Extended-Connectivity Fingerprints

David Rogers*,[†] and Mathew Hahn[‡]

3429 North Mountain View Drive, San Diego, California 92116 and Accelrys, Incorporated, 10188 Telesis
Court, Suite 100, San Diego, California 92121

Extended-connectivity fingerprints (ECFPs) are a novel class of topological fingerprints for molecular characterization. Historically, topological fingerprints were developed for substructure and similarity searching. ECFPs were developed specifically for structure−activity modeling. ECFPs are circular fingerprints with a number of useful qualities: they can be very rapidly calculated; they are not predefined and can represent an essentially infinite number of different molecular features (including stereochemical information); their features represent the presence of particular substructures, allowing easier interpretation of analysis results; and the ECFP algorithm can be tailored to generate different types of circular fingerprints, optimized for different uses. While the use of ECFPs has been widely adopted and validated, a description of their implementation has not previously been presented in the literature.

INTRODUCTION

Molecular fingerprints[1] are representations of chemical structures originally designed to assist in chemical database substructure searching[2] but later used for analysis tasks, such as similarity searching,[3] clustering,[4] and classification.[5] *Extended-connectivity fingerprints* (ECFPs) are a recently developed fingerprint methodology explicitly designed to capture molecular features relevant to molecular activity. While not designed for substructure searching, they are well suited to tasks related to predicting and gaining insight into drug activity.[6] Additionally, ECFPs can be used much like other fingerprints in methods, such as similarity searching, clustering, and virtual screening.

Since their introduction in the first release of Pipeline Pilot[7,8] in the year 2000, ECFPs have been applied to a broad range of scientifically relevant problems, using a wide variety of analysis methods.[9] This paper describes how ECFPs are generated; the contrasts of using ECFPs to other fingerprint methodologies; and lists some of the many scientific application areas in which they have been used and published.

METHODS

Relation to Morgan Algorithm. ECFPs are derived using a variant of the *Morgan algorithm*,[10] which was proposed as a method for solving the molecular isomorphism problem (that is, identify when two molecules, with different atom numberings, are the same). In the Morgan algorithm, an iterative process assigns numeric identifiers to each atom, at first using a rule that encodes the numbering invariant atom information into an initial atom identifier, and later using the identifiers from the previous iteration. Thus, identifiers generated are independent of the original numbering of the atoms. The iteration process is continued until every atom identifier is unique (or as close to "unique" as symmetry allows); the intermediate results are discarded, and the final identifiers provide a canonical numbering scheme for the atoms.

The ECFP algorithm makes several changes to the standard Morgan algorithm. First, ECFP generation terminates after a predetermined number of iterations rather than after identifier uniqueness is achieved. The initial atom identifiers, and all identifiers after each iteration, are collected into a set; it is this set that defines the extended-connectivity fingerprint. Rather than discarding the intermediate atom identifiers, the ECFP algorithm retains them. Indeed, obtaining these partially disambiguated atom identifiers is the goal of the process. This means that the iteration process does not have to proceed to completion (that is, maximum disambiguation) but is performed for a predetermined number of iterations. Second, since perfectly accurate disambiguation is not required, algorithmic optimizations are possible. Consider, for example, that in the standard Morgan process, the identifiers must be carefully recoded after each iteration to avoid mathematical overflow and possible "collision" (where two different atom environments are accidentally given the same identifier). This recoding has the side-effect of creating identifiers that are not comparable between different molecules (i.e., two identical atom environments may be given different identifiers). In the ECFP algorithm, this computationally expensive step is replaced by a fast-hashing scheme. This results in a savings of computational effort when the ECFP algorithm is used for fingerprint generation, as compared to the rigorous Morgan algorithm used for canonicalization. Importantly, the ECFP-hashing scheme generates identifiers that are comparable across molecules.

ECFP Generation Process. The ECFP generation process has three sequential stages:

1. An *initial assignment stage* in which each atom has an integer identifier assigned to it.

* Corresponding author. Telephone: (619) 282-5480. E-mail: drogers@unterhund.com.
[†] 3429 North Mountain View Drive, San Diego, California.
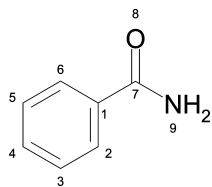[‡] Accelrys, Incorporated.

EXTENDED-CONNECTIVITY FINGERPRINTS

J. Chem. Inf. Model., Vol. 50, No. 5, 2010 **743**

Figure 1. Benzoic acid amide atom numbering (of non-hydrogen atoms).

2. An *iterative updating stage* in which each atom identifier is updated to reflect the identifiers of each atom's neighbors, including identification of whether it is a structural duplicate of other features.

3. A *duplicate identifier removal stage* in which multiple occurrences of the same feature are reduced to a single representative in the final feature list. (The occurrence count may be retained if one requires a set of counts rather than a standard binary fingerprint.)

The above process is further described as follows. First, atoms are assigned integer identifiers (for example, atoms might use their atomic number). These initial atom identifiers are collected into an initial fingerprint set. Next, each atom collects its own identifier and the identifiers of its immediately neighboring atoms, into an array (the neighbors are ordered using their identifiers, and the order of the attaching bonds, to avoid order-dependence). A hash function is applied to reduce this array back into a new, single-integer identifier. Once all atoms have generated their new identifiers, they replace their old identifiers with their new identifiers. The new atom identifiers are added into the fingerprint set. This iteration is repeated a prespecified number of times. When the specified number of iterations is completed, duplicate identifiers in the set are removed, and the remaining integer identifiers in the fingerprint set define the ECFP fingerprint.

The iteration process is illustrated using benzoic acid amide, with atom numbering as shown in Figure 1. The effect of iteratively updating the information around atom 1 in this compound is shown in Figure 2. At the beginning (iteration 0), the initial atom identifier only represents information about the atom itself and its attached bonds; this is shown as the substructure in the lower left corner of the figure (we represent allowed attachment points in the substructure using the "A" atom type). After one iteration, the identifier now contains information about atom 1's immediate neighbors, as shown in the lower center substructure. After two iterations, the represented substructure has grown further, now fully incorporating the amide group as well as much of the aromatic ring. It also captures the absence of substituents either *ortho-* or *meta-* to the ring amide.

This illustrates the power of the Morgan algorithm-based updating process: a strictly *local* operation (that is, each atom collecting identifiers only from its immediate neighbors) results in identifiers that may represent quite large substructures. Given that the process is executed over all atoms in the molecule (and not just a single atom, as shown in the figure), the final set of identifiers will contain substructural information from all parts of the molecule. Also, since the set is generated by collecting all identifiers up to some number of iterations, the final set contains a mixture of substructures of differing size for each atom in the molecule,
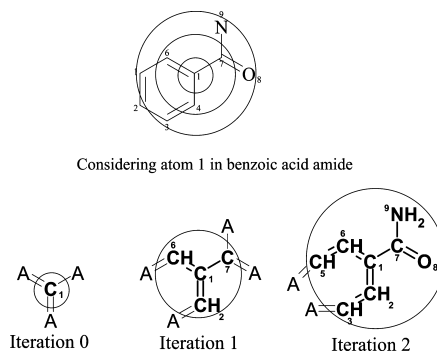

Figure 2. Illustration of the effect of iterative updating on the information represented by an atom identifier. Here, we consider atom 1 in benzoic acid amide. Each iteration has the effect of creating an identifier that represents larger and larger circular substructures around the central atom, as shown at the top of the figure. At iteration 0 (that is, the initial atom identifier), the atom only represents information about atom 1 and its attached bonds and can be represented by the substructure on the bottom left ("A" represents an atom of any type other than hydrogen). After one iteration, the identifier now contains information about atom 1's immediate neighbors, as shown in the bottom center substructure. After two iterations, the represented substructure has grown further, now fully incorporating the amide group as well as much of the aromatic ring, as shown in the bottom right.

some large and quite precise (such as the identifier representing *an aromatic carboxylic acid amide with no ortho- or meta-substitution* obtained starting at atom 1 in the figure after performing two iterations), and some small and relatively common (such as the *substituted aromatic carbon* atom represented by the initial atom identifier of atom 1).

The previous informal discussion avoids many of the details of the ECFP generation process. The following sections describe the process with enough precision to allow reproduction of all key aspects of the algorithm.

Initial Assignment of Atom Identifiers. The generation of ECFPs for a molecule begins with the assignment of initial atom identifiers. Hydrogen atoms and bonds to hydrogen atoms are ignored. In theory, any rule that generates integer values for atoms, and is independent of atom numbering, could be used. In this paper we describe in detail two rules leading to two different fingerprints: a standard ECFP and a variant termed FCFP. ECFPs are intended to capture precise atom environment substructural features, while FCFPs are intended to capture more abstract role-based substructural features. The *ECFP rule* is derived from the properties used in the Daylight atomic invariants rule.[11] The *FCFP rule* is derived from the functional class (i.e., pharmacophore role) of the atoms in a molecule.

