CHAPTER FIVE

Combinatorial Library Design, Molecular Similarity, and Diversity Applications

JONATHAN S. MASON
Pfizer Global Research & Development
Sandwich, United Kingdom

STEPHEN D. PICKETT
GlaxoSmithKline Research
Stevenage, United Kingdom

Contents

1 Introduction, 188
1.1 Scope, 188
1.2 Molecular Similarity/Diversity, 188
1.3 Combinatorial Library Design, 190
1.4 Subset Selection and Screening Set Enrichment, 190
2 Molecular Similarity/Diversity, 191
2.1 Descriptors, 191
2.1.1 2D Substructural and Topological Descriptors, 192
2.1.2 Atomic/Molecular Properties and 2D/3D Structural Descriptors, 193
2.1.2.1 Physicochemical, 193
2.1.2.2 2/3D Structural, 193
2.1.3 3D Properties, 194
2.1.3.1 3D Pharmacophores, 194
2.1.3.2 Shape, 199
2.1.3.3 Field-Based, 201
2.1.4 Analysis, 201
2.1.4.1 Descriptor Transformations, 201
2.1.4.2 Similarity and Distance Measures, 201
2.2 Analysis and Selection Methods, 202
2.2.1 Cell-Based Partitioning Methods, 203
2.2.1.1 Diverse Solutions, 203
2.2.1.2 Pharmacophore Fingerprints, 204
2.2.2 Cluster-Based Methods, 206
2.2.3 Dissimilarity-Based Methods, 206
2.2.4 Biasing to Desired/Desirable Properties, 208

Burger's Medicinal Chemistry and Drug Discovery
Sixth Edition, Volume 1: Drug Discovery
Edited by Donald J. Abraham
ISBN 0-471-27090-3 © 2003 John Wiley & Sons, Inc.

2.2.5 Relative Diversity/Similarity, 209
3 Virtual Screening by Molecular Similarity, 209
3.1 Use of Geometric Atom-Pair Descriptors, 210
3.2 Use of 3D Pharmacophore Fingerprints (Three- and Four-Point), 210
3.3 Validation Studies, 210
4 Combinatorial Library Design, 214
4.1 Combinatorial Libraries, 214
4.2 Combinatorial Library Design, 214
4.3 Optimization Approaches, 217
4.4 Handling Large Virtual Libraries, 220
4.5 Library Comparisons, 221
4.6 Pharmacophore-Based Fingerprints, 223
4.7 Combined Pharmacophore Fingerprints and BCUTs, 223
4.8 Oriented-Substituent Pharmacophores, 224
4.9 Integration, 224
4.10 Structure-Based Library Design, 225
5 Example Approaches, 228
5.1 General Target Class-Focused Approaches, 228
5.1.1 Defining the Chemical/Biological Space, 228
5.1.2 7-Transmembrane G-Protein-Coupled Receptors, 229
5.2 Property-Biased Design, 234
5.3 Site-Based Pharmacophores, 235
6 Conclusions and Future Directions, 237

1 INTRODUCTION

1.1 Scope

This chapter discusses molecular similarity and diversity methods and their main applications to combinatorial library design, the selection of compound subsets, and ligand-based virtual screening. Protein structure-based virtual screening is discussed in chapters 6 and 7. Medicinal chemistry-relevant applications discussed include the design of "diverse," "representative," and "thematic/focused/biased" libraries and subsets. The last application is of particular relevance, in that there is a recent trend to approach drug discovery by a "target class" or "gene family" approach; for example, 7-transmembrane G-protein-coupled receptors (7-TM GPCRs); nuclear hormone receptors (NHRs); ion channels; proteases; kinases; phosphodiesterases.

1.2 Molecular Similarity/Diversity

Molecular similarity and diversity methods have been developed based on the principle that similar molecules exhibit similar activities/properties (1). Molecular similarity is a key concept in the identification of new molecules that have similar biological activity to one or more molecules of known activity. Molecular diversity concepts are used to explore "chemical space," with the scope of application ranging from a particular structure/reaction to a large database of different molecules. The process of evaluating similarity and diversity involves the calculation of descriptors for each structure and the determination of the proximity of compounds within the descriptor (or chemical) space. Virtual screening is the name given to the process by which these computational methods are used to identify a subset of compounds from a database for a specific purpose. The source database may, for example, be compounds in a corporate registry where the goal may be to identify compounds for a biological assay. Alternatively, the source database may be compounds that the chemist believes are synthesizable and the goal of virtual screening is to prioritize compounds for synthesis. Depending on the amount of information available to guide the computational screening, and the method used, different levels of enrichment (number of actives selected in a set relative to a random selection) are obtained. It should be noted also that virtual screening applies not only to the selection of compounds for biological screening but also to the prioritization of compounds based on general properties of biological relevance, for example, selecting compounds more likely to be well absorbed.

Molecules are typically represented by a vector of real-valued properties (molecular weight, log P, etc.) or binary values (e.g., 0 for absence, 1 for presence of a substructure feature) in a bit-string or binary fingerprint. The term ***fingerprint*** or ***key*** or ***signature*** thus refers to an encoding of features/characteristics a molecule exhibits (e.g., substructures present, all possible combinations of 2–4 pharmaco-

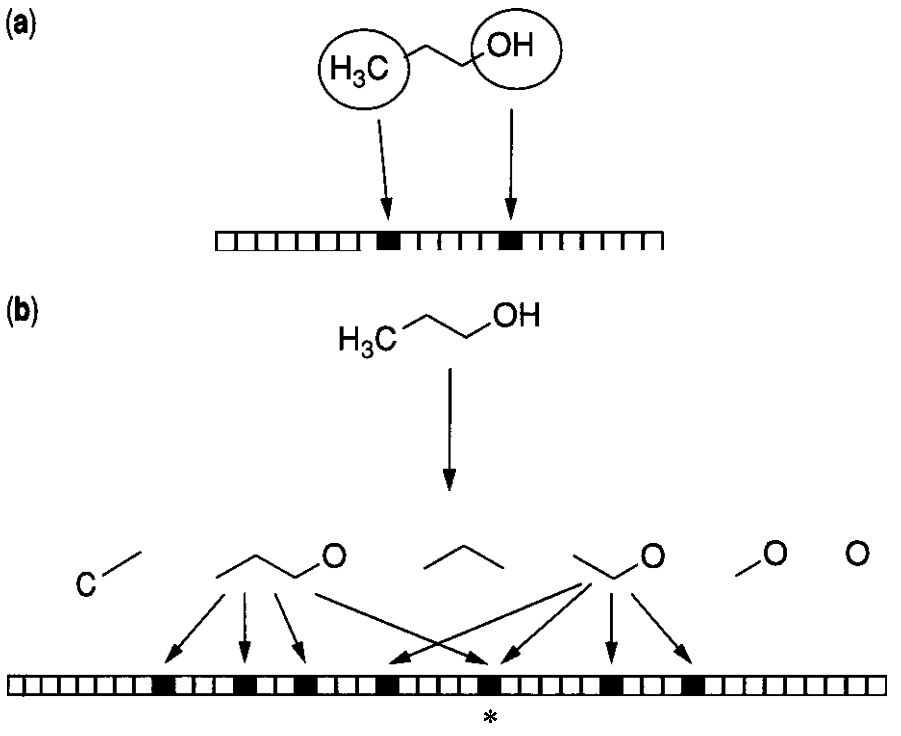


Figure 5.1. A simple illustration of bit-string encoding of chemical structure (7). (a) A fragment dictionary-based approach. (b) Illustration of a hashing scheme using a path-based decomposition of the structure. The asterisk denotes an element in the bit string where a collision has resulted from the hashing procedure.

phoric features) as a string of bits (indicating either the presence or absence of a particular characteristic; see section 2.1.1 and Fig. 5.1), optionally including a count of the number of times the characteristic is exhibited. A wide variety of descriptors is available to evaluate the potential similarity or diversity between structures (2). These range from one-dimensional (1D) descriptors based on molecular properties such as molecular weight, which can be derived from the molecular formula; two-dimensional (2D) substructural fingerprints, topological methods, and atomic/molecular properties [e.g., physicochemical properties such as calculated log P (c log *P*)] that require knowledge of the "flat" or 2D structure, which represents the bonds between the atoms; to three-dimensional (3D) properties (e.g., pharmacophoric fingerprints), requiring knowledge of the full 3D conformational space available to a molecule. A 3D pharmacophoric fingerprint marks the presence or absence of potential pharmacophores [combinations of different features and distances between them, often for three- or four-point pharmacophore fingerprints (i.e., triplets/triangles or quartets/tetrahedra)] within a molecule.

Three-dimensional properties such as the pharmacophore fingerprints can also be calculated for the target protein binding site, being derived from site points complementary to the functional groups in the protein backbone and side chains, thus bridging the ligand-based and protein structure-based universes. The pharmacophore fingerprints also represent a simplified approach to the goal of providing molecular descriptors with 3D shape and property content, while obviating the need for molecular superposition or refined pharmacophore hypothesis generation.

Partitioning methods are widely used. The compounds are grouped using either a cell-based approach, in which each dimension of the chemical space is subdivided or "binned," or by a clustering approach, in which islands of similar compounds are formed. Alternatively, the distance between pairs of molecules can be calculated, and this distance minimized (for similarity) or maximized (for diversity). For diversity the goal is normally not to identify a diverse compound in isolation, but to explore a range of diversity through selection of a diverse subset of compounds. Cell-based methods provide the advantage of a common frame of reference in terms of the multidimensional cell positions. It is possible with a cell-based method to evaluate both what is there and what is missing (in terms of empty cells); clustering, by contrast, is based on exploring what is there. The same method/descriptor may thus be used to evaluate both similarity and "diversity." In practice, "dissimilarity" approaches often provide a more acceptable approach to diversity, ensuring that compounds are not too similar, but avoiding a potential pitfall of exploring too frequently the ex-

tremes of chemical space. Methods and descriptors are discussed for each of these categories.

1.3 Combinatorial Library Design

Combinatorial library design is an important application of molecular similarity and diversity principles and methods. Combinatorial chemistry approaches can exploit automation and robotics to enable the rapid production of large numbers of compounds. Libraries are synthesized for both lead identification and lead optimization purposes. The resultant libraries consist of products formed by combining "reactants" (reagents, monomers) with each other or with a "scaffold" (template, core). The most efficient use of reactants and automation/robotics would use a strictly combinatorial combination of reactants/scaffold, but other constraints, including the issue of generating products that have suitable properties for biological screening and as potential drugs, often lead to sparse arrays. Parallel synthesis, in which multiple analogs are synthesized at a time, is now a standard part of the drug discovery process.

Many molecular diversity and similarity approaches are brought together in the combinatorial library design process. Either the properties of the reactants/scaffolds are used (reactant-based design) or the properties of the resultant enumerated products are used in selecting appropriate reactants (product-based reactant selection). The latter approach requires much greater computational resources, and a preselection of potential reactants may need to be made to control the total size of the "virtual" (potentially synthesizable) library to be analyzed. Regardless of the method, the required deliverable is sets of reactants/scaffolds to be combined. When working with the properties of the products, the constraint that reactants are to be used as efficiently as possible presents a major optimization problem.

Virtual screening, with experimental verification by biological screening, has provided a validation of many of the molecular similarity1 diversity methods used for combinatorial library design, and some ligand-based approaches and examples are discussed in Section 3.

1.4 Subset Selection and Screening Set Enrichment

A related task to combinatorial library design that uses molecular diversity/similarity methods is subset selection of compound screening sets. Initial efforts were focused on small "diverse" or "representative" sets of large corporate compound collections. The increased capabilities of high throughput screening have changed the demand for such sets, and there is a renewed demand for "focused" and "representative" screening subsets of varying sizes; this includes target class (gene family) focus and the identification of "interesting" (e.g., novel) compounds in a large set. Newer biophysical screening methods [e.g., NMR-based screening (3)] still have capacity issues and a need for smaller representative and focused sets. Diverse subset selection can be used to generate sets of compounds to probe a biological assay or to select a subset of reactants to probe the scope of a chemical reaction scheme or screen. However, such methods have a tendency to select compounds at the extremes of chemical space; that is, the selected compounds tend to be less suitable as drug candidates, and hence the approach is less favored for general screening sets. Rather, diversity methods are used to ensure that a random, subset of a screening set contains compounds that are representative of the whole, or, in conjunction with a focused method, to ensure a representative sampling of biologically relevant chemical space.

Compound subsets focused/biased to particular target classes (gene families) have become of greater importance, with application to both lead identification and de-orphaning of new targets from genomics studies. Properties important for the target class of interest are identified, using descriptors used for molecular similarity/diversity. A focused subset can then be selected using a combination of all the possible hypotheses for activity for that target class, including the use of one or more molecular similarity approaches to select compounds similar to any known active compounds. For targets that have structural information available, docking methods (one widely used method for virtual screening) can be used to select compounds that are comple-

mentary to the binding site(s). Applications encompass both high throughput screening (HTS) and therapeutic area screening where only smaller numbers of compounds can be screened. For HTS, smaller thematic studies using these enriched focused sets enable the rapid prosecution of a set of related targets, and make the use of duplicate runs for all compounds feasible. This enables selectivity to be addressed up front, and the duplicate runs provide potentially higher quality information, with the potential for the identification of hits that might otherwise be missed.

General enrichment of the available screening compound set for lead identification is a major application for both combinatorial library design/synthesis and compound acquisition. The goal of ***in silico*** (i.e., computer-based) studies in compound acquisition is to evaluate the interest of compounds that could be purchased to add to the screening file, and to select a subset that meets the same type of physicochemical/"druglikeness" criteria discussed for combinatorial libraries. The "interest" of a compound or compound set is evaluated as in combinatorial library design: diversity relative to existing compound, target, target-class focus, and so forth.

2 MOLECULAR SIMILARITY/DIVERSITY

The field of medicinal chemistry is based on the hypothesis that similar compounds will display similar, but probably not identical, activities in some biological screen, and that potency, selectivity, and properties can thus be modulated by analog synthesis. The challenge facing the computational chemist is how to represent compounds in a computer in such a way that "similar" compounds in the ***in silico*** world are "similar" in the biological world. It is evident that the biological process that is being modeled will influence the nature of the chosen representation. For example, c log P is a useful descriptor for modeling processes involving cell penetration, whereas a pharmacophoric representation would be more appropriate for selecting compounds for screening against a particular protein active site. In this section we review the wide range of representations that have been developed and describe the various methods for applying these representations to real-world problems. The reader is referred to a number of reviews on various aspects covered by this chapter (2, 4–9). A diverse set of perspectives/reminiscences on computational aspects of molecular diversity has been assembled by Martin (10).

2.1 Descriptors

The problem lies in finding a representation of chemical structure that allows a mapping between the chemical structure and its response in a biological or physical process. The representation must be general enough to be applicable to a range of chemical structures but specific enough to capture the differences between structures that account for differences in response. Once found, this representation or set of descriptors can be said to define a ***chemistry space*** (11) for the population of compounds of interest. The similarity between two compounds is their distance within this space. Unfortunately, this simple statement hides a number of difficulties. Many descriptors of choice are correlated and it can be difficult to combine categorical (e.g., acid, base, neutral) and real-valued (charge, dipole, c log *P*) variables. The issue of how to analyze compounds within the chemistry space is covered in Section 2.2.

Methods for describing chemical structures fall into two broad classes. Two-dimensional (2D) methods can be calculated from the 2D graph in which atoms are nodes in the graph and the bonds are the connections between the nodes. Three-dimensional (3D) methods require the generation of a 3D structure (x, y, z coordinates) for a structure. Because a molecule does not exist in a single low energy conformer, the issue of conformer generation also requires addressing with this latter method. Combining the various descriptors, particularly 2D and 3D, is an area of active research.

The potential advantage of 3D descriptors (ligand-protein binding is a 3D spatial/electronic property that can be described only in part using 2D descriptors) (5c) has led many groups to identify 3D descriptors that can handle large numbers of compounds and multiple potential models, and do not require a superimposition in 3D coordinate space (e.g., for re-

view, see Ref. 12). The pharmacophore fingerprints described in Section 2.1.3 are an example of this.

2.1.1 2D Substructural and Topological Descriptors. The principle behind substructural keys or fingerprints is shown in Figure 5.1. A molecule is encoded by the presence or absence of a set of predefined atoms, atom types, and fragments (e.g., S, aromatic nitrogen, CO_2H). The most widely used set of keys is the publicly available ISIS (MACCS) key set provided by MDL (13a). An alternative to the use of predefined fragments is provided by software packages such as Daylight (13b) and UNITY (13c). In this approach, all possible bond paths in a molecule from zero (the atoms) to a specified number of bonds (usually 7) are identified. A hashing procedure is used to store the paths in a bit string of fixed length. Each path will set several bits in the bit string (giving them the value of 1) and there is the possibility of different paths setting some of the same bits. As a result, individual bits lose any meaning.

The origin of the 2D substructural representation lies in the first chemical registration systems where some means was required to enhance the speed of compound retrieval. Thus, if the query molecule contains a particular combination of features, the whole database can be screened very rapidly using the keys to identify compounds that are likely to contain those features before a more exhaustive graph matching is performed to ensure an exact match with the query. The features represented in the keys (ISIS) or the fingerprint length and density (Daylight) were selected to optimize the process of compound retrieval. Nevertheless, they have proved very useful for a variety of similarity-based tasks (14). Despite these successes, issues surrounding their use in diversity-based approaches have been highlighted (15).

