, from which a daemon process takes the jobs and distributes them on a computing cluster that has a high-speed connection to a file server hosting the scaffold database and the corresponding 3D structure and index record files. The server and daemon are written in Python, and the worker processes are Python wrappers for C++ for maximum efficiency.

The Scaffold Hopping web application interface allows the user to upload a 3D query structure and select the part to be replaced and provides options to generate 3D conformer ensembles and to perform automatic fragmentation of the input (see above), each fragment being used as the scaffold, if no scaffold is specified by the user. These features allow many types of queries, ranging from a specific replacement based on a bioactive conformation to a broad chemistry idea generation based on a SMILES representation of a known active compound. The web application shows the parent compounds of a hit scaffold and links the Chemistry Connect database for additional information on the parents.

Example Cases. To the authors' knowledge, there is no de facto standard way of benchmarking a scaffold hopping method. Previous studies have used molecular frameworks, scaffold networks and trees, and scaffold similarity or distance scores[48] to assess performance (see reviews[3,6]). Approaches to benchmark virtual screening methods typically use a set of known active and inactive molecules, which are ranked and the retrieval rate of the active compounds is used as a performance measure. In order to benchmark a scaffold hopping method that performs fragment replacement, one would need large series of matched pairs whose activity was measured in the same biochemical assay. Such data are scarce and/or proprietary. Therefore, matched pairs that preserve biological activity in at least one biochemical assay are used here to make a qualitative assessment of the retrieval performance of the implemented method as follows.

The scaffolds derived from GOSTAR were sorted in descending order according to the frequency of occurrence. The distribution of frequencies can be expected to approximately follow the power law distribution,[49] and the most frequent scaffolds are assumed to be, from a medicinal chemistry point of view, ubiquitous and uninteresting molecular fragments. Therefore, example scaffolds that were *not* among the 100 most frequent ones were manually chosen so that the selected set would cover different sizes, shapes, and electrostatic properties, for 1, 2, and 3 attachment vectors. The selected scaffolds and their SMILES representations are shown in Figure 4. A single 3D conformation was generated for each example scaffold using CORINA (options "-d wh") and used to search the scaffold database using SSD similarity as a basis for including scaffolds in the shortlist of hits. The final pruning of the hit list was made according to the TanimotoCombo score using ROCS. The search was allowed to return a maximum of 1000 hit scaffolds. The hits were sorted in decreasing order according to the ET_combo score assigned using EON.

For comparison, the scaffold database was searched using SSD and USR without final pruning through ROCS, using the example scaffolds with one attachment vector as queries for which a match between attachment vector geometries of the query and hit scaffolds can be trivially guaranteed. A maximum of 1000 hit scaffolds were collected.

Known bioisosteres of each example scaffold were extracted from the SwissBioisostere[50] database (release 1.1.13) that contains medicinal chemistry scaffold replacements from the literature annotated with biochemical assay data extracted from the ChEMBL[51] database. Stereoisomers were generated for each extracted bioisostere as necessary using CORINA (options "-d stergen,preserve"). The set of bioisosteres was then pruned to discard scaffolds that were not present in the search database