Other initial atom identifiers based on different abstraction rules can be used to generate additional fingerprint variants, for example, Sybyl atom types[12](termed SCFPs), or aLog *P* atom codes[13](termed LCFPs). The number of possible variants of the ECFP generation process requires a naming convention to distinguish between related alternatives. We have chosen a convention (used in Pipeline Pilot)[6] that names a fingerprint using a four-character string (e.g., "ECFP"), followed by an underscore, followed by a number. The appended number is the effective diameter of the largest feature and is equal to twice the number of iterations performed; for example, if three iterations are performed, the largest possible fragment will have a width of 6 bonds,
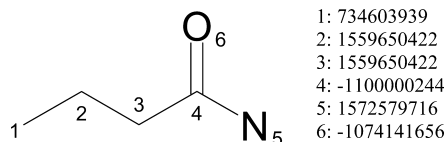
1: 734603939
2: 1559650422
3: 1559650422
4: -1100000244
5: 1572579716
6: -1074141656

Figure 3. The initial atom identifiers for butyramide, calculated using the Daylight atomic invariants-derived rule. (Note that the hash function may return either positive or negative numbers for the identifiers.)

and the fingerprint name will end in "_6" (e.g., "ECFP_6"). Regardless of the choice of initial atom identifier method, the remainder of the algorithm can be executed without change.

As previously stated, the initial atom identifier for the standard "ECFP" fingerprint uses atom information from the Daylight atomic invariants rule. The Daylight atomic invariants are six properties of an atom in a molecule that do not depend on initial atom numbering. These properties are: the number of immediate neighbors who are "heavy" (non-hydrogen) atoms; the valence minus the number of hydrogens; the atomic number; the atomic mass; the atomic charge; and the number of attached hydrogens (both implicit and explicit). We include one additional property: whether the atom is contained in at least one ring. To create an integer identifier from this information, these values are hashed into a single 32-bit integer value. This value is the initial atom identifier. For example, the initial atom identifiers for the atoms in butyramide are shown in Figure 3.

The fingerprint set for the molecule is initialized with the initial atom identifiers. For the above case, that would be the set of identifiers [734603939, 1559650422, 1559650422, −1100000244, 1572579716, and −1074141656]. This set is used as a starting point for saving additional identifiers collected after each iteration of the iterative updating process.

Each atom also keeps an associated set of bonds which define the substructure covered by the current identifier. As no iterations have yet been performed, the bond set for each atom is initialized to the empty set. As the iterations proceed, this bond set will be used to remove structural duplicates, as described in the next section.

Iterative Updating of Identifiers. The iterative updating process generates features that represent each atom within larger and larger circular substructural neighborhoods. Each iteration uses, as input, the atom identifiers from the previous iteration (or, if no iterations have yet been performed, the initial atom identifiers). Once each atom has calculated its new identifier, all atoms simultaneously update their identifier value, which completes the iteration. Any newly generated identifiers are added to the fingerprint set. Once a specified number of iterations is performed, the process proceeds to duplicate identifier removal.

A single iteration for a given atom is performed using the following sequence:

1. An array of integers is initialized to contain the iteration number and the identifier for the given (core) atom.
2. Attached atoms are sorted into a deterministic order using the bond order (single, double, triple, and aromatic) and the current identifier of each attached atom. A standard Hückel $4n + 2$ method is used for calculating aromaticity.
3. For each attachment, the attachment identifier and the bond order are appended to the array.

4. If the atom is a possible stereoatom but is not yet disambiguated, and all attachment atoms have different identifiers, then the atom is marked as disambiguated, and a stereochemical flag is appended to the array, depending on the marked stereochemistry. (Step 4 is only performed if stereochemical fingerprints are requested.)
5. The array is hashed into a single 32-bit integer. This is the new identifier for the atom.

To illustrate this process, consider the carboxylic acid carbon (atom 4, with the initial identifier of "−1100000244") in the molecule in Figure 3. In the first iteration, this atom will update as follows. First, an array will be created, with its first element initialized to "1" (the iteration level) and the second to "−1100000244" (the core atom's identifier). Next, we add two numbers to the array for each non-hydrogen attachment. The first of the two numbers will be the bond order for the bond to that attachment: 1, 2, 3, and 4 for single, double, triple, and aromatic bonds, respectively. The second of the two numbers is the current atom identifier of the attachment atom. To avoid order dependency in the attachment list, the attachments are sorted using their number pairs; in this case, the final order for the pairs is (1, 1559650422), (1, 1572579716), and (2, −1074141656). The final array for this atom contains eight elements and is: [1, −1100000244, 1, 1559650422, 1, 1572579716, 2, −1074141656]. Finally, the array of numbers is hashed to generate a single number, which is the new identifier (in this case, the number "−1708545601"). Repeating this process for each atom, each atom gets a new identifier, as shown in Figure 4.

Once every atom has a new identifier, features representing duplicate substructures are identified and are marked for removal, as described in the next section. Any remaining newly generated identifiers are then added to the current fingerprint array. After one iteration, the array of identifiers (the fingerprint set) for butyramide is: [734603939, 1559650422, 1559650422, −1100000244, 1572579716, −1074141656, 863188371, −1793471910, −1789102870, −1708545601, −932108170, 2099970318].

Conceptually, as the process is repeated, the feature denoted by an atom identifier represents an atom-centered substructure of increasing size. Figure 5 shows the process as applied to the oxygen atom in butyramide. Before the iteration process begins, the feature represented by the initial atom identifier is simply a double-bonded oxygen. After one iteration, the identifier represents a carbonyl group. After two iterations, the identifier represents an aliphatic carboxylic acid amide, with no substituents on the nitrogen atom and exactly one substituent on the α carbon. This shows how even a small number of iterations quickly creates identifiers that represent larger and larger substructures.

What is the appropriate number of iterations? The answer depends on the desired use of the fingerprint. Typically, two iterations is sufficient for fingerprints that will be used for similarity or clustering, while activity learning methods often benefit from the greater structure detail available after three or even four or more iterations. Since there is no objective termination condition, the number of iterations is under the control of the user (though as the number of requested iterations is increased, the number of newly discovered identifiers will diminish, and eventually no new identifiers will be discovered).
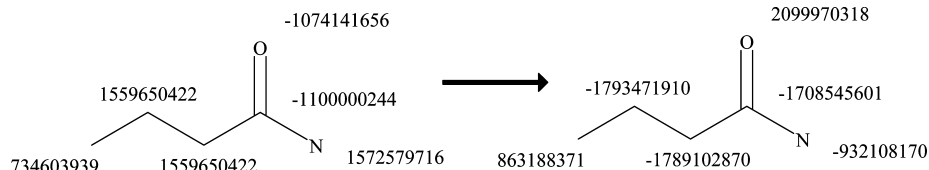
Figure 4. Generation of new identifiers by performing one iteration using butyramide. The initial atom identifiers are shown on the molecule on the left; after the updating process, each atom is given a new identifier, shown on the molecule on the right.
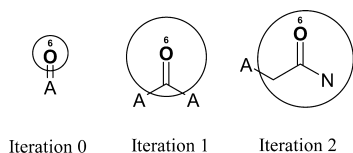

Figure 5. Effect of the iteration process on information stored in the identifier of the oxygen atom in butyramide. (The "A" atom type can map onto any atom type and is the only atom that may have connections that are not specified.) The sphere shows the *feature region* represented in the identifier after the given number of iterations. After zero iterations, only information about the atom itself and its connectivity are available. After one iteration, the identifier contains information from the core atom's immediate neighbors; in this case a carbonyl. After two iterations, atoms within two bonds of the core atom are included. At this point, it represents an aliphatic carboxylic acid amide, with no substituents on the nitrogen atom and exactly one substituent on the α carbon.
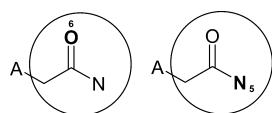

Figure 6. The feature regions after two iterations centered on the oxygen (left) and the nitrogen (right). Since both of these regions contain exactly the same atoms and bonds, they represent duplicate information, even though their hashed identifier values are different.

Duplicate Structure Removal. After several iterations, it is possible for two different atoms to contain information about identical structural regions of a molecule, as defined by the set of bonds covered by the atom-centered environment. This is called *structural duplication*. For example, consider the two substructures of the benzoic acid amide molecule from Figure 1 that were generated starting with the oxygen or nitrogen atoms, respectively, as shown in Figure 6. Whether the feature region is centered on the nitrogen or the oxygen, it represents the same structural unit of the molecule (after two or more iterations). But since the regions started at different atoms, the hashed identifier generated for the two originating atoms will be different, even though they represent the same underlying substructure. Such duplicates need to be identified and then removed to avoid adding useless redundancy to the fingerprint.

To identify such duplicates, each fingerprint feature keeps track of the set of bonds that it represents in a particular molecule. At each iteration step, the set of bonds is updated to include the union of all bonds in the core atom's bond set from the previous iteration, the attachment atom's bond sets from the previous iteration, and all attachment bonds. These bonds define the substructure within the molecule that is covered by the newly generated feature.

