

(51) International Patent Classification:
G06F 19/00 (2006.01)

(21) International Application Number:

PCT/US2010/000125

(22) International Filing Date:

20 January 2010 (20.01.2010)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/146,108 21 January 2009 (21.01.2009) US

(71) Applicant (for all designated States except US): **PRESIDENT AND FELLOWS OF HARVARD COLLEGE** [US/US]; 17 Quincy Street, Cambridge, MA 02138 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **KUTCHUKIAN, Peter** [US/US]; 89 Beacon Street #2, Somerville, MA 02143 (US). **SHAKHNOVICH, Eugene** [US/US]; 81 Shade Street, Lexington, MA 02421 (US).(74) Agent: **OYER, Timothy, J.**; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210-2206 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM,

[Continued on next page]

(54) Title: SYSTEMS AND METHODS FOR GENERATING AND/OR CHARACTERIZING MOLECULES FOR PHARMACEUTICAL AND OTHER USES

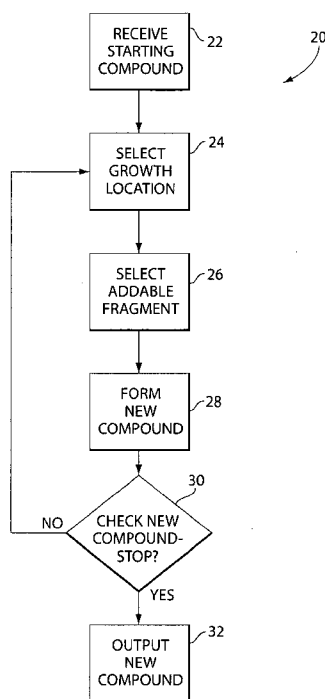


Fig. 1

(57) Abstract: Systems and methods for generating and/or characterizing molecules. One aspect of the invention is generally directed to the generation of new compounds, using a starting compound and "growing" the compound by adding fragments to the starting compound. In one set of embodiments, a set of existing compounds is identified by fragments forming the compounds, and the connectivities between the fragments. A distribution of such connections then is used to build a matrix of probability elements, which can then be used in some cases to generate new compounds. In certain instances, such generated compounds may have some similar characteristics to the set of existing compounds (e.g., drug-like characteristics, ease of synthesis, or the like). Another aspect of the invention generally relates to the characterization of compounds. For example, a set of compounds may be scored for various descriptors (e.g., number of rings, connectivity of fragments forming the compound, or the like), and the scores combined in some manner. For instance, in one set of embodiments, the scores may be combined in a weighted sum, which may then be compared against a standard (e.g., against a threshold score) to characterize the compounds.



TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, **Published:**
ML, MR, NE, SN, TD, TG).

— *without international search report and to be republished
upon receipt of that report (Rule 48.2(g))*

- 1 -

**SYSTEMS AND METHODS FOR GENERATING AND/OR CHARACTERIZING
MOLECULES FOR PHARMACEUTICAL AND OTHER USES**

GOVERNMENT FUNDING

5 Research leading to various aspects of the present invention were sponsored, at least in part, by the National Institutes of Health, grant no. GM52126. The U.S. Government has certain rights in the invention.

RELATED APPLICATIONS

10 This application claims the benefit of U.S. Provisional Patent Application Serial No. 61/146,108, filed 01/21/09, entitled "Systems and Methods for Generating and/or Characterizing Molecules for Pharmaceutical and Other Uses," by Kutchukian, *et al.*, incorporated herein by reference.

FIELD OF INVENTION

15 The present invention generally relates to systems and methods for generating and/or characterizing molecules.

BACKGROUND

20 Prior *de novo* techniques for generating libraries of molecules have often used methods of "fitting" portions of newly-designed molecules to specific binding pockets of targets. However, these techniques generally result in libraries of molecules having poor diversity, since their design is constrained by the pocket fit objective of the target. Moreover, many members of such libraries are often difficult to synthesize, since ease of synthesis is not considered in determining whether a newly generated molecule fits a binding pocket. Libraries of molecules with low diversity and/or difficult synthesis have generally not performed well in drug screening, and this has kept such combinatorial chemistry approaches from being more widely used.

SUMMARY OF THE INVENTION

30 Embodiments of the present invention generally relate to systems and methods for generating and/or characterizing molecules without using a target binding pocket as a primary constraint. The subject matter involves, in some cases, interrelated products, alternative solutions to a particular problem, and/or a plurality of different uses of one or more systems and/or articles.

 One aspect of the invention is generally directed to the generation of new compounds, using a starting compound and "growing" the compound by adding a

- 2 -

fragment(s) to the starting compound. The resulting compound may, in turn, be used as a new "starting compound" and a further fragment(s) to that starting compound by the judicious selection of fragments to be added, as explained below, a set of resulting compounds can be generated which will reflect certain structural properties of a starting set of compounds. In this way, a set of compounds can be generated wherein there is an enhanced or controlled likelihood that at least some of the generated compounds will exhibit a fundamental characteristic of some or all of the compounds in the starting set.

In one set of embodiments, one or more existing compounds are identified by fragments forming the compounds, and the connectivity of the fragments is determined. Such information is used to build a matrix of probability elements (e.g., each element corresponding to a probability of a particular fragment connectivity being present), which matrix is then used in some cases to generate new compounds. The new compounds may be generated by selectively adding fragments in accordance with the probability matrix. In certain instances, such generated compounds may have similar characteristics to the original set of compounds (e.g., drug-like characteristics, ease of synthesis, or the like).

Another aspect of the invention generally relates to the characterization of compounds. For example, a set of compounds may be scored for various descriptors (e.g., number of rings, connectivity of fragments forming the compound, or the like), and the scores combined in some manner. For instance, in one set of embodiments, the scores may be combined in a weighted sum, which may then be compared against a standard (e.g., against a threshold score) to characterize the compounds.

Another aspect of the invention is generally directed to a computer-implemented method. In one set of embodiments, the invention comprises an act of operating a computer to access data comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing potential fragments contained within the growth compounds and a second dimension representing potentially addable fragments; select a fragment contained within the growth compound; expand the growth compound to form a new growth compound by identifying an addable fragment to be added to the growth compound, the addable fragment being selected using the matrix comprising probability elements by identifying the probability elements corresponding to the growth compound, selecting, using the probability elements, an addable fragment from the potentially addable fragments, and adding the addable fragment to the growth

- 3 -

compound; optionally, repeat the step of expanding the growth compound one or more times, using each time the new growth compound as the growth compound; and output the new growth compound.

5 The method, according to another set of embodiments, includes an act of operating a computer to: access a database of compounds; identify fragments forming each compound; for each of the fragments, determining the number of other fragments connected to it within the database; and creating a matrix comprising probability elements for the distribution of fragment connections in the database.

10 In still another set of embodiments, the method includes an act of operating a computer to access a database of compounds, including a first category of compounds and a second category of compounds; access a plurality of descriptors, wherein at least one of the descriptors comprises determining connectivity of two or more fragments, each comprising two or more atoms, within each compound; determine a plurality of scores for each compound, each score for each compound corresponding to one of the
15 descriptors; and select constant weighing factors in a weighted sum of each of the scores of each compound such that at least about 55% of the compounds of the first category exceed a threshold score and/or no more than about 55% of the compounds of the second category do not exceed the threshold score.

20 The method, in yet another set of embodiments, includes an act of operating a computer to: access a database of compounds and one or more categories to which each of the compounds of the database belongs; access a plurality of descriptors, wherein at least one of the descriptors comprises determining connectivity of two or more fragments, each comprising two or more atoms, within each compound; determine a plurality of scores for each compound, each score for each compound corresponding to
25 one of the descriptors; and select constant weighing factors in a weighted sum of each of the scores of each compound such that each of the categories can be identified using a distinguishable range of scores such that at least about 55% of the compounds belonging in each category have a score that falls within the distinguishable range of scores for that respective category.

30 In still another set of embodiments, the method includes an act of operating a computer to: access a representation of a test compound; determine a plurality of scores for the test compound, each score corresponding to a descriptor, wherein at least one of the descriptors comprises determining connectivity of two or more fragments, each

- 4 -

comprising two or more atoms, within each compound; and determine an overall score for the test compound by calculating a weighted sum of each of the scores of the test compound using a predetermined set of weighing factors.

The invention, in another aspect, is generally directed to a computer-readable
5 medium having recorded thereon a data structure for use in generating chemical compounds, the structure comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing potential growth compounds and a second dimension representing potentially addable fragments, the probability elements having values representative of connectivities between the potential growth compounds
10 and the potentially addable fragments.

In one aspect, the invention is directed to a method of inputting, into a computer, a data structure comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing potential growth compounds and a second dimension representing potentially addable fragments. In another aspect, the invention is
15 directed to a method of causing a computer to construct a data structure comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing potential growth compounds and a second dimension representing potentially addable fragment.

Other advantages and novel features of the present invention will become
20 apparent from the following detailed description of various non-limiting embodiments of the invention when considered in conjunction with the accompanying figures. In cases where the present specification and a document incorporated by reference include conflicting and/or inconsistent disclosure, the present specification shall control. If two or more documents incorporated by reference include conflicting and/or inconsistent
25 disclosure with respect to each other, then the document having the later effective date shall control.

BRIEF DESCRIPTION OF THE DRAWINGS

Non-limiting embodiments of the present invention will be described by way of example with reference to the accompanying figures, which are schematic and are not
30 intended to be drawn to scale. In the figures, each identical or nearly identical component illustrated is typically represented by a single numeral. For purposes of clarity, not every component is labeled in every figure, nor is every component of each

- 5 -

embodiment of the invention shown where illustration is not necessary to allow those of ordinary skill in the art to understand the invention. In the figures:

Fig. 1 is a flow chart illustrating a method for generating a new compound, according to one embodiment of the invention;

5 Figs. 2A-2B together show an example of the use of the technique of Fig. 1 for generating a new compound;

Fig. 3 is a flow chart illustrating the generation of a new compound, in accordance with yet another embodiment of the invention;

10 Fig. 4 is an example illustrating the generation of new compounds, in still another embodiment of the invention;

Figs. 5A-5B illustrate the generation of a matrix of probability elements, in one embodiment of the invention;

Figs. 6A-6B illustrate the generation of a matrix of fragments or connectivities, in various embodiments;

15 Figs. 7A-7B are non-limiting examples of matrices of probability elements, according to certain embodiments of the invention;

Figs. 8A-8B are flow charts illustrating categorization systems, in yet other embodiments of the invention;

Fig. 9 illustrates an example growth matrix in one embodiment of the invention;

20 Fig. 10 illustrates an example growth matrix in another embodiment of the invention;

Fig. 11 illustrates examples of double counting corrections, in one embodiment of the invention;

25 Fig. 12 illustrates data showing a correlation of fragment connectivities, in accordance with another embodiment of the invention;

Fig. 13 illustrates examples of user-defined 3-mers, in still another embodiment of the invention;

Fig. 14 illustrates data showing a correlation of fragment connectivities in one embodiment of the invention;

30 Fig. 15 illustrates a matrix of transition probabilities, in accordance with another embodiment of the invention;

Fig. 16 illustrates data showing the probability of growing molecules using a training database, in yet another embodiment of the invention;

- 6 -

Fig. 17 illustrates data showing branching probabilities, in accordance with still another embodiment of the invention;

Fig. 18 illustrates example data showing the relative probabilities of various fragments being categorized as drug or non-drug, in accordance with another
5 embodiment of the invention; and

Figs. 19A-19C illustrate example data showing descriptor scores for categorizing compounds as drug or non-drug, in yet another embodiment of the invention.

DETAILED DESCRIPTION

The present invention generally relates to systems and methods for generating
10 and/or characterizing molecules. One aspect of the invention is generally directed to the generation of new compounds, using a starting compound and “growing” the compound by adding a fragment(s) to the starting compound. The resulting compound may, in turn, be used as a new “starting compound” and a further fragment(s) to that starting compound by the judicious selection of fragments to be added, as explained below, a set
15 of resulting compounds can be generated which will reflect certain structural properties of a starting set of compounds. In this way, a set of compounds can be generated wherein there is an enhanced or controlled likelihood that at least some of the generated compounds will exhibit a fundamental characteristic of some or all of the compounds in the starting set.

20 In one set of embodiments, one or more existing compounds are identified by fragments forming the compounds, and the connectivity of the fragments is determined. Such information is used to build a matrix of probability elements (e.g., each element corresponding to a probability of a particular fragment connectivity being present), which matrix is then used in some cases to generate new compounds. The new
25 compounds may be generated by selectively adding fragments in accordance with the probability matrix. In certain instances, such generated compounds may have similar characteristics to the original set of compounds (e.g., drug-like characteristics, ease of synthesis, or the like).

Another aspect of the invention generally relates to the characterization of
30 compounds. For example, a set of compounds may be scored for various descriptors (e.g., number of rings, connectivity of fragments forming the compound, or the like), and the scores combined in some manner. For instance, in one set of embodiments, the

- 7 -

scores may be combined in a weighted sum, which may then be compared against a standard (e.g., against a threshold score) to characterize the compounds.

A first aspect of the present invention is generally directed to techniques for generating new compounds, optionally (but preferably) on a computer. "Generate," as
5 used herein, refers to the production of a new compound (or more accurately, the production of a representation of a new compound), for instance, virtually on a computer. The generated compound may be encoded using any suitable format, for example, as a structural diagram, in SMILES format, InChI format, SLN format, or the like, which are well-known symbolic representational languages of chemical compounds.
10 The generated compound can be further used or manipulated, for example, for synthesis, testing or assays, or the like, as discussed in detail below.

The compound may be a single molecule (e.g., formed using covalent bonds), an ionic species, a supramolecular assembly (e.g., comprising multiple molecules held together by hydrogen bonding, van der Waals interactions, charge interactions,
15 hydrophobic interactions, etc.), or the like. In some cases, the compound may be a partial chemical structure.

In one set of embodiments, a starting compound is expanded to form a new growth compound by identifying and adding an addable fragment to the starting compound. Then a further addable fragment may be identified and added to the new
20 growth compound, using the latter as a new starting compound, and such step may be repeated, if desired. Each addable fragment can be selected, for example, using a matrix of probability elements, where the probability elements are associated with one or more fragments that are present in the starting or prior new growth compound. That is, for example, each probability element may represent the probability that a specific
25 connection between two fragments is present in a starting set of compounds. Based on those fragments that are present in the starting or prior new growth compound, an addable fragment can be selected from the matrix using the corresponding probability elements. Thus, from a single starting compound, one may generate a plurality of growth compounds in which a probability distribution of fragment connections will be
30 similar to a probability distribution of fragments connections in the population from which the probability matrix was obtained.

Thus, using techniques such as those described herein, a set of new generated molecules will generally "resemble" an original starting set of compounds, although

- 8 -

specific members of the newly generated set may not necessarily be present in the original starting set of compounds. Since the members of the new set may be generated based on the probability distributions of the fragments and their connections from the original starting set of compounds, the new set will also exhibit similar probability
5 distributions of the fragments and their connections, thus creating a resemblance of the new set of compounds to the original starting set of compounds (where the amount of resemblance can be controlled, as discussed below). As examples, a starting set of compounds rich in hydroxide groups will result in a generated set of compounds also rich in hydroxide groups, a starting set of compounds rich having a relatively large
10 number of particular fragment connections will result in a generated set of compounds also having a relatively large number of particular fragment connections, a starting set of compounds chosen to be easy to synthesize will result in a generated set of compounds also relatively easy to synthesize, etc.

A flow chart schematically diagramming an example of such a growth process 20 according to the first aspect of the invention is shown in Fig. 1. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. In this figure, a starting compound is first received in step 22. The compound may be received, for example, from input from a user, from a computer file containing representations of chemical structures (in any
20 suitable format), randomly (e.g., selected from a list of suitable chemical structures), randomly generated, etc. A growth location is then selected on the compound in step 24. The growth location may be selected, for instance, randomly from among the hydrogen atoms within the compound, based on fragments forming the compound (e.g., from the most-recently added fragment or from a randomly selected fragment, etc.), or the like.
25 An addable fragment is then selected in step 26 to be added to the compound at the growth location. The addable fragment can be selected, for example, randomly, from user input, based on a matrix of probability elements, based on a list of structures, or the like. The addable fragment is then added to the compound to produce a new compound in step 28.
30 Next, the new compound is examined in step 30 to determine whether any stated goal has been met. If a goal was reached, then further growth of the compound is stopped and the new compound is outputted and/or recorded in some fashion, as shown in step 32. Examples of goals that can be checked include, but are not limited to, the

- 9 -

number of addable fragments that have been added, the number of rings present within the structure, the molecular weight of the new compound, or the like; other examples are discussed below. If no goals have been met, then the compound may continue to be grown, i.e., enlarging the compound. That is, the new compound may be used as a starting compound and the growth process repeated, e.g., starting with step 24 by
5 selecting a new growth location on the new starting compound.