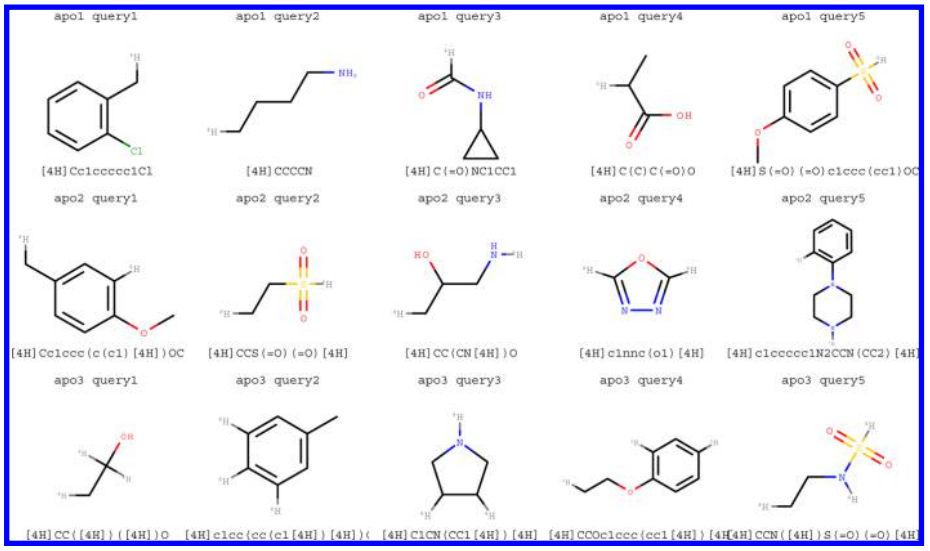


Figure 4. Example scaffolds used as queries in benchmark runs. The scaffolds were selected to have different sizes, shapes, and electrostatic properties. The SMILES representation is given below each depicted structure. Apo stands for the number attachment points (vectors) in the scaffold. The isotope-4 hydrogen atoms represent the tips of attachment vectors.

Table 2. Results of Scaffold Hopping Runs on Example Scaffolds

		enrichment factors[b]					bioisosteres			
attachment vectors	SMILES[a]	EF^{10}	EF^{20}	EF^{50}	EF^{100}	EF^{1000}	known[c]	retrieved[d]	hits[e]	run time
1	[4H]Cc1ccccc1Cl	0	0	312	214	76	776	39	1000	0:05:00
	[4H]CCCCN	273	273	164	137	71	553	26	1000	0:06:10
	[4H]C(=O)NC1CC1	2380	1388	1031	714	143	381	36	1000	0:05:41
	[4H]C(C)C(=O)O	1698	1274	934	679	166	356	39	1000	0:04:10
	[4H]S(=O)(=O)c1ccc(cc1)OC	0	0	0	0	13	118	1	1000	0:07:22
2	[4H]Cc1ccc(c(c1)[4H])OC	0	1287	515	687	240	304	28	1000	2:46:40
	[4H]CCS(=O)(=O)[4H]	1115	558	223	112	100	234	9	1000	2:59:31
	[4H]CC(CN[4H])O	3034	1517	1214	1214	273	86	9	1000	2:13:03
	[4H]c1nnc(o1)[4H]	2899	2174	870	1160	333	180	23	1000	8:10:28
	[4H]c1ccccc1N2CCN(CC2)[4H]	0	0	0	0	0	38	0	1000	2:07:18
3	[4H]CC([4H])([4H])O	7022	5852	2341	1873	304	79	13	1000	0:10:26
	[4H]c1 cm[3](cc(c1[4H])[4H])C	3879	3233	2069	1422	336	143	26	1000	0:35:37
	[4H]C1CN(CC1[4H])[4H]	0	0	0	0	0	33	0	1000	0:08:56
	[4H]CCOc1ccc(cc1[4H])[4H]	0	0	0	0	0	17	0	1000	0:12:26
	[4H]CCN([4H])S(=O)(=O)[4H]	0	0	0	0	0	26	0	78	0:09:16

[a]The structures are depicted in Figure 4. [b]EF^n is the enrichment factor at the top n hits. [c]The total number of known bioisosteres in the search database. [d]The number of known bioisosteres in the returned hit list in total. [e]The number of scaffolds in the returned hit list in total.

or did not preserve activity to within 0.5 log units at least in one assay.

The known bioisosteres were compared with the hit scaffolds from screening and the number of known bioisosteres H^n in the top $n \in \{10, 20, 50, 10, 1000\}$ hits was recorded. The values of n were selected to reflect the number of hits that could be manually inspected, and n = 1000 represents a situation where the hit list is further scored using, for example, docking or pharmacophore modeling, to rescue the known bioisosteres. The enrichment factor EF^n at a given number of top hits n can then be calculated from H^n, the total number of known bioisosteres in the search database A, and the total number of scaffolds with the given number of attachment vectors in the search database N (as given in Table 1) using the formula $EF^n = NH^n/An$.[52]