Molecular connectivity indices were first proposed by Randić in 1975 (16) as a means of estimating physical properties of alkanes. This formalism was quickly extended to other types of molecules (17) and, since then, a wide range of indices has been proposed, as reviewed by Hall and Kier (18) and Randić (19). The indices are derived from a graph theoretical representation of the structure where bonds are represented by the edges between nodes (atoms). They provide a direct representation of the topological structure of a molecule encoding information such as the degree of branching ($^1\chi$) and the adjacency of the branch points ($^3\chi$), flexibility, and shape (20a). The superscript describes the number of bonds in the path between atoms used to calculate the index. The software package MOLCONN-Z (20b) was developed specifically for generating these descriptors. A number of authors have included topological indices or variants thereof in their description of molecules for describing compound collections (21) or large combinatorial libraries, often allied to a dimensionality-reduction algorithm such as principal components analysis (PCA) (6a, 23).

Cahart et al. (24) introduced the concept of atom-pairs, where the topological distance (number of bonds) between atoms of specified element type are encoded in a bit string. This was extended to the topological torsion (25), where elements on all paths of length four are encoded. Kearsley et al. (26) extended this approach to use more generic atom-type properties in place of element type. They termed these types ***binding property classes*** because they represent key features of intermolecular interactions (positive and negative charge; hydrogen bond donor, hydrogen bond acceptor, and groups that are both of these, such as hydroxyl; hydrophobic atoms; and all others). These descriptors have been used widely for similarity- and diversity-related tasks. The CATS descriptors of Schneider et al. (27a) are a variant on this approach. All topological distances (number of bonds) between a pair of binding property classes (e.g., acid-base) in a molecule are recorded with count information in a correlation vector; that is, how often that topological distance occurs between a specified pair of features in the molecule of interest. Similarity is calculated as the Euclidian distance between the correlation vectors. These CATS descriptors were shown to be useful in scaffold-hopping, identifying actives with a structural type distinct from that of the initial lead structure, and have also been used as the basis for a ***de novo*** design program, TOPAS (27b).

Functional diversity requirements of com-

pound libraries have been reviewed (28), for which molecular descriptors that relate to both structure and properties are needed, as well as their evaluation in terms of biological relevance.

2.1.2 Atomic/Molecular Properties and 2D/3D Structural Descriptors

2.1.2.1 Physicochemical. The descriptors in the previous section focus largely on the structure of the compound. The binding property classes generalize this to some extent by replacing relationships between elements or atom types with a broader definition, still within the framework of an atoms-and-bonds description of the molecule. An alternative approach would be to describe compounds by whole molecule properties, such as molecular weight and log P. Indeed such properties have been related to important pharmacological and physical properties such as absorption across cell membranes, distribution, and solubility. These properties are represented, in part, by the well-known Lipinski Rule-of-5 based on molecular weight, calculated log P, and hydrogen bond donor and acceptor counts (29). Thus, such properties have an important role in drug design, and in general assessments of "druggability." However, their use as descriptors for tasks related to similarity or diversity in the context of receptor affinity is less clear and has been questioned (11b). A primary concern is that such properties do not reflect sufficient information regarding chemical structure to enable their use for lead follow-up or similar purposes. For example, a steroid and a benzodiazepine can have identical log P values but are clearly dissimilar from a medicinal chemistry perspective. Another major problem is that many properties (e.g., log P, molecular weight, surface area, volume, molar refractivity, molecular polarizability) are correlated, making it difficult to find a reasonable set of orthogonal descriptors for the calculation of meaningful distances or for cell-based partitioning (see Section 2.2.1). Such whole molecule properties are best used as constraints on a design, to define boundaries of a pharmacologically relevant chemical space or to define a distribution to match. The challenge is then how to combine the measures of diversity while simultaneously maintaining suitable physicochemical properties. This is addressed in later sections. Such properties can also be used to identify particular combinations that are preferred for different gene families, and these are used to focus a design.

2.1.2.2 2/3D Structural. The issues with whole molecule descriptors mentioned above led Pearlman (11) and colleagues to look at an alternative representation ("BCUT" descriptors/metrics) based on atomic properties and on how atoms are connected. The approach stems from original work of Burden **(30)** to derive a unique signature for a molecule. Pearlman extended the concept to develop the BCUT descriptors suitable for diversity- and similarity-related tasks. Each molecule is described by a series of square matrices with atom labels defining the rows and columns. In a given matrix, the diagonal represents an atomic property such as charge, hydrogen bonding ability (donor/acceptor), or polarizability, with optional weighting by accessible surface area; the off-diagonal terms represent topological or Cartesian interatomic distance or other such property. Molecular descriptors are generated from the lowest and highest eigenvalues of these matrices. and describe the molecular surface distributions of positive or negative charge, H-bond donors, H-bond acceptors, and high or low polarizability.

A number of such matrices can be calculated based on the nature of the diagonal and off-diagonal properties and the scaling between them. An "auto-choose" algorithm [see the DiverseSolutions (DVS) program below] typically finds a 5D or 6D orthogonal chemistry space that best represents the diversity of a given population. This ability to identify relevant (to drug-receptor interactions and reflecting molecular substructure) and orthogonal (noncorrelated) descriptors is critical for the effective use of both distance-based and cell-based methods. Three-dimensional properties may be included by the use of a single conformer to represent atom-atom distances or the inclusion of quantum mechanical properties (bond order or overlap-squared) from semiempirical molecular orbital (MO) calculation. However, the inclusion of 3D/MO information significantly slows down descriptor calculation and does not appear to offer any

practical advantage. DVS, a suite of programs, has been written to calculate and manipulate the BCUT (and other) descriptors for a variety of library design, similarity- and diversity-related tasks (31) (see Section 2.2.1.1). DVS uses the power of a cell-based method as a rapid means to derive a chemistry space relevant to the representation of the diversity of large populations of compounds and methods to pick diverse subsets and compare large data sets rapidly. The BCUTs provide an excellent diversity metric based on electronic properties directly related to ligand-receptor interaction that should also relate to biological activity. Indeed, BCUT metrics appear to reflect pharmacophorically important information, albeit in a relatively crude (low dimensional) fashion. They have proven useful for quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) analyses (32a,b), classification of pharmacologically active compounds (32j), diverse and focused combinatorial library design (11d, 32c-e), rational compound acquisition strategies (11c), and various other diversity-related tasks:

2.1.3 **3D Properties.** The properties and descriptors above are essentially 2D in nature, in that they can be generated from the compound connectivity table, that is, from a knowledge of the bonding pattern within a molecule. There are many advantages to this, not the least of which is the speed of descriptor calculation. Nevertheless, compound interactions with most biological targets are largely 3D in nature. That is, it is the disposition of key functional groups in the molecule in relation to complementary groups within the enzyme or receptor that is important. Thus, there has been much active research into how best to represent the spatial properties of molecules. A particular issue that needs to be handled is conformational flexibility, given that most compounds have rotatable bonds that will change the 3D properties and there is no means ***apriori*** of identifying which particular conformation is the bioactive conformation (i.e., the conformation of the ligand bound to the biomolecular receptor).

The methods presented below tackle this in one of three ways:

1. A single fixed conformer is used.
2. A relatively small number of representative conformers is generated.
3. An exhaustive enumeration of conformers is used.

3D BCUTs are an example of case (1). They reflect confonnational differences, but only to a limited extent because they are inherently low dimensional. This is actually somewhat advantageous because the single low energy conformation from which they are computed may or may not be similar to the bound conformation for a particular receptor. Pearlman (11) has noted that 3D BCUTs appear to be advantageous when the population of interest is a single combinatorial library but that, on average, 2D and 3D BCUTs appear to be equally advantageous when the population of interest is much more diverse. In cases (2) and (3), the descriptors need to be accumulated over all conformers. In the case of bit strings this means "ORing" them over all conformations (combine using logical OR). Herein lies a potential issue with such techniques, in that data from multiple conformations could obscure the signal from the particular bound conformation relevant to a particular target.

2.1.3.1 *3D* ***Pharmacophores.*** The representation of a set of active compounds by a single or small set of pharmacophores that is necessary for that activity was first proposed many years ago and is an excellent model for lead optimization. The development of database systems capable of handling three-dimensional structures in the late 1980s enabled the further exploitation of such methods through giving the ability to search a corporate collection for molecules containing a particular pharmacophore. This approach to lead generation has proved highly successful (e.g., for reviews, see Ref. 33). In particular it is possible to identify active compounds that contain a different core structure from that of the compounds used to generate the model (lead-hopping). This success and the importance of the pharmacophore hypothesis in understanding the interaction of a ligand with a protein target prompted groups to look for ways to use pharmacophores to generate a molecular descriptor for similarity- and diversity-related

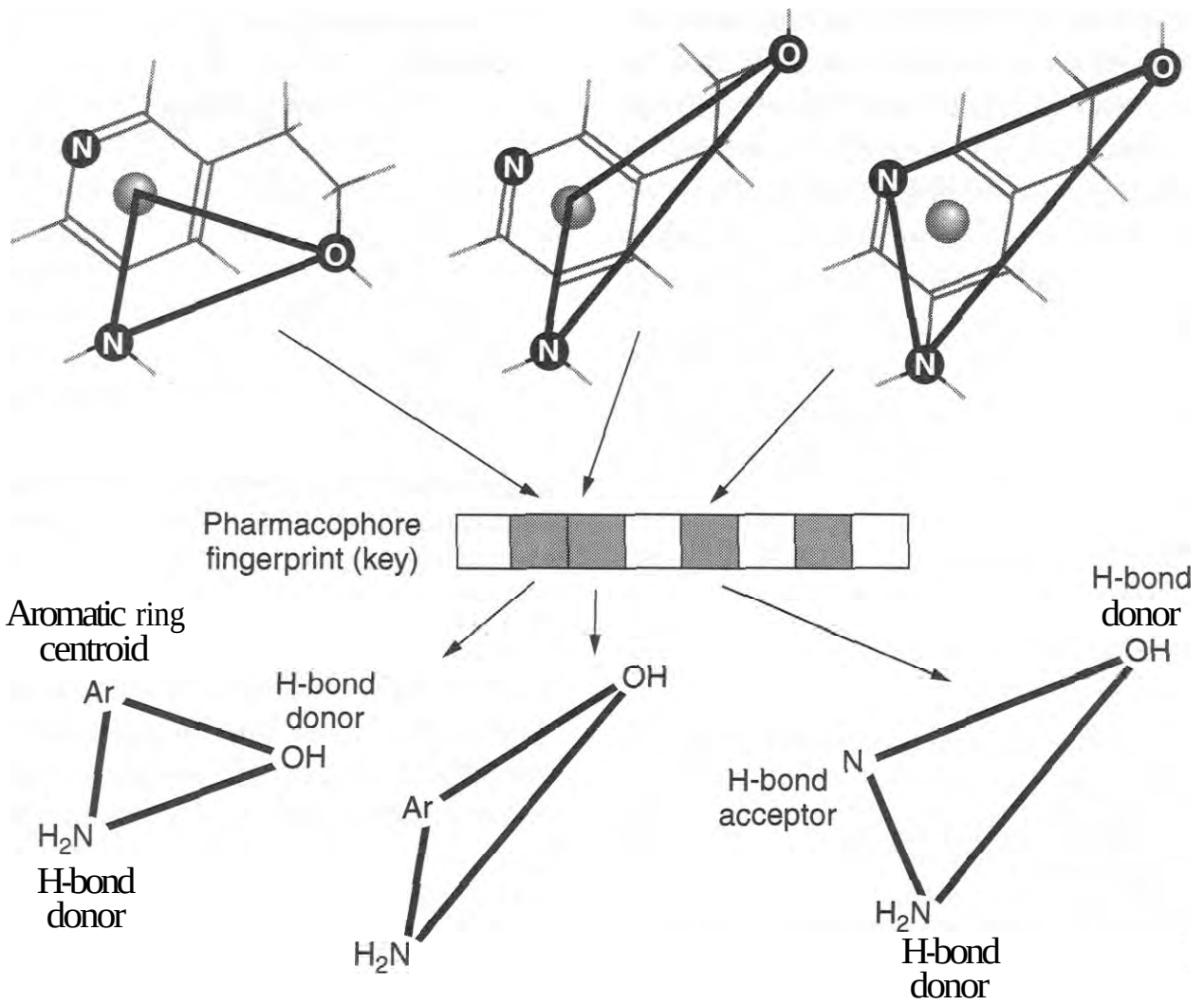


Figure 5.2. Illustration of the creation of a pharmacophore key. As the conformation of a molecule changes, so do the distances between the pharmacophoric groups, shown as spheres. The two different three-point pharmacophores shown each set their own particular bit in the pharmacophore key.

tasks. The diversity-related use was based on the hypothesis that sampling over all potential pharmacophores leads to diversity in a biologically relevant space, in contrast to some other methods that focus on chemical diversity. The descriptor thus generated identifies in a systematic way all the potential pharmacophores that a molecule could exhibit. Triplet (three-point) and quartet (four-point) pharmacophore representations have been extensively used (in addition to two-point/2D approaches), with a variety of features sampled at each point and interfeature distances considered in a discrete set of ranges ("bins") (see Fig. 5.2). The ability of pharmacophores to divorce the three-dimensional structural requirements for biological activity from the two-dimensional chemical makeup of a ligand has been highlighted in a recent review (34).

In an initial implementation from the authors (35), a set of 5916 three-point pharmacophore queries was generated and used to search a database. Compounds were characterized by the pharmacophores that they matched. This method was powerful because it gave precise control over the queries that were generated and ensured that the compounds matched the query, as opposed to satisfying a set of distance constraints; however, it was slow in execution. The Chem-X/Chem-Diverse implementation (36) generates a pharmacophore fingerprint during the course of a single systematic conformational search, with a bump-check and/or rules to eliminate high energy conformers. The details of the conformational search and the definitions of the pharmacophoric features are key components of the system and this methodology has been used extensively for a range of library design and both diversity-and similarity-based tasks (e.g., see Ref. 37). The use of 3D pharmacophores in drug design applications has recently been reviewed (12, 34).

To perform the necessary analyses to generate the pharmacophore fingerprint, relevant features in a molecule need to be identified.

Either substructural definitions to find pharmacophoric features are applied at search time or atom types [and, optionally, additional centroid "dummy" atoms (35)] are used. These can be preassigned (e.g., on database registration) or assigned at search time; a variety of approaches is used (including the use of substructures and connectivity, and of more sophisticated computational approaches). Six properties (features) have commonly been used to describe the potential pharrnacophoric features of a structure:

1. Hydrogen bond donor (e.g., amide NH, aromatic amine, and hydroxyl)
2. Hydrogen bond acceptor (e.g., carbonyl, ether, and hydroxyl)
3. Basic ionizable center (positively charged at physiological pH of about 7) (e.g., aliphatic amines, amidines/guanidines, and 4-amino pyridine)
4. Acidic ionizable center (negatively charged at physiological pH of about 7) (e.g., carboxylic acid, unsubstituted tetrazole, acyl sulfonamide)
5. Aromatic rings (ring centroids often used)
6. Hydrophobic regions (e.g., isopropyl, butyl, cyclopentyl, and certain aromatic rings)

It has also been useful to define a seventh feature type in some situations. For example, it may be beneficial to classify separately the groups that can be both hydrogen bond donor and acceptor such as hydroxyl groups or imidazole nitrogens. Alternatively, the seventh feature provides a mechanism to identify an anchor point to substructures of particular interest (see Section 2.2.5).

All combinations of three or four pharmacophoric points (forming triangles or tetrahedra), for all accessible conformations of a given molecule, can be analyzed, with the resultant descriptor bit-string fingerprint (key) containing the pharmacophores from the whole conformational ensemble of the molecule (see Fig. 5.2). Each bit represents a particular combination of pharmacophore points (Donor-Aromatic-Acceptor, Donor-Aromatic-Basic, etc.) and distances between them (defined using discrete ranges, or bins).

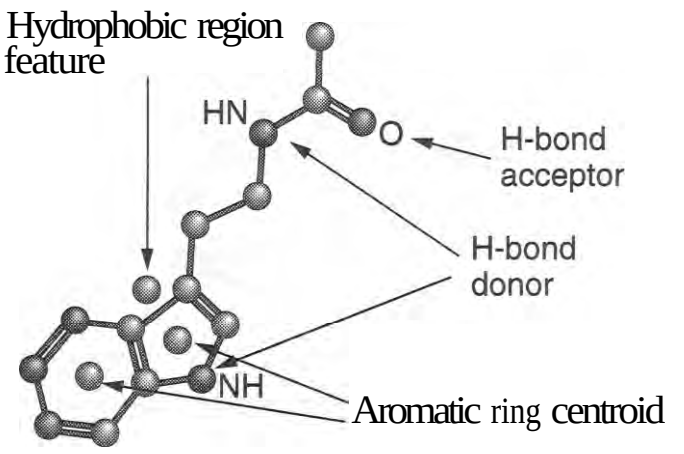


Figure 5.3. Example of how a 3D molecular structure can be broken down into its constituent pharmacophoric elements.

Figure 5.3 illustrates how a molecule can be broken down into pharmacophoric elements. The atom types can be assigned using substructural fragments, taking into account the environment (e.g., a NH group attached to a conjugating group such as C=O is not basic or a H-bond acceptor). Atom types can be automatically assigned when reading a molecule, such as through a customizable substructural fragment database and parameterization file (e.g., Chem-X/ChemDiverse software, 37a). The fragments identify the environment of an atom or group, enabling the correct assignment of a designed feature type. Different options can be set (e.g., a hydroxyl group can be assigned to be both a hydrogen bond donor and acceptor and/or can be assigned to a special feature type for atoms that have both characteristics), and reassignment is possible at search time for structures stored in a database. The identification and representation of hydrophobic regions is one of the most difficult yet critical tasks. Dummy atoms can be used to represent the hydrophobic regions, as a centroid of a group of relevant atoms. This limits the number of hydrophobic features to comparable numbers to other features. An automatic method to add them that uses bond polarities (hydrophobic regions defined for groups of three or more atoms that are not bonded to atoms with a large electronegativity difference) has been implemented in Chem-X/ChemDiverse. Other pharmacophore atom types have also been developed (26).