Before the newly generated features from an iteration are appended to the fingerprint set, they are checked to see if any structural duplicates exist to either previously generated
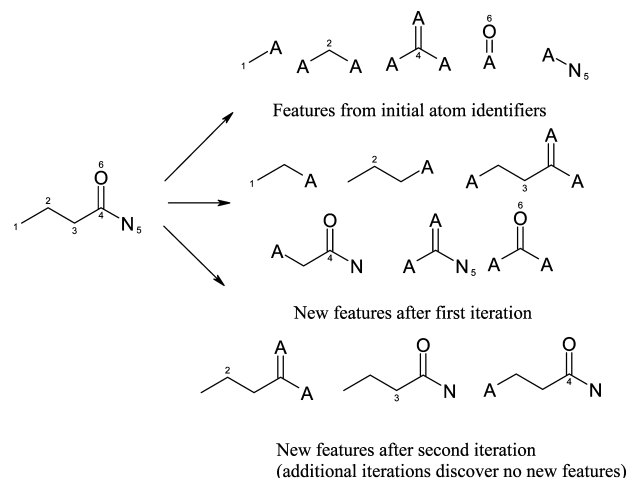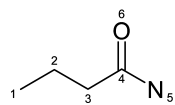

Figure 7. The structures represented by the features of butyramide after duplicate removal. The molecule generates five features (that is, unique atom identifiers) in the initial assignment stage, six additional features after one iteration, and three features after two iterations. Further iterations generate no new features. (The central atom for the figure is denoted by having its atom number shown.)

features or newly discovered features. When two features are discovered to be from equivalent bond sets, the following rules are used to remove one:

1. If the features were generated from a *different* number of iterations, the feature from the larger number of iterations is rejected.
2. If the features were generated from the *same* number of iterations, then the larger hashed identifier value (interpreted as an integer) is rejected.

The removal of duplicates has the additional effect that, at some number of iterations, fewer features will be generated than at the previous iteration level, and at some larger number of iterations, no more new features will be generated. Figure 7 shows the total features generated from butyramide. (The central generating atom for each feature is denoted by having its atom number shown.) Butyramide generates five different features (unique initial atom identifiers) in the initial assignment stage, even though there are six atoms. This is because there is identifier duplication with atoms 2 and 3, because at this level of abstraction, they represent the same feature. After the first iteration, there are six new features found. After the second iteration, however, structural duplication begins to occur, which results in the removal of three of the possible six features. No additional features are discovered in subsequent iterations. This can also be seen in Figure 8, where ECFP is calculated to diameters of 0, 2, 4, and 6. At each diameter, the fingerprint is the combination of all features from the previous diameter, plus any new features discovered at that step. Because butyramide is only 4 bonds wide, all possible structures are discovered by diameter 4, so ECFP_4 and ECFP_6 are identical.

```
> <ECFP_0>        > <ECFP_2>        > <ECFP_4>        > <ECFP_6>
734603939         734603939         734603939         734603939
1559650422        1559650422        1559650422        1559650422
-1100000244       -1100000244       -1100000244       -1100000244
1572579716        1572579716        1572579716        1572579716
-1074141656       -1074141656       -1074141656       -1074141656
                  863188371         863188371         863188371
                  -1793471910       -1793471910       -1793471910
                  -1789102870       -1789102870       -1789102870
                  -1708545601       -1708545601       -1708545601
                  -932108170        -932108170        -932108170
                  2099970318        2099970318        2099970318
                                    -87618679         -87618679
                                    1112638790        1112638790
                                    -627599602        -627599602
```

Figure 8. Fingerprints for butyramide with different diameters. Note that higher diameters contain all the fingerprint bits of lower diameters, possibly with new identifiers appended at the end. Also, note that ECFP_4 and ECFP_6 contain the same list. This is because the final iteration did not discover any new identifiers, where "new" is determined by the set of bonds that underlay a particular feature. By the time we have gone to a maximum diameter of four bonds, the entire molecule has been covered, and there is nothing new to discover.

Duplicate Identifier Removal. Sometimes, the final list of features contains duplicate identifiers. This results from equivalent substructures appearing in the molecule in more than one place. For example, a feature may represent a methyl group, and there may be more than one methyl group in the molecule. These different methyl groups will produce the same initial hashed identifier (and may even produce the same identifiers for some number of successive iterations, until a unique substructure is discovered). This kind of duplication can be eliminated by the optional removal of the redundant identifiers. Alternatively, duplicate identifiers can be kept and represented in the fingerprint. In our terminology, a *fingerprint* contains no duplicate identifiers, while a *fingerprint with counts* retains information about multiple occurrences by leaving the duplicates in the set.

The presence or absence of a particular ECFP 'bit' in an ECFP indicates presence or absence, respectively, of that feature. Because multiple features may 'collide', that is, be represented by the same bit code, the absence of a code is determinative (i.e., the feature is not present), but the presence of a code is only suggestive (i.e., the feature is likely present).

The iteration process, combined with duplicate identifier removal, provides a rich set of structural primitives to describe a molecule and can represent a very large number of different molecular features (over 4 billion). (The section Number of ECFP Features in Typical Libraries will discuss how many different features are generated from a typical compound library.)

Choice of Hash Function. We do not describe the particular hash function used in our calculation because any "reasonable" hash function can be used, and the scientific validity of the results is equivalent. What is most important is to have the hash function map arrays of integers randomly and uniformly into the 2^{32}-size space of all possible integers; without uniform coverage, the collision rate may increase, leading to a loss of information. One could also choose a hash function that mapped into a larger address space; for example, 64-bit integers would provide a 2^{64}-size space, further reducing the collision rate. From a practical perspective, we find that using 64-bit integers leads to no measurable improvement in analysis, and the manipulation of 64-bit integers is slower than 32-bit integers on current 32-bit computing hardware.

Functional-Class Fingerprints. The highly specific atom information contained in the initial atom identifiers for ECFPs allows the generation process to rapidly discover identifiers that represent a broad set of precisely defined structural features. However, for some purposes, this specificity may be undesirable, and some level of abstraction useful. For example, a chlorine or a bromine substituent on a ring may be functionally equivalent but would be distinguished by the ECFP process. We might prefer to have all halogens appear as equivalent atom types in the fingerprinting process. Similarly, we may want to represent all hydrogen-bond acceptors as equivalent types. This kind of abstraction is achieved with *functional-class fingerprints* (FCFPs).

FCFPs are generated using a more abstract, pharmacophoric set of initial atom identifiers similar to the catalyst pharmacophore identifiers.[14] Each atom is identified by a six-bit code, where a given bit is "on" if the atom plays the associated role. The atom roles are: hydrogen-bond acceptor and donor; negatively and positively ionizable; aromatic; and halogen. (Note that an atom may have more than one role or no role at all.) The six-bit code for each atom is the initial atom identifier. Once the initial identifiers are calculated, the process proceeds identically to the ECFP process.

Interpretation of Identifiers. One way to conceptualize the identifiers generated by the ECFP is as indices of bits in a large (2^{32}) bitset. For example, a molecule containing an atom with identifier "1559650422" could be considered to have bit 1 559 650 422 "on" in the bitset. (Negative identifiers are treated as "unsigned integers".) However, this is only an analogy, not a requirement; the identifiers themselves can be of any type, and the hash space any size. A more precise analogy would be of a hash table with the identifiers as keys. In either case, since the number of possible identifiers is much larger than the number present in any particular molecule (typically up to a few hundred features), the fingerprint is stored as a variable-length list of "on" bits, rather than as actual "on" bits in a large, fixed-length, nonvirtual, bitset. In any case, the size of the space of identifiers is an artifact of the hash function; a more traditionally sized fingerprint (e.g., 1024 bits) could be created by hashing into that smaller, fixed-length, space. There is some evidence that only a small amount of information is lost by this 'folding' operation,[18] but as the collision rate (two different substructures being represented
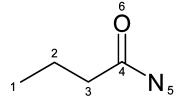
EXTENDED-CONNECTIVITY FINGERPRINTS

J. Chem. Inf. Model., Vol. 50, No. 5, 2010 **747**

> <ECFP_6>	> <ECFP_6#F>						> <ECFP_6#X>
734603939	1						[A]C
1559650422	2						[A]C[A]
-1100000244	4						[A]C(=[A])[A]
1572579716	5						[A]N
-1074141656	6						[A]=O
863188371	1	2					[A]CC
-1793471910	1	2	3				[A]CCC
-1789102870	2	3	4				[A]CCC(=[A])[A]
-1708545601	3	4	5	6			[A]CC(=O)N
-932108170	4	5					[A]C(=[A])N
2099970318	4	6					[A]C(=O)[A]
-87618679	1	2	3	4			[A]C(=[A])CCC
1112638790	1	2	3	4	5	6	CCCC(=O)N
-627599602	2	3	4	5	6		[A]CCC(=O)N

Figure 9. At the time of fingerprint calculation, additional information about the substructure represented by each identifier is available. This example uses butyramide, shown on the left. The bits of the ECFP_6 fingerprint are shown in the first column. The second column contains a set of atoms that were represented by the corresponding ECFP_6 bit. (Sets of atoms leading to duplicate features are not included in the list.) The final column shows a SMARTS representation of the extracted substructure, with the attachment points denoted using the "A" atoms. These are the legal attachment points. (Note: While Pipeline Pilot was used to generate these results and uses the "*" atom rather than the "A" atom for attachment points; we use "A" here for consistency with the text.) One can create a table of features and corresponding SMARTS for a particular training set, then use that information to see a substructural example of any feature of interest.

by the same identifier) is higher, both the quality and interpretability of the results will suffer.

While there is no way to directly decode the integer identifiers of ECFPs, it is still possible to ascertain the substructure a particular feature represents. While the identifier itself does not contain any structural information, at the time an ECFP is generated, each feature has access to the set of atoms and bonds that defines its substructure within the parent molecule. The identifier and substructure pairs can be stored so that one can look up the substructure corresponding to a particular feature of interest. The substructure information can be regenerated from the original data on demand, or saved up-front as a table mapping features to SMARTS[20] strings. This is illustrated for butyramide in Figure 9. The identifiers of the ECFP_6 fingerprint are shown in the first column. The second column contains the set of atoms that are represented by the corresponding ECFP_6 feature. The final column shows the SMARTS representation of the extracted substructure, with the attachment points denoted using the "A" atoms.