In many techniques described herein, various "fragments" of compounds are mentioned, and fragments are sometimes used as "building blocks" to generate new compounds. For example, a compound may be identified as being formed from one or more fragments, and/or a compound may be grown by adding one or more fragments to the compound. A "fragment," as used herein, is a molecular structure composed of one or more interconnected atoms (e.g., connected by covalent bonds). For example, a fragment may be a single atom (e.g., a halogen, such as fluorine, chlorine, bromine, iodine, etc.), or an organic moiety. The fragments are often identifiable functional
10 groups (e.g., amide, amine, carbonyl, carboxylic acid, benzene, methyl, ethyl, etc.). Further examples of fragments are shown in Table 1. Note that "interconnected," as used herein, does not necessarily require every atom within the fragment to be connected to every other atom within the fragment (although they can be in some cases), but rather requires that each atom within the fragment be bonded (e.g., covalently) to at least one other bond within the fragment. Thus, in methyl (-CH_3), each of the hydrogen atoms is
15 connected to the carbon atom, although the hydrogen atoms are not directly connected to each other.

The fragments used herein may be predetermined fragments, or the fragments may be selected using a database of compounds (and/or portions of compounds) and
25 identifying fragments present within the database. The database may be, for example, provided or identified by a user. Examples are discussed in detail below. In one set of embodiments, a database of compounds is analyzed in some fashion, e.g., by using commercially-available programs, and fragments are identified within the database. In another set of embodiments, fragments may be identified from a database based on
30 commonly-occurring functional groups. For example, fragments within a database may be identified by identifying common organic groups such as alkyl moieties (methyl, ethyl, propyl, isopropyl, etc.), alkenyls, alkynyls, cyclic groups (cyclopropyl, cyclobutyl, cyclopentyl, etc.), halogens (e.g., fluorine, chlorine, bromine, iodine, etc.), alcohols,

- 10 -

ethers, esters, carbonyls (e.g., aldehydes or ketones), carboxylic acids, carboxylic acid halides, nitrogen-containing groups (e.g., amines, amides, azides, nitriles, nitros, etc.), aromatic groups (e.g., benzene), sulfur-containing groups (e.g., sulfides, sulfones, sulfoxides, thiols, etc.), or the like. The fragments may be connected by one or more
5 covalent bonds, or in some cases, non-covalent bonds such as through coordinative or dative bonds, hydrophobic interactions, hydrogen bonds, van der Waals forces, or the like.

In one set of embodiments, fragments may be selected using a set of compounds (and/or portions of compounds) by using commercially-available programs able to
10 identify substructural patterns within a database of compounds. One non-limiting example of such a system is the SMILES Arbitrary Target Specification system (SMARTS), an extension of the SMILES system (simplified molecular input line entry specification). Both of these systems have been widely reported in the literature and are readily known to those of ordinary skill in the art. In one embodiment, a database of
15 compounds is analyzed using SMARTS software to identify substructures forming the molecules, and a certain number or percentage of the substructures that are found can then be used as fragments. Thus, for example, at least 20%, at least 40%, at least 60%, at least 80%, or 100% (i.e., all of the identified substructures) of the substructures identified in a database of compounds using SMARTS may be used as fragments in some of the
20 techniques discussed herein. As other examples, some minimum number (e.g., at least about 20, at least about 40, at least about 1000, at least about 3000, etc.) of the substructures identified in a database of compounds using a SMARTS tool may be used as fragments in some of the techniques discussed herein. Programs that are able to perform SMARTS operations are commercially available from a number of different
25 sources, including OpenEye Scientific Software (Santa Fe, NM) and Daylight Chemical Information Systems (Aliso Viejo, CA).

In some embodiments, a matrix of probability elements (or other suitable data structure) is used to identify and select addable fragments for growing compounds. The matrix can contain data representative of fragments that may be present within a growth
30 compound and/or may be added to the growth compound, and in some cases may contain probability information relating such fragments in some manner. For instance, in one set of embodiments, a matrix of elements may be arranged according to a first dimension representing fragments that may be present within a growth compound (e.g., as the rows

- 11 -

or columns of the matrix), and according to a second dimension representing fragments that could be added to the growth compound (e.g., as the columns or rows of the matrix). The corresponding individual elements or cells may then represent a probability or frequency within the database that the two fragments are connected to each other.

5 Accordingly, it should be understood that such a matrix can be used to generate a new set of compounds that generally resembles an original starting set of compounds. The new set of compounds will generally resemble the original starting set of compounds, for example, with respect to the distributions of the fragments and their connections within each of the sets. For instance, the original set of compounds may
10 exhibit a first probability distribution of fragments and/or connections between fragments, and the newly generated set of compounds may exhibit a second probability distribution of fragments and/or connections between fragments that is similar to the first probability distribution. Thus, as a specific non-limiting example, if fragment "A" appears in 10% of the original set of compounds, fragment "A" may also appear in about
15 10% of the new set of compounds. As another example, if fragments "A" and "B" are connected in 30% of the members of the original set of compounds, then fragments "A" and "B" may be connected in about 30% of the members of the new set of compounds, etc. As yet another example, if in the original set of compounds "A" is connected to "B" with a frequency of 30%, relative to the likelihood that "A" is connected to other
20 fragments, than in a grown set of compounds the likelihood that "A" is connected to "B" will also be the same, or approximately the same.

 It should be noted that such a new set of compounds also may reflect inherent features of the original starting set of compounds, in addition to the distributions of fragments and/or connections between fragments. for example, if the original starting set
25 of compounds were relatively easy to synthesize, then the new set of compounds also will often be generally easy to synthesize due to their structural similarity. As another example, if the original starting set of compounds had a fairly narrow distribution of melting points, boiling points, etc., then the new set of compounds also may exhibit a fairly narrow distribution of melting points, boiling points, etc., due to their structural
30 similarity.

 In some cases, the matrix has more than two dimensions. For instance, the matrix may have three dimensions, with two of the dimensions representing, respectively, a first fragment and a second fragment present within a growth compound, while a third

- 12 -

dimension may represent addable fragments that could be added to the growth compound. Higher-order matrices (e.g., four dimensions, five dimensions, etc.), may also be used in some cases.

In addition, other elements may be present within the matrix, other than specific fragments. For example, a row or a column within a matrix may represent other structural or physical/chemical information, for example, molecular weight, the number of rings, the number of charge groups, the number of double and/or triple bonds, or the like. Thus, as a particular example, a first row in a matrix may represent compounds having no rings and a second row may represent compounds having one ring, while columns within the matrix may represent addable fragments (or categories of fragments, e.g., ring or non-ring) that could be selected, with the elements at their intersections representing probabilities, e.g., a probability that two fragments are connected to each other via a covalent bond or the like.

In one set of embodiments, such a matrix may be created by starting with a database of compounds, identifying fragments forming some or all of the compounds within the database, and determining the connectivity of the fragments forming the compounds. This may be implemented on a computer in some cases, and in some cases, the matrix may be stored on a computer-readable medium. The fragments, as previously discussed, may be composed of one or more interconnected atoms, such as organic moieties, functional groups, or the like, and can be identified using techniques such as those described above, e.g., using SMARTS.

An example flow chart of this technique is now discussed with reference to Fig. 5. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. In Fig. 5A, a starting set of compounds is initially received, e.g., from input from a user or an entity's recording (e.g., a computerized database) exhibiting some property (e.g., a functional activity), in step 40. The set of compounds may be extracted from and/or recorded in any suitable database, e.g., ones that are commercially available. For instance, a database may contain various drugs and/or drug candidates, biologically active molecules, various catalysts, or the like. Non-limiting examples of such compound databases include the ChemBank Bioactives Database, the NCI Open Database, or the like. The database containing the starting set of compounds may be user-provided (e.g., a weblink), a database stored on a local drive, etc.

- 13 -

The starting set is scanned and fragments forming each member of the set are identified, as shown in step 42. In addition, the connectivities of these fragments for each member, or element of the set, may also be identified, as is shown in step 44. For example, benzoic acid, if present, may be identified as being formed of benzene ring and a –COOH fragment. These fragments and their connectivities are then counted in some fashion in step 46. For instance, in one embodiment, a two-dimensional matrix may be formed using each fragment as an element having a first dimension and a second dimension. The first dimension may represent a starting fragment and the second dimension may represent an addable fragment. The fragments within the database of compounds, and their connectivities, can then be counted to populate the matrix, each element of the matrix representing the count, or fraction, of connections made between the relevant pairs of starting and addable fragments. Optionally, the matrix may then be stored, e.g., saved to a disk, as shown in step 48.

As a specific example, in the matrix shown in Fig. 5B, a series of fragments are shown on the rows and the columns of the matrix, with the rows representing starting fragment and the columns representing addable fragments. The cells within the matrices then may represent the probability that the respective starting and addable fragments are connected within the starting set of compounds. Such a matrix of probabilities may be generated, in one set of embodiments, by identifying each fragment and each connection between the fragments within the starting set of compounds. For example, each connection between fragments within the starting set of compounds may be counted and tallied within a matrix. After counting each of the fragments and each of the connectivities, the matrix may optionally be normalized in some cases to form a percentage or ratio, or otherwise processed to form a plurality of probability elements. In some cases, the matrix may also be modified, e.g., to account for various duplicate counts (i.e., redundancies), as discussed below.

As discussed above, in some cases, the matrix may have more than two dimensions. For instance, the matrix may have three dimensions, with the dimensions indicating connectivity of a first fragment with a neighboring fragment, and a penultimate fragment (e.g., one that connects to the neighboring fragment, but does not connect to the first fragment). Higher-order matrices (e.g., four dimensions, five dimensions, etc.), may also be used in some cases.

- 14 -

Thus, in some cases, other structural information about the fragments may also be recorded in the matrix. For example, growth locations of the fragments may be recorded, the fragments may be identified as being ring or non-ring fragments, charged or uncharged groups, hydrogen bond donors or acceptors, or the like.

5 Alternatively, other data structures may be employed. For example, in a three-dimensional matrix, the third dimension may comprise a pointer to a separate table, such as a table in which related attributes are recorded.

A specific non-limiting example of generating a compound is now discussed with reference to Figs. 2A-2B. In Fig. 2A, process 20 starts with a starting compound 12
10 (benzene, in this example) to be used as a growth compound. This compound may be input by a user, preselected by a computer-implemented process, randomly selected by a computer program (e.g., from among various starting options, for instance, based on the probability of finding it in a database of compounds of interest), or the like. A growth location within the growth compound may be selected, e.g., selected randomly from the
15 hydrogen atoms present within the compound. This growth location is indicated in this example by arrow 14 in Fig. 2A. An addable fragment is then selected. An example of such a matrix 28 is shown in Fig. 2B, produced using techniques such as those described above. For instance, the first row of the example matrix is cyclohexane, and the first column of the matrix is chlorine. At the intersection of that row and column is an
20 element, or cell, containing the number 0.07, indicating (based on available information) a 7% probability that cyclohexane and chlorine may be connected. That is, in a starting set of compounds selected for some reason such as the fact that all exhibit some desired biological activity on animal subjects, 7% of the fragment connections were cyclohexane to chlorine. As noted, this matrix is an example only; other matrices will contain
25 different numbers of fragments and different probabilities; thus, different numbers of rows and columns can be generated using techniques such as those discussed below.

In some cases, the matrix may also be modified, e.g., to account for various redundancies in the resulting compounds. For instance, the matrix may be corrected for fragments that are substructures of other fragments that can be combined to form other
30 fragments within the matrix, moieties that are more likely to be grown because they result from the combinations of more than two fragments, or the like. As a schematic, non-limiting example, adding fragment "A" + "B" may result in a fragment "C" that has been previously counted (e.g., if A is a carbonyl, B is an amine, and C is an amide), the

- 15 -

matrix may be corrected to eliminate this duplicate counting, as adding fragments A and B to a compound may produce the same result as adding fragment C to the compound. Thus, "B" can be disallowed from being added to "A" in this example. Furthermore, as another example, if a resulting 2-mer substructures can be grown in more than one way, e.g. adding fragment "E" to fragment "D" results in an identical 2-mer substructure as adding fragment "D" to fragment "E," these can be corrected to offset the inherent bias towards their growth.

Similarly, using benzene (the second row of Fig. 2B), fragments that can be added to the growth compound, and their connection probabilities, are then identified. Thus, in this example, -Cl, -Br, -COOH, -CH₃, etc., are all potentially addable fragments to benzene (but not -OH, which has a probability of 0 in this example). One of these addable fragments is then chosen, e.g., randomly or using sampling based by the probability elements from the matrix (e.g., -COOH has a 24% chance of being selected), and is then added to the benzene ring to form a new growth compound, benzoic acid 16, as is shown in Fig. 2A. This process can be repeated any number of times, ending when a certain goal is reached (e.g., adding a certain number of fragments, reaching a certain molecular weight, including a certain number of rings, etc.).

An example of further additions of addable fragments to this compound is shown in Fig. 4, producing more complicated compounds. Starting with a database of compounds (e.g., provided by any of the above-discussed techniques), compounds having structural connections (and a distribution of such similar to those present within the database of compounds are generated, as discussed below. In this figure, benzene is received as a starting compound by this process. The starting compound may be selected, for example, based on the number or frequency of fragments appearing within the database. A growth location is then selected on the benzene ring (since all six hydrogen atoms in benzene are chemically equivalent, this step is, of course, trivial for this example).

Next, a decision is made to select a non-ring structure as the addable fragment. This decision can be made, for example, on the basis of a fixed percentage (for example, the percentage of ring and non-ring fragments that appears in the database). In this example, a non-ring fragment is chosen. Then, the addable fragment is selected, for example, with reference to a matrix of probabilities previously generated using the database of compounds (e.g., matrix 28 of Fig. 2B). For instance, based on the benzene

- 16 -

ring, -COOH is chosen as an addable fragment, and added to the growth compound to form a new growth compound (benzoic acid, $\text{C}_6\text{H}_5\text{-COOH}$). The new growth compound is then analyzed to determine if certain criteria are met—for example, if a certain molecular weight has been exceeded. If not, a new mode of growth is then selected for the next addable fragment, where the “mode of growth” defines the number of available growth locations within a fragment. This mode of growth can be selected, for example, on the basis of a fixed percentage (for example, the percentage of linear and/or branched fragments that appears in the database). For instance, a fragment may have no available locations for growth, one available location (a “linear” growth mode), or more than one location (a “branch” growth mode). In this example, a linear mode of growth is chosen.

This process can then be repeated any number of times, for example, until a certain molecular weight is reached. For example, Fig. 4 shows that benzoic acid is chosen as the new growth compound. In this process, a growth location is selected on the benzene ring, e.g., chosen randomly from the available hydrogen atoms present within benzoic acid, or chosen from among the various fragments forming benzoic acid (i.e., the benzene ring and the carboxylic acid moiety). In this example, the para position on the benzene ring is chosen to be the next growth location. A non-ring fragment is chosen to be added to the growth location, which is selected to be methyl (-CH_3). The methyl group is then added to the growth compound to form 4-methylbenzoic acid. Further iterations of this process can be used to produce more complex compounds (e.g., producing 4-methylphthalamic acid, and more complicated structures, etc.), as is shown in Fig. 4.

After growth, the generated compound may be outputted in some fashion, for example, to a screen or other display device, to a printer, or to a file for further use. The generated compound may be outputted in any suitable representation, including as a structural diagram, in SMILES (simplified molecular input line entry specification) format, InChI (IUPAC International Chemical Identifier) format, SLN (SYBYL Line Notation) format, or the like. For example, the compound may be stored in a file, e.g., a library, for later analysis.

In one set of embodiments, a starting compound is first received and used to generate new compounds. The starting compound may be relatively complex, or as simple as a single atom, depending on the application. As used herein, “received,” includes situations in which a human selects or inputs a compound (or a range of

- 17 -

compounds), as well as situations in which a computer selects a compound (or range of compounds) using any suitable technique, and situations in which an entire intact database of compounds is input without subselection of its contents. For example, a user may input a compound into a computer. The user may input the compound using any
5 suitable technique or format. For instance, the compound may be inputted as a structural formula or representation, e.g., using accepted techniques such as SMILES format, InChI format, SLN format, etc.

In some cases, however, a computer-implemented process may generate a starting compound. For example, in one embodiment, a starting compound may be
10 pseudorandomly generated from a set of candidate starting compounds. In some cases, the set of candidate starting compounds may be initially biased in some way, e.g., based on the probabilities that the candidate starting compounds are found in a database of compounds. In another embodiment, a computer-implemented process may select a starting compound in an orderly manner from a list or a range of potential starting
15 compounds. The list may be one generated or inputted by a user, or one that is generated by the same or a different computer program that performs the (optional) selection process. The computer-implemented process may then select a starting compound from the list or range of compounds using any suitable method, for example, randomly or sequentially. The list may be presented, for example, as or in a computer file(s)
20 containing representations of chemical structures in any suitable format, such as those described above. In some embodiments, the list of potential starting compounds may be generated from a list of fragments contained within a database of compounds, e.g., as discussed above. As a specific example, if fragments contained within a database of compounds include those fragments shown in Fig. 2A, one of the fragments (e.g.,
25 cyclohexane, benzene, etc.) may be randomly selected to start growth.