The extension to four-point pharmacophores enables chirality to be handled and enables some elements of volume/shape linked to

electronic properties to be included. This can give a much better performance in similarity searching. It also increases enormously the number of potential pharmacophores that need to be considered. To analyze pharmacophoric patterns in molecules, the distances between pharmacophoric features are divided into a finite number of ranges using a predefined binning scheme (e.g., 0–2, 2–3, 3–5, 5–8 Å, etc.), up to a maximum distance normally between 15 and 20 Å [a nonuniform binning is often used because this mirrors the tolerances (e.g., ±20%) used in 3D database searching that can be more appropriate than fixed increments, given the limited conformational sampling that is possible]. The additional pharmacophoric combinations created in moving from a three- to four-point description provides additional shape information, thus increasing molecular separation in similarity and diversity studies.

Separation has a central role in determining the final result of such calculations, with too little separation resulting in a noisy descriptor and too many molecules being defined as similar, whereas when too large a separation exists, trivial differences can have a disproportionately negative effect on the similarity value. Conformational sampling is necessary, and the granularity of this affects the useful resolution that can be used, as defined by the number and size of the distance bins. The sampling is generally performed by torsional sampling of rotatable bonds.

Thus fewer ranges are generally considered with four-point pharmacophores while concomitantly maintaining or improving on the performance of three-point pharmacophore methods. For example, by the use of 32 distances for three-point pharmacophores with seven different features possible for each of the points, there are about one million possibilities (35). Expanding to four-point pharmacophores, just 15 distance bins generate about 350 million geometrically valid possibilities. Therefore for pragmatic reasons of both memory/disk space, and the limited resolution of the conformational sampling that is normally applied, seven or 10 distance ranges for four-point pharmacophore fingerprints have been used by Mason et al. (37) and recommended for combinatorial library design and virtual screening applications. Around 2–10 million different potential pharmacophores are resolved in such a fingerprint. A limited sampling of conformations has generally been used to achieve reasonable times (in seconds) for descriptor calculation. For example, Mason et al. (37) use two (conjugated), three (single bonds), or four (sp^2-sp^3 and some conjugated) increments with large data sets, using a systematic analysis for less flexible molecules and random sampling for flexible molecules. See Fig. 5.4 for a comparison of three- and four-point fingerprints. Software companies such as Accelerys, Tripos, the Chemical Computing Group (MOE, http:llwww.chemcomp.com), and Treweren Consultants (THINK) are developing their versions of pharmacophore fingerprinting methods, with three-point pharmacophore fingerprints already implemented. The automatic assignment of pharmacophore features such as hydrophobes, acids and bases, conformational sampling, and other key options discussed above for the Chem-X software (now no longer supported; owned by Accelerys) such as nonuniform binning are challenges that have variable levels of current implementation; other options and extensions such as overlapping bins are becoming available.

Others have developed similar approaches for library design (38, 39). Horvath (40) generates an autocorrelogram of feature-feature distances for conformers and calculates a dissimilarity score that takes into account separate weightings for each feature and allows fuzziness between the distance bins. These 3D pharmacophoric descriptors were termed *fuzzy* bipolar pharmacophore *autocorrelograms* (FBPAs), and the use of fuzzy logic to build up and compare the fingerprints avoids the "all-or-nothing" bitwise match of bitstring representations in which sampling artifacts can cause significant differences. The method has been shown useful in library design and for analyzing selectivity profiles in terms of pharmacophore similarity (41).

It is possible to represent not only a ligand by the potential pharmacophores it possesses but also a protein target. In this case the pharmacophore points are identified by the positions where a ligand atom of a particular type (donor, acceptor, acid, base, hydrophobic, aro-

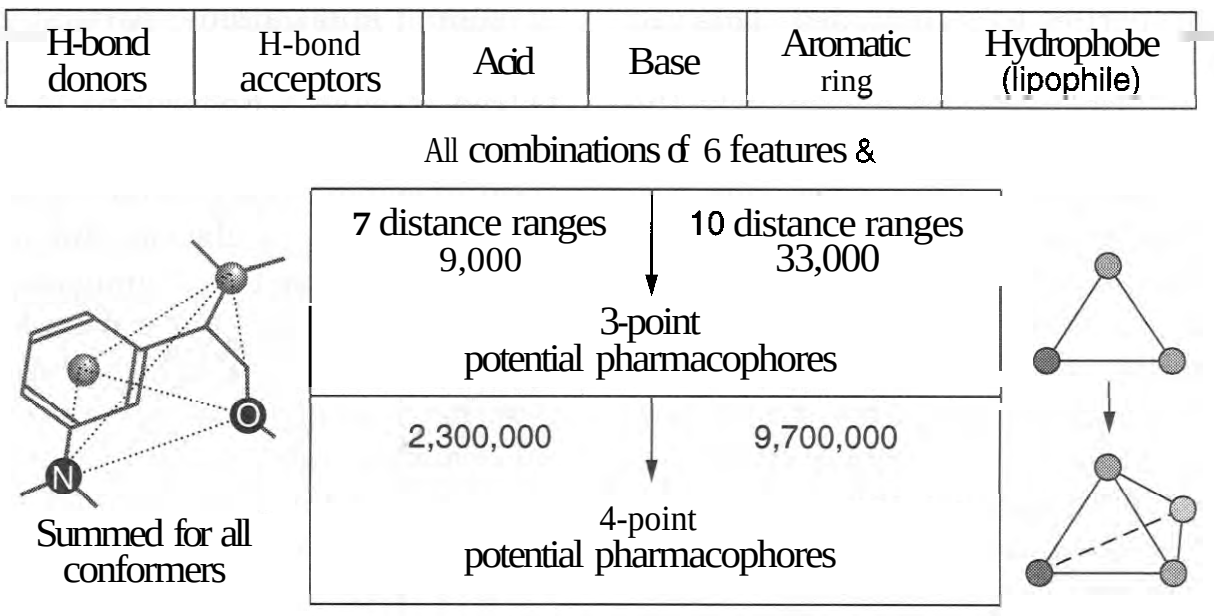


Figure 5.4. Three- and four-point (triplet/quartet) pharmacophore fingerprint creation. Assignment is often binary (on or off), although a count can be kept, and has been used in more recent studies. The large difference in bin numbers between three- and four-point pharmacophores provides additional shape information, thus increasing molecular separation in similarity and diversity studies.

matic centroid) is likely to bind and so provide a complementary interaction with the adjacent protein residue side chain. The pharmacophore fingerprints are thus generated from these complementary site points. The site points can be positioned in the active site using methods such as GRID (42), in which an energetic survey of the site is made using a variety of functional groups. Figure 5.5 illustrates the favorable energy contours for a variety of pharmacophoric probes for the Factor Xa serine protease active site. Atoms (with associated pharmacophore features) are then added in the positions for the most favorable interaction (also shown in Fig. 5.5).

The resultant ensemble of atoms represents a hypothetical molecule that interacts at all favorable positions in the binding site, and

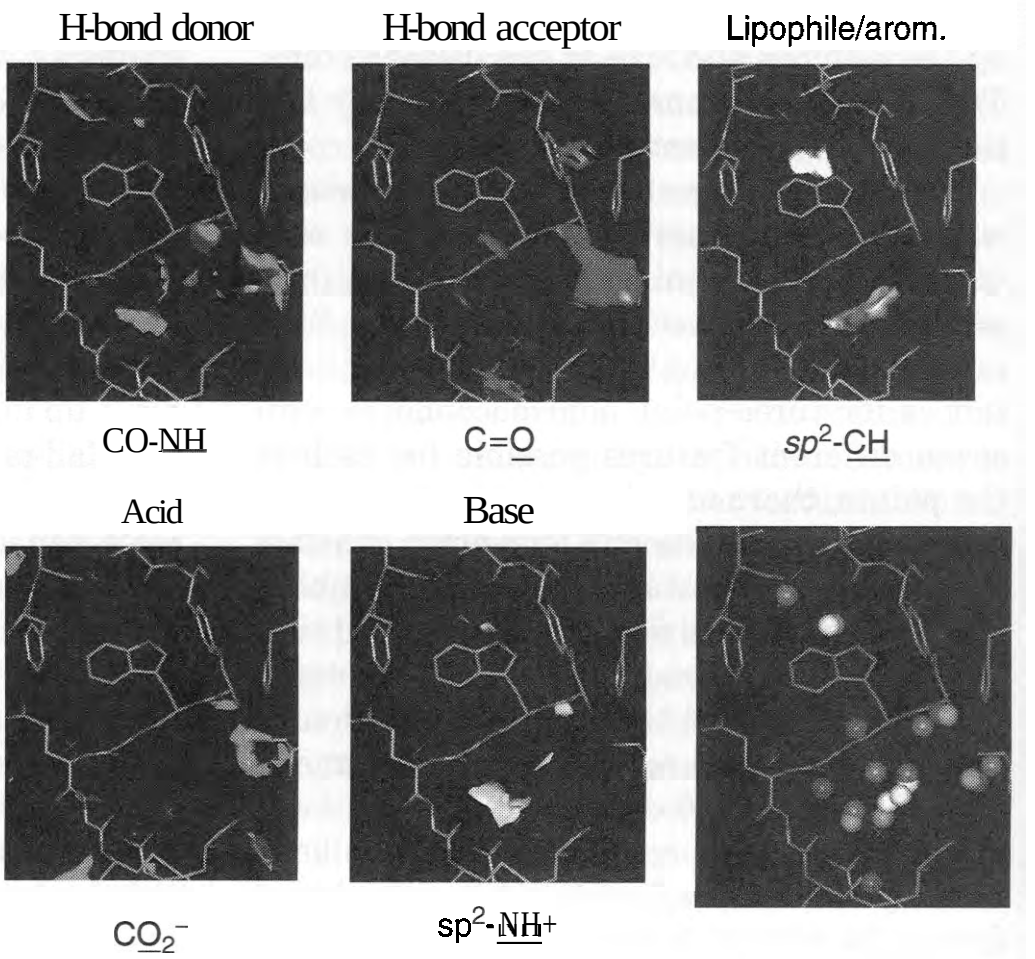


Figure 5.5. GRID probes on Factor Xa site and the combined resultant complementary site points that can be used for pharmacophore fingerprint calculations (lower right). See color insert.

a pharmacophore fingerprint is calculated from this. This fingerprint represents a form of "protein structure-based diversity," quantifying the range of different pharmacophoric shapes complementary to a target protein binding site. For example, for the Factor Xa serine protease active site, 13 complementary site points generated a fingerprint of 2103 four-point pharmacophore shapes, of which 354 were the same as the 2062 found for the serine protease thrombin, generated from 13 site points. Only 11 significant complementary site points were found for the serine protease trypsin, which has a less defined S4 pocket. Of the 1233 total pharmacophore shapes, 363 were in common with Factor Xa, with 120 in common for all three serine proteases. It is thus possible to identify ensembles of pharmacophores that can be used to both differentiate the sites (selectivity) and identify common features. Comparison of these protein-derived pharmacophore fingerprints with known ligands, using four-point fingerprints, shows that they can be used for searching for novel ligands within a database and that they are specific enough to capture ligand selectivity between similar proteins such as the serine proteases thrombin, Factor Xa, and trypsin (37). With three-point fingerprints, the comparison of ligand- and site-derived fingerprints could identify common binding motifs, although selectivity was not captured (37b).

Pharmacophore fingerprints are relatively slow to calculate, however. Thus, their application to very large virtual libraries requires a great deal of computer power. Researchers at Chiron (12, 43) have developed a pharmacophore-based methodology applicable to reactants, OSPPREYS (Oriented-Substituent Pharmacophore *PROPERTY* Space). In this approach, reactant pharmacophores are calculated with respect to the reactant attachment atom and combinations of up to nine pharmacophore centers are considered (see Section 4.8). In the Gridding and Partitioning (GaP) approach, developed at GlaxoWellcome (44), reactants are aligned such that the bond between the attachment atom and the first nonhydrogen atom is along the x-axis with the attachment atom at the origin. A conformational analysis is then performed and the pharmacophore features are mapped to a 1-Å grid (Fig. 5.6). Cells occupied by a particular feature are recorded in a bit string. This descriptor is ideally suited to monomer acquisition and reactant diversity.

Topomer shape similarity, developed by Cramer (45) at Tripos, has been used for similarity searching and targeted library design (using Tripos' proprietary software, "ChemSpace"), building on earlier work on steric fields of single "topomeric" conformers, clustering reactants by their 3D steric fields into "bioisosteric" clusters. The descriptor was considered to be useful in describing variations about a fixed molecular core. defining a single, unambiguous, aligned conformation for any nonchiral molecule.

Approaches such as the GaP program that exploit 3D descriptors for monomer selection address a need for an easily accessible set of in-house monomers available for library generation. Such monomers need to be diverse in nature and able to probe regions of space through attachment to known leads, while producing compounds with druglike properties. More detailed conformational searching paradigms can be used for the smaller monomer compounds, and approaches such as GaP and OSPPREYS exploit this opportunity.

For the selection of diverse compound subsets, studies (46a) have compared three-point pharmacophore descriptors and 2D fingerprints. These have highlighted benefits of the different approaches, and the improved performance of some combined descriptors. The use of clustering for the rational selection of compounds for acquisition and for in-house compound collections used for screening has also been investigated (46b), with comparable results obtained with 3D pharmacophore-derived fingerprints to the typically used 2D fingerprints.

2.1.3.2 ***Shape.*** Pharmacophores capture the key features of intermolecular interactions. However, they do not explicitly capture the shape and volume of the ligand, even if this is crudely implied by the largest four-point pharmacophore exhibited, and the totality of potential pharmacophores exhibited across a range of conformations encodes shape fragments. Hahn (47) has described a method for three-dimensional shape-based searching implemented in the Catalyst program. Seven

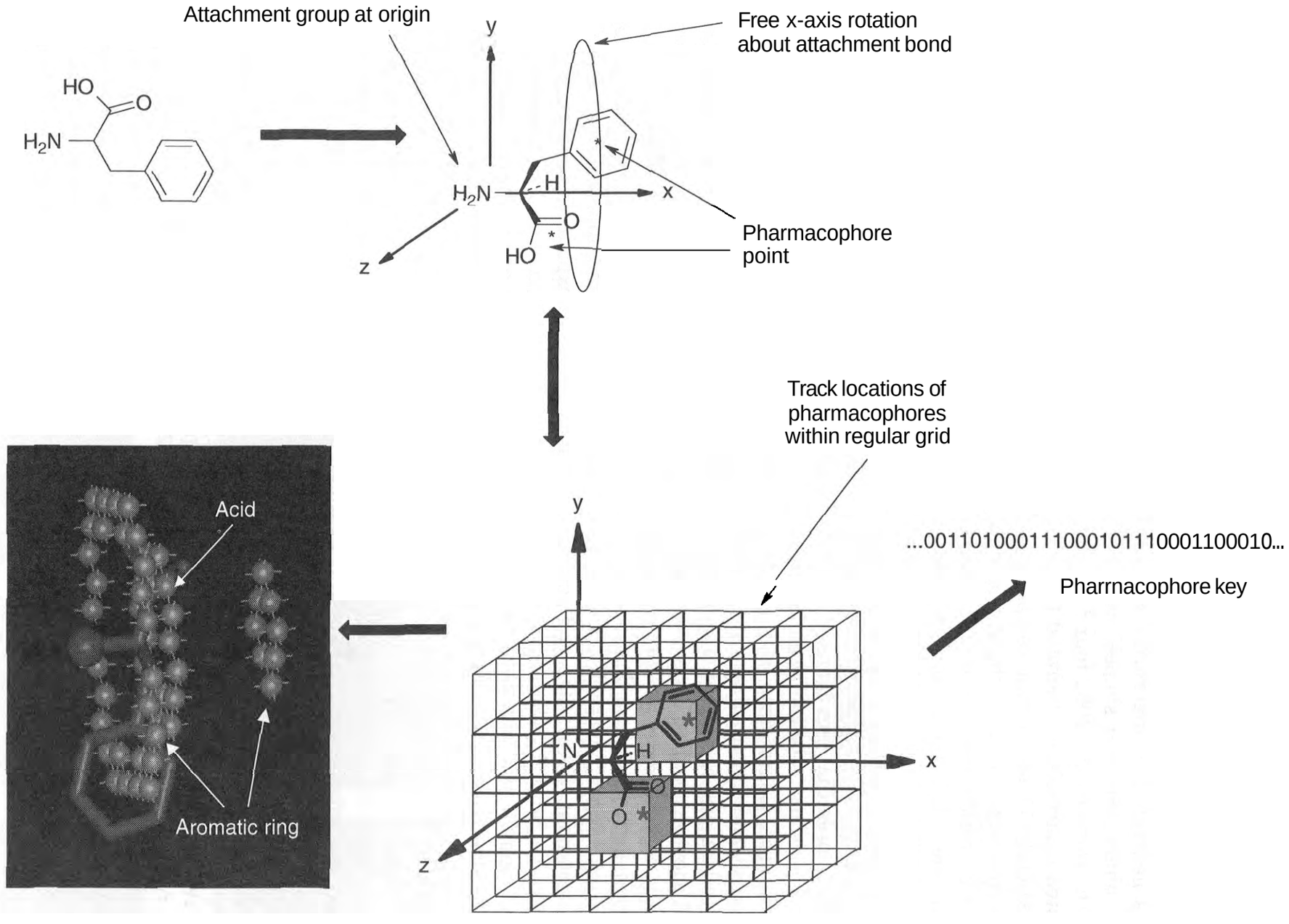


Figure 5.6. Overview of the Gridding and Partitioning (GaP) procedure as applied to monomers, exemplified using phenylalanine as a potential primary amine. This molecule thus contains two pharmacophoric groups (the aromatic ring and the carboxylic acid). During the conformational analysis the locations of these pharmacophoric groups are tracked within a regular grid. See color **insert. [Reproduced from A. R. Leach and M. M. Hann, *Drug Discovery* Today, *5, 326-336* (2000),** with permission of Elsevier Science]

shape indices, positive and negative extents along the three principal axes from the molecular centroid, and the volume of that conformer are computed and stored in a database. These indices can then be used for rapid comparison with a query shape derived from active structures. Conformers passing this filter are then aligned with the query and the similarity is assessed from the volume overlap. Shape-based searching can be used independently, in which case it will complement a 2D similarity search. The method can also be employed in conjunction with a 3D pharmacophore search; however, it is not clear that results are improved in this case (48).