Note that the relationship between fingerprint features and the substructures may *not* always be one-to-one, that is, different substructural representations may share the same identifier (and, more rarely, different identifiers may represent the same underlying substructural representation). This is especially true when using FCFPs, since the initial atom identifiers are already abstracted and do not contain any connectivity or hydrogen information as do ECFPs. For example, Figure 10 shows a sampling of the underlying substructures that contained FCFP feature "1499521844" in a set of 7500 random drug molecules from the National Cancer Institute (NCI) AIDS data set.

There are additional effects that may lead to multiple structural representations of a particular bit that are seen with both ECFPs and FCFPs. One is that bond types to attachment "A" atoms may be of any type (as that information is not considered until the following iteration step). Also, rings that contain an attachment "A" atom may be opened at the attachment atom with no change in the identifier. This combination of effects leads to a multiplicity of substructures; that said, the similarity among the structures is striking, and
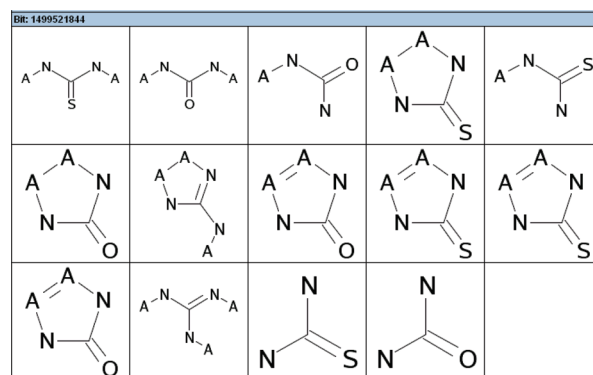

Figure 10. A set of underlying substructures that all generate the FCFP_6 feature "1499521844". (Legal attachment atoms are denoted using the "A" atom.) There are a number of reasons for the multiplicity of substructures. First, in FCFPs, the initial atom identifiers are abstracted and do not contain any connectivity information, hydrogen information, or explicit atom types, as do ECFPs. Second, information about the bond types to attachment atoms is not included in the feature. Third, rings that contain an attachment "A" atom may be opened at the attachment atom with no change of feature value. This combination of effects leads to a multiplicity of substructures; that said, the similarity among the structures is striking, and an abstraction that treats them the same is not without interest.

an abstraction process that treats them the same is chemically reasonable. Using ECFPs rather than FCFPs reduces, but does not eliminate, the one-to-many representational effect. The best solution, if one discovers a feature of interest, is to simply search for that feature in the calculated fingerprint, rather than generating a substructural query and using it to identify candidates. While the structural representation can assist the chemist in interpretation, it is not definitive.

RESULTS AND DISCUSSION

Comparison to Other Fingerprint Methods. There are numerous fingerprinting methods described in the literature. One of the nearest relatives to ECFPs is also one of the oldest. This is the concept of a fragment reduced to an environment that is limited (FREL) of the DARC substructure search system,[15,16] one of the earliest molecular

databases. The FREL describes two concentric layers of atoms around a central focus (typically, an atom or bond), creating a tree structure with the central atom at the root. This information is approximately the same as the information in the atom identifiers during the calculation of an ECFP after two iterations. These FRELs can be calculated for all molecules in a database and matched against abstracted query FRELs (termed "fuzzy-FRELs") to speed-up molecular database substructure search. ECFPs are different in that they do not explicitly retain the connectivity but hash the information into a single numeric identifier. The DARC system needed the structures themselves to perform matching operations, a requirement not applicable to ECFPs. Also, FRELs are defined to have a four-bond diameter, while the size of the circular substructures represented by ECFPs may include a mixture of many different diameters.

A later development in molecular database optimization led to the first explicitly binary fingerprint for substructure search. This was the set of 960 predefined substructure-based keys used by the MACCS system from Molecular Design Limited (MDL).[17] ECFPs have several advantages when contrasted to predefined substructural keys. First, since the set of keys is fixed, the system may not contain keys appropriate to the novel structural variation in a given compound library. Next, the keys are designed for substructure search, which limits their utility for activity modeling and categorization (see the Hydrogen-Filled Substructural Features Section for this discussion). Finally, even the most intelligently selected set of keys[18] is limited in coverage, making it unlikely that more detailed and specific substructures particular to some unusual activity class would be represented. The hundreds of thousands of different features generated by the ECFP process for a typical HTS data set of 50 000 molecules are more likely to expose larger and more detailed structural information critical to activity.

The concept of predefined substructural keys has been expanded by companies such as LeadScope,[19] which use a fingerprint based upon a large (>50 000) fragment dictionary, but in this case, use the fingerprint for the analysis of biological activity rather than database searching. While the much larger set of features is better suited to activity analysis, the predefined features still may not reflect structural variants in novel activity classes. Even the larger set of features is much smaller than the typical number of features that would be generated by computing ECFPs against all the molecules in a vendor library.

Another commonly used class of fingerprints is available through Daylight Chemical Information Systems.[20] It uses features based upon the presence of paths of varying lengths containing specific atom types. This generates a sparse binary vector, which is commonly "folded" to a bitset of specified size (e.g., 1024 bits) to reduce its size for ease of manipulation. ECFPs have several advantages when contrasted to path-based fingerprints. Like the MDL keys, Daylight keys were also designed for substructure and similarity search, which limits their effectiveness for activity modeling. If some path feature is found to be useful during an analysis, then interpretation can still be difficult, as a given path may not correspond to an easily identifiable function group or substructural entity. Further complicated but often relevant chemical attachment patterns, such as quaternary centers or cyclic patterns, cannot be directly represented by a path-

based description. Path-based schemes also cannot capture atom-based stereochemistry.

Another recently developed fingerprint type is termed *atom environment* fingerprints, first described by Xing and Glen.[21] In this fingerprint, called MOLPRINT_2D fingerprints, the initial atom identifiers are strings, and are lists of Sybyl atom types.[12] For any particular atom, all atoms within a given distance are given a "level", which is the shortest distance, in bonds, to the originating atom. For each level, a string is generated by concatenating all the initial atom identifiers of the atoms at that level. Finally, a string is generated by concatenating all levels into a single string. Bond types are not used explicitly but are implicit in the Sybyl atom types. Unlike ECFPs, there is no explicit use of connectivity beyond the assignment of levels to the atoms in the substructures. However, the string representation, while less efficient in memory use than the hashed integer identifier of ECFPs, has an advantage of being interpretable without further work. (If one desires the efficiency of numeric identifiers, one could hash the strings into an integer, yielding a representation computationally equivalent of ECFPs.) Similar to ECFPs, the final representation is a circular substructure around each central atom; studies have shown them to be an effective descriptor for activity prediction in concert with Naïve Bayesian methods.[22,23] This may be due to the quality of the Sybyl atom types in encoding useful atom neighborhood information and led to the development of SCFP fingerprints using Sybyl atom types for the initial atom identifiers. Studies comparing these to other variants of ECFP are ongoing; however, recent work by Li et al.[24] tasked with building "drug-like" models using support-vector machines concluded that, in this case, "ECFP_4 fingerprints gave a consistently superior performance compared to MOLPRINT_2D on all four performance measures (accuracy, sensitivity, specificity, and Matthews correlation coefficient)".

Another closely related fingerprint to atom environment fingerprints is the *signature molecular descriptor* developed by Faulon and co-workers for molecular enumeration.[25−28] In this work, a canonical spanning tree is constructed to cover all atoms in the subgraph of atoms contained with a given radii of a central atom, similar to the FRELs of the DARC system[15,16] but without the four-bond diameter limitation. This procedure has the benefit of preserving more of the connectivity information of the subgraph, but for our purposes, the speed and efficiency of the generation process outweigh the benefit of the canonicalization procedure, since our goal is to use features in analysis rather than enumeration.

Molecular holograms[29] are an extension of topological fingerprints that explicitly track the counts of multiply occurring features. The molecular hologram for a molecule is defined by generating all possible fragments of a molecule containing between n and m atoms. A canonicalization procedure is used to uniquely represent each fragment as a string, and the strings are assigned an integer value through a deterministic process (either hashing or using a lookup table). A separate count is kept of the number of occurrences of each value. Typically, fragments with 2−7 atoms are considered, leading to holograms containing 900−2400 counts. These holograms have been used in a process known as HQSAR,[30] in which a table is created, containing one column for each integer value, and whose cell contains the count. Partial-least-squares (PLS) regression is used to find

EXTENDED-CONNECTIVITY FINGERPRINTS

J. Chem. Inf. Model., Vol. 50, No. 5, 2010 **749**

a relationship between these columns and a user-supplied dependent variable.

The fingerprints used for molecular holograms have some key differences compared with ECFPs. In ECFPs, canonicalization is avoided, sparing significant computational cost. Also, the fragments underlying an ECFP feature are always circular around the central atom, while molecular hologram fragments may be of any shape. This may appear, at first, to be an advantage for the molecular hologram fingerprint, but because of the combinatoric explosion that occurs as the number of atoms in the fragment grows (typically, seven atoms is the limit), ECFP fingerprints can contain much larger features than is possible using molecular hologram fingerprints.