Starting from a growth compound (which may be a starting compound, such as described above, or a growth compound previously grown through one or more rounds of iteration), a growth location within the compound is then selected. The growth location is the next location on the growing compound where an addable fragment is
30 added to the compound. For instance, a growth location may be a hydrogen atom on the compound that is replaced by another moiety, e.g., a halogen atom, a carbon-containing organic group, or the like. As a specific, non-limiting example, a hydrogen atom within the benzene compound (C₆H₆) shown in Fig. 2A may be replaced with –COOH as an

- 18 -

addable fragment, resulting in benzoic acid ($\text{C}_6\text{H}_5\text{--COOH}$). It should be noted, however, that other atoms besides hydrogen may be chosen as the growth location, in some embodiments. For example, oxygen atoms, nitrogen atoms, halogen atoms, or other atoms may be chosen as growth locations.

5 The growth location may be selected within the compound using any suitable technique. For example, in one embodiment, the growth location is selected randomly from one or more of the hydrogen atoms present within the molecule. In another embodiment, the growth compound is selected randomly from a location within the most-recently added fragment to the growth compound (e.g., from a hydrogen atom or
10 other atom within this fragment). In yet another embodiment, the growth location is selected by randomly selecting one of the fragments forming the molecule, and then selecting a location within the fragment as the growth location based on a matrix of probabilities. The matrix can contain, for instance, multiple fragments, and the matrix also may contain probability elements regarding growth locations within those
15 fragments, e.g., such that a growth location can be selected based on those probabilities. For instance, in a fragment $\text{--CH}_2\text{OH}$, the methylene hydrogen atoms may have a first probability for being selected as a growth location, while the hydroxide hydrogen atom may have a second probability for being selected as a growth location, which may or may not be equal to the probability for selecting the methylene hydrogen atoms.

20 An addable fragment can then be added to the growth compound at the growth location. The addable fragment can be selected using any suitable technique. For instance, the addable fragment may be selected by a user, selected randomly, selected from a list or a range of potential addable fragments, or in some cases, selected using a matrix of probability elements. In one set of embodiments, the addable fragment may be
25 selected (e.g., randomly) from among the fragments identified using a database of compounds, for example, provided by a user, such as described above. In another set of embodiments, an addable fragment may be identified from a database of compounds based on commonly-occurring functional groups, such as those discussed above. As a specific, non-limiting example, in Fig. 2A, the addable fragment selected to be added to
30 benzene ring 12, at growth location 14, is --COOH .

As mentioned, in one set of embodiments, the addable fragment may be chosen using a matrix of probability elements. The matrix may include, for example, a table of fragments in two (or more) dimensions, with probability elements (e.g., numbers ranging

- 19 -

from 0 to 1, or other representations of probabilities, such as whole numbers which are then normalized in some fashion to yield probabilities, ratios of numbers, etc.) as the elements of the matrix. Thus, for example, a suitable matrix may include a first dimension representing fragments that may be present within a growth compound (e.g., as the rows or columns of the matrix), and a second dimension representing fragments that could be added to the growth compound (e.g., as the columns or rows of the matrix). Additional examples of matrices, as well as the generation of such matrices, are discussed in more detail, below.

A non-limiting example of such a matrix is shown in Fig. 2B. In this figure, the rows of the matrix (showing cyclohexane, benzene, butane, isopentane, etc.) may indicate fragments that may be present within a growth compound, while the columns of the matrix (showing -Cl, -Br, -OH, -COOH, -CH₃, etc.) may indicate addable fragments that could be added to a growth location on the growth compound. The intersections of each of the rows and columns comprise cells showing probability elements (in this example, numbers ranging between 0 and 1) that represent the probability or likelihood of selecting and adding the respective addable fragment to the growth compound. For instance, a growth compound such as benzene may be evaluated using this matrix, and the probabilities for each addable fragment can be determined by examining the row representing benzene on the matrix. Thus, chlorine (-Cl) has a 22% probability of being selected, bromine (-Br) has a 12% probability of being selected, hydroxide (-OH) has a 0% probability of being selected, etc.

The addable fragments chosen to be added to the growth compound can be selected using any number of techniques. For example, in one embodiment, a fragment within the growth compound is selected, and the probabilities of adding an addable fragment to that fragment are then identified. As a specific example, again referring to Fig. 2B, if a growth compound contained both a benzene ring and a butane moiety (e.g., butylbenzene), then one of the fragments can be chosen (for example, the butane moiety), and the probabilities in the row representing butane may be identified. Thus, chlorine (-Cl) has a 16% probability of being selected, bromine (-Br) has a 4% probability of being selected, hydroxide (-OH) has a 28% probability of being selected, etc. In another embodiment, however, each of the fragments within the growth compound may be selected, and the probabilities corresponding to some or all of those fragments may be combined (e.g., by adding the probabilities, calculating root-mean-

- 20 -

squares, or the like), and optionally normalized, to determine the probabilities of choosing an addable fragment for the growth compound. Thus, referring to Fig. 2B, if a growth compound contained both a benzene ring and a butane moiety (e.g., butylbenzene), then the probabilities of the row representing benzene and the row representing butane can be combined, and those probabilities used to determine the addable fragment to be added to the growth compound.

In some cases, additional criteria may be used to select an addable fragment. For example, in one set of embodiments, there may be two (or more) groups of addable fragments that could be added to a growth compound (for example, ring or non-ring fragments, charged or uncharged groups, hydrogen bond donors or acceptors, etc.), and certain criteria used to determine which of the groups of addable fragments are selected. For instance, by selecting a ring fragment, a first group of addable fragments may be used (e.g., a first matrix of probability elements containing ring fragments), and by selecting a non-ring fragment, a second group of addable fragments may be used (e.g., a second matrix of probability elements containing non-ring fragments).

As another non-limiting example, the fragments may be divided in accordance with the number of growth locations available within the fragments, for instance, none, one (a "linear" fragment), or more than one (a "branch" fragment). Thus, for instance, as a growth compound is generated, the number of fragments having "open" growth locations is determined, and the growth sequence may be ended if there are no more open growth locations present. Conversely, one (or more) of the "open" growth locations may be selected for further growth based on various criteria. For example, a user may desire that there be a certain distribution of branch fragments relative to linear fragments, or a user may desire that there be no more than a certain number of fragments of a particular type within the generated compounds (e.g., no more than 3 branches, no more than 2 rings, or the like).

The growth of a molecule, using techniques such as those described above, may be ended for any number of reasons (depending on the application); or the growth compound, after addition of the addable fragment, may then be used as a new (starting) growth compound and the above steps repeated. One non-limiting example of a goal that may be used to end growth of the compound include the addition of a certain number of fragments (e.g., 1, 2, 3, 4, 5, 6, 7, etc.). In some cases, these may be referred to as "mers," by analogy to oligomers, although in this context, a 3-mer is a molecule formed

- 21 -

from three fragments, a 4-mer is a molecule formed from four fragments, etc. Other non-limiting examples of goals include, but are not limited to, the growth compound reaching or exceeding a certain molecular weight or a certain number of atoms, reaching a certain maximum number of specific types of fragments having been added (e.g., a certain number of ring fragments, branch fragments, etc.), there being a certain number of functional groups present (e.g., a certain number of rings, a certain number of double bonds and/or triple bonds, a certain number of hydrogen bond donors or acceptors, a certain number of electron-donating or electron-withdrawing moieties, or the like), no more growth locations being present within the growth compound, etc. Those of ordinary skill in the art will be aware of hydrogen bond donors, for instance, comprising a hydrogen atom attached to a relatively electronegative atom, such as fluorine, oxygen, or nitrogen, and hydrogen bond acceptors, for instance, comprising an electronegative atom, such as fluorine, oxygen, or nitrogen (regardless of whether it is bonded to a hydrogen atom or not). Those of ordinary skill in the art will also be aware of electron-withdrawing groups, for example, moieties which can draw electrons therein, such as halogens, nitriles, carboxylic acids, or carbonyls; and electron-donating groups, for example, moieties which can release electrons, such as alkyl groups, alcohol groups, amino groups, or the like. In some cases, a compound, after formation, may also be rejected for reasons similar to those discussed above, for example, if the compound contains too many hydrogen bond donors or acceptors, too many halogen atoms, or the like. As another example, a compound may be rejected if it is formed from certain combinations of fragments not present in the original database of compounds. For instance, if a 3-mer compound is formed, or a 3-mer substructure of a larger compound is formed, using certain fragments, and that combination of fragments does not appear in a single compound in the original database, the newly-formed 3-mer compound may be subsequently rejected.

A non-limiting example flow chart using some of these additional criteria is now described with reference to Fig. 3. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. In this flow chart, starting compound 22 is received using any suitable method. A growth location is then selected on the compound in step 24 for the addition of the addable fragment. This location may be selected using any suitable technique, such as those described above. Next, a desired type of addable fragment is selected from

- 22 -

various groups. As is illustrated in this flow chart at step 23, the addable fragment is selected as being either a ring or a non-ring fragment, for example, based on the growth compound (e.g., based on the number of rings present in the growth compound, or other similar criteria). The appropriate addable fragment can then be selected in respective
5 step 25 or 27 to be added to the compound. The addable fragment can then be added to the growth compound, producing a new growth compound in step 28. The new growth compound may be checked to see if one or more goals have been met, as is shown in step 31, and if a (or any) goal has been met (e.g., a certain number of fragments have been added, a certain molecular weight has been reached or exceeded, etc.), growth is stopped
10 and the new compound is outputted or recorded, as shown in step 32.

However, if no goals have been met, then the new growth compound may be designated in step 31 to grow in a linear fashion or in a branch fashion. In linear growth, a growth fragment is chosen that only has one current connection to another fragment. In contrast, in branched growth, a growth fragment is selected that has at least two other
15 fragments connected to it. In addition, depending on the selection of linear or branch growth, different sets of addable fragments may be selected for potential addition to the molecule in some cases, e.g., selecting a fragment from a first set of fragments for linear growth, and selecting a fragment from a second set of fragments for branch growth. The mode of growth may be selected using any suitable method, for example, based on the
20 number of fragments and/or to the type of fragments already added to the molecule. As non-limiting examples, some addable fragments (e.g., -Cl, -Br, or -OH) may not be suitable for further growth, while other fragments (e.g., -COOH, -OH, or -CH₃) may be suitable for further growth. (Note that some compounds, such as hydroxide, may be suitable for either or both categories, depending on the application, i.e., the hydroxide
25 moiety may signal the end of growth, or be modified to form an ether bond, etc.) Then, growth of the compound is continued by going back to step 23 and selecting a desired addable fragment, as previously discussed.

In some cases, the compound may then be synthesized, for example, by a human or by a suitable robot, depending on the compound. In other cases, the compound may
30 be given to a person, or an entity, with instructions to synthesize the compound. The compound may also be used in other applications. For example, the compound can be used in other synthesis, testing, or assay procedures, or the compound may be grown or modified using other techniques, or the like.

- 23 -

The present invention is directed to systems and methods for categorizing and/or separating compounds, according to another aspect, and these may be implemented on a computer in some cases. For example, a set of compounds may be scored using various descriptors (e.g., determining the number of rings, connectivity of fragments forming the compounds, or the like), and the scores weighted in some manner. For instance, in one set of embodiments, the weighted sum may be compared against a standard (e.g., against a threshold score) to characterize or separate the compounds.

As used herein, a “descriptor” is a test, such as an assay or an analysis based on structural information or data, that can be applied to a compound to produce a result. The result may be quantitative or qualitative, depending on the test. In one set of embodiments, the descriptor may produce a result based on structural information about the compound. For example, a descriptor may be used to determine the number of rings within a molecular structure, i.e., the result may be 0, 1, 2, 3, etc. Other non-limiting examples of descriptors based, at least partially, on structural information about the compound include determining the number of atoms within a structure, the number of carbon atoms within a structure, the number of heteroatoms within a structure, the number of halogens within a structure, the number of charge groups within the molecule, the overall charge of the molecule, or the number of rotatable bonds within a structure. For instance, the descriptor may be the number of atoms, etc. within the structure, or the descriptor may be a function that includes the number of atoms, etc. within the structure. As a specific non-limiting example, structures having up to a certain number of hydrogen bonds acceptors or donors may receive a first score, while structures having more than that number of hydrogen bond acceptors or donors may receive a second score.

A “rotatable bond,” as used herein, is an atomic bond between two atoms that is free to rotate about the axis joining the two atoms. For example, the C-C atoms within ethanol are free to rotate with about the central axis joining the carbon atoms, but the C-C atoms within cyclohexane are not free to rotate, due to the constraints imposed by the cyclohexane ring structure.

As examples of descriptors, in one set of embodiments, each fragment within a compound may be given a numerical score, and the result of the descriptor is the sum (or other combination) of the scores of each of the fragments forming the compound. Such scores may be determined, for example, by determining the fragments found within a given database of compounds, and applying weights to them according to some method

- 24 -

(e.g., based on the frequency at which each fragment appears within the database). To use such a descriptor, a test compound is analyzed by determining the fragments forming the compound. The scores of each fragment can then be summed (or otherwise combined, e.g., multiplicatively) to yield a final score for the descriptor.

5 A flow chart for constructing such a matrix of weighted fragments is shown in Fig. 6A. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. In this figure, in step 61, a database of compounds is received, for instance, from input from a user or selected by a computer. The database is then scanned and fragments forming
10 each member of the database are identified, as shown in step 63. The fragments are then counted and tabulated in step 67, e.g., within a matrix, as is shown in Fig. 6A. A specific example is illustrated in Fig. 7A. In this example, a matrix of fragments is illustrated in Fig. 7A. In this example, the relative percentage that each fragment appears is shown, although in other examples, other methods can be used, such as the number of times each
15 fragment appears. Thus, for example, a benzene ring has a probability of 0.38, chloride has a probability of 0.89, methyl has a probability of 0.13, etc. To score a compound such as benzoic acid (C_6H_5-COOH), the compound is first analyzed to determine the fragments forming the compound (in this example, benzene and a $-COOH$). The scores for each fragment are then determined using the matrix of Fig. 7A (i.e., benzene has a
20 score of 0.38 and $-COOH$ has a score of 0.57), and the scores may be combined (e.g., added) to produce a final score for the descriptor (for example, 0.95, which is $0.38 + 0.57$).

 In another set of embodiments, a descriptor may include the connectivities of fragments. For instance, each of the connectivities of the fragments forming a compound
25 may be given a numerical score, and the result of the descriptor is the sum (or other combination) of the scores of each of the fragments forming the compound. Such scores may be determined, for example, by determining the fragments found within a given database of compounds, determining the connectivities of those fragments, and applying weights to the connectivities according to some method (for example, based on the
30 frequency at which each of the connections appears within the database). To use such a descriptor, a test compound is analyzed by determining the fragments forming the compound, and the connectivities of those fragments. The scores of the connectivities

- 25 -

can then be summed (or otherwise combined, e.g., multiplicatively) to yield a final score for this descriptor.

In still another set of embodiments, a descriptor may be scored based on differences between more than one database of compounds, e.g., a first database
5 containing desired compounds (for example, drugs) and a second database containing undesired compounds (for example, non-drugs). For example, numerical scores may be calculated for the compounds in each database (e.g., as discussed above), and the differences between the scores used to determine the descriptor. As a specific, non-limiting example, a first database may contain a first probability of a specific fragment
10 (or other feature, such as the number of rings, the number of hydrogen bond acceptors or donors, the number of rotatable bonds, or the like), and a second database may contain a second probability of the specific fragment; a descriptor may be formed based on the differences of probabilities between these two databases. Thus, if the fragment or other feature is present, the descriptor is given a first value (e.g., the difference in probability),
15 and if the fragment or other feature is not present, the descriptor is given a second value (e.g., 0).

A flow chart for constructing such a matrix of weighted connectivities is shown in Fig. 6B. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. In this
20 figure, in step 61, a database of compounds is received, for instance, from input from a user or selected by a computer. The database is then scanned and fragments forming each member of the database are identified, as shown in step 63. The connectivities between the fragments forming the compounds is next identified, as shown in step 65, and these connectivities are counted and tabulated, e.g., to form a matrix, as is shown in
25 step 67. A non-limiting example of this is shown with reference to Fig. 7B. In this example, a matrix is shown with various fragments appearing on the rows and the columns of the matrix (for example, benzene rings, chloride, methyl, etc.). At the intersections of the rows and columns are numbers indicating the relative percentage of each connection. In other embodiments, however, other metrics can be used, such as the
30 number of times each connection appears in the database. Thus, for example, the probability that a benzene ring is connected to a chloride atom is 0.18, the probability that a benzene ring is connected to a bromide atom is 0.08, the probability that a benzene ring is connected to -COOH is 0.49, etc. To score a compound such as 4-methylbenzoic

- 26 -

acid, the compound is first analyzed to determine the fragments forming the compound (in this example, between benzene and -COOH , and between benzene and methyl). The scores indicating these connectivities are then determined using the matrix of Fig. 7B (i.e., 0.49 and 0.71), and the scores are then added or otherwise combined to produce a
5 final score for the descriptor (for example, 1.20, which is $0.49 + 0.71$).