2.1.3.3 Field-Based. A receptor site recognizes the surface properties of a molecule. These can be represented by different types of molecular fields, electrostatic, steric, and hydrophobic, that can be calculated from the atomic composition of the molecule and compared using a measure such as the Carbo index (49). A gaussian representation of the field allows for a more rapid alignment of the molecules (50). Willett's group has developed a program FBSS (51), which uses a genetic algorithm for the alignment of the molecular fields. They have compared the performance of this method with a 2D structural fingerprint (UNITY software, (13c), in searching the WDI, a collection of drug molecules and compounds in development, and the BIOSTER database, a database of functional groups that have been used to replace other groups and retain biological function (e.g., a carboxylic acid and a tetrazole). Although the 2D measure will tend to find more bioactive molecules, the 3D measure gives a greater structural diversity in the hits (52). This seems to be the case for most 3D methods. In these examples conformational flexibility can be considered during the alignment stage but will slow the search down considerably and may also lead to the algorithm becoming stuck in local minima.

An alternative to using the molecule composition in calculating the fields is to use molecular fragments as probes to represent protein side chains. The interaction energy between the probe and the molecule is calculated on a grid surrounding the molecule. These grid fields can then be used in conjunction with PLS as in the CoMFA (comparative molecular field analysis) 3D-QSAR methodology (53). More recently, these fields have been further transformed to generate 3D molecular descriptors. The VolSurf program (54) calculates a wide range of descriptors from the grid energies [calculated with the program GRID (42)]. These have been shown to correlate to a range of properties such as membrane penetration and solubility (55). The Almond program (56) uses a transform known as the Maximum Auto-Cross Correlation (MACC) between pairs of grid nodes, to give a type of two-point pharmacophoric representation of the fields. Such descriptors have been useful in QSAR studies because they are alignment free; that is, they are independent of the position within the defining grid, and have also been used in reactant selection (Pickett, unpublished results, 1999). However, the limitations of the lack of conformational flexibility have so far precluded their use in more general database searching and diversity applications.

2.1.4 Analysis

2.1.4.1 Descriptor Transformations. A large number of potential descriptors are available and this presents a number of issues. Many descriptors will tend to be correlated with one another to a greater or lesser degree. There is the question of the scale of the descriptors and also the difficulty of combining, say, a fingerprint with a calculated property. Thus the descriptors must first be transformed in some way. A key study in this regard was the work of the Chiron group (57). Groups of similar descriptors were combined using principal components analysis (PCA) and multidimensional scaling (MDS), to give a total of 16 composite descriptors. D-optimal design was then used to further analyze a data set. Also of interest was the use of a "flower plot" to visualize the results. In the DPD (diverse-property derived) methodology (21a), the search was for six noncorrelated descriptors. The selection of relevant BCUT descriptors using a χ^2 test is mentioned below.

2.1.4.2 Similarity and Distance Measures. A variety of measures exist for assessing the similarity or distance between molecules in a given descriptor space (2a), as described above. Similarity measures give a direct mea-

sure of similarity between molecules in some property space and give values in the range of 0 to 1, with 1 being identical. Typical examples are the Tanimoto coefficient and the Cosine coefficient. For real-valued properties the Tanimoto is defined as

$$\text{Tanimoto} = \frac{\sum_{i=1}^{i=N} x_{iA}x_{iB}}{\sum_{i=1}^{i=N} (x_{iA})^2 + \sum_{i=1}^{i=N} (x_{iB})^2 - \sum_{i=1}^{i=N} x_{iA}x_{iB}}$$

where x_{iA} is the value of property i of molecule A. When i can take values of only 0 or 1 as in a bit string, then this reduces to

$$\text{Tanimoto} = ab/(a + b - c)$$

where a is the number of on-bits in A and c is the number of bits in common between A and B. The Cosine coefficient can be defined as

$$\text{Cosine} = \frac{\sum_{i=1}^{i=N} x_{iA}x_{iB}}{\sqrt{\sum_{i=1}^{i=N} (x_{iA})^2 \sum_{i=1}^{i=N} (x_{iB})^2}} \Rightarrow \frac{c}{\sqrt{ab}}$$

For field-based measures and overlap of electron density functions then the Carbo index can be used (49), which is equivalent to the Cosine coefficient.

Distance measures give 0 for identical structures and have an upper bound defined by the property space. The Euclidean and Hamming distances are the most common:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{i=N} (x_{iA} - x_{iB})^2} \Rightarrow \sqrt{a + b - 2c}$$

$$\text{Hamming distance} = \sum_{i=1}^{i=N} |x_{iA} - x_{iB}| \Rightarrow a + b - 2c$$

The fundamental difference between similarity and distance measures is that the latter expressly include the absence of a feature (or low values for real-valued properties) in the measure of similarity. This has led to the suggestion (58) that, in the chemical domain at least, such measures are best for relative similarity; that is, ranking the similarity of two molecules to a target, as opposed to measuring the absolute similarity of molecules for which similarity measures, are preferred.

Similarity and distance measures form the basis for most of the analysis and selection methods described in the next section and the reader is referred to the reviews by Willett et al. (2, and references therein) for a fuller discussion of the characteristics and specific properties of these measures.

2.2 Analysis and Selection Methods

In this section we describe some general methods for analyzing and partitioning large data sets, with particular reference to selecting representative or diverse subsets. Library design also employs many of the strategies described here and is discussed in more detail in Section 4. The methods fall into two broad categories: cell-based or partitioning methods and distance-based methods. Partitioning methods use the population to define the limits for cells into which the compounds are divided. Adding or comparing to other compound sets requires identifying the cells into which the new compounds would fall based on their descriptors. This is very rapid and the partitioning process provides a frame of reference for many design tasks; for example, compounds can be readily identified to fill empty or poorly represented cells. Potential issues are where to place the cell boundaries and the handling of compounds that fall near to a cell boundary. Also, new compounds may fall outside the range of properties of the initial population. Distance-based methods, such as clustering and dissimilarity-based methods, require the calculation of similarity between members of the population and are thus population dependent. Adding new members to the population requires recalculating similarities and could change the distribution of compounds between the clusters. Identifying poorly represented or empty areas of property space is not possible. Each of these methods is further described below with examples of their application.

2.2.1 Cell-Based Partitioning Methods. Partitioning methods divide chemistry space into hyperdimensional "cells" by "binning" the axes (descriptors) that define the chemistry vector space, just as the eight divisions on the x- and y-axes of a two-dimensional checker board divide the board into 64 squares. A chemical compound occupies a position in chemistry space determined by the descriptors (coordinates) computed based on its structure. Once the compounds have been partitioned, selecting diverse or representative sets of compounds involves selecting a small number of compounds from each occupied cell, either in proportion to the number of compounds in the cell or a specified number from each occupied cell. For focused sets, compounds are sampled from cells neighboring the population of actives. The real advantage of partitioning methods, however, lies in their ability to readily identify underpopulated regions of property space. Selections can then be made from a second population of molecules—a virtual library for instance—to increase the occupancy of underpopulated cells. Usually, such methods require a low dimensional representation of the space, although the pharmacophore methods are a notable exception to this. The low dimensional space may be the result of a dimensionality-reduction algorithm, as described earlier. Alternatively, a small number of descriptors may be judiciously selected. This latter approach was taken by Lewis et al. in their DPD methodology (21a), which is a good example of partition-based selection. The aim was to select a representative set of compounds based on molecular and physicochemical properties for screening. Six properties were chosen from among 49, based on their low pairwise correlation: number of H-bond acceptors, number of H-bond donors, molecular flexibility, an electrotopological state index, c log P, and a measure of aromatic density. Each descriptor (axis) was divided into two to four partitions, to give a total of 576 bins. A major issue was in identifying six relevant and reasonably noncorrelated (orthogonal) descriptors, leading to the definition of a new descriptor. The chosen ranges covered more than 85% of a 150,000 subset of the corporate collection and approximately three compounds were taken from each bin. Follow-up of initial hits involves the screening of additional compounds from the cells containing hit molecules. Several leads were identified using this approach (7).

2.2.1.1 Diverse Solutions. DiverseSolutions (DVS) is software developed by Pearlman et al. (11, 31) to generate and use the BCUT descriptors in addition to other DVS-computed or user-provided low dimensional descriptors. (DiverseSolutions is also designed to work with high dimensional metrics such as 2D fingerprints, and includes some novel algorithms for such distance-based work.) DVS uses a χ^2-based "auto-choose" algorithm (11c) to identify the combination of low-D descriptors, which are mutually orthogonal and which most uniformly distribute a given large population of compounds among the cells of the resulting chemistry space. Originally, the binning was performed in a uniform manner along each axis, with a given percentage of outliers to avoid sampling the extremes of space. This could be useful for large sets of diverse compounds where the extremes tend to be undesirable compounds. However, for large (virtual) libraries initial filtering can remove these before the analysis, and thus a nonuniform binning scheme was suggested (59), so that acceptable compounds are not lost as outliers, and is now the preferred option. Often, the large population of compounds used as the basis for defining a chemistry space is the entire compound collection available to a pharmaceutical company for its drug discovery efforts, together optionally with structures from commercial databases of biologically active compounds. The resulting chemistry space can be regarded as the "corporate standard chemistry space" and provides an ideal basis for comparing large sets of compounds such as alternative commercially available compound collections or alternative combinatorial libraries. It is also a good basis for comparing small sets of compounds such as compounds with reasonable affinity for various bioreceptors.

The axes of a corporate standard chemistry space are intended to represent all aspects of molecular structure. Thus, all axes of the corporate chemistry space must be considered for purposes such as general diverse subset selection or rational compound acquisition. How-

ever, not all aspects of molecular structure may be important for understanding structure-activity relationships (SARs) for a particular receptor. This led Pearlman and Smith (11d) to introduce the concept of a receptor-relevant subspace (RRSS) of a full chemistry space. For example, starting with a chemistry space of six dimensions, defined to best represent the diversity of all druglike compounds in the MDDR (MDL Drug Data Report) database (13a), they showed how to perceive the three-dimensional subspace that conveys information that is particularly relevant for affinity to the ACE (angiotension converting enzyme) receptor. ACE inhibitors of diverse structure were tightly clustered with respect to the receptor-relevant metrics, thereby providing an obvious near-neighbor strategy for lead follow-up. They (11d) also emphasized the importance of not considering metrics that are not "receptor-relevant" when computing distances for such near-neighbor-based discovery efforts. This also enables diversity in these other dimensions to be explored (e.g., with combinatorial libraries), to obtain compounds with a modified profile for other properties such as bioavailability.

Work on the design and diversity analysis of large combinatorial libraries at Pharmacopeia using BCUT metrics and DiverseSolutions was reported by Schnur (32). A cell-based analysis of synthon-derived libraries was performed, using full product libraries, including library comparisons. Active molecules in these libraries, which involved multiple scaffolds, were found to cluster in various three-dimensional subspaces of the diversity spaces. The utility of a simple property-based reactant/synthon selection tool was also described, targeted at the synthetic chemists, with reactants binned according to patterns based on the ranges of a set of user-selected properties that form a diversity hypothesis.

Chemistry space metrics have been used at Rhône-Poulenc Rorer for diversity analysis, library design, and compound selection (59, 80) using DiverseSolutions to generate a "universal" chemistry space for use as a standard for profiling structural sets of interest. The complementarity of three different diversity measures for comparing and profiling compound collections (a corporate database, combinatorial libraries, and the MDDR drugs database) was also shown. The methods used were a 2D structural characterization (Daylight fingerprints), DiverseSolutions, and 3D pharmacophore fingerprints. A combinatorial library of 100,000 structures appeared structurally different from the other databases by the Daylight fingerprint clustering, yet the bulk of its compounds overlapped with druglike compounds (MDDR) in DiverseSolutions BCUT chemistry space and 3D pharmacophore space ("cells" in fingerprints). It was shown and "quantified" that new diversity relative to the company database was explored, with much of this new diversity in desirable areas occupied by MDDR compounds. The nonuniform binning scheme was developed to enable the use of chemistry spaces scaled to include all structures within a set, while maintaining a reasonable distribution of compounds within cells. The method was used to select a subset for initial screening of a large set of combinatorial libraries designed for 7-TM GPCR targets.

2.2.1.2 Pharmacophore Fingerprints. Pharmacophore fingerprints can also be considered as a high dimensional partitioning of the compound space (35). Underrepresented pharmacophores within a population can be identified and act as a possible focus for library design or compound acquisition. Using six feature types (hydrogen bond acceptor, donor, acid, base, hydrophobe, and aromatic ring centroid) with four-point pharmacophores and 7–10 binned distance ranges, it is possible to resolve about 2–10 million different phannacophoric shapes. Different databases can be compared using this fingerprint, and differences identified. For example, by comparing a corporate screening file (100,000 structures) with the MDDR database (62,000 structures) of biologically active compounds (as discussed above for DiverseSolutions, Refs. 62, 80) "holes" could be identified, in terms of about 1 million 3D pharmacophores exhibited only by MDDR compounds (about 2.7 million were in common and 0.2 million unique to the corporate set). This provides a design space for which combinatorial libraries were designed and synthesized. A total of 100,000 combinatorial library compounds were able to match about 40% (0.4 million) of the pharmacophore "holes" (i.e.,

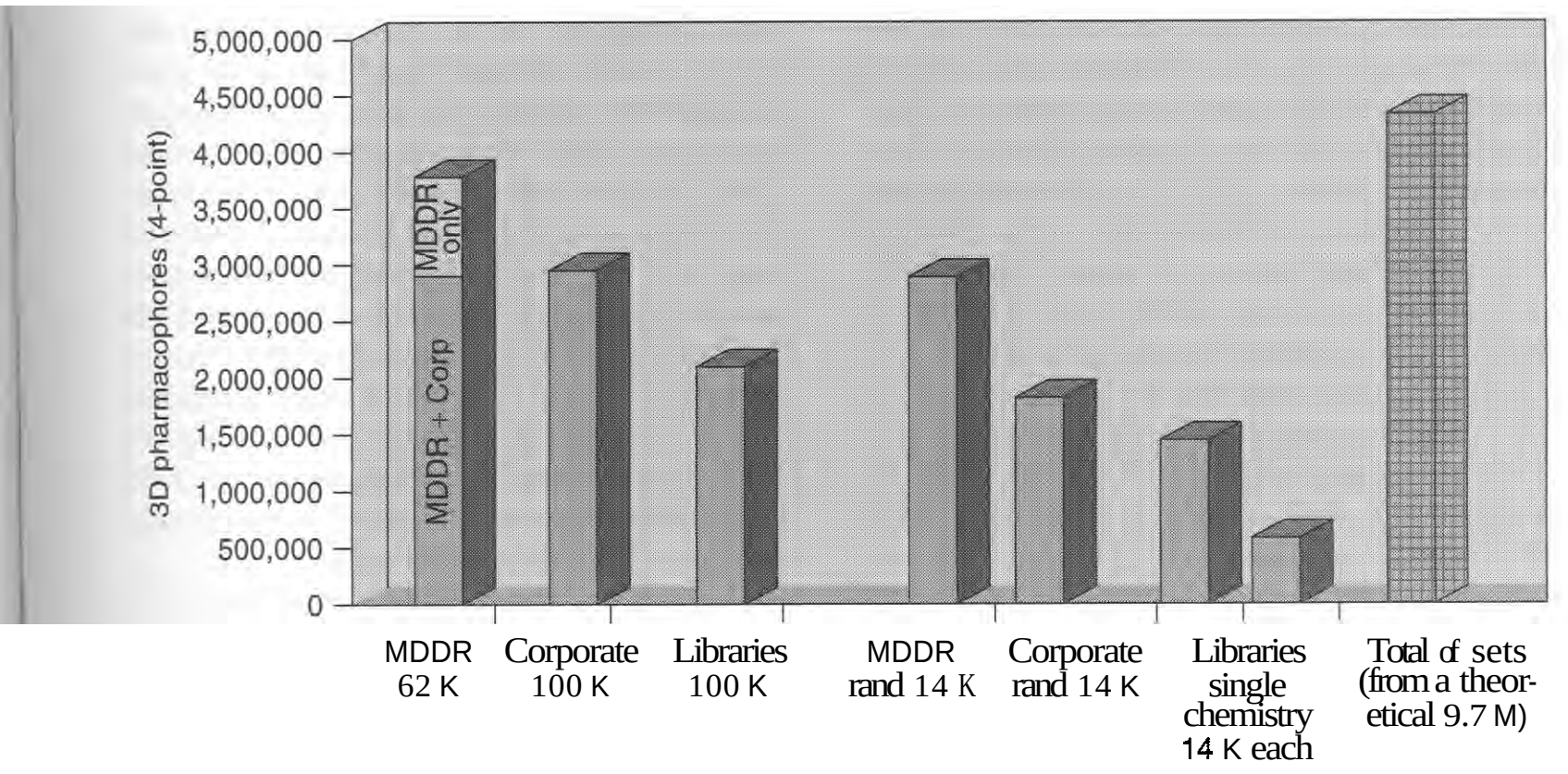


Figure 5.7. Comparisons of the **3D** four-point pharmacophore fingerprints exhibited by several sets [MDDR database of 62,000 biologically active compounds, a corporate registry database of 100,000 compounds used for screening, 100,000 compounds from combinatorial libraries (from a four-component Ugi condensation reaction), and 14,000 compound random subsets (MDDR, corporate) or individual libraries]. The four-point potential pharmacophores were calculated using 10 distance range bins and the standard six pharmacophore features.