Other types of fingerprints can be derived from the 3D conformation (or conformations) of a molecule, rather than just the topology of a molecule. In 3D fingerprints, the presence of a feature in the fingerprint represents a particular 3D atom arrangement or pharmacophoric pattern. Akin to the initial implementation of topological fingerprints for substructure searching in molecular databases, early work on 3D fingerprints was inspired by the need to perform 3D database searching. An early example is the 3DSEARCH program developed by Sheridan and cohorts.[31] Later, 3D pharmacophoric fingerprints were developed for use in structure−activity studies in programs, such as Catalyst.[32,33]

Rather than representing the 3D arrangements of atoms or pharmacophores in small molecules, *affinity fingerprints*[34] are 3D fingerprints where each element is the binding affinity to a reference protein. The assumption is that two molecules with similar binding to a panel of uncorrelated proteins will likely be correlated in their binding to other proteins. A variant that depends on computational binding rather than experimental results is called *virtual affinity fingerprints*.[35,36] Virtual affinity methods consider the 3D structure of the ligand and also take the structure of the receptor into account.

ECFPs, being a topological method, do not directly represent 3D information. However, for many purposes, topological methods like ECFPs have advantages over 3D methods. High-throughput data analysis requires processing a large number of compounds; 3D fingerprints are expensive to generate because of the need to generate 3D conformations, restricting their use to smaller data sets. The generation of representative conformations is an area of ongoing research, and different conformational generation methods may result in vastly different 3D fingerprints. 3D fingerprints, such as affinity fingerprints, that reply on experimental data are also expensive to generate and are unavailable for virtual compounds. Often, the phenomena under study may not depend on the 3D conformation but only on general structural features, making the extra effort of generating 3D information unnecessary.

Since the 3D conformation of molecules depends on the topological structure, topological information contains much of the same useful information as the 3D information. Indeed, in most published analyses of topological vs 3D descriptors, the authors came to the conclusion that topological descriptors are as good or superior to 3D descriptors for molecular tasks like similarity searching[36,37] and activity prediction. There is, however, ongoing debate as to whether 3D fingerprints are better than topological fingerprints for "scaffold-hopping" between structural classes.[38−40]
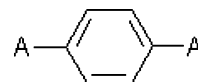

Figure 11. A *para*-substituted benzene query. If this feature were contained in a standard substructural fingerprint, then it could map onto *para*-substituted structures also containing substituents at other positions on the ring. If this feature is from an ECFP, then no substituents other than the *para* substituents are allowed. This is the difference between a hydrogen-filled substructural feature and a standard substructural query feature.
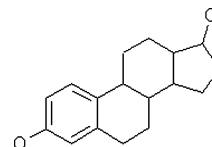

Figure 12. An estrogenic target molecule. If the feature in Figure 11 were represented as a standard substructural fingerprint, then it would be marked present. If the feature were from an ECFP, then it would not be present, as the aromatic ring has three substituents, not just the two *para* substituents.

Hydrogen-Filled Substructural Features. In ECFPs, each feature represents the presence of a *hydrogen-filled* substructure and not a standard *query* substructure. The difference is that a hydrogen-filled substructural representation contains hydrogens at all positions except where explicit attachment points are marked, and at those points, attachments are *required*. In a standard query substructure (without hydrogen counts explicitly specified on all atoms), heavy atom attachments may occur at any position in the fragment with unfilled valence. The usefulness of this distinction is our claim that the structural features of ECFPs are better at representing "negative" structure information (that is, the *absence* of substitution) when compared against standard substructurally based fingerprints (such as MDL substructure search keys,[17] Lead-Scope features,[19] or Daylight path-based fingerprints[20]) and that such information is important for molecular activity analysis.

The subtle distinction is best explained with an example. Assume there are features representing a *para*-substituted benzene ring in both a standard substructural fingerprint and in an ECFP. Such a query is shown in Figure 11.

If this *para*-substituted feature is from a standard substructural fingerprint, then it can map as a substructure somewhere in the target molecule. For example, the estrogenic structure shown in Figure 12 would have that feature.

If this feature is from an ECFP, the estrogen would *not* contain the feature, as there is an additional *ortho*-substitution on the ring in addition to the *para*-specified attachment atoms marked "A". Thus, an ECFP feature represents an *exact* substructure with limited, specified attachment points. However, the related molecule in Figure 13 would have the ECFP feature, as there is only a *para*-substitution present.

The rationale for this difference between ECFPs and standard substructure-based fingerprints is two-fold. First, the latter fingerprints were developed specifically for substructure searching. Substructural fingerprints have the property that all the features contained within a query must also be contained within a target, if the query can map
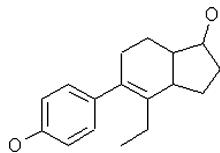
Figure 13. Another estrogenic molecule. In this case, the feature from Figure 11, if represented in an ECFP, would be present. Thus, that feature in an ECFP can distinguish subtle changes in connectivity that is more difficult to represent using substructural features.
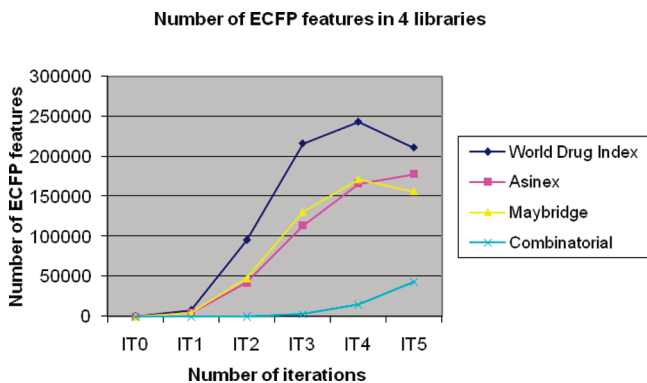

Figure 14. The number of ECFP features generated at a given iteration level for four different molecular libraries. The four libraries are the Derwent World Drug Index, the Asinex vendor catalog, the Maybridge vendor catalog, and a combinatorial library based upon an indole core. From each library, 50,000 compounds were randomly selected.

onto that target. This allows the fingerprint to be used to rapidly eliminate molecules from consideration when performing a substructure search against a database. For ECFPs, there was no requirement that they be useful for database search optimization.

Second, a primary goal in ECFP development was to represent the *absence* of functionality as well as the *presence* of functionality, as both were deemed equally important to characterizing activity. The *para*-substituted benzene ring in Figure 11 is a good example of this; by restricting any bonds to parts of the structure that are not represented by explicit attachment points "A", it is more restrictive than a substructural feature, and able to represent more subtle variations in a structure. (Of course, in most database search systems, one could add explicit hydrogens or hydrogen counts to substructural queries to force a similar effect, at the cost of losing most of their utility as substructural keys.)

Number of ECFP Features in Typical Libraries. ECFPs can represent a much larger set of features than is common for other fingerprints. The virtual size of the fingerprint is 2^{32} features, which is greater than four billion (4 × 10^9) different features. For a given molecule, only a small subset of those features will be present.

A typical molecule generates fingerprints containing tens or hundreds of features; a typical molecular catalog or library may, in aggregate, contain hundreds of thousands to millions of different features. Figure 14 shows the number of different ECFP features generated at increasing iteration levels from four different molecular libraries: the Derwent World Drug Index[41] (WDI), the Asinex catalog,[42] the Maybridge catalog,[43] and a combinatorial library defined as an indole core

Figure 15. Percentage of features found in both the 50 000 compound subset of Asinex and (from top to bottom) Maybridge, the WDI, and the combinatorial library. As the number of iterations increases, the size of the overlap rapidly shrinks. This shows that ECFPs provide a rich supply of features that can aid in discriminating molecules.

Figure 16. The number of FCFP features generated at a given iteration level for four different molecular libraries. Compared with the ECFP features shown in Figure 9, the shapes of the curves are quite similar, but FCFPs appear to give a smaller number of features, especially after few iterations. This is expected, as FCFPs are more abstract than ECFPs.
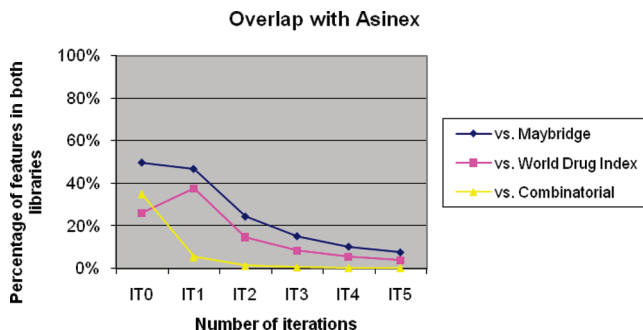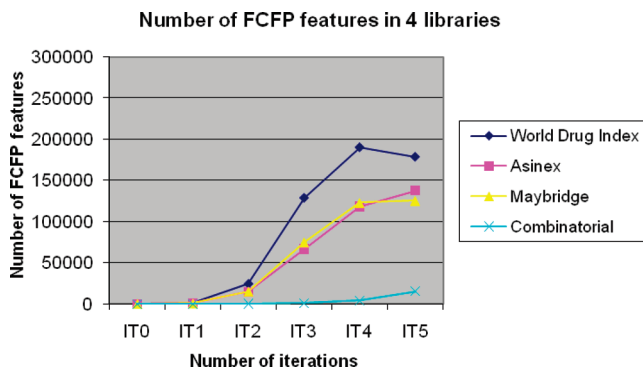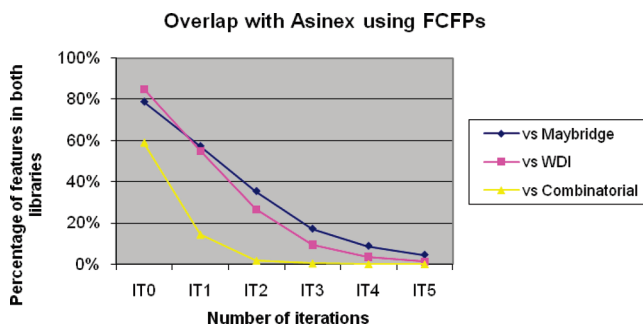

Figure 17. Percentage of FCFP features found in both the 50 000 compound subset of Asinex and (from top to bottom) Maybridge, the WDI, and the combinatorial library. Compared with the overlaps shown in Figure 15, FCFP overlaps are consistently larger, especially after several iterations.

and attachments in an RG-format file. To make library to library comparisons consistent, 50 000 molecules were randomly chosen from each library.