Other chemical properties may also be used as descriptors. For example, the descriptor may include the number of hydrogen bond donors or acceptors within the compound. For instance, the number of hydrogen bond donors or acceptors can be calculated based on the structure of the compound, in some cases, using methods such as
10 cxcalc calculations (commercially available from ChemAxon), and this can be scored as a descriptor. The descriptor, in another set of embodiments, may also use predictive methods, e.g., in conjunction with structural information. For example, the melting point or boiling point of a compound, the diffusivity, etc., may be predicted using structural information about the compound, for instance, using estimation techniques such as those
15 known to those of ordinary skill in the art.

In some cases, a plurality of descriptors may be combined together (e.g., added, multiplied, etc.) to yield a final score, and in some cases, the descriptors may be weighted in some manner to produce a final score, e.g., using a series of constant weighing factors for each of the descriptors. For example, a final score may be
20 calculated such that 3, 4, 5, 6, etc. descriptors are summed together using appropriate constant weighing factors. For instance, a plurality of descriptors may be used, including the fragment scores, the connectivity scores, the number of hydrogen bond donors or acceptors, the number of atoms, the number of rings, the number of rotatable bonds, the number of carbon atoms, the number of oxygen atoms, the number of nitrogen atoms, the
25 number of halogen atoms, the number of double and/or triple bonds, or the like.

The final score, in some embodiments of the invention, may be used to categorize and/or separate compounds into one of two or more categories, for example, drug and non-drug, biological and non-biological, catalytic and non-catalytic, or the like. The categories can be objective, or even relatively subjective in some cases, for example,
30 easy to synthesize versus hard to synthesize. A compound can be assessed to determine which category it falls into by calculating a final score, as discussed above, and comparing the final score to a standard (e.g., against a threshold score) to determine which of several potential categories the compound belongs to.

- 27 -

In one embodiment, a categorization system may be prepared using a database having its members divided into two or more categories, for example, drug and non-drug. (Or equivalently, starting with two or more databases, e.g., one for drug and one for non-drug compound.) A number of descriptors are then selected, and the scores for
5 each of the compounds for each of the descriptors can be determined. This may be performed and optimized relatively quickly, for example, with a computer and/or using a certain number of "test compounds" selected from the database(s).

The descriptors for each compound can then be combined together in a weighted sum, for instance, by optimizing the weights such that at least a certain number of
10 members of a first category falls within a certain range, a certain number of members of a second category falls within another range, etc. For instance, the weights may be set up such that a first category of compounds (e.g., drug compounds) tend to have a positive score, while a second category of compounds (e.g., non-drug compounds) tend to have a negative score. These standards or threshold scores can be arbitrarily chosen in
15 some cases. The weights applied to each of the descriptors can then be optimized such that at least a certain number or percentage of members of a first category are scored to be within the first category and/or such that a certain number or percentage of members of a second category are scored to be within the second category. For instance, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about
20 75%, at least about 80%, etc., of the members of the first category may be weighted such that they are scored to be within the first category, and/or at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, etc., of the members of the second category may be weighted such that they are scored to be within the second category. The optimized weights for each descriptor can
25 be determined, for example, using matrix optimization techniques known to those of ordinary skill in the art, such as linear least-squares fitting or non-linear least squares fitting techniques (e.g., Gauss-Newton algorithms, the Levenberg-Marquardt algorithm, etc.).

A flow chart schematically diagramming an example of this process is now
30 discussed with reference to Fig. 8A. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. In this figure, a data set containing two or more categories (or equivalently, two or more databases) is received in step 70. A number of descriptors are

- 28 -

then selected in step 72, and the compounds within each category (or a subset of such compounds) are then scored using each of these descriptors, as is shown in step 74. The scores for each of the descriptors are then combined in a weighted manner in step 76, and the weights optimized using matrix optimization techniques known to those of ordinary skill in the art, for instance, such that the compounds from the first category tend to have scores within a first range (e.g., greater than 0) while the second compounds from the second category tend to have scores within a second range (e.g., less than 0). In step 78, the final weights are then recorded. It should be understood that this flow chart is by of example only, and other methods may be used in other embodiments of the invention.

For example, in some cases, the scores of each of the descriptors may be directly compared, e.g., without computing a final score.

In Fig. 8B, a flow chart illustrating the determination of a new compound, e.g., as belonging to a first category or a second category. Each of the steps shown in this flow chart may be prepared as a module or other suitable coded representation within a computer, in some embodiments. A test compound is received in step 80, and the descriptors (e.g., chosen as discussed above with reference to Fig. 8A) are then applied and scores calculated for each of these in step 82. The scores for each of the descriptors are then combined to produce a final score in step 84, and this final score may be compared against a standard to determine whether the test compound belongs to the first category or the second category. It should be noted that the use of two categories with reference to Figs. 8A and 8B is by way of example only, and in other embodiments, more than two categories may be used, each with various distinguishable ranges of scores.

U.S. Provisional Patent Application Serial No. 61/146,108, filed 01/21/09, entitled "Systems and Methods for Generating and/or Characterizing Molecules for Pharmaceutical and Other Uses," by Kutchukian, *et al.*, is incorporated herein by reference.

The following examples are intended to illustrate certain embodiments of the present invention, but do not exemplify the full scope of the invention.

EXAMPLE 1

This example illustrates an algorithm that was performed, in which new compounds were grown by adding fragments to a nascent compound in a statistically biased manner, using techniques such as those described herein. In addition, it is capable

- 29 -

of being trained to grow molecules that are substantially identical in their chemical and topological features to specific classes of chemicals, such as natural products (NP), diversity-oriented synthesis (DOS) products, drugs, or the like. In order to validate the latter finding, an algorithm capable of separating and identifying classes of compounds, such as drugs from non-drugs, was developed (see Example 2). As discussed below, this algorithm exploits the statistical bias of fragments and fragment connections (2D metrics), as well as coupled 1D metrics, such as number of atoms and rotatable bonds. The separation algorithm is also transparent in the features that it classifies compounds by, which helps bring to light salient features of drug-like compounds.

Initially, connectivity statistics were collected for fragments of interest from a database of small molecules. The statistics were then converted into transition probabilities ($P_{i \rightarrow j}$), the probability that a fragment i will transition (be connected to) fragment j during small molecule growth. In the simplest scenario, where each fragment is a single atom (represented by a letter in Fig. 9, showing an example of a simple growth scheme where fragments A through C can be joined to fragments A through C) and only linear growth is allowed, it can be seen that in order to convert counts (N_{ij}) into transition probabilities that will reproduce those counts, the counts should be divided by the total connection fragment i makes to all other fragments ($\sum_s (N_{is}^D)$).

In the following equations, superscripts are used to denote the database that is being searched. In this case, D is used to denote that the linkages are being counted in a first, training database. Also, on-diagonal elements (such as A-A) are half as likely to grow as the off-diagonal elements (such as A-B), and this must also be taken into account. This can be accomplished by dividing the counts for each linkage by the number of times the linkage occurs in an exhaustive 2-mer (where each “mer” represents a fragment in this context) database of fragments (N_{is}^{2mer}), giving:

$$P_{i \rightarrow j} = \left(\frac{N_{ij}^D}{\sum_s (N_{is}^D)} \right) \left(\frac{1}{N_{ij}^{2mer}} \right)$$

(1).

The next step was to account for larger fragments that include smaller fragments as part of their substructure (Fig. 10, showing an example of a simple growth scheme

- 30 -

where fragments A through CB can be joined to fragments A through CB). For example, an amide fragment could be divided into a carbonyl and an amine fragment. It was found that by accounting for how often linkages occur in an exhaustive 3-mer library N_{ij}^{3mer} (every way three fragments can be placed in positions i , $i + 1$, and $i + 2$),

- 5 connectivity statistics (i.e., the probability that a given fragment is connected to another fragment, that is, the first part of the following equation) could be produced that were in agreement with the training database, giving:

$$P_{i \rightarrow j} = \left(\frac{N_{ij}^D}{\sum_s (N_{is}^D)} \right) \left(\frac{1}{N_{ij}^{3mer}} \right)$$

(2).

- 10 Also, note that fragments cannot be joined in such a way that they grow fragments that already exist (e.g., B-C and "BC" linked fragments in Fig. 10 are essentially the same compound). This can be done by setting the corresponding transition probabilities to zero. This rule was used because by growing fragments already in the database, one could effectively lose information, since the transition probabilities solely depend on the growth fragment, and not on what it is connected to.
- 15 For instance, consider a fragment pool that included amide, amine, and carbonyl fragments. If the amine were allowed to combine with the carbonyl, and then grow from the nitrogen, the growth algorithm would be unaware that the nitrogen is now better represented by an amide, and would attach moieties to it that might be more likely
- 20 bonded to an amine nitrogen rather than an amide nitrogen.

- Next, the techniques discussed above were applied to a set of organic fragments. In all database searches that were performed, SMARTS strings were used in order to query the desired fragment or substructure. The jcsearch program from ChemAxon was used in all searches. Keeping with the sequential growth of fragments joined by single
- 25 bonds employed by SMOG as well as a number of other *de novo* algorithms, fragments were attached to each other by removing a hydrogen from each fragment, and subsequently connecting the atoms that were attached to those hydrogens to each other. The fragments were added, however, in a statistically biased way, depending on the growth fragment. Certain corrections (C) were made for double counting, due to search
- 30 strings being able to match a database molecule in more than one orientation, although

- 31 -

the algorithm would only be able to grow one of these linkages if it were to reproduce the molecule.

The corrections (C) that were made for double counting, due to search strings being able to match a database molecule in more than one orientation, although this algorithm would only be able to grow one of these linkages if it were to reproduce the molecule, are now discussed. These errors are shown in Fig. 11. In experiment 1 (Fig. 11, first row), if the search string that corresponds to the schematic representation was used (Fig. 11, column 2), and how often an amide carbonyl is bonded to a carbonyl fragment was counted, two linkages would be counted for the depicted database molecule. The algorithm, however, would only grow this linkage once if it attempted to reproduce the substructure (Fig. 11, column 5). The algorithm subtracts from the original count (N) the correction count (C) that is obtained by searching the molecule with the search string depicted in Fig. 11 (column 6). In experiment 2 (Fig. 11, second row), when a carboxyl carbonyl is attached to ethene, the search string matches the database molecule twice. In this case, a count of two was acceptable since when the structure is grown, two ethene to carbonyl carboxyl linkages might be employed (Fig. 11, column 5). The algorithm was sensitive to double-counting errors such as these, and was able to correct these errors when necessary in an automated fashion. In particular, Fig. 11 shows examples of double counting that either need to be corrected (Ex. 1) or are permissible (Ex. 2). X can be any atom including hydrogen. Corrected counts ($N-C$) were obtained by subtracting "Correction Search String" counts (C) from the original "Search String" counts (N).

Taking double counting errors into account when necessary:

$$P_{i \rightarrow j} = \left(\frac{N_{ij}^D - C_{ij}^D}{\sum_s (N_{is}^D - C_{is}^D)} \right) \left(\frac{1}{N_{ij}^{3mer} - C_{ij}^{3mer}} \right)$$

(3).

Using the above equation to obtain transition probabilities, connectivity statistics were reproduced with a small pool of fragments (methyl, ethyl, amine, hydroxyl, thiol, carbonyl, carboxyl, and amide). Some fragments were represented more than once using this method, since each unique growth site on a fragment had its own transition probabilities associated with it. The algorithm was then trained on the ChemBank

- 32 -

Bioactives Database, and a library of 10,000 linear molecules was grown, each 5 fragments long.

During the growth of each new compound, a random fragment was initially selected. Then, based on the growth fragment's transition probabilities, a second
5 fragment was added to it. The newly added fragment subsequently became the growth fragment, and the process was repeated until 4 fragments have been added to the initial fragment. This is in a sense a Markov chain, where the selection of the next state (fragment j), depends on the current state (fragment i). Fragments that were more likely to be connected to the growth fragment in the database were also more likely to be
10 selected during growth. This technique was able to reproduce the connectivity statistics in the grown library ($R^2=0.99$, Fig. 12). This figure shows a correlation of fragment connectivity propensities between linear molecules grown with a Markov Chain and ChemBank Bioactives used in training Markov Chain, represented by the following equation:

$$15 \quad P_{ij} = \left(\frac{N_{ij}^D - C_{ij}^D}{\sum_s (N_{is}^D - C_{is}^D)} \right)$$

(4).

Only eight fragments were used for growth. As expected, libraries grown with no statistical biases as a control were found to correlate poorly with the connectivity statistics obtained from the Chembank Bioactives Database ($R^2=0.25$).

20 More fragments were subsequently to the algorithm (Table 1), and the compounds allowed to branch. The branching probability, or the likelihood that a fragment will grow off of a fragment that already has at least two fragments connected to it, could be controlled by the user, allowing one to bias the growth of small molecules ranging from linear to highly branched. With the branching mode present, it was not as
25 straightforward to correct the counts by dividing them by the counts from an exhaustive library (as in the previous two cases where a 2-mer or 3-mer library, respectively, were searched). Thus, a 5-mer normalization library (10,000 molecules) was grown, employing the rules of growth (for example, fragments could not be connected if they grew a fragment already present in the pool of fragments) but without any statistical
30 biases and with a branching probability $P(B)$ of 0.5. This library was then searched for the likelihood that linkages will occur when no statistical biases are present

- 33 -

$(N_{ij}^{5mer} - C_{ij}^{5mer})$. Then, to obtain transition probabilities:

$$P_{i \rightarrow j} = \left(\frac{N_{ij}^D - C_{ij}^D}{\sum_s (N_{is}^D - C_{is}^D)} \right) \left(\frac{1}{N_{ij}^{5mer} - C_{ij}^{5mer}} \right)$$

(5)

5

Table 1

Fragments			
methane	tetrahydrofuran	sulfonyl	carboxyl
cyclopropane	morpholine	sulfoxide	amide
cyclobutane	benzene	sulfonamide	fluorine
cyclopentane	naphthalene	oxime	chlorine
cyclohexane	pyrrole	carbonyl	bromine
cycloheptane	furan	hydroxyl	iodine
cyclooctane	thiophene	amine	cyano
pyrrolidine	pyridine	thiol	nitro
alkene	alkyne		

During growth of a new compound, an initial fragment is chosen, either randomly or based on how often it is observed in the training database. Subsequent fragments may be added as follows. Throughout growth, three lists of fragments were maintained based on what type of growth is available from that fragment: linear (only one existing connection to another fragment), branch (at least two connections to other fragments), or none (all growth sites have been filled). The population of these three lists determined the possible growth modes that are available (linear, branch, or none). If both the linear and branching growth modes were available, one of the modes was selected based on a user defined branching probability. If only one of the modes was detected, then that mode was automatically selected. If all growth sites were saturated, then growth was terminated and the computer outputs the molecule. Unless stated otherwise, the branching probability $P(B)$ was set to 0.5 for the experiments in this example, although other probabilities could be used. A growth fragment was then selected from the appropriate list depending on the growth mode, and growth site on that fragment was randomly selected.

- 34 -

The selection of the fragment that will be connected to the current growth fragment was then made. This could be done, for instance, by using the transition probability of the growth fragment to select the subsequent fragment, or by deciding to select a ring or non-ring fragment prior to using transition probabilities to select the next
5 fragment. When the latter method was used, a ring non-ring decision was made based on how often the growth point of the fragment was connected to a ring or non-ring in the training database. This value can also be set by the user to be a given value for all fragments. Once a decision to grow to a ring or non-ring was made, the correct type of fragment was then selected based on the growth fragments transition probabilities. It
10 should be noted that the transition probability matrix in this case was split into two matrices, one for transitions to rings, and one for transitions to non-rings, and was normalized accordingly.

This process was then repeated until all growth sites were saturated, a user defined maximum number of fragments have been added, or a maximum molar mass had
15 been obtained. The molecule was outputted to a file as a SMILES string.

After a library of molecules has been grown, optionally, the library was screened for disallowed 3-mers. These are sequences of three fragments that were disallowed. There are two sources for identifying disallowed 3-mers. First the training database was searched for all 3-mer sequences that could be composed of the fragments, i.e., any 3-
20 mer sequences not present in the training database were considered to be disallowed. In order to avoid being too stringent, rings were treated very generally. For example, a SMARTS string representing a ring carbon would match all sp^3 carbons that are in a ring, but it would not be sensitive to the type of ring that the sp^3 carbon belongs to. Any 3-mer sequence that was not observed in the training database was considered
25 disallowed. In addition, a user could also disallow additional 3-mers, e.g., to avoid chemically unstable moieties such as acetals, ketals, amins, and iminals (Fig. 13). Examples of user defined disallowed 3-mers are shown in this figure. R is any ring or non-ring sp^3 carbon. Hydrogens can also be occupied by any non-hydrogen atom.

This example illustrates an algorithm that generates new compounds in a
30 chemical space that is similar to the compounds that it was trained on, whether they are drugs, natural products, or DOS compounds. The algorithm used the sequential growth of small molecules constrained by the transition probabilities of the growth fragment. Importantly, molecules grown with these transition probabilities reproduced the

- 35 -

frequencies in connections between fragments. This algorithm can easily be incorporated into de novo design programs that employ sequential growth of fragments, or it can be used as a stand-alone program to generate a virtual library of compounds of specific classes.