MDDR pharmacophores not in corporate set), and acldditionally explore about 0.3 million new pharmiacophores. Figure 5.7 illustrates the number of pharmacophores found in these sets, together with those for the ACD (Available Chemicals Directory), random 14,000 subsets of the database sets and some of the combinatorial libraries (−14,000 each, from a four-component Ugi condensation reaction, 12 X 12 × 12 × 8 reactants). The relative richness and diversity of the MDDR database, which includes structures from a large number of' companies, is clear from the comparisons. 'The contributions, and eventual diminishing; return, of successive libraries using the same chemistry is discussed in Section 5.1.2 (see Fig. 5.24 below).

An example of the use of 3D pharmacophore fingerprints for the design of GPCR libraries (37a) using "relative" fingerprints focused around privileged substructures is described in Section 5.1.2. An approach that combines an optimization of a four-point pharmnacophore fingerprint and BCUT chemistry space diversity, using simulated annealing, has been described (37d; see Section 4.7). Simulated annealing is a widely used optimization methodology whereby the "temperature" of the system is used to control the degree of sampling of solution space. The "temperature" is cooled or annealed as the run progresses so that the system moves into a minimum for the function at low "temperature." In the classical sense, temperature controls the kinetic energy of the system; in a more general sense, the "temperature" has no physical meaning and is a parameter to control the sampling of solution space. Diversity was the goal (function to be optimized) of the studies reported, but the approach can equally be applied to optimize to a desired distribution of properties (e.g., from sets of biologically active compounds). The power of this pharmacophoric approach has been exemplified by Leach et al. in their GaP protocol for monomer acquisition (44).

Pharmacophore fingerprints derived from complementary site points to a target binding site have been used as a quantification of "biological diversity"/structure-based diversity

(37), defining a measure of the intersection between chemical and biological space. They can be compared to the pharmacophore fingerprints calculated from ligands, and the pharmacophore fingerprints of different target binding sites can also be compared to identify similarities (e.g., common binding motifs) and differences (e.g., for selectivity). The four-point pharmacophore fingerprint of a serine protease binding site was used to quantify all the possible binding modes. An example was given of how a combinatorial library could be designed to match as many as possible of these site pharmacophores, with the idea that the biological screening of the resultant library would provide information as to which hypotheses lead to (the best) binding. The site points can be generated by both geometric methods (as implemented in Chem-X/Chem-Protein; see Ref. 133) or through energetic surveys of the site [e.g., by using a variety of probe atoms (as implemented and used for pharmacophore fingerprint generation) (37); see Section 2.1.3.1].

The pharmacophore fingerprinting method thus provides a novel method to measure similarity when comparing ligands to their binding site targets, with applications such as virtual screening and structure-based combinatorial library design, as well as to compare binding sites themselves. Flexibility of the binding site can also be explicitly accounted for by using a composite fingerprint generated from several different binding site conformations.

2.2.2 Cluster-Based Methods. Clustering methods have a long history of application in chemical information (60). Any set of descriptors can be used in the clustering, but most typically some form of structural fingerprint is used in conjunction with a similarity measure such as the Tanimoto coefficient (see Section 2.1.4.1). The methods fall into two broad classes, hierarchical and nonhierarchical.

Nonhierarchical methods such as that described by Jarvis and Patrick (61) have been widely used for compound selection from large databases (62). The principle behind the Jarvis-Patrick method is to group together compounds that have a large number of nearest neighbors in common. However, the method requires the user to specify the number of clusters desired, and tends to be prone to singletons (clusters of one) and/or a small number of very large clusters. The cascade clustering methodology (59b) was developed to address some of these issues. Parameters were selected to produce an acceptable size distribution for the largest clusters and the small clusters were then reclustered. Doman et al. (63) have developed a fuzzy clustering technique, also based around the Jarvis-Patrick algorithm but which has no user-defined parameters and allows a compound to belong to more than one cluster.

Hierarchical methods can be further subdivided into agglomerative and divisive methods. Agglomerative methods start with each compound in a separate cluster and iteratively join the closest clusters together. Divisive methods start with a single cluster and iteratively subdivide until each compound is a singleton. Hierarchical clustering methods generate a dendrogram showing the relationship between the compounds, the issue being the level at which to cut the hierarchy (i.e., how many clusters to generate). Although heuristics exist, there is no automated method. Such algorithms, however, at best scale to order (N^2) in time, where N is the number of compounds, and so are limited in application to a few hundred thousand compounds at most (64). Nevertheless, they have been shown to be superior to nonhierarchical methods for clustering of chemical compounds (65). Ward's method was shown (5) to be the most effective at separating active from inactive compounds by clustering bit strings that describe the presence or absence of 153 small generic and specific fragments (ISIS structural key descriptors). Even better performance was obtained with the inclusion of pharmacophore distances between site points complementary to hydrogen bonding and charged groups combined with distances between centers of aromatic rings and attachment points for hydrophobic groups.

2.2.3 Dissimilarity-Based Methods. The methods for compound selection described above essentially group compounds either by partitioning into cells or by clustering. Dissimilarity-based methods (66) avoid this step.

The subset selection can be performed iteratively. The first compound is chosen at random and the next compound is selected to be maximally dissimilar to the first; the third is then selected to be maximally dissimilar to the first two, and so on. The selection stops when a prespecified number of compounds have been selected or no more compounds can be chosen that are below a given similarity or above a certain distance to another compound in the selected set. Pearlman (11b,c) refers to such methods as "addition" algorithms because they add compounds to a diverse set of increasing size. He notes that such algorithms are quite satisfactory when the size of the desired subset is relatively modest but, given that the time required for such algorithms is proportional to the size of the total population and the square of the size of the desired diverse subset, they are far less satisfactory when, for example, selecting a subset of 10,000 from a population of 1,000,000.

Alternatively, the number of desired compounds can be predefined and a stochastic algorithm used to maximize the diversity of the selected set, although these methods are even slower than addition methods. Sphere-exclusion methods, which Pearlman calls "elimination" algorithms because the diverse subset is created by eliminating compounds from the superset, have been implemented in Diverse-Solutions (31) (see Section 2.2.1.1), providing a rapid distance-based diverse subset selection method. The minimum distance between nearest neighbors within the diverse subset is first defined (D_{min}), a compound is chosen at random, all compounds within D_{min} are removed, a second compound is chosen at random, all compounds within D_{min} are removed, and this is repeated until no more compounds can be chosen. The algorithm controls the size of the resulting diverse subset by automatically repeating the process with a larger or smaller value of D_{min} as necessary. Because the time required for each elimination subset is proportional to the size of the superset and size of the subset (not size of subset squared), the elimination method, despite the need for (typically) four or five automatic repetitions, is far faster than the addition method and yields subsets of essentially equal diversity.

Maximum dissimilarity-based methods tend to give diverse selections, including many outliers of less potential interest to a medicinal chemist. By contrast, methods such as sphere exclusion (minimum dissimilarity selection) tend to give representative selections that mimic the underlying distribution of compounds. The OptiSim method developed by Clark (66c) attempts to achieve a compromise between these two extremes. Three parameters are required: the first two, radius or similarity cutoff and maximum number of selections are common to the other algorithms. A third parameter, K, is required to define a subsample size. Up to K selections are added to the subsample at each iteration and the best compound from the subsample added to the selected set. At the limit of $K = 1$, this is equivalent to minimum dissimilarity selection, whereas at the limit of K = N (total number of compounds) the algorithm is equivalent to maximum dissimilarity selection. By altering the value of K the user can thus achieve a compromise between the diversity and representativeness of the selected set. Tests of the algorithm suggest that it is possible to achieve selections similar to those achieved from hierarchical clustering methods at a greatly reduced computational cost. Maximum dissimilarity methods were shown (66d) to lead to more stable QSAR models with higher predictive power, based on a comparative mean field analysis of angiotensin-converting enzyme inhibitors.

Hudson et al. (67) describe two parameter-based methods for compound selection. The most descriptive compound (MDC) method is aimed at selecting compounds that represent the population as a whole. An information vector is accumulated from the ranked Euclidean distance of each compound in the data set to all others. The most descriptive compound is that with the largest information, which equates to the compound with the smallest overall distance to all other compounds. The next compound is chosen to give the greatest additional information and so on. The sphere-exclusion method used attempts to select compounds that most effectively cover the property space. A compound is selected, say the MDC, and all compounds are removed that are closer to it than a user-defined radius. The

next selection is that compound in the remaining set that is closest to the one already selected. This process is repeated until no compounds are left for selection. The methods were applied to the selection of standard sets for biological screening at Wellcome.

The maxmin approach (23c) uses the shortest nearest-neighbor distance as a measure of diversity in the sample:

$$D_{\max\min} = 1 - \frac{1}{N}\sum_{i=1}^{N} \max_{i}[\min_{j\neq i}(d_{ij})]$$

This measure is particularly useful in selecting diverse compound sets from a corporate collection, as exemplified by Higgs et al. (68). They also introduce the concept of a coverage design for lead follow-up, in which compounds are selected to be maximally similar to a set of leads.

An alternative approach is to use the sum of pairwise similarities in the maxsum approach:

$$D_{\max\text{ sum}} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \text{sim}(i,j)}{N^2}$$

This approach is particularly efficient when combined with the Cosine coefficient (69) and was used by Pickett et al. in combination with pharmacophore descriptors (70). In lower dimensional spaces the maxsum measure tends to force selection from the corners of diversity space (6b, 71) and hence maxmin is the preferred function in these cases. A similar conclusion was drawn from a comparison of algorithms for dissimilarity-based compound selection (72).

An excellent discussion of different diversity functions has been given by Waldman et al. (73). A set of ideal behaviors for a diversity function was defined. These are particularly relevant to library design tasks. Thus, although maxmin is suitable for selecting highly diverse molecules, it is less well suited to library design optimization processes. An alternative function was defined based on minimum spanning trees, which had previously been used by Mount et al. (71) and a gaussian error function, erf( ):

$$D = \sum \text{erf}(\alpha d_i)$$

The reader is referred to reviews (e.g., see Refs. 2, 6, 7) for further detail and discussion of other related measures and methods for diversity-based selection.

2.2.4 Biasing to Desired/Desirable Properties. Any of the methods above can be used to bias the compound selection toward a particular region of property space, for example, by restricting the selection to cells or clusters containing known actives (in the latter case it may mean reclustering if the active is from an external source). However, it is a common experience when applying diversity-based selection to large databases to see a number of compounds that are undesirable for a number of reasons. They may be too large, too flexible, too lipophilic, contain too many acid groups, and so forth. Thus, it is general practice to apply filters during the selection process. These can include limits on property values such as calculated log P and molecular weight (68) and the application of substructure filters to remove undesirable or reactive compounds (21a, 66a, 74). Lipinski (29) formalized some of these ideas in the Rule-of-5, derived from analysis of orally absorbed drugs, with problems more likely if two or more of MW > 500, c log P > 5, sum NH, OH (~ H-bond donors) > 5, sum N,O (~ H-bond acceptors) > 10. Variants of this, often stricter (e.g., only one violation and/or lower values for hits/leads), are now widely used in conjunction with other methods for classification of compounds as likely to be orally absorbed or to penetrate the CNS. The reader is directed to several reviews on this topic (75). The usefulness of such approaches has been shown by the work of Pickett et al. (76), where a library was designed using simple descriptors such as polar surface area for oral absorption (77). The designed library showed improved absorption in a Caco-2 system over a previous related library where the products had not been formally designed to these criteria.

In a more general sense, compounds can be selected to reproduce a given set of property profiles for calculated log P, molecular weight, and so forth derived from, say, a set of known

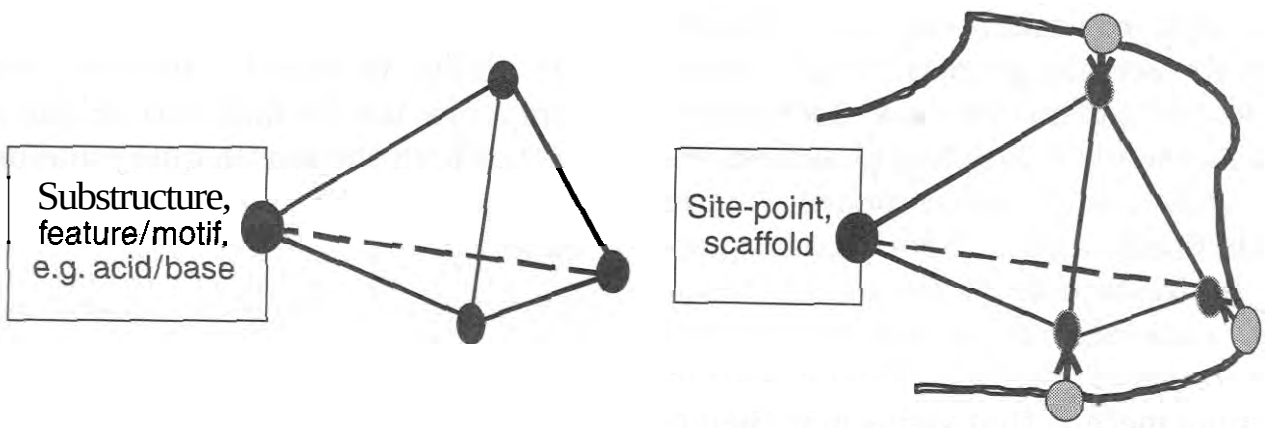


Figure 5.8. Example of privileged four-point pharmacophores, either created from a ligand using a particular feature (e.g., the centroid of a "privileged" substructure) or complementary to a protein site using a site point or attachment point of a docked scaffold. Only pharmacophores that include this special feature are included in the fingerprint, thus providing a relative measure of diversity1 similarity with respect to the privileged feature.

drugs. Such an approach is most widely used as an additional constraint in library design algorithms (78) and is further discussed below

An interesting example of biasing in compound selection is provided by Grassy et al. (79). Lead compounds were used to derive a range of acceptable values for topological indices and other molecular descriptors. These were used to filter a large virtual library and led to an active compound being synthesized.

2.2.5 Relative Diversity/Similarity. This describes an approach that measures "relative" similarity and diversity between chemical objects, in contrast to the use of the concept of a total or "absolute" reference space (80). The ability of 3D pharmacophoric fingerprint descriptors to separate ligand-binding properties from chemical structure has enabled a useful modification to the way the descriptor is evaluated (37). It is possible to identify one of the points of a pharmacophoric description such as a triplet or quartet with a special feature, such as a "privileged" substructure deemed important for binding or a pharmacophore group. A fingerprint can be generated that describes the possible pharmacophoric shapes from the viewpoint of that special point/substructure (see Fig. 5.8). This creates a "relative" or "internally referenced" measure of diversity, enabling new design and analysis methods. The technique has been extensively used to design combinatorial libraries that contain "privileged" substructures focused on GPCRs (37a), and this is described further in Section 5.1.2. The use of "receptor-relevant" BCUT chemistry spaces from DiverseSolutions provides a different approach to a focused similarity/diversity measure (11d, 32e-h).

3 VIRTUAL SCREENING BY MOLECULAR SIMILARITY

The use of molecular similarity to analyze large databases of structures using information derived from one or several ligands provides a powerful ligand-based virtual screening method (protein structure-based virtual screening methods are by comparison based on docking structures into a binding site). Virtual screening requires that a set of structures is ranked, with the goal of identifying new structures that have similar biological activity, with top-scoring compounds sent for evaluation in a biological assay. Usually, the requirement is to provide a small subset of compounds (10–1000) from a large set (100,000–1,000,000) of possible compounds for screening that is enriched in actives (i.e., contains a greater proportion of actives than that of the full compound set). In this context, enrichment involves identifying the highest number of new chemotypes as opposed to analogs of the query structure(s). Pharmacophoric methods have been found to be particularly effective for this, building on the successful use of 3D database searching for lead generation. Other similarity methods such as the use of 2D descriptors (Section

2.1.1) are also commonly used to identify structures for screening based on the structure of a known ligand. The use of similarity searching in chemical databases has been reviewed by Willett et al. (2a), comparing newer types of similarity measure with existing approaches. In this section the focus is on the use of the 3D pharmacophoric methods, which have been shown to provide a ligand-based virtual screening method that yields new chemotypes.