EF measures how well a method picks the known bioisosteres relative to random selection: EF equal to 1 would be random selection. EF equal to, say, 10 means the hit list contains 10 times more bioisosteres than with random selection. EF < 1 indicates a performance worse than random selection.[52]

For illustrative purposes, a 3D structure of rofecoxib, a cyclooxygenase-2 (COX-2) inhibitor shown in Figure 1, was generated using CORINA and used as query with the five-member ring scaffold selected to be replaced. The search used the SSD descriptor to rank the shortlist of hits and at most 100 hits were allowed. The returned set of virtual compounds contained 3-(4-methylsulfonylphenyl)-2-phenylcyclopent-2-en-1-one that is a known active COX-2 inhibitor.[53]

RESULTS

The enrichment factors, the number of retrieved known bioisosteres, and the total number of known bioisosteres in the search database for Scaffold Hopping are listed in Table 2.

dx.doi.org/10.1021/ci4001019 | J. Chem. Inf. Model. 2013, 53, 1825−1835

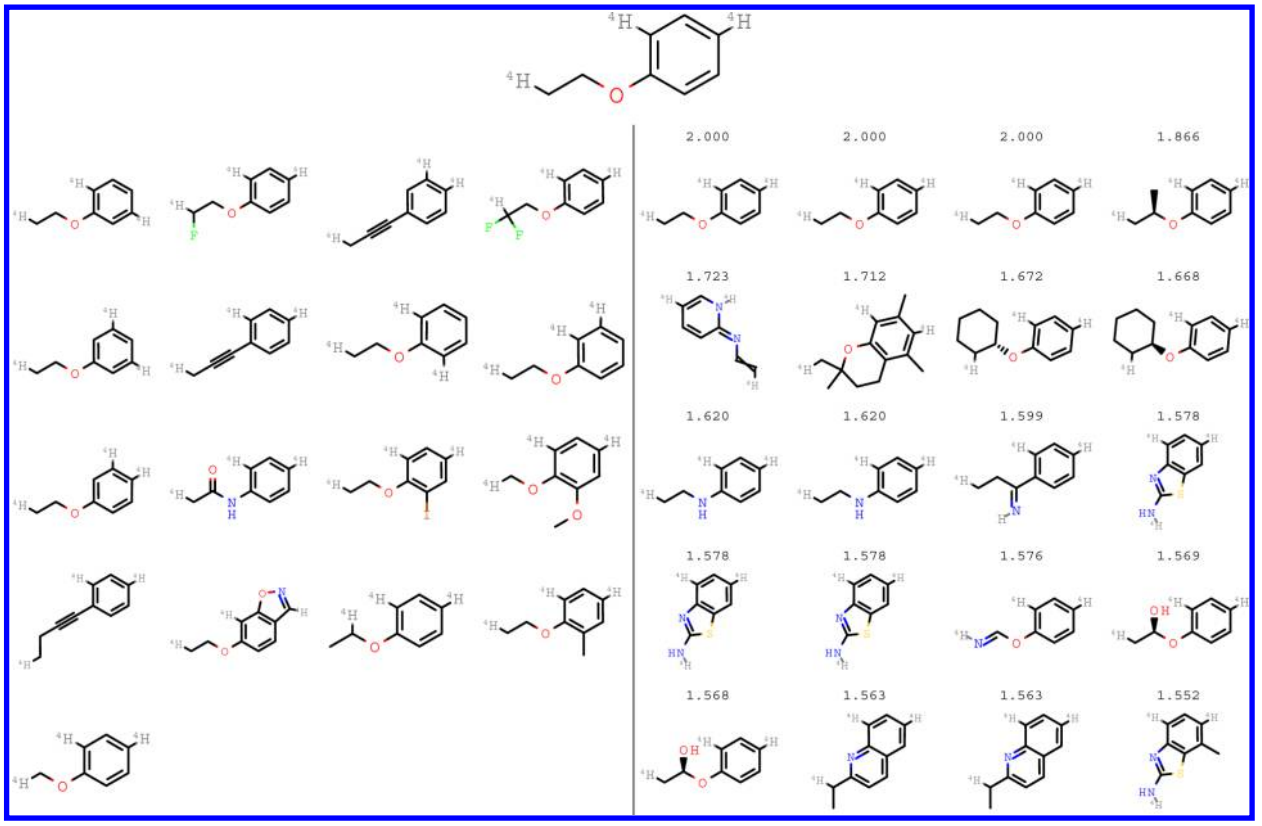


Figure 5. Search results for the example scaffold apo3 query4 in Figure 3, shown on top. The search failed to retrieve any of the known bioisosteres, shown on left. The highest-ranking hit scaffolds are shown on right with their EON ET_combo score (that obtains a maximum value of 2.0) for the overlap with the query scaffold. The duplicate hit scaffolds result from matching to scaffolds with varying number of attachment vectors. The unmatched attachment vector tip atoms are suppressed in the output. See text for details.

The known bioisosteres of the example ethoxybenzene scaffold ("apo3 query4" in Figure 3) and the highest scoring hit scaffolds are shown in Figure 5.

The number of retrieved known bioisosteres within the 1000 highest ranking hits for Scaffold Hopping, SSD, and USR are depicted in Figure 6. The total number of known bioisosteres in the search database remains the same as listed in Table 2. The runtimes for SSD and USR were approximately 2 min on average.

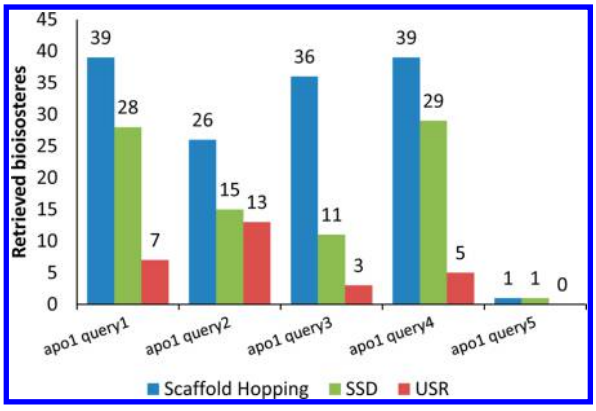