5

EXAMPLE 2

In order to evaluate the output of the growth algorithm of Example 1, a separation algorithm was developed. The separation algorithm scores molecules based on a number of individual components that score different features of a molecule. Each measure or descriptor was based on the difference in probabilities or score of observing some feature in a given database versus another database. The descriptors were designed to return a positive or negative value, depending on whether the scored molecule was deemed more representative of one class of molecules or another. The final score was a linear, weighted summation of the individual scores, and a compound was classified based on the sign of the final score (i.e., positive or negative). For example, the probability of finding a fragment in a database, as well as the log odds score of fragment connections, could be computed by ascertaining the probability of fragments and fragment connections, in addition to atoms and atom connections.

The probability of finding fragment i in a given database of compounds was defined as the number of occurrences of that fragment N_i^D divided by the sum of the counts of all fragments:

$$p_i = \left(\frac{N_i^D}{\sum_s (N_s^D)} \right)$$

(6).

The difference in probabilities of observing fragment i in database A compared to database B is thus:

$$D_i = p_i^A - p_i^B$$

(7).

The full fragment frequency score is then:

$$L_1 = \frac{1}{M_1} \sum_i^n \begin{cases} D_i & \text{if } i \text{ in molecule} \\ 0 & \text{else} \end{cases}$$

(8),

- 36 -

where M_1 is the total number of fragments in molecule. Initially, the same fragments were used in growth to generate this score. It was quickly observed however, that by adding fragments that were more representative of given databases (DOS, NP, etc.), more accurate classification of compounds could be obtained.

- 5 The probability that fragment i is connected to fragment j in a database was computed as the total times i was found connected to j (N_{ij}^D) divided by the sum of the counts of all pairing combinations of the fragments:

$$q_{ij} = \left(\frac{N_{ij}^D}{\sum_s \sum_k (N_{sk}^D)} \right)$$

(9).

- 10 The probability that a fragment i is found connected to any other fragment in the pool as all connections i makes to other fragments $\sum_s N_{is}^D$, divided by all possible pairing combinations of the fragments, is:

$$p_{is} = \left(\frac{\sum_s N_{is}^D}{\sum_s \sum_k (N_{sk}^D)} \right)$$

(10).

- 15 Note this is not the probability of finding i , but rather the frequency that was observed that it was bonded to a specific fragment, as opposed to other fragments belonging to the fragments of interest.

The relative probability that i is connected to j was then computed by dividing the frequency of finding that pairing q_{ij} by the sum of the individual frequencies of finding i

- 20 or j connected to other fragments:

$$S'_{ij} = \left(\frac{q_{ij}}{p_{is} p_{js}} \right) = \frac{N_{ij}^D}{\sum_s N_{is}^D \sum_s N_{js}^D}$$

(11).

It is more convenient to convert this value to the log odds score by taking the logarithm of the relative probability:

- 37 -

$$S_{ij} = \ln \left(\frac{q_{ij}}{p_{is} p_{js}} \right)$$

(12).

Since it is possible that a certain fragment pairing is not observed, or that a specific fragment is never observed, the relative probability was assigned a minimum value of 0.0001. The log odds matrix was then renormalized. The difference in log odds scores between database A and database B for a specific fragment pairing is then:

$$D_{ij} = S_{ij}^A - S_{ij}^B$$

(13).

The full fragment connection score is then:

$$L_2 = \frac{1}{M_2} \sum_i^n \sum_j^n \left\{ \begin{array}{l} D_{ij} \text{ if } ij \text{ in molecule} \\ 0 \text{ else} \end{array} \right\}$$

(14),

where M_2 is the total number of fragment pairs in molecule.

For each 1D topology metric descriptor, the joint probability of two variables $P(a,b)$ that would yield more information when looked at jointly, rather than as two single probabilities $P(a)$ and $P(b)$, was computed. For example, one of the joint probabilities computed was hydrogen bond acceptors and hydrogen bond donors per molecule $P(don,acc)$. Hydrogen bond interactions allow drugs to specifically interact with a macromolecule, and it would be highly unlikely that a drug-like molecule would lack both hydrogen bond donor and hydrogen bond acceptor sites. This is not a requirement for nondrugs, however, so it was suspected molecules lacking both hydrogen bond donors and hydrogen bond acceptors would be scored as nondrug-like.

A second topology metric descriptor that was used was the joint probability that a molecule has a certain number of atoms as well as a certain number of rings $P(atoms,rings)$. It was reasoned that conformationally rigid rings would be favored in drugs, and that a molecule with a high atom count but low ring count would probably be scored as a nondrug. In a similar vein, the joint probability that a molecule has a certain number of atoms as well as a certain number of rotatable bonds $P(atoms,bonds)$ was computed.

- 38 -

Each of the joint probabilities was computed as the total observances of both variables $N_{a,b}^D$ divided by the total number of molecules in the training database N_{total}^D :

$$P(a,b) = \left(\frac{N_{a,b}^D}{N_{total}^D} \right)$$

(15).

- 5 The difference of the joint probabilities between two database A and B was then computed for each possible joint value:

$$D_{P(a,b)} = P(a,b)^A - P(a,b)^B$$

(16).

Thus:

10 $D_{P(don,acc)} = P(don,acc)^A - P(don,acc)^B$

(17)

$$D_{P(atoms,rbonds)} = P(atoms,rbonds)^A - P(atoms,rbonds)^B$$

(18)

$$D_{P(atoms,rings)} = P(atoms,rings)^A - P(atoms,rings)^B$$

15 (19).

Each of the descriptors (donor, acceptor, rings, rotatable bonds, and atoms) was computed for each molecule using cxcalc (commercially available from ChemAxon), although other commercially available systems could also be used. A linear summation of each individual score yielded the total score. Thus, if differences in fragment

20 frequencies and fragment connections are considered:

$$L = \alpha_1 L_1 + \alpha_2 L_2$$

(20).

Taking into account the coupled 1D topology metrics as well:

$$L = \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 D_{P(don,acc)} + \alpha_4 D_{P(atoms,rbonds)} + \alpha_5 D_{P(atoms,rings)}$$

25 (21).

The coefficients ($\alpha_1 - \alpha_5$) were chosen in order to yield the best separation, without over-fitting to the training set. This is achieved by optimizing their values on a validation set, assembled by randomly selecting ~10% of the training set, and then

- 39 -

applying these optimized values to the test set. These values were varied in 0.05

increments with the constraint that $\sum_{i=1}^5 \alpha_i = 1$.

EXAMPLE 3

To develop an algorithm that generates novel small molecules that are similar to
5 known compounds, a Markov Chain approach was used in this example with branching,
treating each growth fragment as the current state, and selecting subsequent fragments
based on transition probabilities. These probabilities for a diverse set of fragments
(Table 1) were initially trained on the ChemBank Bioactives. Using such an approach,
10,000 molecules were grown, and their connectivity statistics were compared (the
10 probability that a given fragment i is connected to fragment j) to those of the ChemBank
Bioactives Database. The excellent agreement observed ($R^2=0.90$, Fig. 14, showing a
correlation of fragment connectivity propensities between molecules grown with a
Markov Chain and ChemBank Bioactives used in training Markov Chain), suggested that
the molecules that were grown were to some degree similar to the database that the
15 algorithm had been trained on. The same agreement was observed when a larger training
database (NCI Open Database, $R^2=0.92$) was employed. These correlations were not
observed for unbiased growth.

The transition probability matrix (Fig. 15) that was obtained was in good
agreement with chemical intuition. Transitions from ring fragments to other ring
20 fragments were low, as might be expected since these connections are often difficult to
synthesize. The matrix was also quite sparse, revealing that many fragments are never
connected in the training database. This may have helped focus combinatorial growth.
It is also apparent that transitions to specific fragments were especially high, most
notably the methyl and benzene fragments. Since sp^3 carbons often serve as part of the
25 framework of organic compounds, it is not surprising that the methyl group is prominent.
Benzene chemistry is very mature, and facile substitutions and transformations of
appendages allows for diverse groups being connected to benzene.

In particular, in Fig. 15, the transition probability matrix used in Markov Chain
growth is shown. High probability transitions are depicted as black, while low
30 probability transitions are clear. Certain transitions are highlighted as indicated (i.e., ring
-> ring, ring -> non-ring, non-ring -> non-ring, and non-ring -> ring). Transitions from
methyl to other fragments (bottom row) and from other fragments to methyl (first

- 40 -

column) are not highlighted. Asterisks are used to denote the columns representing transitions to methyl (left *) and benzene (right *).

To further investigate the behavior of this algorithm, the number of fragments added and the branching probability influence of similarly grown molecules to the training database were studied. Molecules of various sizes (1-11 fragments) were used and grown with different branching probabilities ($P(B) = 0.0-1.0$) as substructure search strings on the original training database (Fig. 16). This figure shows the probability of a grown molecule to be a substructure hit of a compound in the training database. Molecules are either grown with no statistical bias (No Bias), with a Markov Chain (MC), or with a Markov Chain employ a ring/non-ring transition probability as well as a disallowed 3-mer screen (MC+).

The branching probability did not have a significant effect on the percentage of substructure hits (Fig. 17). As the number of fragments increased, on the other hand, the number of hits fell quite rapidly (Fig. 16, MC). It was reassuring that the number of substructure hits was much higher in the molecules that were grown with a statistical bias compared to the unbiased control (Fig. 16, No Bias). This data suggested that when a few fragments are added with the algorithm, it was likely that they yield a substructure of a molecule in the original database. As more fragments were added, an entirely new molecule was accessed, but it is likely that it is composed of one or more substructures that can be found in the database.

EXAMPLE 4

This example shows a separation algorithm capable of classifying molecules. Initially, an algorithm was used that classified compounds based on statistical biases in the fragments that they were composed of, and how they were connected (Equation 20). Using such an algorithm, separate authentic DOS products could be separated from natural products (Table 2). The growth algorithm was then trained on either DOS or NP compounds, and generated libraries of putatively DOS and NP-like molecules, respectively. The classification algorithm then demonstrated that the grown DOS compounds (100%) and natural product compounds (88%) were indeed classified as the molecules they were trained on.

Table 2 shows that a separation algorithm was assessed using two models. Fragment and fragment connection biases (2D) were used as well as coupled 1D metrics such as hydrogen bond donor/hydrogen bond acceptors in addition to the 2D descriptors.

- 41 -

Test sets were evaluated (DOS vs. NP, drug vs. nondrug), as well as molecules grown with the FOG (fragment-optimized growth) algorithm described above in Example 3 (DOS grown, NP grown).

5

Table 2

Compound Set	Compounds	Accuracy	Method
DOS test	673	79	2D
NP test	230	90	2D
DOS grown	100	100	2D
NP grown	100	88	2D
drugs test	218	89	2D
nondrugs test	110	40	2D
drugs test	218	81	2D + coupled 1D
nondrugs test	110	63	2D + coupled 1D

Authentic drugs from nondrugs were then tested. The initial results for identifying drugs (89%) and nondrugs (40%) were superior to accuracies reported in literature for the same databases. Thus, this method allowed the inspection the
 10 overrepresentation of certain fragments in drugs versus nondrugs (Fig. 18). For example, the top three fragments overrepresented in drugs were methyl, amide, and non-ring tri-substituted sp^3 carbon, while the alkene, non-ring sp^2 oxygens and fused benzene rings (as in naphthalene) were overrepresented in nondrugs.

The next step was to show improved separation accuracy by adding three 1D
 15 coupled topology metric descriptors to the classification algorithm. These metric descriptors scored the differences in joint probabilities for two variables between two databases. They are depicted as 2D matrices in Fig. 19, and investigation of the plots yields valuable information concerning drug-like features. Interestingly the drug-like region (marked "D") of the donor-acceptor plot resides above the nondrug-like (marked
 20 "N") region. From these plots, it can be seen that molecules with ~3-7 more acceptors than donors were scored drug-like (Fig. 19A). It was also observed that the bulk of the nondrug-like region lies within the Lipinski cutoffs (<10 acceptors, <5 donors), while some of the drug-like region lies outside of these cutoffs.

- 42 -

From the atoms-rings plot, highly fused structures (high ring:atom ratio) as well as large molecules without any rings (low ring:atom ratio) lie in the nondrug-like region (Fig. 19B). Drug and nondrug-like regions were also separated for rotatable bonds versus atoms (Fig. 19C). Molecules with ~30-40 atoms that were both highly flexible or
5 extremely rigid tend to be non-drugs, but those with intermediate flexibility tend to be scored as drugs at that size. Molecules with >40 atoms are predominantly scored as drugs, while those with <25 are predominantly scored as nondrugs.

In order to ascertain how enriched in drug-likeness the grown molecules were, as compared to those grown with no bias, the following two step screen was used. First, the
10 separation algorithm was used to classify molecules as "random" or drugs. This was done by first training the classification algorithm with authentic drugs and molecules grown with no bias. Compounds that passed the first screen were then classified as drugs or non-drugs. When 200 compounds grown with no bias were subjected to the first screen, not a single molecule was classified as drug-like. When molecules grown
15 with a statistical bias were classified, however, 84% remained after the first screen, and 75% of the initial 200 remained after both screens.

Two popular ADMET screens were also used to assess the drug-likeness of the grown molecules. To assess the accuracy of the Lipinski and Verber screens, they were first applied to the authentic drugs and nondrugs (Table 3). The majority of drugs passed
20 these screens, but unfortunately so do the majority of nondrugs. It has been demonstrated that the Lipinski rule of 5 is a poor discriminator of drugs versus nondrugs, and this further supports this. Even so, to demonstrate the weaknesses inherent in relying on these screens during *de novo* design, they were applied to the grown molecules. A good amount of compounds grown with a bias pass the Lipinski (55%)
25 and Verber (80%) screens. 80% of the compounds grown with no bias pass either the Lipinski screen or the Verber screen even though none of them passed the two step screen.

Table 3

		Screen		
Compound Set	Compounds	2 steps	Lipinski	Verber
Drugs test	218		85	85

- 43 -

nondrugs test	110		88	94
Drugs grown	100	75	55	80
no bias grown	100	0	80	80

These examples showed a linear scoring algorithm that separates classes of compounds, such as drugs from nondrugs. This algorithm is transparent, and allows the investigation of interesting features that distinguish drugs from nondrugs. For example, the hydrogen bond donor and acceptor plot (Fig. 19) revealed that drugs tended to have ~3-7 more acceptors than donors. This observation lead one to question whether binding sites of proteins have a higher ratio of donors than acceptors, or whether it is because of some other physical or biological reason. Nondrugs, on the other hand tended to have the same number of donors and acceptors, so it does not seem to be a synthetic bias.

Using the separation algorithm, these examples show that the generated molecules did indeed occupy the chemical space that was intended. Generating drug-like libraries is not a trivial task. It was found that with the growth algorithm, 75% of the generated compounds were scored as drugs with the two-step method.

While several embodiments of the present invention have been described and illustrated herein, those of ordinary skill in the art will readily envision a variety of other means and/or structures for performing the functions and/or obtaining the results and/or one or more of the advantages described herein, and each of such variations and/or modifications is deemed to be within the scope of the present invention. More generally, those skilled in the art will readily appreciate that all parameters, dimensions, materials, and configurations described herein are meant to be exemplary and that the actual parameters, dimensions, materials, and/or configurations will depend upon the specific application or applications for which the teachings of the present invention is/are used. Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. It is, therefore, to be understood that the foregoing embodiments are presented by way of example only and that, within the scope of the appended claims and equivalents thereto, the invention may be practiced otherwise than as specifically described and claimed. The present invention is directed to each individual feature, system, article, material, kit, and/or method described herein. In addition, any

- 44 -

combination of two or more such features, systems, articles, materials, kits, and/or methods, if such features, systems, articles, materials, kits, and/or methods are not mutually inconsistent, is included within the scope of the present invention.

5 All definitions, as defined and used herein, should be understood to control over dictionary definitions, definitions in documents incorporated by reference, and/or ordinary meanings of the defined terms.

The indefinite articles “a” and “an,” as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to mean “at least one.”

10 The phrase “and/or,” as used herein in the specification and in the claims, should be understood to mean “either or both” of the elements so conjoined, i.e., elements that are conjunctively present in some cases and disjunctively present in other cases. Multiple elements listed with “and/or” should be construed in the same fashion, i.e., “one or more” of the elements so conjoined. Other elements may optionally be present other
15 than the elements specifically identified by the “and/or” clause, whether related or unrelated to those elements specifically identified. Thus, as a non-limiting example, a reference to “A and/or B”, when used in conjunction with open-ended language such as “comprising” can refer, in one embodiment, to A only (optionally including elements other than B); in another embodiment, to B only (optionally including elements other
20 than A); in yet another embodiment, to both A and B (optionally including other elements); etc.

As used herein in the specification and in the claims, “or” should be understood to have the same meaning as “and/or” as defined above. For example, when separating items in a list, “or” or “and/or” shall be interpreted as being inclusive, i.e., the inclusion
25 of at least one, but also including more than one, of a number or list of elements, and, optionally, additional unlisted items. Only terms clearly indicated to the contrary, such as “only one of” or “exactly one of,” or, when used in the claims, “consisting of,” will refer to the inclusion of exactly one element of a number or list of elements. In general, the term “or” as used herein shall only be interpreted as indicating exclusive alternatives
30 (i.e. “one or the other but not both”) when preceded by terms of exclusivity, such as “either,” “one of,” “only one of,” or “exactly one of.” “Consisting essentially of,” when used in the claims, shall have its ordinary meaning as used in the field of patent law.