3.1 Use of Geometric Atom-Pair Descriptors

The topological atom-pair descriptors (24) have been extended by Sheridan and coworkers to geometric atom pairs (26), and shown to be effective at generating hit lists enriched in active molecules of different chemotypes. A set of precalculated conformations (–10–25) is used for each molecule, and each atom is assigned two different atom types: (1)a binding property (donor, acceptor, acid, base, hydrophobic, polar, and other); (2)a combination of element type, number of neighbors and π-electron count. All combinations of atom pairs are analyzed, for each conformation, and resultant histograms of each probe and database molecule conformation are compared. The technique was compared with its topological equivalent (counting bond connections between atoms to estimate interatomic "distance"). This demonstrated that, although both methods were able to significantly enrich the highest ranking structures with other active molecules for the same target (–20- to 30-fold enhancement over random in the top 300 compounds), the 3D structure-derived descriptors were able to show their advantage by picking out active chemotypes with greater structural variation relative to those from the 2D searches. The analysis used about 30,000 structures from the Derwent Standard Drug File (SDF; version 6, developed and distributed by Derwent Information Ltd., London, England, 1991, now known as the World Drug Index) using probe molecules with known activity against a particular target to rank the database. Sheridan et al. (26c) have also shown how a single combined atom-pair descriptor from a set of molecules can be used in a single fast search to provide results similar to those from the slower process of individual molecule-by-molecule searches. This provides the ability to search mixtures, which some companies use for high throughput screening, in that both the search query and/or the database being searched can be mixtures of structures.

3.2 Use of 3D Pharmacophore Fingerprints (Three- and Four-Point)

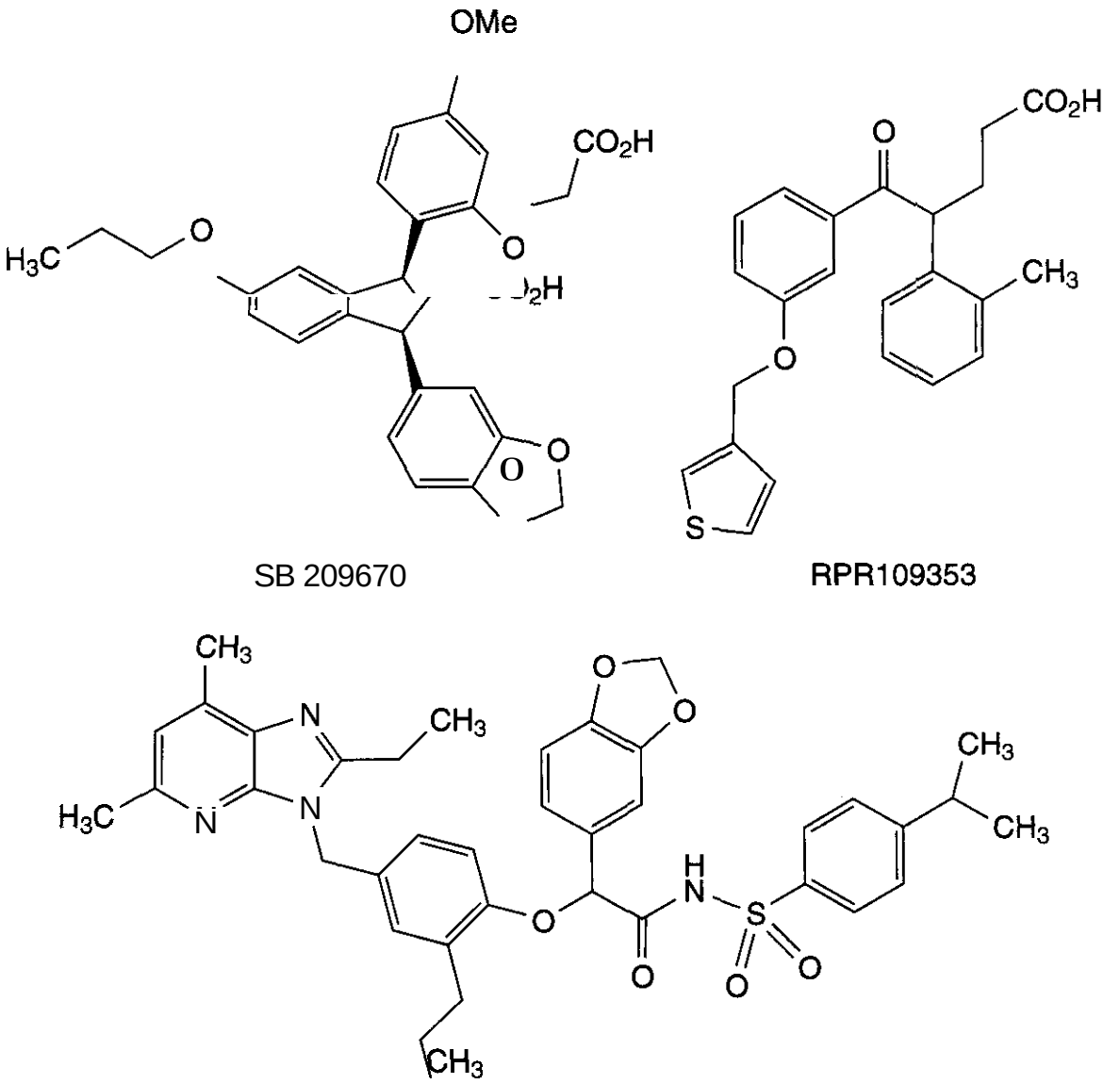
Some research groups have extended the atom-pair descriptors to three-point (triplets) and four-point (quartets) pharmacophore descriptors (35, 37, 76, 81) as described in section 2. These descriptors have a potentially superior descriptive power, and a perceived advantage over atom pairs is the increased "shape" information (intrapharmacophore distance relationships) content of the individual descriptors (37a). The quartet (tetrahedral) four-point descriptors offer further potential 3D content by including information on volume and chirality (37a, 82), compared with the triplets that are components of the quartets and represent planes or "slices" through the 3D shapes.

The fingerprints can be precalculated for database compounds, with conformational sampling, and stored in an efficient format (e.g., four-point pharmacophore fingerprints, where one line of encoded information uses, about 11 kilobytes of space for 1000 pharmacophores). Probe fingerprints from one or more structures can be rapidly compared against such databases at speeds of >100,000 compounds/min, even for large four-point pharmacophore fingerprints, representing about 10 million different pharmacophoric shapes. Similarity is measured using potential pharmacophore overlap and similarity indices such as the modified Tanimoto index (37a).

The relative merits of two-, three- and four-point pharmacophore descriptors for different applications is an area of ongoing study (37, 83). Figure 5.9 shows some structurally diverse endothelin antagonists that exhibit low 2D similarity, but maintain significant overlap of their four-point pharmacophore fingerprints (37a).

3.3 Validation Studies

The validation issue for 1D, 2D, and 3D descriptors for similarity searching and virtual



L-746,072

Figure 5.9. Structurally diverse endothelin antagonists exhibiting low 2D similarity while maintaining common pharmacophoric elements crucial to activity.

screening has been addressed in several publications (5d, 14a, 45, 72, 84, 85). Conflicting results have been reported, probably because of the way the different descriptors were used and biases in the test sets. Two primary concepts have been applied to the analysis of biological data. The concept of "neighborhood" behavior (84) as a measure of descriptor utility has been promoted, based on the idea that if a descriptor is able to cluster molecules with a particular biological activity, the descriptor encodes information regarding the requirements for that activity, and by extension is a useful measure for molecular similarity/diversity. Comparisons using 2D fingerprints with pharmacophore fingerprints with this approach led to the conclusion that 2D descriptors performed better than their 1D and 3D counterparts (14a, 45). However, issues with the studies undertaken have been raised (85). These relate to bias in the data sets arising from the presence of closely related analogs, which by their nature have high 2D substructural similarities, and the way the 3D pharmacophoric descriptors were generated (single conformation only) and used (bin setting, Tanimoto index).

Some comparative studies of ligand-based virtual screening methods have been undertaken within Bristol-Myers Squibb (85) using more optimum settings for pharmacophore fingerprint generation [four-point pharmacophores, 7 distance bins, and full conformational analysis (37a)], which gave quite different results. An example using melatonin as a probe molecule to search against a database of about 150,000 compounds containing about 250 known melatonin antagonists is shown in Fig. 5.10. The graph shows the hit rates obtained by similarity ranking in terms of the

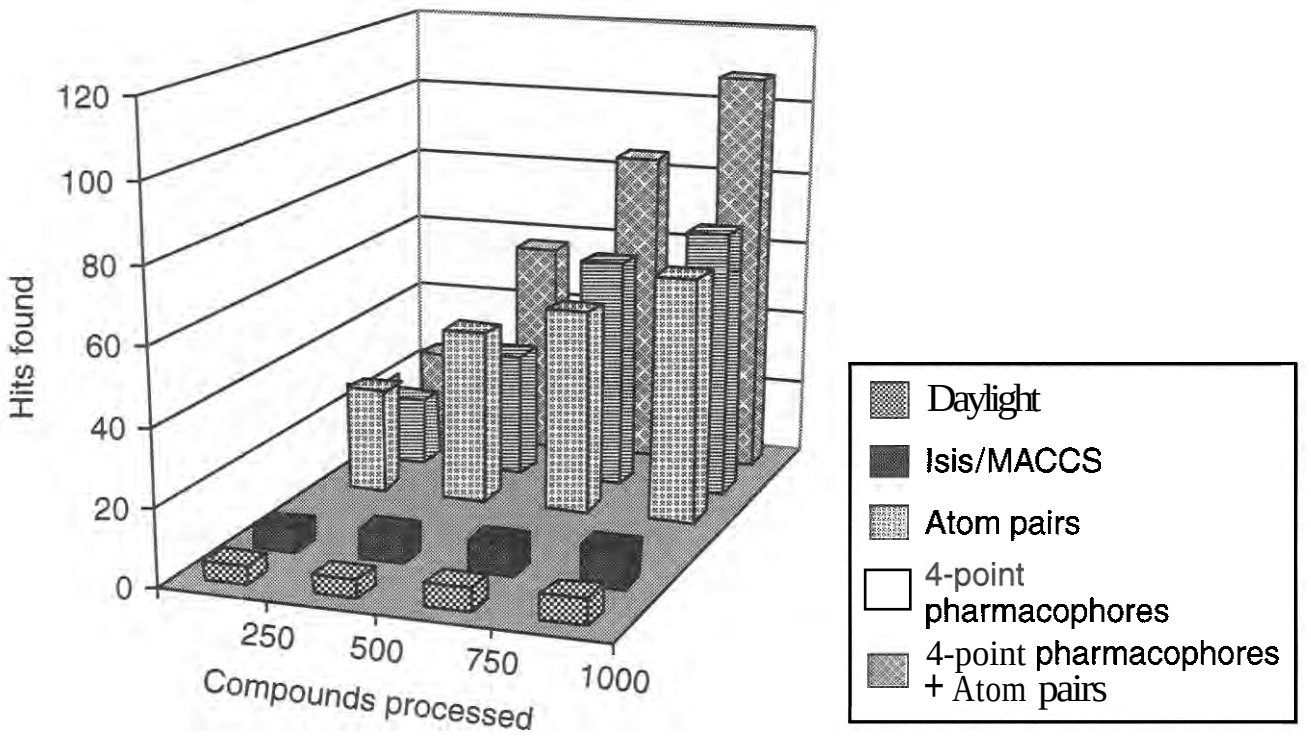


Figure 5.10. Ligand-based virtual screening/similarity analysis of a 150,000-compound database containing about 250 known melatonin antagonists, showing the strong performance of the pharmacophore descriptors, and the complementarity of atom pairs and pharmacophore descriptors when combined.

number of active compounds located across the top-ranking 1000 compounds. In this case, for the 2D descriptors shown, only the atom-pair (26a) descriptor (which has elements of a two-point pharmacophore fingerprint with intercenter distance replaced by bond count) produces comparable results. A 2D similarity search using the UNITY 2D fingerprint (13c), with a very low 50% similarity cutoff, produced a hit list of 1669 compounds containing only 10 melatonin actives (an enrichment of 3.6 relative to random screening of 1669 compounds, for which 2.8 actives would be expected to be found). In contrast, the pharmacophore fingerprint similarity search finds 93 actives in the first 1669 compounds, a total enrichment of >33 relative to random, and a further >ninefold relative to the 2D similarity search. Preliminary studies on systems with a much wider structural diversity in active ligand chemotypes suggested hit rates even more favorable to pharmacophore fingerprints, with cases where no 2D methods were able to improve on random hit rates. It is of interest that averaging the four-point pharmacophore/atom-pair rankings leads to even better results in the melatonin investigation, highlighting (45) the potential advantages of combined descriptors.

The advantage of 3D pharmacophore-based geometric descriptors over topological descriptors in being able to pick up new chemotypes with major 2D structural variations is of particular importance when exploiting a peptide lead. In such cases, the goal of screening is normally a nonpeptidic molecule. Using the pharmacophore fingerprint from relatively large and flexible peptidic molecules (e.g., tetrapeptides), it is possible to identify structures that match just part of the pharmacophoric information; a modified Tanimoto coefficient can be used to reduce the penalty for only a partial match. It is possible to identify a set of reasonable structures that, as an ensemble sample most of the potential pharmacophores exhibited by the peptide. With 2D methods the high ranking molecules will tend to be similarly peptidic, rather than more druglike molecules exhibiting 3D properties of the peptide.

An example of the use of peptidic information comes from the work of Pickett et al. (76) using the known tripeptide RGD (Arg-Gly-Asp) (see Fig. 5.11) motif fibrinogen uses for receptor binding (86). A database of 100,000 compounds, which had been seeded with fibrinogen receptor antagonists covering a wide range of structural classes (see Fig. 5.12), was

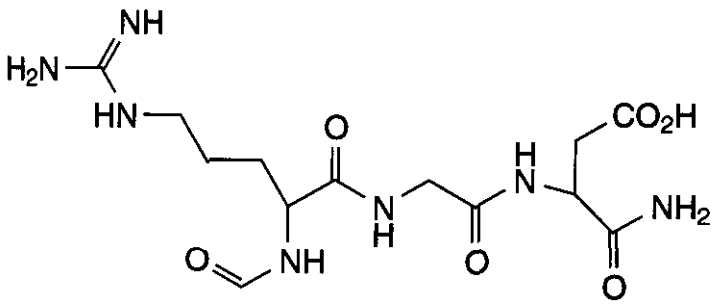
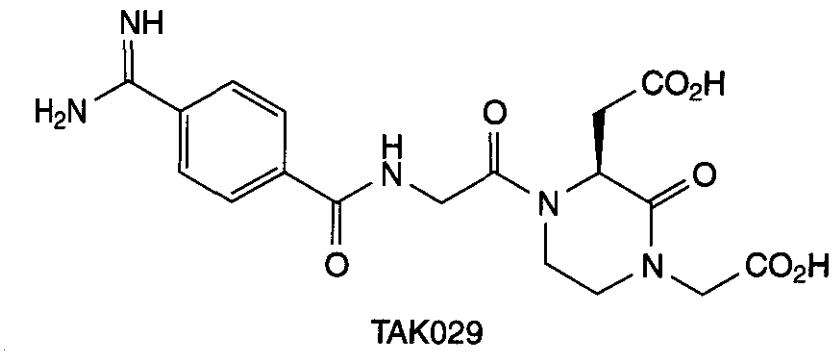
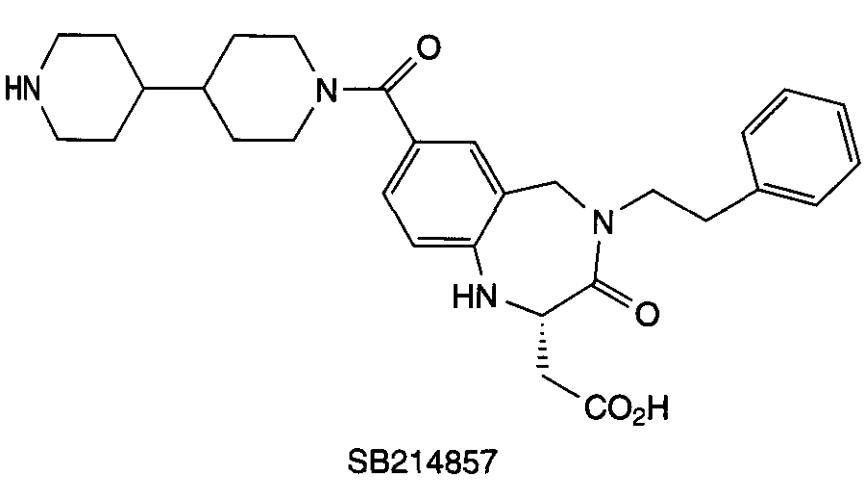
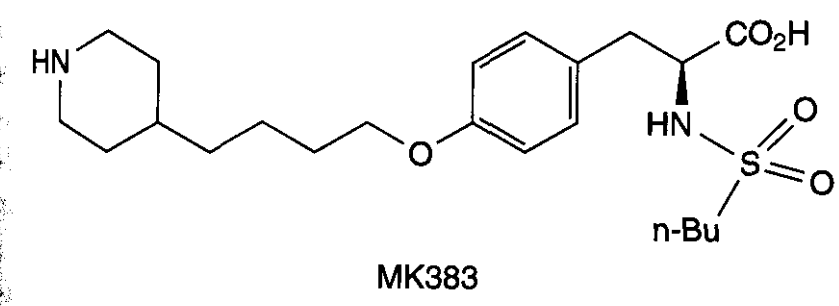


Figure 5.11. Structure of the RGD motif.

screened using a customized (35, 37, 70) version of Chem-X/ChemDiverse with four-point pharmacophore fingerprints. A degree of flexibility of the RGD motif loop region was suggested by structural biology studies, highlighting an advantage of the descriptor encoding a conformational ensemble and not requiring a bioactive conformation to be known. Virtual screening can thus be undertaken even when only limited structural information is available. The four-point pharmacophore fingerprint of the RGD tripeptide probe, N- and C-capped with amide groups, was generated with a full conformational analysis and used for the search. All the actives appeared in the top 3% of the data set, with a reasonable diversity of hits and little 2D resemblance to the probe peptide.