The graphs show that the drug library contained the most structural diversity, followed by the vendor catalogs, then the combinatorial library. The combinatorial library showed little structural diversity until many iterations were performed, and even then the diversity was still much less than the other libraries. This reflects a typical combinatorial library, which is constructed from one or a small

Extended-Connectivity Fingerprints

J. Chem. Inf. Model., Vol. 50, No. 5, 2010 **751**

number of scaffold cores. There is little structural variation until the neighborhoods get large enough to encompass multiple substituents surrounding the core scaffold.

Figure 14 also illustrates the large number of ECFP features in typical vendor libraries or drug compendia. For a particular fingerprint (say, ECFP_6) the total number of features is the sum of the features discovered at all iteration steps (for ECFP_6, IT0 + IT1 + IT2 + IT3 + IT4 + IT5). Thus, considering all features up to iteration 5, the 50 000 compound subset of the WDI generates over 750 000 different ECFP features.

The ECFP feature count for a library can be used as a rapid estimate of the structural diversity of a library. It can also be used to compare the structural diversity found in different libraries. For example, Figure 14 shows that WDI has about twice as many substructural features four bonds in diameter (that is, after two iterations) than either Maybridge or Asinex. Four bonds is slightly larger than the size of a typical functional group and, thus, may indicate that important structural classes are contained in WDI that are missing from Maybridge or Asinex. Differences at larger diameters may be less interesting as a measure of diversity, as the ECFP iteration process increasingly discovers features that are effectively unique for any particular molecule.

Overlap between Libraries. The ECFP feature count can also be used to understand the amount of feature overlap between two different libraries. This is illustrated in Figure 15, which shows the percentage of features found in both the 50 000 compound subset of Asinex and (from top to bottom) the subsets of Maybridge, the Derwent WDI, and the combinatorial library.

It is surprising how small the overlap region is, even after a small number of iterations. This reflects the ability of the ECFP process to rapidly capture unique structural variations that may be used to distinguish molecular libraries. Thus, ECFPs can provide downstream analysis methods with a rich supply of information to discover relationships that may be missed if only whole-molecule properties are used for library comparison.

Feature Counts and Overlap with Functional-Class Fingerprints. Figure 16 shows the number of different FCFP (rather than ECFP) features generated at increasing iteration levels from the same four molecular libraries. It is interesting to compare this to the results for ECFPs shown in Figure 14. The number of FCFP features is lower, especially after few iterations. This shows that the generalized atom types used in FCFPs do lead to feature abstraction. As the number of iterations increases, FCFPs ability to provide feature abstraction decreases because graph complexity begins to dominate in the creation of new features.

Similarly, the overlap regions between the libraries when considering FCFPs can be studied, as shown in Figure 17 (the corresponding figure for ECFPs is Figure 15). It shows much greater overlap in all cases, as would be expected from the abstraction process used to create FCFPs. The extra abstraction of FCFPs can be useful when the fingerprints are used in analysis methods, such as clustering or modeling abstract classes.

Collision between Numeric Identifiers. The *collision rate* in fingerprint generation is the number of times two different structural features generate the same hashed identifier. Collisions can add noise to analysis and can make interpretation of important features more difficult. In this light, we discuss the estimated collision rate during ECFP generation.

The collision rate depends on the size of the library being considered (more precisely, the number of different features in the library). For example, assume we are given a library containing 1×10^6 features. The probability of a particular identifier out of the 4×10^9 possible identifiers being in the set of 1×10^6 features is approximately $1 \times 10^6/4 \times 10^9 = 1/4000$. Using this value as an upper bound on the collision probability per feature added (the true collision rate depends on the number of features already in the set when we add a new feature), then an estimate of the percentage of features that collide is $\sim 1/4,000 = 0.025\%$, and the number of features involved is $\sim 1 \times 10^6 \times (1/4,000) = 250$. (Note that this minor collision rate could be reduced to nearly zero by increasing the size of the hash result from an integer with 2^{32} features into a 64 bit integer with 2^{64} features.)

APPLICATIONS

ECFPs have been available for nearly a decade and in that time have been applied to a broad range of scientifically relevant problems, using a wide variety of analysis methods. A sampling of publications illustrating the breadth of applications and methods that have used ECFPs are described below. These applications are grouped into three areas: virtual screening, structure−activity relationship modeling, and compound library analysis. (This grouping is somewhat arbitrary, as a particular application may fit in more than one area.)

Virtual Screening. The initial applications of ECFPs have been in the area of high-throughput screening (HTS). In particular, when combined with Bayesian-based analysis, ECFPs have proven to be very effective at categorizing actives from nonactives in the processing of noisy, high-volume, HTS data.[44]

Structure−Activity Relationship Modeling. This is the construction of models given a set of molecules with some annotation of bioactivity. A wide variety of modeling methods are possible. The high dimensionality of ECFPs is a particular advantage for Bayesian analysis[45] or Tanimoto (and related) similarty methods,[46] as they make good use of the wide variety and large number of ECFP features. In contrast, methods based upon "fitting" a model to the data (such as linear regression[47,48] or neural networks)[49] can overfit the data when confronted with large numbers of features so should be used with high-dimensional descriptors, such as ECFPs, with care. [The separation of a group of applications into the subheading absorbtion, distribution, metabolism, excretion, and toxicity (ADMET) is somewhat arbitrary; we tended to accept the self-classifications of the original authors.]

Compound Library Analysis. This broad category contains techniques whose common feature is a goal to analyze some aspect of a library of molecules other than simple structure−activity modeling. Examples include 3D docking,[50] diversity analysis,[51] and visualization methods.[52] The final row lists a number of papers that describe comparisons of ECFPs with other descriptors toward a variety of tasks.