Figure 6. Number of retrieved known bioisosteres within 1000 highest ranking hits from screening the scaffold database using Scaffold Hopping, SSD, and USR. The query structures are depicted in Figure 3.

DISCUSSION

One should not overinterpret the results of the benchmark test in Table 2 for two reasons. First, EF is sensitive to the number of positive and negative samples in the test set and not very good at discriminating between different virtual screening methods unless several values at different percentage of screened database are used.[54,55] Second, the negative samples are not all confirmed nonbioisosteres, but their activity is either unknown or they were not measured in an assay where they would be active. Therefore, the use of the term enrichment factor should be taken with a grain of salt in this context and the results should not be taken as quantitative measures but rather as a qualitative indication of the level of performance against the background of currently available public bioactivity data.

The retrieval of bioisosteres for scaffolds with four attachment vectors was not benchmarked mainly because the scaffolds in SwissBioisostere contain one to three attachment vectors; therefore no benchmark data was available. Moreover, the probability of matching the relative geometry of attachment vectors of two scaffolds falls rapidly with the increasing number of vectors, which is easily seen from the number of vector pairs and bytes that arise in the hashing step of the indexing procedure. The drop in the probability is seen in the results of the example runs with three attachment vectors in Table 2, where known bioisosteres were retrieved only in two out of the five cases. For the last case, only 78 hits were returned of the allowed maximum of 1000, which indicates that the arrangement of the attachment vectors in the query is a rare one. For a query with four attachment vectors, the existence of relevant hit

scaffolds with a matching geometry is not at all guaranteed. Inclusion of more flexible (or single occurrence) scaffolds into the search database would improve the odds, with the cost of increased size of the database.