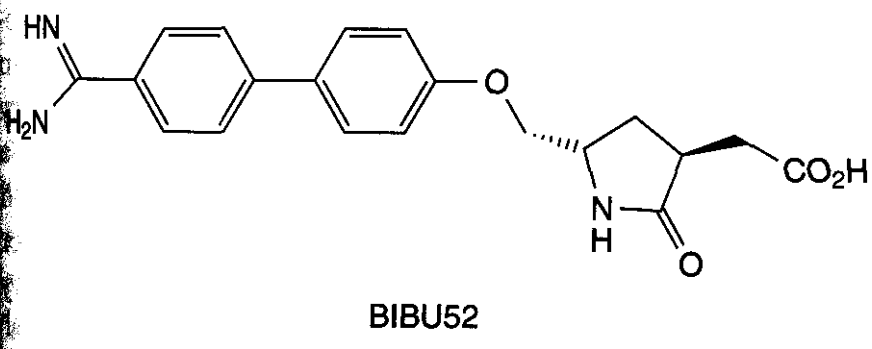


Figure 5.12. Some structurally diverse RGD antagonists matched through virtual screening using the RGD motif and 3D pharmacophore fingerprints.

4 COMBINATORIAL LIBRARY DESIGN

4.1 Combinatorial Libraries

Normally, the term ***combinatorial library*** implies a library of a few hundred to many thousands of products produced using high throughput robotics in a facility dedicated for such purposes. In contrast, the term ***parallel library*** implies a library of less than 10 to a few hundred products produced using more or less traditional medicinal chemical synthetic procedures or increasingly common low throughput robotics. In either case, lists of reactants are combined in a combinatorial fashion to yield an array of products. The methods described in this section are applicable equally to both cases. A strictly combinatorial combination of reactants (or reactants and scaffold) produces the most efficient use of reactants and automation/robotics for a library synthesis. However, the issue of generating products that have suitable properties for biological screening and as hit/lead material is key, and constraints discussed below are often used, resulting in not all combinations of reactants being used.

4.2 Combinatorial Library Design

The process of combinatorial library design brings together many molecular diversity and similarity approaches with the aim of identifying a set of reactants that are to be combined (reacted) to form products. Combinatorial library design is, inevitably, an iterative process: software is used to suggest lists of reactants; chemists accept some suggestions but reject others (for various reasons ranging from cost or availability to poor synthetic yield). If software is to be used to suggest replacements for rejected reactants, it must be designed to accommodate this iterative process.

Although the objective is always to identify which reactants should be used to make the products, there are two fundamental approaches to library design: reactant-based methods and product-based methods. Purely product-based methods, which select (or cherry pick) desired products without regard for the number of reactants required to form those products (as in standard similarity searching or diverse set selection), lead to noncombinatorial synthetic schemes and are clearly at odds with the efficiency objectives of a combinatorial synthesis. Reactant-based methods suggest lists of reactants based solely on comparisons of reactant properties without regard for the properties of the resultant virtual products. Thus, reactant-based methods avoid the need for enumerating what could be very large numbers of virtual products and making even greater numbers of comparisons of product properties. Compromise solutions, discussed later, that approximate certain product properties from the reactants without enumeration have been developed. By directly selecting a desired number of each type of reactant, the chemist can ensure an efficient, full-combinatorial array design, and the expediencies of reactant-based methods led to their widespread use for the design of both large and small combinatorial libraries. However, the growing awareness of a need to maintain druglike properties, if only for practical issues such as solubility, has led to the use of reactant-biased methods that consider the properties of the products from enumerated structures, for which the full array of ligand- and structure-based design methods can be applied, and the resultant synthesis of sparse arrays (see later). In addition, the assumption that optimal product diversity can be approximated by using diversity-optimized reactants has been questioned (87), and several product-based diversity methods for selecting reactants have been developed that consider the need for full or sparse combinatorial arrays in the design process. The rationale behind this observation can be understood when one considers that "constraints" based on whole molecule properties and some form of molecular similarity/diversity are usually required.

Pearlman (11, 87c) has made this argument quite convincingly by comparing the results of alternative diverse library design methods in a low dimensional chemistry space, as illustrated in Fig. 5.13. Figure 5.13a depicts a virtual library of 634,721 allowed combinatorial AB products (remaining after optional filtering of the full virtual library based on Lipinski's Rule-of-5 "druglikeness" criteria) in a chemistry space specifically cho-

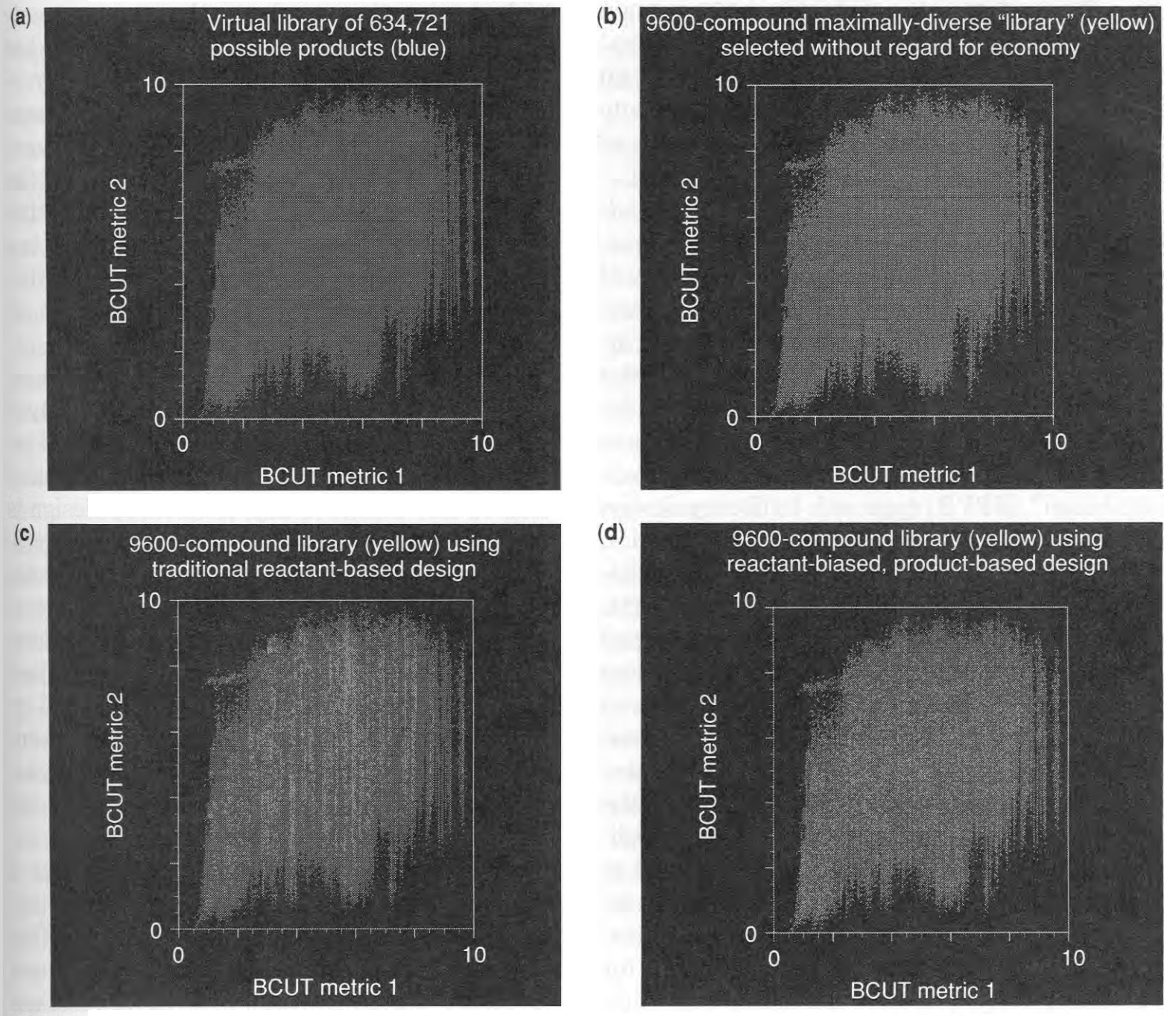


Figure 5.13. (a) **A** virtual library of 634,721 allowed combinatorial AB products (after filtering out products that failed Lipinski's Rule of **5** "druglike" criteria) shown in a BCUT chemistry space specifically chosen to best represent the diversity of the virtual library. (b) The maximally diverse 9613-compound subset of the virtual library, illustrating the results of purely product-based "library design." Although providing the maximal diversity, synthesis of these 9600 AB products would require the use of 347 A's and 1024 B's-clearly unacceptable from the perspective of synthetic economy (numbers of reactants and robotic control). (c) The 9600-compound library resulting from the traditional, purely reactant-based library design strategy of selecting the 80 most diverse A's and the 120 most diverse B's. Although providing user-selected synthetic economy, the diversity of these 9610 AB products is clearly quite poor. (d) The 9600-compound library resulting from the reactant-biased, product-based (RBPB) algorithm developed by Pearlman and Smith (see Refs. 31, 87c and text). The algorithm selected a different set of 80 A's and a different set of 120 B's, thus providing the same level of user-selected synthetic economy, while also providing substantially greater diversity than could be achieved using a purely reactant-based library design strategy. See color insert.

sen to best represent the diversity of that virtual library. Figure 5.13b illustrates an optimally diverse "library" of 9600 products selected by using cell-based diverse subset selection to cherry pick the 9600 most diverse products without regard for synthetic economy. Although the diversity of these products is clearly optimal, the fact that 347 A's and 1024 B's would be required to make the 9600 AB products provides an equally clear indication of why purely product-based methods are unsatisfactory from an economical perspec-

4 COMBINATORIAL LIBRARY DESIGN

4.1 Combinatorial Libraries

Normally, the term ***combinatorial library*** implies a library of a few hundred to many thousands of products produced using high throughput robotics in a facility dedicated for such purposes. In contrast, the term ***parallel library*** implies a library of less than 10 to a few hundred products produced using more or less traditional medicinal chemical synthetic procedures or increasingly common low throughput robotics. In either case, lists of reactants are combined in a combinatorial fashion to yield an array of products. The methods described in this section are applicable equally to both cases. A strictly combinatorial combination of reactants (or reactants and scaffold) produces the most efficient use of reactants and automation/robotics for a library synthesis. However, the issue of generating products that have suitable properties for biological screening and as hit/lead material is key, and constraints discussed below are often used, resulting in not all combinations of reactants being used.

4.2 Combinatorial Library Design

The process of combinatorial library design brings together many molecular diversity and similarity approaches with the aim of identifying a set of reactants that are to be combined (reacted) to form products. Combinatorial library design is, inevitably, an iterative process: software is used to suggest lists of reactants; chemists accept some suggestions but reject others (for various reasons ranging from cost or availability to poor synthetic yield). If software is to be used to suggest replacements for rejected reactants, it must be designed to accommodate this iterative process.

Although the objective is always to identify which reactants should be used to make the products, there are two fundamental approaches to library design: reactant-based methods and product-based methods. Purely product-based methods, which select (or cherry pick) desired products without regard for the number of reactants required to form those products (as in standard similarity searching or diverse set selection), lead to non-combinatorial synthetic schemes and are clearly at odds with the efficiency objectives of a combinatorial synthesis. Reactant-based methods suggest lists of reactants based solely on comparisons of reactant properties without regard for the properties of the resultant virtual products. Thus, reactant-based methods avoid the need for enumerating what could be very large numbers of virtual products and making even greater numbers of comparisons of product properties. Compromise solutions, discussed later, that approximate certain product properties from the reactants without enumeration have been developed. By directly selecting a desired number of each type of reactant, the chemist can ensure an efficient, full-combinatorial array design, and the expediencies of reactant-based methods led to their widespread use for the design of both large and small combinatorial libraries. However, the growing awareness of a need to maintain druglike properties, if only for practical issues such as solubility, has led to the use of reactant-biased methods that consider the properties of the products from enumerated structures, for which the full array of ligand- and structure-based design methods can be applied, and the resultant synthesis of sparse arrays (see later). In addition, the assumption that optimal product diversity can be approximated by using diversity-optimized reactants has been questioned (87), and several product-based diversity methods for selecting reactants have been developed that consider the need for full or sparse combinatorial arrays in the design process. The rationale behind this observation can be understood when one considers that "constraints" based on whole molecule properties and some form of molecular similarity/diversity are usually required.

Pearlman (11, 87c) has made this argument quite convincingly by comparing the results of alternative diverse library design methods in a low dimensional chemistry space, as illustrated in Fig. 5.13. Figure 5.13a depicts a virtual library of 634,721 allowed combinatorial AB products (remaining after optional filtering of the full virtual library based on Lipinski's Rule-of-5 "druglikeness" criteria) in a chemistry space specifically cho-

thesis of a library of compounds with a high degree of control over associated properties.

Thus, the combinatorial library design process brings together many of the methods already described for molecular similarity and molecular diversity coupled to synthetic feasibility considerations. Diversity-based and structure-based approaches to the design of virtual libraries have been reviewed (7, 91a). Both ligand-based and protein structure-based virtual screening methods can be used, with the combinatorial nature of the virtual compounds being exploited to increase the speed of the analysis. Some properties of the products can be estimated rapidly on the fly from the reactants, and products can be generated in the active site. The CombiDOCK approach that can rapidly analyze very large virtual databases in a binding site by connecting reactants to scaffolds docked in multiple orientations is discussed in Section 4.10. A genetic algorithm-based method for the combinatorial docking of reactants has been described by Jones et al. (92), with the application of a ligand-docking genetic algorithm to screening combinatorial libraries.

A challenge in the design of small- and medium-sized focused combinatorial libraries is to harness for use in library design the experience and knowledge gained in generating structure-activity relationships (91b). Screening libraries biased for pharmaceutical discovery are often designed to augment the structural diversity of a chemical library. The approach used in the LASSOO algorithm (93) is based on the identification of compounds from a virtual library that are most different from those already present in a screening set and to a reference set of undesirable compounds, while being simultaneously most similar to a set of compounds with desirable characteristics. An illustration of the method using bit-string structure descriptors is given.

Combinatorial library design approaches have been discussed (94), with the design of library subsets that simultaneously optimize the diversity or similarity of a library to a target, properties (such as druglikeness) of the library members, properties (such as cost or availability) of the reactants required to make them, and the efficiency for array synthesis. They showed that libraries can be designed to contain molecules constrained to certain druglike properties with only a small trade-off in terms of the maximum possible diversity.

The design of leadlike combinatorial libraries is an approach of more recent interest. A lower molecular weight starting point is advantageous, in that bulk can be added for potency/selectivity/properties without exceeding "rule of 5" parameters for orally absorbed drugs; otherwise a more labor-intensive step may be needed to identify a smaller active part of the hit. The properties required of library compounds intended to provide leads suitable for further optimization, that may be rather different from final optimized leads, has been reviewed (95).

Thus, library design is a complex optimization problem with often competing constraints, including requirements to have combinatorial efficiency and/or several specified product properties (both desired and nondesired). Methods such as genetic algorithms, simulated annealing, and Monte Carlo optimization have been used, and iterative cyclic approaches applied. The next section describes the application of these methods within the context of library design but the reader should note that some of these methods are applicable only for the design of diverse libraries.

4.3 Optimization Approaches

The most basic product-based selection process used in library design is an order-dependent analysis of products, selecting a compound if it exhibits sufficient "diversity" to products already selected. This approach was used in the Chem-X/ChemDiverse software with three- and four-point pharmacophore fingerprints. A compound was selected if the overlap with the ensemble fingerprint of already selected compounds was less than a user-defined amount; that is, the molecule contains a significant number of pharmacophores not already exhibited in selected compounds. This cherry-picking process is an efficient method for ensuring a high diversity library, but can be a combinatorially inefficient selection for synthesis, with no explicit reference to the constituent reactants being made (see Section 4.2 above for further examples). A preferred selection for combinatorial efficiency is arrays of reactants, in which all

reactants from one component of a combinatorial library are reacted with all the reactants in the other components, or sparse arrays, in which subsets of reactants are combined. Additional constraints such as physicochemical properties and flexibility are addressed implicitly by assigning upper and lower bounds for given properties, or controlling the order in which molecules are processed.

To address the issue of using pharmacophore fingerprints in a way that enabled a combinatorially efficient selection of reactants to be selected, and the explicit inclusion of additional molecular properties such as a balance of druglike physicochemical properties and shape descriptors, the HARPick program (78a,b) was created. A stochastic optimization technique [Monte Carlo simulated annealing (96)] was used to enable selections in reactant space, whereas diversity is still calculated in product space. User-defined flexibility for the reactant array sizes was possible, and additional descriptors could be used (e.g., to address the selection of non-drug-like compounds). The pharmacophore fingerprint (three-point, triplets) was used in a nonbinary mode (the frequency of occurrence of each pharmacophore was calculated), and the HARPick diversity measure was tuned to include a term (Conscore) to force molecules to occupy relative rather than absolute voids in pharmacophore space. This avoids the problem of saturation of the fingerprint with large databases in a binary mode, particularly a problem with the three-point pharmacophore descriptors. It was thus possible to design combinatorial libraries that exhibited pharmacophores that were poorly represented in a reference set of compounds. The Conscore constraint score sums the product of the number of times pharmacophore i has been hit for molecules selected from the current data set with the score associated with pharmacophore i for the constraining library. The Conscore term can be inverted, enabling focused designs, in which the selection of products that occupy the more highly occupied bins (e.g., from a set of active compounds) is desired. The flexibility and success of this kind of stochastic optimization methodology has led to its use by many other researchers for library design (5c, 6b, 23d, 78c, 97c,d). Simulated annealing has been used to perform reactant selection for combinatorial libraries based on three-point pharmacophores (78a,b), as described above, and other metrics (6b, 23d, 97c,d).