- 45 -

As used herein in the specification and in the claims, the phrase “at least one,” in reference to a list of one or more elements, should be understood to mean at least one element selected from any one or more of the elements in the list of elements, but not necessarily including at least one of each and every element specifically listed within the list of elements and not excluding any combinations of elements in the list of elements. This definition also allows that elements may optionally be present other than the elements specifically identified within the list of elements to which the phrase “at least one” refers, whether related or unrelated to those elements specifically identified. Thus, as a non-limiting example, “at least one of A and B” (or, equivalently, “at least one of A or B,” or, equivalently “at least one of A and/or B”) can refer, in one embodiment, to at least one, optionally including more than one, A, with no B present (and optionally including elements other than B); in another embodiment, to at least one, optionally including more than one, B, with no A present (and optionally including elements other than A); in yet another embodiment, to at least one, optionally including more than one, A, and at least one, optionally including more than one, B (and optionally including other elements); etc.

It should also be understood that, unless clearly indicated to the contrary, in any methods claimed herein that include more than one step or act, the order of the steps or acts of the method is not necessarily limited to the order in which the steps or acts of the method are recited.

In the claims, as well as in the specification above, all transitional phrases such as “comprising,” “including,” “carrying,” “having,” “containing,” “involving,” “holding,” “composed of,” and the like are to be understood to be open-ended, i.e., to mean including but not limited to. Only the transitional phrases “consisting of” and “consisting essentially of” shall be closed or semi-closed transitional phrases, respectively, as set forth in the United States Patent Office Manual of Patent Examining Procedures, Section 2111.03.

What is claimed is:

- 46 -
CLAIMS

1. A computer-implemented method, comprising an act of operating a computer to:
access data comprising a matrix of probability elements, addressable in at
5 least two dimensions, a first dimension representing potential fragments
contained within the growth compounds and a second dimension representing
potentially addable fragments;
select a fragment contained within the growth compound;
expand the growth compound to form a new growth compound by
10 identifying an addable fragment to be added to the growth compound, the addable
fragment being selected using the matrix comprising probability elements by
identifying the probability elements corresponding to the growth compound,
selecting, using the probability elements, an addable fragment from the
potentially addable fragments, and adding the addable fragment to the growth
15 compound;
optionally, repeat the step of expanding the growth compound one or
more times, using each time the new growth compound as the growth compound;
and
output the new growth compound.
20
2. The method of claim 1, wherein the matrix has two dimensions.
3. The method of claim 1, wherein the growth compound is chosen by a user.
- 25 4. The method of claim 1, wherein the growth compound is chosen by the computer.
5. The method of claim 1, wherein at least some of the addable fragments are
organic.
- 30 6. The method of claim 1, wherein at least some of the addable fragments are
identified using SMARTS.

- 47 -

7. The method of claim 1, wherein the growth compound is selected from a database of potentially addable fragments.
8. The method of claim 1, wherein the act to expand the growth compound
5 comprises selecting, on the growth compound, a growth location from which to add the addable fragment.
9. The method of claim 8, wherein the growth location is selected randomly from among at least some of the hydrogen atoms present within the growth compound.
10
10. The method of claim 8, wherein the growth location is selected on the most-recently added addable fragment within the growth compound.
11. The method of claim 8, wherein the growth location is selected by randomly
15 selecting one of the fragments forming the growth compound, and selecting a location within the selected fragment as the growth location based on a matrix of probabilities
12. The method of claim 8, wherein the act to expand the growth compound
20 comprises determining whether to add a ring or a non-ring structure as an addable fragment to the growth compound.
13. The method of claim 12, wherein the determination of whether to add a ring or a non-ring is determined based on its frequency within the matrix of probability
25 elements.
14. The method of claim 8, wherein the act to expand the growth compound
30 comprises determining whether to add a linear fragment or a branch fragment as an addable fragment to the growth compound.
15. The method of claim 1, comprising growing the growth compound until a predetermined molecular weight is met and/or exceeded.

- 48 -

16. The method of claim 1, comprising growing the growth compound until a predetermined number of fragments has been added to the growth compound.
17. The method of claim 1, comprising growing the growth compound until a
5 predetermined number of rings has been added to the growth compound.
18. The method of claim 1, comprising growing the growth compound until a predetermined number of hydrogen bond donors has been added to the growth compound.
- 10 19. The method of claim 1, comprising growing the growth compound until a predetermined number of hydrogen bond acceptors has been added to the growth compound.
- 15 20. The method of claim 1, comprising growing the growth compound until a predetermined number of electron-donating moieties has been added to the growth compound.
- 20 21. The method of claim 1, comprising growing the growth compound until a predetermined number of electron-withdrawing moieties has been added to the growth compound.
22. The method of claim 1, comprising repeating the step of expanding the growth compound.
- 25 23. The method of claim 1, comprising repeating the step of expanding the growth compound no more than 3 times.
24. The method of claim 1, comprising repeating the step of expanding the growth
30 compound no more than 2 times.
25. The method of claim 1, further comprising, after growing the growth compound, comparing the fragments of the growth compound to a database, and eliminating

- 49 -

the growth compound if at least some of the fragments connected within the growth compound are not connected within the database.

26. The method of claim 1, wherein the act of outputting the new growth compound
5 comprises outputting the new growth compound in SMILES format.
27. The method of claim 1, wherein the act of outputting the new growth compound
comprises outputting the new growth compound in InChI format.
- 10 28. The method of claim 1, wherein the act of outputting the new growth compound
comprises outputting the new growth compound in SLN format.
29. The method of claim 1, wherein act of outputting the new growth compound
comprises displaying the new growth compound on a display device.
15
30. The method of claim 1, wherein the matrix of probability elements is produced
by operating a computer to:
access a database of organic compounds;
identify fragments forming each compound;
20 for each of the fragments, determining the number of other fragments
connected to it within the database; and
create the matrix comprising probability elements for the distribution of
fragment connections in the database.
- 25 31. The method of claim 1, wherein at least some of the fragments comprise more
than one atom.
32. The method of claim 1, further comprising synthesizing the new growth
compound.
30
33. The method of claim 1, further comprising instructing a person to synthesize the
new growth compound.

- 50 -

34. The method of claim 1, further comprising programming a robot to synthesize the new growth compound.
35. The method of claim 1, wherein the data comprising the matrix of probability elements is selected by a user.
36. The method of claim 1, wherein the data comprising the matrix of probability elements is selectable by a user.
37. A computer-implemented method, comprising an act of operating a computer to:
access a database of compounds;
identify fragments forming each compound;
for each of the fragments, determining the number of other fragments connected to it within the database; and
creating a matrix comprising probability elements for the distribution of fragment connections in the database.
38. The method of claim 37, wherein the fragments that are identified within the database are selected by applying a system for determining fragments to the database, and designating as the fragments to be identified as the fragments that appear the most frequently.
39. The method of claim 38, wherein the fragments are selected using SMARTS.
40. The method of claim 37, wherein at least some of the fragments comprise more than one atom.
41. The method of claim 37, further comprising, for each fragment, identifying neighboring fragments that are connected to it, and recording the identity of the neighboring fragments in the matrix.

- 51 -

42. The method of claim 41, further comprising, for each fragment, identifying penultimate fragments that are connected to the neighboring fragments, and recording the identity of the penultimate fragments in the matrix.
- 5 43. The method of claim 37, wherein the database of compounds comprises biologically active molecules.
44. The method of claim 37, wherein the database of compounds comprises catalytically active molecules.
- 10 45. The method of claim 37, wherein the matrix has at least two dimensions.
46. The method of claim 37, wherein the matrix has only two dimensions.
- 15 47. The method of claim 37, further comprising identifying the fragments forming each compound of the database as ring or non-ring fragments.
48. The method of claim 37, further comprising identifying growth locations on the fragments.
- 20 49. The method of claim 37, further comprising outputting at least a portion of the matrix.
50. The method of claim 46, wherein act of outputting at least a portion of the matrix comprises displaying at least a portion of the matrix on a display device.
- 25 51. The method of claim 46, wherein act of outputting at least a portion of the matrix comprises storing the matrix in a file.
- 30 52. The method of claim 37, further comprising:
selecting a growth compound;
expanding the growth compound to form a new growth compound by
identifying an addable fragment to be added to the growth compound and adding

- 52 -

the addable fragment to the growth compound, the addable fragment being selected using the matrix comprising probability elements by identifying the probability elements corresponding to the growth compound, and selecting, using the probability elements, an addable fragment from the potentially addable fragments;

optionally, repeating the step of expanding the growth compound one or more times, using each time the new growth compound as the growth compound; and

outputting the new growth compound

10

53. The method of claim 52, further comprising synthesizing the new growth compound.

15

54. The method of claim 52, further comprising instructing a person to synthesize the new growth compound.

55. The method of claim 52, further comprising programming a robot to synthesize the new growth compound.

20 56. The method of claim 37, wherein the database of compounds is selected by a user.

57. The method of claim 37, wherein the database of compounds is selectable by a user.

25

58. A computer-implemented method, comprising an act of operating a computer to:
access a database of compounds, including a first category of compounds and a second category of compounds;

30

access a plurality of descriptors, wherein at least one of the descriptors comprises determining connectivity of two or more fragments, each comprising two or more atoms, within each compound;

determine a plurality of scores for each compound, each score for each compound corresponding to one of the descriptors; and

- 53 -

select constant weighing factors in a weighted sum of each of the scores of each compound such that at least about 55% of the compounds of the first category exceed a threshold score and/or no more than about 55% of the compounds of the second category do not exceed the threshold score.

5

59. The method of claim 58, wherein the first category of compounds are biologically active compounds and the second category of compounds are not biologically active compounds.

10 60. The method of claim 58, wherein the first category of compounds comprises drug compounds.

15 61. The method of claim 58, wherein the constant weighing factors are selected such that at least about 65% of the compounds of the first category exceed a threshold score and/or no more than about 65% of the compounds of the second category do not exceed the threshold score

20 62. The method of claim 58, wherein the constant weighing factors are selected such that at least about 75% of the compounds of the first category exceed a threshold score and/or no more than about 75% of the compounds of the second category do not exceed the threshold score

63. The method of claim 58, comprising accessing at least two descriptors.

25 64. The method of claim 58, comprising accessing at least three descriptors.

65. The method of claim 58, comprising accessing at least five descriptors.

30 66. The method of claim 58, wherein at least one descriptor comprises determining hydrogen bond donors and/or hydrogen bond acceptors of each compound.

67. The method of claim 58, wherein at least one descriptor comprises determining the number of atoms within each compound.

- 54 -

68. The method of claim 58, wherein at least one descriptor comprises determining the number of rings within each compound.
- 5 69. The method of claim 58, wherein at least one descriptor comprises determining the number of rotatable bonds within each compound.
70. The method of claim 58, further comprising outputting the constant weighing factors.
- 10 71. The method of claim 70, wherein the act of outputting the constant weighing factors comprises displaying the constant weighing factors on a display device.
72. The method of claim 70, wherein the act of outputting the constant weighing factors comprises storing the constant weighing factors in a file.
- 15 73. The method of claim 58, further comprising operating the computer to:
access a representation of a test compound;
determine a plurality of scores for the test compound, each score
20 corresponding to one of the descriptors; and
determine membership of the test compound in the first category of compounds or the second category of compounds by determining a weighted sum using the constant weighing factors applied to the scores of the test compound.
- 25 74. The method of claim 73, further comprising synthesizing the test compound based on the membership of the test compound in the first category of compounds or the second category of compounds.
75. The method of claim 73, further comprising synthesizing the test compound.
- 30 76. The method of claim 73, further comprising instructing a person to synthesize the test compound.

- 55 -

77. A computer-implemented method, comprising an act of operating a computer to:
- access a database of compounds and one or more categories to which each of the compounds of the database belongs;
 - access a plurality of descriptors, wherein at least one of the descriptors comprises determining connectivity of two or more fragments, each comprising two or more atoms, within each compound;
 - determine a plurality of scores for each compound, each score for each compound corresponding to one of the descriptors; and
 - select constant weighing factors in a weighted sum of each of the scores of each compound such that each of the categories can be identified using a distinguishable range of scores such that at least about 55% of the compounds belonging in each category have a score that falls within the distinguishable range of scores for that respective category.
78. The method of claim 77, further comprising operating the computer to:
- access a representation of a test compound;
 - determine a plurality of scores for the test compound, each score corresponding to one of the descriptors; and
 - determine membership of the test compound in one of the categories of compounds by determining a weighted sum using the constant weighing factors applied to the scores of the test compound.
79. The method of claim 77, wherein at least one of the categories of compounds are biologically active compounds.
80. The method of claim 77, wherein at least one of the categories of compounds comprises drug compounds.
81. The method of claim 77, wherein the constant weighing factors are selected such that at least about 65% of the compounds of the first category exceed a threshold score and/or no more than about 65% of the compounds of the second category do not exceed the threshold score

- 56 -

82. The method of claim 77, wherein the constant weighing factors are selected such that at least about 75% of the compounds of the first category exceed a threshold score and/or no more than about 75% of the compounds of the second category do not exceed the threshold score
- 5
83. The method of claim 77, further comprising operating the computer to:
- access a representation of a test compound;
 - determine a plurality of scores for the test compound, each score corresponding to one of the descriptors; and
- 10
- determine membership of the test compound in the first category of compounds or the second category of compounds by determining a weighted sum using the constant weighing factors applied to the scores of the test compound.
84. The method of claim 83, further comprising synthesizing the test compound
- 15
- based on the membership of the test compound in the one or more categories.
85. The method of claim 83, further comprising synthesizing the test compound.
86. The method of claim 83, further comprising instructing a person to synthesize the
- 20
- test compound.
87. The method of claim 77, wherein the database of compounds is selected by a user.
- 25
88. The method of claim 77, wherein the database of compounds is selectable by a user.
89. A computer-implemented method, comprising an act of operating a computer to:
- access a representation of a test compound;
 - determine a plurality of scores for the test compound, each score
- 30
- corresponding to a descriptor, wherein at least one of the descriptors comprises determining connectivity of two or more fragments, each comprising two or more atoms, within each compound; and
 - determine an overall score for the test compound by calculating a

- 57 -

weighted sum of each of the scores of the test compound using a predetermined set of weighing factors.

- 5 90. The method of claim 89, further comprising synthesizing the test compound based on the overall score.
91. The method of claim 89, further comprising synthesizing the test compound.
- 10 92. The method of claim 89, further comprising instructing a person to synthesize the test compound.
93. The method of claim 89, further comprising outputting the overall score.
- 15 94. The method of claim 93, wherein the act of outputting the overall score comprises displaying the overall score on a display device.
95. The method of claim 93, wherein the act of outputting the overall score comprises storing the overall score in a file.
- 20 96. The method of claim 89, wherein the test compound is input into the computer by a user.
97. The method of claim 89, wherein the test compound is selected by a user.
- 25 98. The method of claim 89, wherein the test compound is selectable by a user.
99. A computer-readable medium having recorded thereon signals defining operations for performing the method of any one of claims 1, 37, 58, 77, or 89, when executed on a processor.
- 30 100. A computer-readable medium having recorded thereon a data structure for use in generating chemical compounds, the structure comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing

- 58 -

potential growth compounds and a second dimension representing potentially addable fragments, the probability elements having values representative of connectivities between the potential growth compounds and the potentially addable fragments.

5

101. A method, comprising:

inputting, into a computer, a data structure comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing potential growth compounds and a second dimension representing potentially addable fragments.

10

102. A method, comprising:

causing a computer to construct a data structure comprising a matrix of probability elements, addressable in at least two dimensions, a first dimension representing potential growth compounds and a second dimension representing potentially addable fragment.