The EF values in Table 2, where not zero, are rather high (hundreds or thousands) compared to what is usually seen for virtual screening benchmarks, where EF is usually of order 1−100 (see for example refs 56 and 57). A probable explanation for the high EF values is that the known bioisosteres may be "easy" in many cases, that is, the bioisosteres would be picked up by any virtual screening method or a chemist would intuitively suggest the modification by just looking at the query structure. At the other end of the scale, the zero enrichment values are for cases where there are a relatively small number of known bioisosteres that have attachment vector geometry different from that in the query scaffold. For example, most of the missed bioisosteres in the left panel of Figure 5 have a substitution pattern different from that in the query, shown on top. Many of the highest-scoring hit scaffolds do preserve the topological distance pattern (number of bonds) between the attachment vector atoms, which is in line with the previously reported observation that the attachment vector geometry is quantized.[27] The duplicate hit scaffolds are a result of matching to scaffolds with four attachment vectors, of which the unmatched ones are suppressed (the isotope flags of the tip hydrogen atoms are removed). The duplicates are kept in the output because they have different parent compounds that can be used for backtracking and finding possible synthesis protocols. Indeed, Scaffold Hopping is often used at AstraZeneca in combination with the Virtual Library[58] application that performs fast 2D fingerprint similarity search in a large virtual library of compounds annotated with a synthesis protocol. Virtual Library can therefore suggest routes for chemistry follow up, especially if the hit scaffold is derived from AstraZeneca compound libraries.

The performance of SSD and USR alone on retrieving known bioisosteres for the example scaffolds with one attachment point is depicted in Figure 6. The number of retrieved known bioisosteres is higher for Scaffold Hopping than for the shape descriptor based methods alone in all cases. SSD tends to retrieve more bioisosteres than USR, which was expected given that SSD explicitly takes into account the attachment vector and may therefore be less prone to false positives where the overall shapes of the fragments are similar but the attachment vector locations do not coincide. The shape descriptor methods are faster than Scaffold Hopping, but the increased runtime (from 2 min to 5−6 min) can be justified by the increased retrieval performance of Scaffold Hopping. The differences in runtime and performance are attributable to the postfiltering of SSD hits through ROCS. However, running ROCS on the database was estimated to take approximately 50 h (assuming ROCS needs 1 ms per conformer), which prohibits using ROCS alone in the current use case scenario.

The runtime of a virtual screening program is of practical interest to the users. The runtime of the method implemented here scales approximately linearly with respect to the number of query scaffolds and the number of database scaffolds whose hash value coincides with any of the queries. The number of atoms or bonds in the query scaffold or the database scaffolds should not have a measurable effect on the runtime (or storage) requirements, because only the attachment vector geometry is used. In case the scaffold shape descriptor is used for ranking, the calculation of the descriptor formally scales linearly with the number of non-hydrogen atoms in the database scaffolds, but the practical effect of the varying size of the scaffolds can be expected to be very small because the typical scaffolds are from a few atoms to a few tens of atoms in size.

The SSD file for the scaffolds with one attachment vector is processed from start to end and the runtimes for queries should be constant. Runtimes of 5 min were observed on average for single attachment vector queries, which includes the pruning of the hit list using ROCS and scoring using EON. The relatively large variation of the runtimes for single attachment vector queries seen in Table 2 is due to the varying levels of traffic accessing the network disk used to store the files. Indeed, hardware considerations are important for a data-intensive application such as the one described in this study. The lowering cost of solid state hard drives will likely bring these low-latency high-throughput storage devices within reach for scientists with a moderate budget and improve the runtime of many applications. Scaffold Hopping balances between storage and runtime requirements by calculating the SSDs at screen time, which considerably reduces the size of the scaffold index files but increases the runtime mostly due to the time spent in seeking to the 3D structure file corresponding to an index record being considered. The use of a solid state hard drive can be expected to speed up the seeking considerably.