Table 1

application area	method(s) used	references
Virtual Screening		
HTS prioritization	Naïve Bayes	Rogers et al.;[44] Klon[53]
HTS prioritization	Tanimoto similarity using nearest neighbor	Willet et al.[54]
HTS sequential screening	Naïve Bayes	Cloutier and Sirois[55]
HTS false positive detection	Naïve Bayes	Metz et al.[56]
HTS analysis of mixtures	Naïve Bayes	Glick et al.[57,58]
HTS analysis of noisy data	support vector machines, recursive partitioning, Naïve Bayes	Glick et al.[8,58]
ligand-based virtual screening	multiple machine learning methods	Willett et al.[59]
combinatorial library prediction	multicategory Naïve Bayes	Van Hoorn and Bell[97]
Structure Activity Relationship Modeling		
modeling drug activity	Naïve Bayes	Verkman et al.[52]
modeling drug activity	multicategory Naïve Bayes	Willett et al.[60]
'druglike' models	Naïve Bayes	Costache et al.;[61] Good and Hermsmeier[62]
multidrug resistance modeling	multicategory Naïve Bayes	Sun[63]
chelating agent models	Naïve Bayes	Morao et al.[64]
biological target prediction	multicategory Naïve Bayes	Jenkins et al.;[65] Nettles et al.[39]
biological target prediction	Winnow and Naïve Bayes	Nigsch et. at.[94]
biological target prediction	similarity ensemble approach	Hert et al.[95]
biological target prediction	support vector machines	Wale and Karypis[104]
QSAR	PLS regression	Gedeck et al.[66]
QSAR; antitubercular activity	Naïve Bayes	Prathipati[98]
QSPR; log *P* prediction	PLS regression	Liu and Zhou[101]
QSPR; melting point and aqueous solubility	PLS regression	Liu et al.[67]
ADMET modeling; BBB and SPB	Naïve Bayes	Klon et al.[68]
ADMET modeling; hERG and CYP2D6	Naïve Bayes and neural net consensus modeling	O'Brien and de Groot[69]
ADMET modeling; hERG	Naïve Bayes	Sun[70]
ADMET modeling; CYP2D6	PLS regression	Sciabola et al.[71]
ADMET modeling; CYP3A4	support vector machine	Zhou et al.[72]
ADMET modeling; CYP2D6 and CYP3A4	Gaussian kernel weighted *k*-nearest neighbor	Refsgaard et al.[73]
ADMET modeling; animal clearance	Naïve Bayes	McIntyre et al.[74]
ADMET modeling; phospholipidosis	Naïve Bayes	Pelletier et al.[75]
ADMET modeling; adverse drug reactions	multicategory Naïve Bayes	Scheiber et al.[76]
ADMET modeling; liver microsomal stability	random forest and Naïve Bayes	Lee et al.[77]
drug side effect prediction	similarity ensemble approach	Keiser et al.[93]
similarity searching	Tanimoto similarity	Kellenberger et al.;[78] Bender et al.;[79] Clark et al.[37]
similarity searching	clique-based searching	Lounkine et al.[105]
similarity searching	Tanimoto similarity with compressed fingerprints	Baldi et al.[106]
similarity searching	data fusion using multiple similarity coefficients	Whittle et al.;[92] Willett et al.[45]
similarity searching in target bioactivity space	multicategory Naïve Bayes with PCA	Bender et al.[80]
similarity searching; lead hopping	Tanimoto similarity	Martin and Muchmore,[81] Muchmore et al.[100]
similarity searching; lead hopping	multiple methods	Wale et al.[102]
Compound Library Analysis		
molecular complexity analysis	scaffold self-similarity	Selzer et al.;[82] Krier et al.[83]
diversity analysis	maximal dissimilarity clustering	Langer et al.;[51] Steindl et al.[84]
diversity analysis	clustering with Tanimoto similarity	Schuffenhauer et al.[85]
biological classification	clustering with Tanimoto similarity	Schuffenhauer et al.[86]
selection of diverse 3D docking decoys	Tanimoto similarity	Rognan et al.[87]
improving high-throughput 3D docking	Naïve Bayes	Davies et al.;[88] Klon et al.;[89] Filikov et al.;[50] Klon et al.[90]
improving high-throughput 3D docking	Tanimoto similarity	de Graaf and Rognan[91]
fast 2D models of 3D docking results	PLS regression	Sullivan and Martin[103]
visualization of favorable and unfavorable structural features	Naïve Bayes	Verkman et al.[52]
prediction of key example compound in patent	Tanimoto similarity	Hattori et al.[107]
comparison of ECFPs to other fingerprints	multiple methods	Hert et al.;[45] Hert et al.;[60] Hert et al.;[95] Schuffenhauer et al.;[85] Schuffenhauer et al.;[86] Li et al.;[24] Nisius et al.;[96] Papadatos et al.;[99] Bender et al.;[79] Rhodes et al.;[37] Good and Hermsmeier;[62] Hattori et al.;[107] Kellenberger et al.;[78] Nettles et al.;[39] Sciabola et al.;[71] Whittle et al.[92]

CONCLUSION

Extended-connectivity fingerprints (ECFPs) have a number of strengths that make them useful for a wide variety of applications in computational chemistry. They can be generated quickly using an easily understood method. Because they are not defined *a priori*, they can represent novel structural classes. The features are defined to contain both positive and negative structural information (that is, both what *is* and what *is not* present), crucial for analyzing molecular activity. They are a highly effective representation of topological structural information. Finally, their usefulness in the representation of molecular information is reflected in their widespread adoption and use across a broad range of applications and methodologies, as reported in a large number of published articles.

REFERENCES AND NOTES

(1) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, Germany, 2000.

(2) Christie, B. D.; Leland, B. A.; Nourse, J. G. Structure Searching in Chemical Databases by Direct Lookup Methods. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 545–547.

(3) *Concepts and Applications of Molecular Similarity*; Johnson, M.; Maggiora, G., Eds.; Wiley: New York, NY, 1990.

(4) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL 'keys' as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

(5) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth and Brooks/Cole: Monterey, CA, 1984.

(6) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463–4470.

(7) *Pipeline Pilot*, version 7.5; Accelrys, Inc.: San Diego, CA, 2000.

Extended-Connectivity Fingerprints

J. Chem. Inf. Model., Vol. 50, No. 5, 2010 **753**

(8) Hassan, M.; Brown, R. D.; Varma-O'Brian, S.; Rogers, D. Cheminformatics Analysis and Learning in a Data Pipelining Environment. *Mol. Diversity* **2006**, *10*, 283–299.

(9) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of High-Throughput Screening Data with Increasing Levels of Noise Using Support-Vector Machines, Recursive Partitioning, and Laplacian-Modified Naïve Bayesian Classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193–200.

(10) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–112.

(11) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(12) Clark, R. D.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(13) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragment Methods: An Analysis of AlogP and CLogP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.

(14) Greene, J.; Kahn, S.; Savoy, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297–1308.

(15) Dubios, J. E. French National Policy for Chemical Information and the DARC System as a Potential Tool for this Policy. *J. Chem. Doc.* **1973**, *13*, 8–13.

(16) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

(17) *MACCS-II Database System*, version 1; Molecular Design Limited: San Leandro, CA, 1984.

(18) Durant, J.; Leland, B.; Henry, D.; Nourse, J. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(19) *LeadScope*, version 1, LeadScope, Inc.: Columbus, OH, 1997.

(20) *Daylight Toolkit*, version 1, Daylight Chemical Information Systems: Mission Viejo, CA, 1987.

(21) Xing, L.; Glen, R. C. Novel Methods for the Prediction of log P, pKa, and log D. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(22) Bender, A.; Mussa, H.; Glen, R.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

(23) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs* **2006**, *9*, 199–204.

(24) 24. Li, Q.; Bender, A.; Pei, J.; Lai, L. A Large Descriptor Set and a Probabilistic Kernel-Based Classifier Significantly Improve Druglikeness Classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776–1786.

(25) Faulon, J. L. Stochastic Generator of Chemical Structure: 1. Application to the Structure Elucidation of Large Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1204–1218.

(26) Visco, D. P., Jr.; Pophale, R. S.; Rintoul, M. D.; Faulon, J. L. Developing a Methodology for an Inverse Quantitative Structure-Activity Relationship Using the Signature Molecular Descriptor. *J. Mol. Graphics Modell.* **2002**, *20*, 429–438.

(27) Faulon, J. L.; Visco, D. P., Jr.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.

(28) Faulon, J. L.; Churchwell, C. J.; Visco, D. P., Jr. The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721–734.

(29) Lowis, D. R. *HQSAR. A New, Highly Predictive QSAR Technique*; Tripos Technical Notes, Tripos: St. Louis, MO, 1998; Vol. 1, no. 5, p 3.

(30) Hurst, J. R.; Heritage, T. W. *Molecular Hologram QSAR*. United States Patent 5751605, May 12, 1998.

(31) Sheridan, R. P.; Nilakantan, R.; Rusinko, A., III; Bauman, N.; Haraki, K.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Structure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.

(32) Sprague, P. W. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. *Perspect. Drug Discovery Des.* **1995**, *3*, 1–20.

(33) Norinder, U. The Alignment problem in 3D-QSAR: A combined approach using Catalyst and a 3D-QSAR technique. In *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds; Prous Science Publishers: Barcelona, 1995; pp 433–438.

(34) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; Paull, K. D. Neural Computing in Cancer Drug Development: Predicting Mechanism of Action. *Science* **1992**, *258*, 447–451.

(35) Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK Generated Fingerprints. *J. Med. Chem.* **1996**, *39*, 3401–3408.

(36) Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231–244.

(37) Rhodes, N.; Clark, D. E.; Willett, P. Similarity Searching in Databases of Flexible 3D Structures Using Autocorrelation Vectors Derived from Smoothed Bounded Distance Matrices. *J. Chem. Inf. Model.* **2006**, *46*, 615–619.

(38) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.

(39) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.

(40) Sheridan, R. Chemical similarity searches: when is complexity justified. *Expert Opin. Drug Discovery* **2007**, *2*, 423–430.

(41) World Drug Index; Derwent Information Americas: Alexandria, VA; http://www.derwent.com. Accessed December 14, 2006.

(42) Asinex catalog; Asinex, Inc.: Moscow, Russia; http://www.asinex.com. Accessed March 25, 2005.

(43) Maybridge catalog; Maybridge plc: Trevillett, England; http://www.maybridge.com. Accessed March 25, 2005.

(44) Rogers, D.; Brown, R.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up. *J. Biomol. Screening* **2005**, *10*, 682–686.

(45) Hert, J.; Willett, P.; Wilton, D. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470, See appendix for an excellent overview of different Bayesian methods in drug discovery.

(46) Carhart, R. E.; Smith, D. H.; Venkararaghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(47) Freedman, D.; Pisani, R.; Purves, R.; Adhikari., A. *Statistics*, W.W. Norton: New York, NY, 1991.

(48) Herman, W. In *Partial Least Squares*; Kotz, S., Johnson, N. L. , Eds.; Wiley: New York, NY, 1985; pp 581–591.

(49) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*; Rumelhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, MA, 1986.

(50) Yoon, S.; Smellie, A.; Hartsough, D.; Filikov, A. Surrogate Docking: structure-based virtual screening at high-throughput speed. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 483–497.

(51) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel Screening: A Novel Concept in Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2146–2157.

(52) Yang, H.; Shelat, A. A.; Guy, R. K.; Gopinath, V. S.; Ma, T.; Du, K.; Lukacs, G. L.; Taddei, A.; Folli, C.; Pedemonte, N.; Gallietta, L. J. V.; Verkman, A. S. Nanomolar Affinity Small Molecule Correctors of Defective ΔF508-CFTR Chloride Channel Gating. *J. Biol. Chem.* **2003**, *278*, 35079–35085.