15

1/16

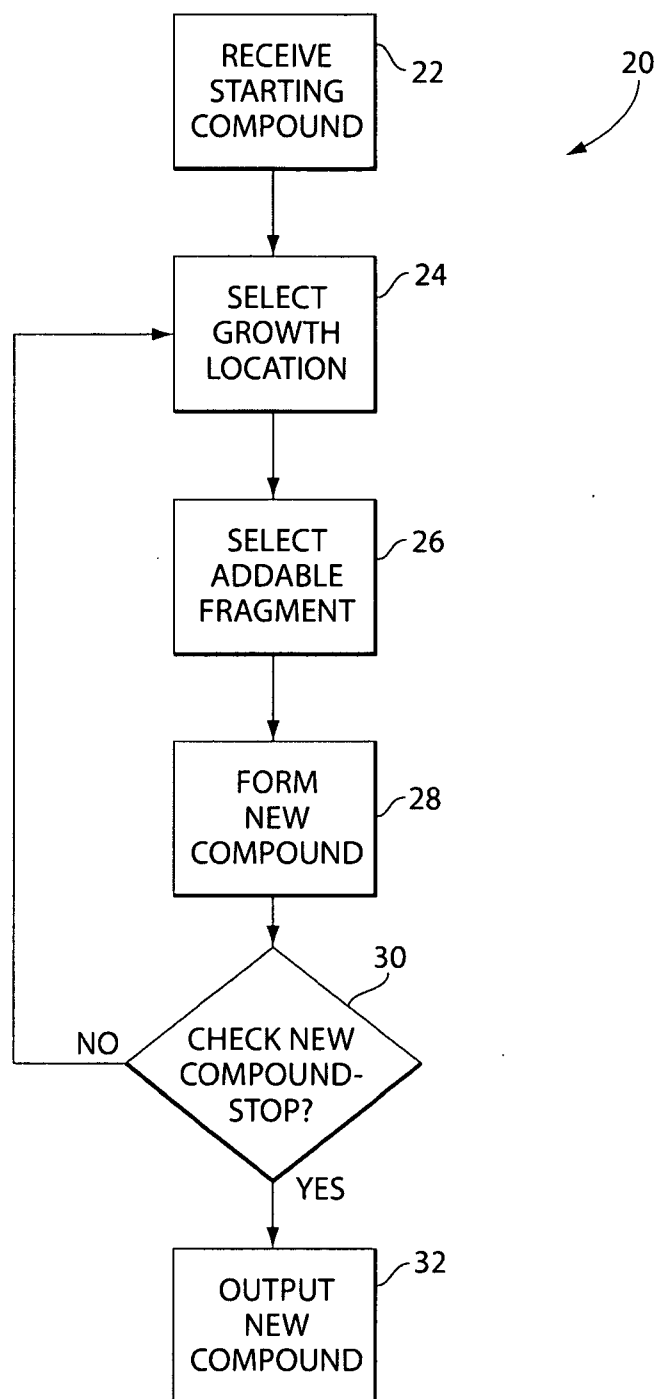


Fig. 1

2/16

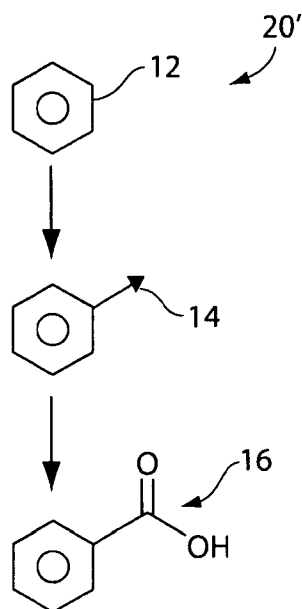


Fig. 2A

28

	-Cl	-Br	-OH	-COOH	-CH ₃	...
	.07	.19	.06	.20	.05	
	.22	.12	0	.24	.02	
	.16	.04	.28	.08	.39	...
	0	.03	.14	.26	.18	
⋮			⋮			⋮

Fig. 2B

3/16

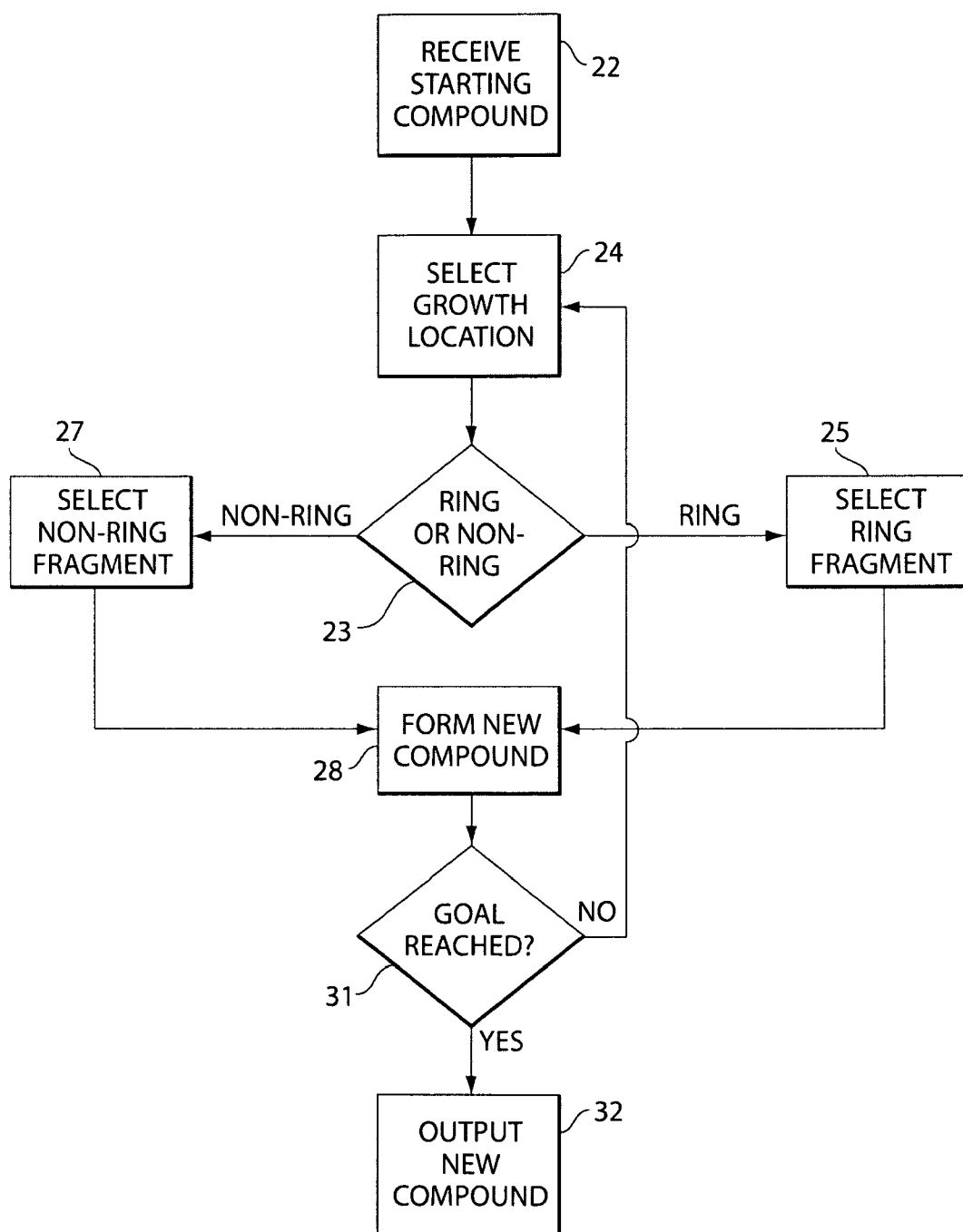


Fig. 3

4/16

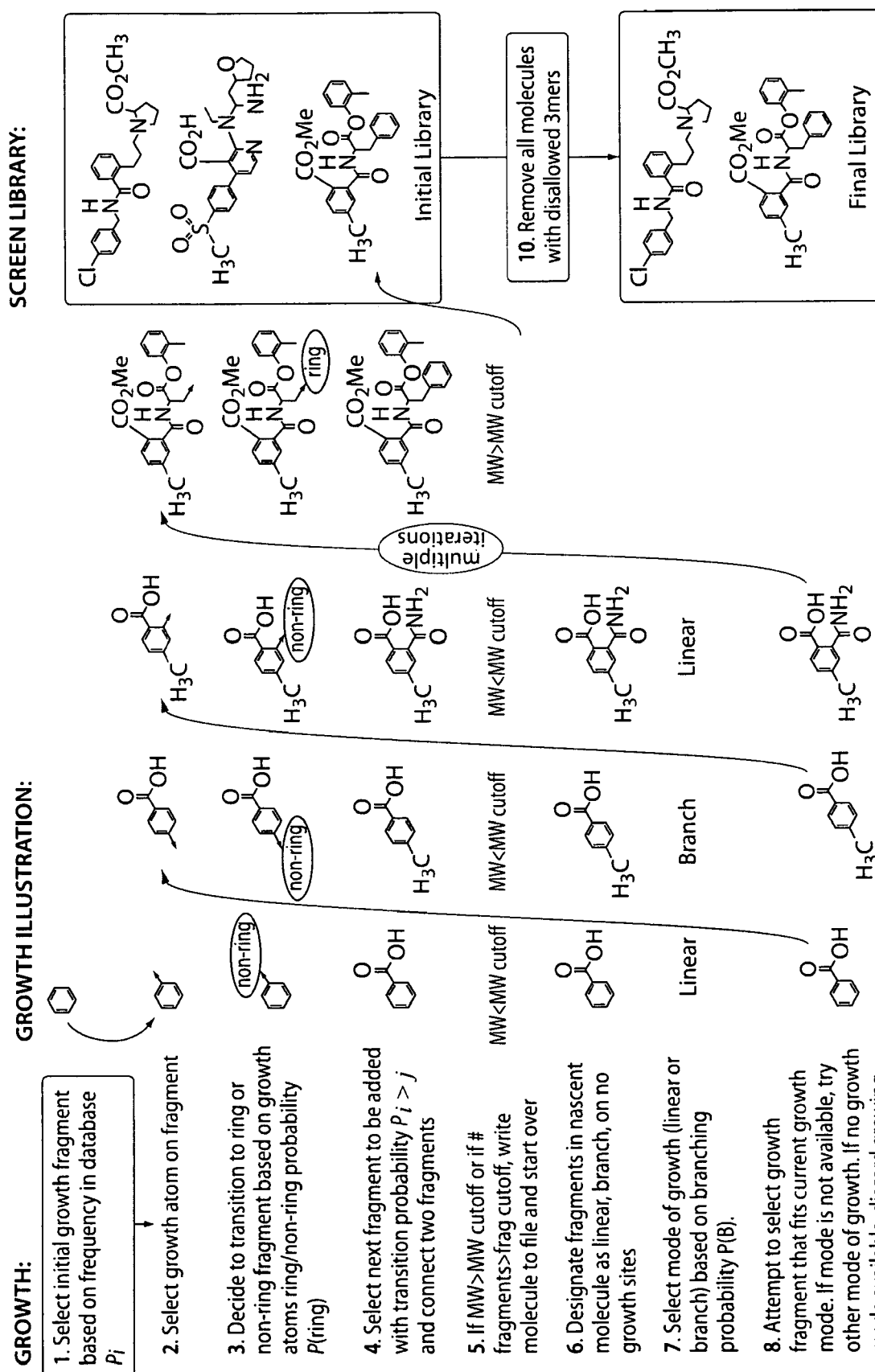


Fig. 4

5/16

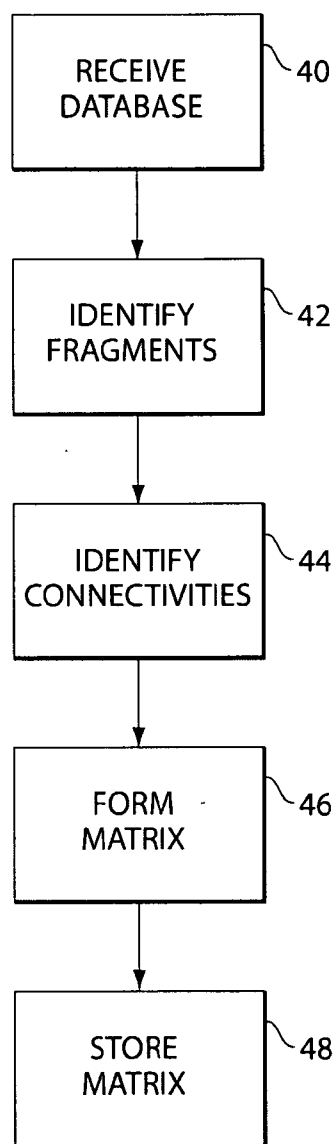


Fig. 5A

6/16

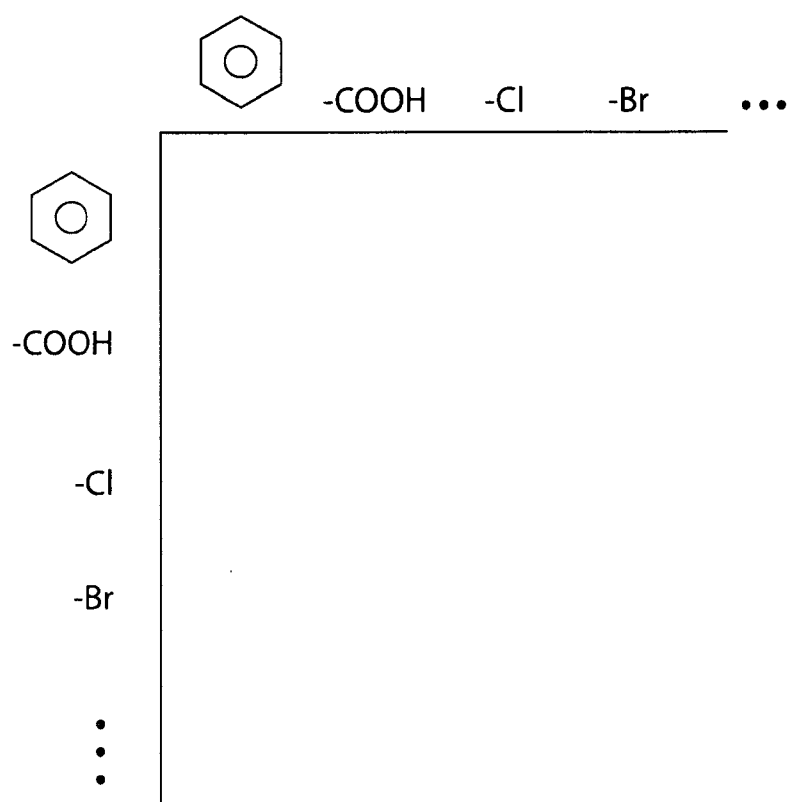


Fig. 5B

7/16

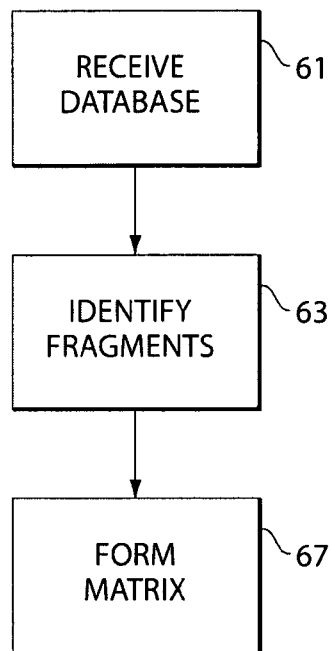


Fig. 6A

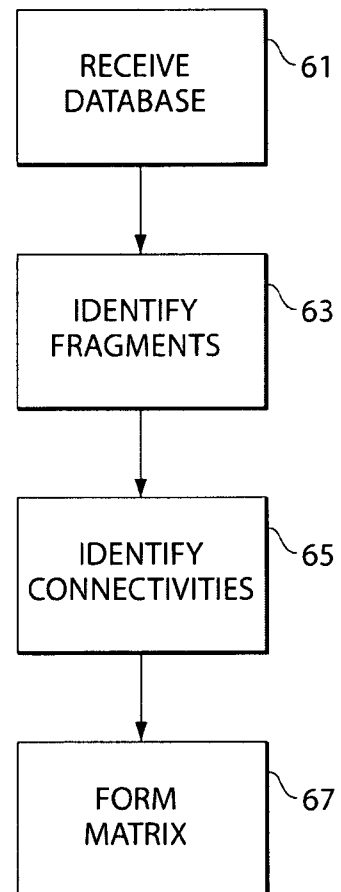


Fig. 6B

8/16

	.38
-Cl	.89
-CH ₃	.13
-CHO	.22
-Br	.04
-COOH	.57
-NH ₂	.14
⋮	⋮

Fig. 7A

	-Cl	-Br	-COOH	-CHO	-CH ₃	-NH ₂	...
	.18	.08	.49	.52	.71	.23	
	.80	0	.24	.16	.52	.03	
CH ₄	0	.77	0	.23	.21	0	...
	.94	.12	0	.33	.48	.24	
⋮		⋮					⋮