■ CONCLUSIONS

A scaffold hopping method, called Scaffold Hopping, was implemented and tested. The method uses an indexing scheme based on the relative geometry of attachment vectors that enables fast pruning of the search database and a novel shape descriptor that is specific to scaffold structures, for the further pruning of the list of candidate hit scaffolds. Scaffold Hopping was tested for the retrieval of known bioisosteres on a small but diverse set of example scaffolds. The method was found capable of retrieving verified bioisosteric replacement scaffolds from a large search database in a reasonable time frame for practical purposes. In comparison to alternative shape descriptor based approaches, Scaffold Hopping retrieved more known bioisosteres with the expense of increased runtime. The method was designed and found to be sensitive for the relative orientation of the attachment vectors; therefore the method is likely to be particularly useful in drug discovery programs where the bioactive conformation of a query template compound is known. In further development of the method, chemical and spatial information about the query structure and its host biomolecule, such as pharmacophoric features and exclusion volumes, may be included in the postprocessing steps.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mikko.vainio@abo.fi.

Present Addresses

†M.J.V.: Varian Medical Systems Finland Oy, Paciuksenkatu 21, 00270 Helsinki, Finland.

‡F.R.: BYK-Chemie GmbH, Abelstrasse 45, 46483 Wesel, Germany.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Roger Sayle for allowing us to reuse the ECFP code he wrote for the HazELNut suite of tools, Dr.

David Cosgrove for helpful discussions, Dr. Péter Várkonyi for help with database related issues, and the AstraZeneca computational chemistry community for feedback.

REFERENCES

(1) Sun, H.; Tawa, G.; Wallqvist, A. Classification of scaffold-hopping approaches. *Drug Discovery Today* **2012**, *17*, 310−324.
(2) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discovery Today* **2004**, *1*, 217−224.
(3) Schuffenhauer, A. Computational methods for scaffold hopping. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 842−867.
(4) Lima, L. M.; Barreiro, E. J. Bioisosterism: a useful strategy for molecular modification and drug design. *Curr. Med. Chem.* **2005**, *12*, 23−49.
(5) *Bioisosteres in Medicinal Chemistry*; Brown, N., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2012.
(6) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inf.* **2010**, *29*, 366−385.
(7) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Model* **2002**, *42*, 103−108.
(8) Giordanetto, F.; Boström, J.; Tyrchan, C. Follow-on drugs: How far should chemists look? *Drug Discovery Today* **2011**, *16*, 722−732.
(9) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. In *Chemoinformatics in drug discovery*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2005; pp 271−285.
(10) Lauri, G.; Bartlett, P. A. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51−66.
(11) *Brood*, version 2.0; OpenEye Scientific Software Inc: Santa Fe, NM, 2010.
(12) *sparkV10*; Cresset BioMolecular Discovery Limited, BioPark Hertfordshire: Welwyn Garden City, Herts, United Kingdom, 2012.
(13) *Suite 2012: Core Hopping*; Schrödinger, LLC, New York, NY, 2012.
(14) *Scaffold Replacement in MOE*; Chemical Computing Group Inc: Montreal, Quebec, 2012.
(15) Bergmann, R.; Linusson, A.; Zamora, I. SHOP: scaffold HOPping by GRID-based similarity searches. *J. Med. Chem.* **2007**, *50*, 2708−2717.
(16) Jakobi, A.-J.; Mauser, H.; Clark, T. ParaFrag – an approach for surface-based similarity comparison of molecular fragments. *J. Mol. Model* **2008**, *14*, 547−558.
(17) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model* **2007**, *47*, 390−399.
(18) *GOSTAR Online Structure-Activity Relationship Database*; GVK Biosciences Private Limited: Hyderabad, India, 2012.
(19) eMolecules. http://www.emolecules.com/ (accessed Dec 18, 2012).
(20) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today* **2011**, *16*, 1019−1030.
(21) Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed May 26, 2013).
(22) 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.
(23) Sayle, R.; Xie, P. H.; Muresan, S. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. *J. Chem. Inf. Model* **2012**, *52*, 51−62.
(24) 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.
(25) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(26) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model* **2010**, *50*, 572−584.
(27) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; McLay, I. M.; Bradshaw, J. Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* **2003**, *46*, 3257−3274.
(28) Theobald, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr., Sect. A* **2005**, *61*, 478−480.
(29) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.
(30) Cannon, E. O.; Nigsch, F.; Mitchell, J. B. O. A Novel Hybrid Ultrafast Shape Descriptor Method for use in Virtual Screening. *Chem. Cent. J.* [Online] **2008**, *2*, Article 3. http://journal.chemistrycentral.com/content/2/1/3 (accessed Dec 18, 2012).
(31) Schreyer, A. M.; Blundell, T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminf.* [Online] **2012**, *4*, Article 27. http://www.jcheminf.com/content/4/1/27 (accessed Dec 18, 2012).
(32) Armstrong, M. S.; Finn, P. W.; Morris, G. M.; Richards, W. G. Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 785−790.
(33) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 789−801.
(34) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Richards, W. G. Molecular similarity including chirality. *J. Mol. Graphics Modell.* **2009**, *28*, 368−370.
(35) Zhou, T.; Lafleur, K.; Caflisch, A. Complementing ultrafast shape recognition with an optical isomerism descriptor. *J. Mol. Graphics Modell.* **2010**, *29*, 443−449.
(36) Ballester, P. J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W. G. Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. *J. R. Soc. Interface* **2010**, *7*, 335−342.
(37) Ballester, P. J.; Mangold, M.; Howard, N. I.; Robinson, R. L. M.; Abell, C.; Blumberger, J.; Mitchell, J. B. O. Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *J. R. Soc. Interface* **2012**, *9*, 3196−3207.
(38) Ballester, P. J. Ultrafast shape recognition: method and applications. *Future Med. Chem.* **2011**, *3*, 65−78.
(39) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(40) *EON*, version 2.1.0; OpenEye Scientific Software Inc.: Santa Fe, NM, 2011.
(41) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(42) Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. A simple formula for dielectric polarisation energies: The Sheffield Solvation Model. *Chem. Phys. Lett.* **2007**, *441*, 163−166.
(43) *Szybki*, version 1.7.2; OpenEye Scientific Software Inc.: Santa Fe, NM, 2012.
(44) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and