(53) Klon, A. E. Bayesian Models in Virtual High-Throughput Screening. *Comb. Chem. High Throughput Screening* **2009**, *12*, 469–483.

(54) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jackoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049–7054.

(55) Cloutier, L. M.; Sirois, S. Bayesian versus Frequentist statistical modeling: A debate for hit selection from HTS campaigns. *Drug Discovery Today* **2008**, *13*, 536–542.

(56) Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 139–144.

(57) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Prioritization of high throughput screening data of compound mixtures using molecular similarity. *Mol. Phys.* **2003**, *101*, 1325–1328.

(58) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. *J. Biomol. Screening* **2004**, *9*, 32–36.

(59) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 51–62.

(60) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K. Comparison of topological descriptors for similarity-based virtual screening using

multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

(61) Costache, A. D.; Trawick, D.; Bohl, D.; Sem, D. S. AmineDB: Large scale docking of amines with CYP2D6 and scoring for druglike properties - towards defining the scope of the chemical defense against foreign amines in humans. *Xenobiotica* **2007**, *37*, 221–245.

(62) Good, A. C.; Hermsmeier, M. A. Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models. *J. Chem. Inf. Model.* **2007**, *47*, 110–114.

(63) Sun, H. A Naïve Bayes Classifier for Prediction of Multidrug Resistance Reversal Activity on the Basis of Atom Typing. *J. Med. Chem.* **2005**, *48*, 4031–4039.

(64) Flahive, E.; Ewanicki, B.; Yu, S.; Higgenson, P. D.; Sach, N. W.; Morao, I. A High-Throughput Methodology for Screening Solution-Based Chelating Agents for Efficient Palladium Removal. *QSAR Comb. Sci.* **2007**, *26*, 679–685.

(65) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.

(66) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924–1936.

(67) Zhou, D.; Alelynas, Y.; Liu, R. Scores of Extended Connectivity Fingerprint as Descriptors in QSPR Study of Melting Point and Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *48*, 981–987.

(68) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved naïve Bayesian Modeling of Numerical Data for Absorbtion, Distribution, Metabolism, and Excretion (ADME) Property Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1945–1956.

(69) O'Brien, S. E.; de Groot, M. J. Greater Than the Sum of Its Parts: Combining Models for Useful ADMET Prediction. *J. Med. Chem.* **2005**, *48*, 1287–1291.

(70) Sun, H. An Accurate and Interpretable Bayesian Clasification Model for Prediction of hERG Liability. *ChemMedChem* **2006**, *1*, 315–322.

(71) Sciabola, S.; Morao, I.; de Groot, M. J. Pharmacophoric Fingerprint Method (TOPP) for 3D-QSAR Modeling: Application to CYP2D6 Metabolic Stability. *J. Chem. Inf. Model.* **2007**, *47*, 76–84.

(72) Zhou, D.; Ruifeng, L.; Otmani, S. A.; Grimm, S. W.; Zauhar, R. J.; Zamora, I. Rapid Classification of CYP3A4 Inhibition Potential Using Support Vector Machine Approach. *Lett. Drug Des. Discovery* **2007**, *4*, 192–200.

(73) Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-Nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511.

(74) McIntyre, T. A.; Han, C.; Davis, D. B. Prediction of Animal Clearance using naïve Bayesian classification and extended connectivity fingerprints. *Xenobiotica* **2009**, *39*, 1–8.

(75) Pelletier, D. J.; Gehlhaar, D.; Tilloy-Ellul, A.; Johnson, T. O.; Greene, N. Evaluation of a Published in Silico Model and Construction of a Novel Bayesian Model for Predicting Phospholipidosis Inducing Potential. *J. Chem. Inf. Model.* **2007**, *47*, 1196–1205.

(76) Scheiber, J.; Jenkins, J. L.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Milik, M.; Azzaoui, K.; Whitebread, S.; Hamon, J.; Urban, L.; Glick, M.; Davies, J. W. Mapping Adverse Drug Reactions in Chemical Space. *J. Med. Chem.* **2009**, *52*, 3103–3107.

(77) Lee, P. H.; Cucurull-Sanchez, L.; Lu, J.; Du, Y. J. Development of in silico models for human liver microsomal stability. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 665–673.

(78) Kellenberger, E.; Springael, J.-Y.; Parmentier, M.; Hachet-Haas, M.; Galzi, J.-L.; Rognan, D. Identification of Nonpeptide CCR5 Receptor Agonists by Structure-based Virtual Screening. *J. Med. Chem.* **2007**, *50*, 1294–1303.

(79) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(80) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When are Multitarget Drugs a Feasible Concept. *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456.

(81) Martin, Y.; Muchmore, S. Beyond QSAR: Lead Hopping to Different Structures. *QSAR Comb. Sci.* **2009**, *28*, 797–801.

(82) Selzer, P.; Roth, H.-J.; Ertl, P.; Schuffenhauer, A. Complex molecules: do they add value. *Curr. Opin. Chem. Biol.* **2005**, *9*, 310–316.

(83) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

(84) Steindl, T. M.; Schuster, D.; Laggner, C.; Chuang, K.; Hoffmann, R. D.; Langer, T. Parallel Screening and Activity Profiling with HIV Protease Inhibitor Pharmacophore Models. *J. Chem. Inf. Model.* **2007**, *47*, 563–571.

(85) Schuffenhauer, A.; Brown, N.; Selzer, P.; Ertl, P.; Jacoby, E. Relationship Between Molecular Complexity, Biological Activity, and Structural Diversity. *J. Chem. Inf. Model.* **2006**, *46*, 525–535.

(86) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space. *J. Chem. Inf. Model.* **2007**, *47*, 325–336.

(87) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(88) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 2743–2749.

(89) Klon, A. E.; Glick, M.; Davies, J. W. Application of Machine Learning to Improve the results of High-Throughput Docking Against the HIV-1 Protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216–2224.

(90) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naïve Bayes Classifier with Consensus Scoring Improves Enrichment of High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 4356–4359.

(91) de Graaf, C.; Rognan, D. Selective Structure-Based Virtual Screening for Full and Partial Agonists of the $\beta 2$ Adrenergic Receptor. *J. Med. Chem.* **2008**, *51*, 4978–4985.

(92) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion. *J. Chem. Inf. Model.* **2006**, *46*, 2206–2219.

(93) Keiser, M.; Setola, V.; Irwin, J.; Laggner, C.; Abbas, A.; Hufeisen, S.; Jenson, N.; Kuijer, M.; Matos, R.; Tran, T.; Whaley, R.; Glennon, R.; Hert, J.; Thomas, K.; Edwards, D.; Shoichet, B.; Roth, B. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–182.

(94) Nigsch, F.; Bender, A.; Jenkins, J.; Mitchell, J. Ligand-Target Prediction Using Winnow and Naïve Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313–2325.

(95) Hert, J.; Keiser, M.; Irwin, J.; Oprea, T.; Shoichet, B. Quantifying the Relationships among Drug Classes. *J. Chem. Inf. Model.* **2008**, *48*, 755–765.

(96) Nisius, B.; Vogt, M.; Bajorath, J. Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback-Leibler Divergence Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347–1358.

(97) van Hoorn, W. P.; Bell, A. S. Searching Chemical Space with the Bayesian Idea Generator. *J. Chem. Inf. Model.* **2009**, *49*, 2211–2220.

(98) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian Models for the Prioritization of Antitubercular Agents. *J. Chem. Inf. Model.* **2008**, *48*, 2362–2370.

(99) Papadatos, G.; Cooper, A. W. J.; Kadirkamanathan, V.; Macdonald, S. J. F.; McLay, I. M.; Pickett, S. D.; Pritchard, J. M.; Willett, P.; Gillet, V. J. Analysis of Neighborhood Behavior in Lead Optimization and Array Design. *J. Chem. Inf. Model.* **2009**, *49*, 195–208.

(100) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. B.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.

(101) Liu, R.; Zhou, D. Using Molecular Fingerprint as Descriptors in the QSPR Study of Lipophilicity. *J. Chem. Inf. Model.* **2008**, *48*, 542–549.

(102) Wale, N.; Watson, I. A.; Karypis, G. Indirect Similarity Based Methods for Effective Scaffold-Hopping in Chemical Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 730–741.

(103) Sullivan, D. C.; Martin, E. J. Exploiting Structure-Activity Relationships in Docking. *J. Chem. Inf. Model.* **2008**, *48*, 817–830.

(104) Wale, N.; Karypis, G. Target Fishing for Chemical Compounds Using Target-Ligand Activity Data and Ranking-Based Methods. *J. Chem. Inf. Model.* **2009**, *49*, 2190–2201.

(105) Lounkine, E.; Hu, Y.; Batista, J.; Bajorath, J. Relevance of Feature Combinations for Similarity Searching Using General or Activity Class-Directed Molecular Fingerprints. *J. Chem. Inf. Model.* **2009**, *49*, 561–570.

(106) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098–2109.

(107) Hattori, K.; Wakabayashi, H.; Tamaki, K. Predicting Key Example Compounds in Competitors Patent Applications Using Structural Information Alone. *J. Chem. Inf. Model.* **2008**, *48*, 135–142.

CI100050T