Fig. 7B

9/16

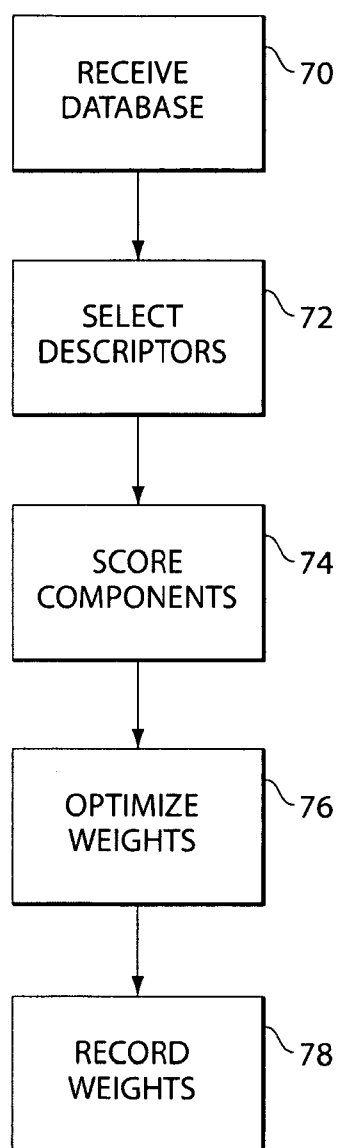


Fig. 8A

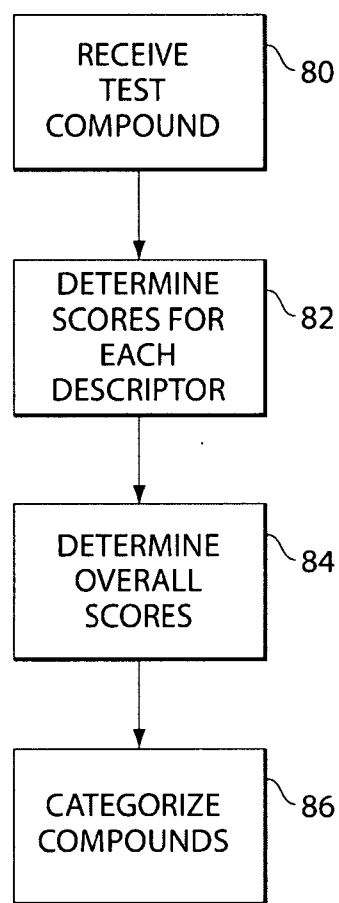


Fig. 8B

10/16

		<i>j</i>		
		A	B	C
<i>i</i>	A	A-A	A-B	A-C
	B	B-A	B-B	B-C
	C	C-A	C-B	C-C

Fig. 9

		<i>j</i>				
		A	B	C	CB	BC
<i>i</i>	A	A-A	A-B	A-C	A-CB	A-BC
	B	B-A	B-B	B-C	B-CB	B-BC
	C	C-A	C-B	C-C	C-CB	C-BC
	BC	BC-A	BC-B	BC-C	BC-CB	BC-BC
	CB	CB-A	CB-B	CB-C	CB-CB	CB-BC

Fig. 10

11/16

Ex.	Search String	Database Molecule	N	Possible Growth	Correction Search String	C	N-C
1			2			1	1
2			2		N/A	0	2

Fig. 11

12/16

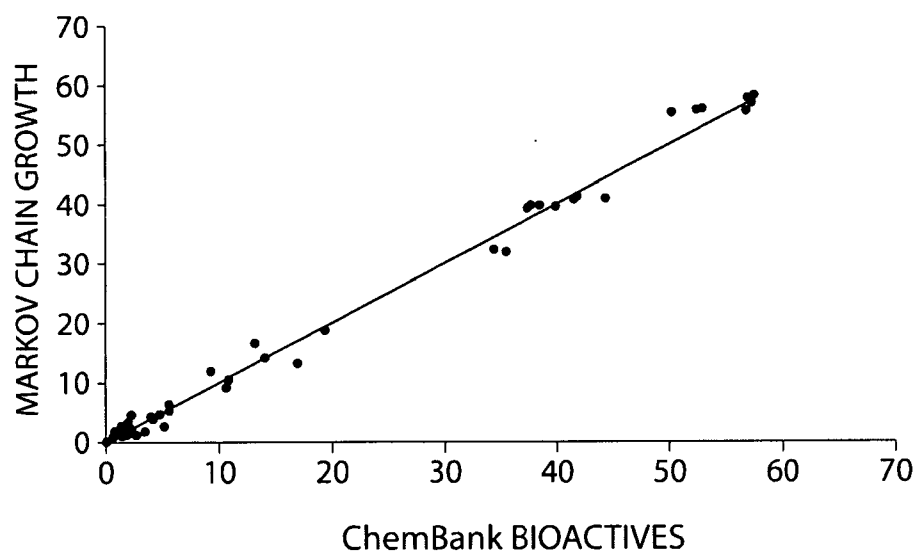


Fig. 12

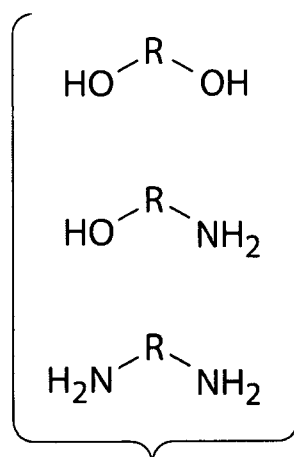


Fig. 13

13/16

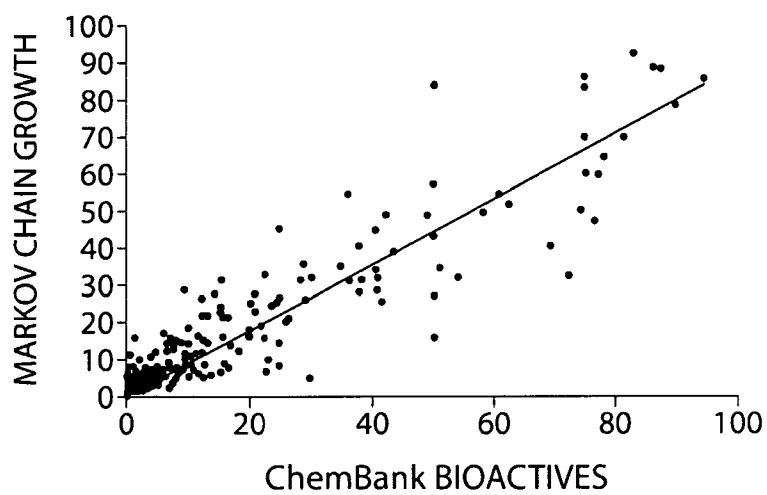


Fig. 14

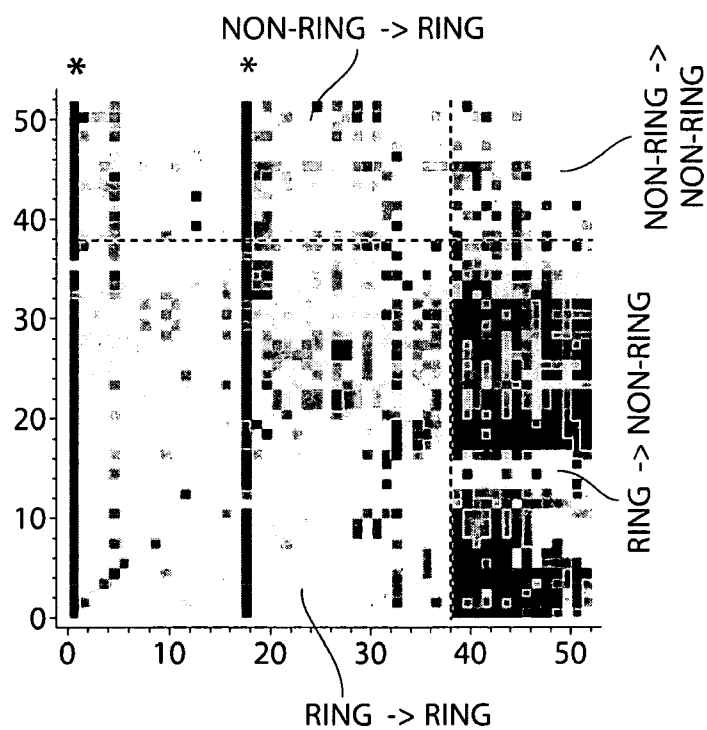


Fig. 15

14/16

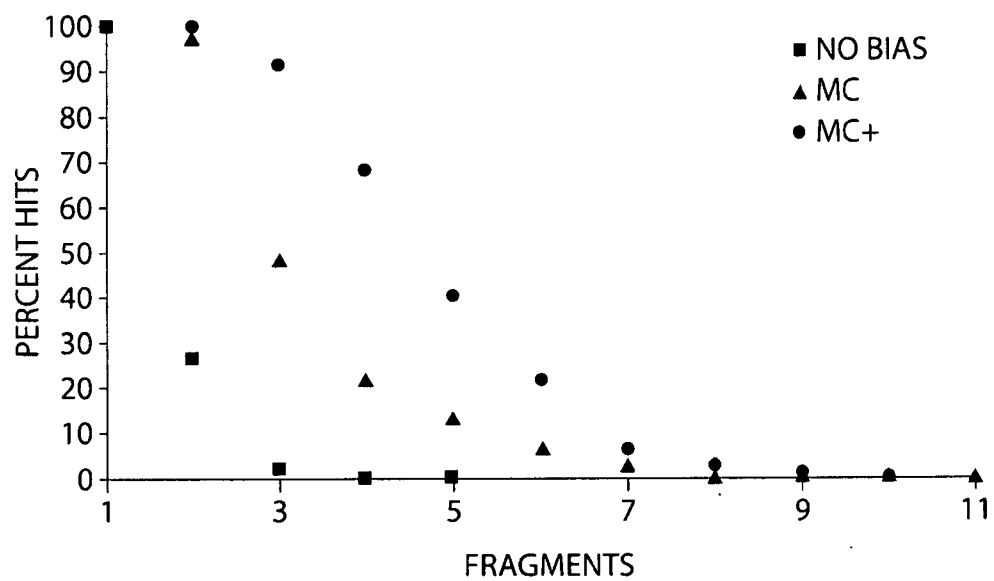


Fig. 16

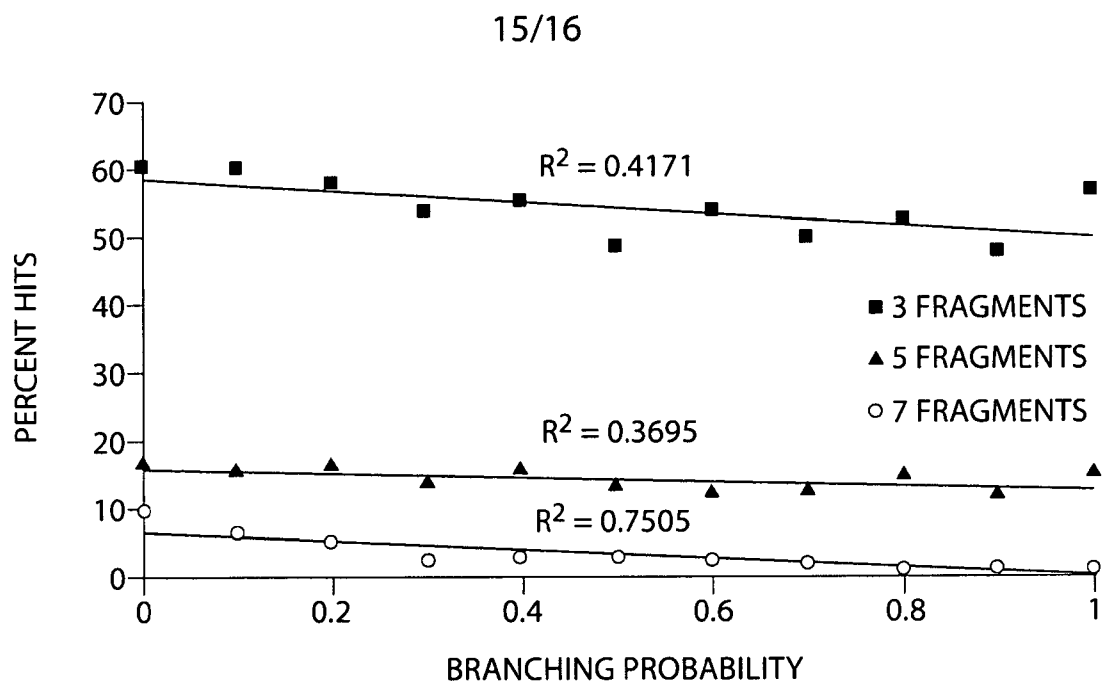


Fig. 17

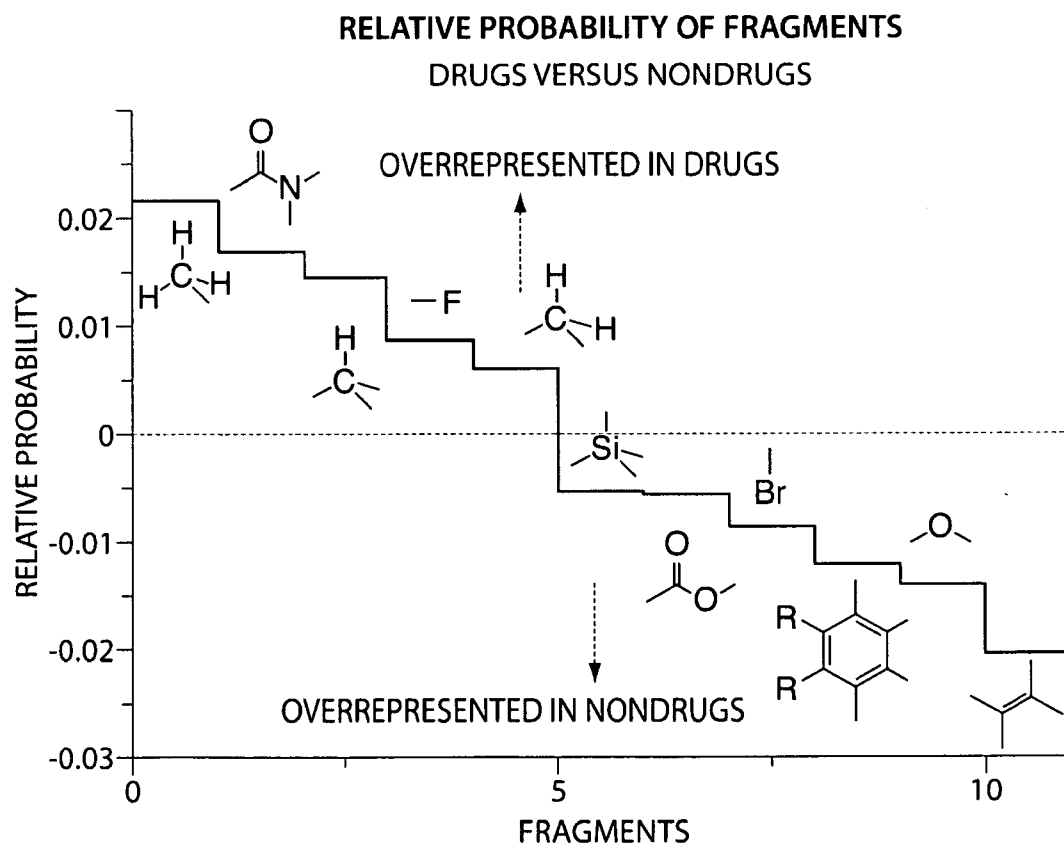


Fig. 18

16/16

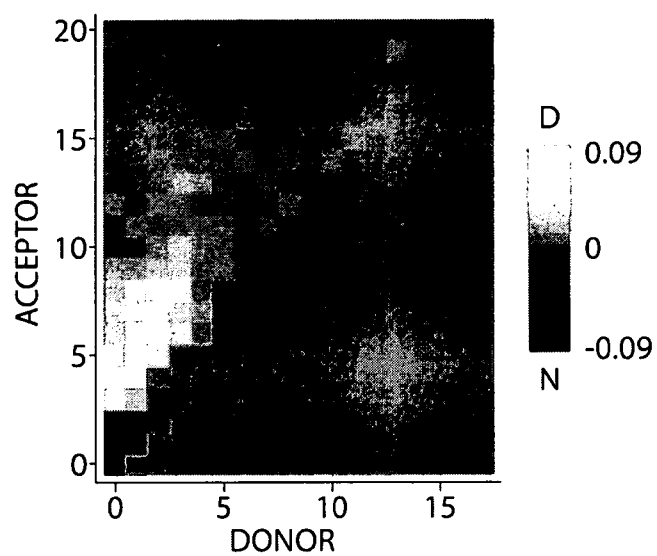


Fig. 19A

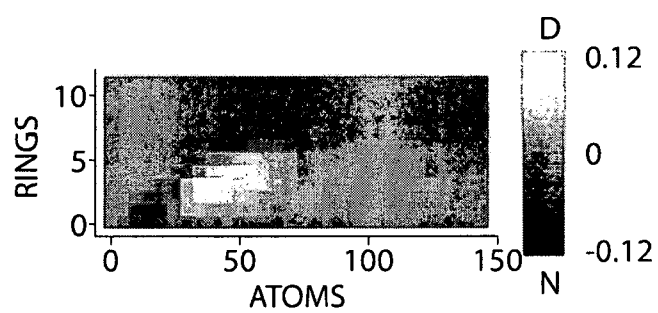


Fig. 19B

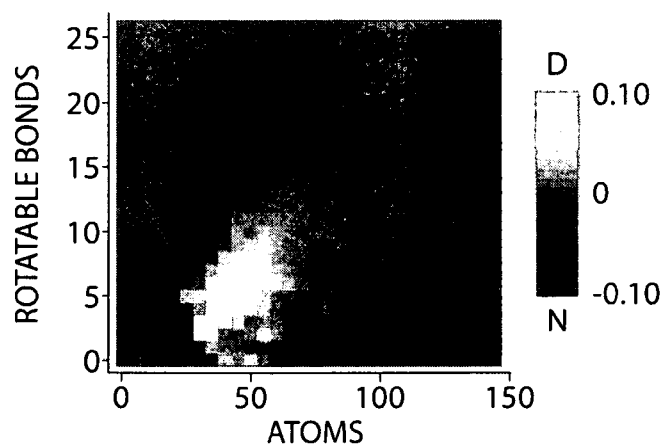


Fig. 19C