fragment contributions. *J. Cheminf.* [Online] **2009**, *1*, Article 8. http://www.jcheminf.com/content/1/1/8 (accessed Dec 18, 2012).
(45) Ertl, P.; Lewis, R. IADE: a system for intelligent automatic design of bioisosteric analogs. *J. Comput.-Aided Mol. Des.* **2012**, 1−9.
(46) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742−754.
(47) *OEChem*, version 1.9.0; OpenEye Scientific Software Inc.: Santa Fe, NM, 2012.
(48) Li, R.; Stumpfe, D.; Vogt, M.; Geppert, H.; Bajorath, J. Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential. *J. Chem. Inf. Model* **2011**, *51*, 2507−2514.
(49) Karakoc, E.; Sahinalp, S. C.; Cherkasov, A. Comparative QSAR-and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model* **2006**, *46*, 2167−2182.
(50) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41*, D1137−1143.
(51) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.
(52) Pearlman, D. A.; Charifson, P. S. Improved Scoring of Ligand−Protein Interactions Using OWFEG Free Energy Grids. *J. Med. Chem.* **2001**, *44*, 502−511.
(53) Black, W. C.; Brideau, C.; Chan, C. C.; Charleson, S.; Chauret, N.; Claveau, D.; Ethier, D.; Gordon, R.; Greig, G.; Guay, J.; Hughes, G.; Jolicoeur, P.; Leblanc, Y.; Nicoll-Griffith, D.; Ouimet, N.; Riendeau, D.; Visco, D.; Wang, Z.; Xu, L.; Prasit, P. 2,3-Diarylcyclopentenones as orally active, highly selective cyclooxygenase-2 inhibitors. *J. Med. Chem.* **1999**, *42*, 1274−1281.
(54) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model* **2007**, *47*, 488−508.
(55) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.
(56) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178.
(57) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical comparison of virtual screening methods against the MUV data set. *J. Chem. Inf. Model* **2009**, *49*, 2168−2178.
(58) Vainio, M. J.; Kogej, T.; Raubacher, F. Automated Recycling of Chemistry for Virtual Screening and Library Design. *J. Chem. Inf. Model* **2012**, *52*, 1777−1786.