Genetic algorithms (GA) are another class of optimization techniques widely used within chemistry (98) that have been explored for library design. A GA is an attempt to utilize the Darwinian process of evolution in an optimization procedure. A solution is represented by a string of fixed length, the chromosome, and is evaluated according to some criterion to give the fitness score, for example, the pharmacophore coverage of the solution (78b). The GA maintains a number of chromosomes (potential solutions) that are ranked on their fitness and are then modified according to operators including mutation, where one element of the string is changed, and crossover, where the string is cut at some position and swapped with equivalent portions of another solution. These new solutions are evaluated and the process is repeated for a defined number of iterations or until all (or most) solutions converge on one result. For library design, the string represents the selected monomers at each variable position of the library. Evaluation involves enumerating the sublibrary defined by the solution and calculating the score associated with the products. The stochastic nature of the process means that the GA is run several times to ensure good convergence.

A GA was used by Sheridan and Kearsley (99) to design peptoid libraries focused to cholecystokinin by scoring on similarity to two peptide leads. Biological activity, rather than a computed fitness, has been used as the score in a directed combinatorial synthesis program (100). Brown and Martin developed GALOPED (101) as a way to design combinatorial mixtures. The SELECT program (78c) combines measures of diversity and the physical properties of the designed library. The library can be designed to be both internally diverse and diverse with respect to a reference population. Physical properties are optimized by comparing to a user-defined profile for the property of interest, c log P for example. As for the HARPick approach (78a,b), however, it is necessary to define a weighting scheme between the different elements of the score, which leads to a number of difficulties. Selec-

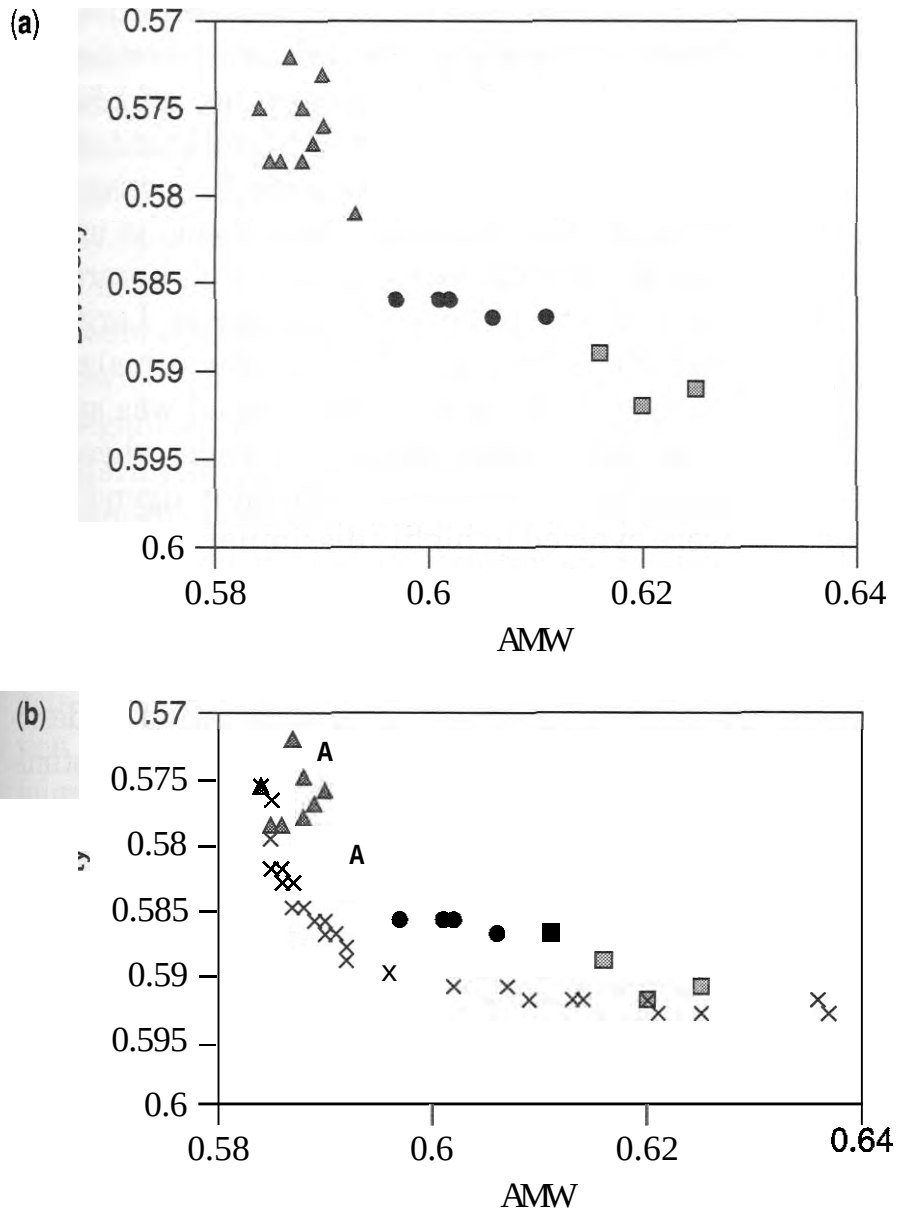


Figure 5.14. (a) Results from multiple SELECT runs with alternative weightings for molecular weight vs. diversity. Filled triangles, 1.0×Div and 1.0×MW; filled circles, 1.0×Div and 0.5×MW; filled squares, 10.0×Div and 1.0×MW. (b) As in a, with results of a single MOGA run shown as crosses. [Reproduced from V. Gillet, et al., **J. *Chem. Inf. Comput. Sci.*, 42, 375–385** (2002) with permission of the American Chemical Society.]

tion of the weights is nonintuitive when comparing different properties or concepts (e.g., diversity and c log P) and the use of weights constrains the search space. Figure 5.14a illustrates how changing the function weight in a SELECT run alters the final solution; note also that the two objectives, molecular weight and diversity, are competing in this example. Given these limitations, a novel modification has been made to the SELECT methodology. The GA in SELECT is replaced by a multiobjective GA (MOGA) (102) that eliminates the need for a weighting scheme. Instead, each element of the scoring function is optimized independently and solutions scored according to the idea of dominance (see Fig. 5.15). The solutions of rank 0, the nondominated solutions, are those solutions for which there is no superior solution when considering all objectives; solutions of rank 1 would be dominated in one objective and so on. It is these ranks that are used to describe the fitness of the solution. Solutions of rank 0 are said to define the pareto surface, as displayed in Fig. 5.14b, which overlays the MOGA results onto the SELECT results. Thus, the MOGA has several advantages over a traditional GA. In one run it generates multiple solutions that are equally valid, more fully explore solution space, and gives the designer an understanding of the relationships between the different objective functions.

The RBPB algorithm of Pearlman and Smith, described above (Section 4.2), considers all possible candidate libraries, which satisfy the user's constraints regarding economy. These include min/max range constraints regarding library size (e.g., number of AB products) and the number of each type of reactant (e.g., number of A's and number of B's). These also include specification of the minimal unit dimensions (MUDs), which define the smallest combinatorial array that the user is willing to address on the robotic table. Each candidate library corresponds to a different way of, at least conceptually, arranging the required

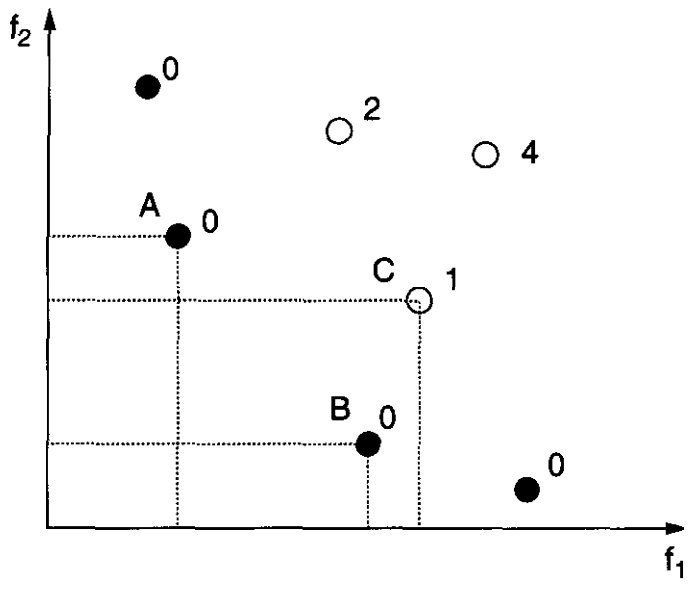


Figure 5.15. Pareto optimality. The filled circles represent rank zero or nondominated solutions for functions *f*1 and *f*2. Point C is rank 1 because it is dominated by point B. (Permission as in Fig. **5.14.)**

number of MUDs to construct a library within the user's specified range limits. Each candidate library is scored based on an appropriate function of the scores of the individual products it contains divided by its size; hence, an average product score. The reactants used to make each candidate library are determined by reactant scores, which are functions of product scores and which are updated at each step of the design of that particular candidate library. For example, at a given step in the design process, the reactant score for reactant A_i depends on the scores of the products actually accessible, given the current choices of B-type reactants. The score also depends on the scores of the products that could be made using B-type reactants, which may be selected at a subsequent step in the process. The candidate libraries with the highest scores are output for the user's final decision. In addition to be being remarkably thorough yet fast, the RBPB also makes it very easy to address the iterative nature of library design and to suggest replacements for previously suggested reactants that had to be rejected for one reason or another.

A rapid computational method for lead evolution has been described by the CombiChem (now DeltaGen) group (39). Their 3D computational approach for lead evolution is based on a pharmacophore fingerprint approach using multiple pharmacophore hypotheses. A set or ***ensemble*** of hypotheses is generated that is most able to discriminate between active and inactive molecules. The ensemble comes from an analysis of a large number of pharmacophore hypotheses, with full conformational sampling for both active and inactive compounds. The ensemble hypothesis is used to rapidly search virtual chemical libraries to identify compounds for synthesis. Large virtual libraries (e.g., a million structures) can be analyzed efficiently. The method was applied to a,-adrenergic receptor ligands, where heterocyclic a,-adrenergic receptor ligand leads were evolved to highly dissimilar active N-substituted glycine structures.

LiBrain (103) is a collection of software modules for automated combinatorial library design, including the incorporation of desirable pharmacophoric features and the optimization of the diversity of designed libraries. A Chemistry Simulation Engine module is trained by chemists to determine the suitability of reactants for a specified reaction, to recognize the risk of undesirable side reactions, and to predict the structures of the most likely reaction products, so as to circumvent major bottlenecks associated with automating the process.

Legion and Selector (66c,d, 104) are software from Tripos (13c) for characterization, comparison, and sampling of sets of compounds, including a combinatorial builder (104), with available descriptors including fingerprints and atom-pair distances. Clustering tools (Hierarchical, Jarvis-Patrick, and Reciprocal Nearest Neighbor) and compound selection and diversity comparison methods available include Tanimoto Dissimilarity, the Reciprocal Nearest Neighbor approach, and the OptiSim algorithm (see Section 2.2.3).

4.4 Handling Large Virtual Libraries

The rate-limiting step in a product-based library design process is often the calculation of molecular descriptors. This becomes particularly acute as one moves into the 3D arena, of course, but even the simplest 2D descriptors take a finite time to calculate. In addition, there are the logistics associated with storing virtual libraries of potentially tens of millions of compounds. The ability to search within the possible chemical space of a particular chemistry, as opposed to the limited space of syn-

thesized compounds, is an important component of lead identification because this allows a weak hit from primary screening to be rapidly expanded into a more potent lead. Existing chemical database systems can be used or readily modified to benefit from the combinatorial nature of libraries (64a) but they do not overcome the fundamental issues.

Downs and Barnard (105a) have proposed an elegant solution to these problems using the Markush representation commonly used in chemical patents. The key component of their approach is that descriptor calculation and diversity analysis can be performed without the need for full enumeration of the products. In other words, both storage and calculation will tend to scale as the sum of the number of building blocks in the library rather than the product as in techniques requiring enumeration. The method has been developed into a software suite and released commercially as the LibEngine module of the Cerius^2 suite for combinatorial library analysis and design (105d).

The background and theory behind the approach have been published (105b). In summary, the algorithm relies on identifying a *core* and associated R-groups that define the library. This may or may not be directly related to the manner of synthesis. For example, imagine a tripeptide library synthesized from $20 \times 20 \times 20$ amino acids. The algorithm detects the tripeptide backbone as the core and the amino acid side chains as the R-groups. Partial substructure fingerprints are calculated on a fragment basis representing the core and R-groups taking full account of the attachment to the core and the possibility that a particular path may extend between two R-groups. The partial fingerprints are then combined into the full fingerprint, a relatively fast exercise. The approach is a couple of orders of magnitudes faster than calculating fingerprints from fully enumerated products. Additive molecular properties such as molecular weight, hydrogen-bond donor and acceptor counts, and log *P* can be calculated in a similar manner as well as topological indices. Fingerprints or property data can also be calculated on demand for use with clustering algorithms, thus avoiding the overhead of storing and retrieving them. There is also interest in extending the approach to 3D property calculation (105c).

An alternative approach has been taken by Agrafiotis and colleagues. In a conference presentation (106) they show how a neural network can be trained on a small sample of enumerated combinatorial products to reproduce 2D molecular descriptors and properties for all library members without the need to construct their connection tables.

A method for rapid similarity searching in large combinatorial spaces using a new algorithm Ftrees-FS was published by Rarey and Stahl (135). The similarity search is based on the feature tree similarity measure representing molecules by tree structures. Combinatorial chemistry spaces are handled as a whole rather than looking at subsets of enumerated compounds. A set of 17,000 fragments of known drugs was used, which could be combined to 10^{18} compounds of reasonable size. A novel ChemSpace approach (45a) for searching large virtual libraries that does not require enumeration has also been developed by Tripos, using shape descriptors (topomeric fingerprints) on the monomers, and has been used for targeted library design (45b).

4.5 Library Comparisons

In the previous sections we described the design of libraries based on a number of user-defined criteria, whether they were focused or whether they were of a more general nature. So far, these designs have been undertaken, treating the library in isolation, with the inclusion of property profiles in methods such as HARPick and SELECT to ensure that the synthesized compounds are of a suitable physical nature. In this sense, the designed library can be said to be internally diverse; that is, the selected compounds are diverse within the limited chemistry space of all virtual products. Even for very large virtual libraries, the chemistry space is still small with respect to the possible chemistry universe. It is very difficult, *a* priori, to address how "diverse" a designed library is compared to a library generated with another set of reactions without having to go through the computationally expensive process of computing all pairwise similarities between members of the libraries.

Nevertheless, questions such as "How diverse is the library compared to the screening collection?" or "Which of the following chemistries should I choose for a library?" are often posed and methods are required to answer them.

Distance-based methods such as clustering can be and have been used but suffer from a number of drawbacks both in terms of speed and the fact that the exercise needs to be repeated for every additional library (i.e., there is no common frame of reference). In addition, all pairwise comparisons would need to be performed. Thus, Shemetulskis et al. (107) used clustering methods to compare the Parke-Davis corporate collection (117,000 compounds) with external compounds from Chemical Abstracts Service (380,000) and Maybridge (42,000). Even today, clustering half a million compounds is a daunting task and interpreting the results is not straightforward. The Jarvis-Patrick method employed by Shemetulskis et al. has several input parameters, including the need to predefine the number of clusters. Voigt et al. (108) compared the National Cancer Institute (NCI) database, a publicly available database of compounds used in the NCI screening program, to a number of compound databases. The diversity of each collection was estimated by the number of compounds selected by use of a diversity-selection algorithm as a function of database size. The similarity overlap between two databases has been determined by calculating the percentage of compounds of the first database for which a compound exists in the second database with a similarity greater or equal to a specified cutoff (109). Such an approach necessitates the calculation of the Tanimoto similarity coefficient of all compounds in a database with all compounds in the other databases. As indicated before, the largest drawback of distance-based methods is that they give no indication of where the voids are within the chemistry space, and searching an additional compound source for interesting compounds would require reclustering.

Therefore, partition/cell-based methods are preferred for such library comparison tasks. They provide a common frame of reference in which it is possible to identify voids within the chemistry space of a population. It must be emphasized that the chemistry space is still defined with respect to a reference population. By comparing the libraries with reference to a population (REFDB), such as a corporate database or a combination of known drug databases, one can make statements such as, library A shows the greatest overlap with REFDB, whereas library B fills the greatest number of empty or low occupancy cells. Cummins et al. (22) used a cell-based approach to compare five databases, including the Wellcome Registry, to select screening sets of diverse compounds. Topological indices and a measure of free energy of solvation were taken as the descriptors and factor analysis was used to combine them and define a four-dimensional chemistry space that was then partitioned. Outliers were removed to allow the partitioning to focus on the most densely populated region. The use of pharmacophore descriptors in such a task was illustrated by Mason and Pickett (4), where the pharmacophore overlap between three libraries was calculated. It was possible to identify the library covering regions of pharmacophore space not covered by the other two. Alternatively, given that library A is synthesized and gives hits in screening, then presumably the library that overlaps best with A should be made. Pearlman and Smith (11d) have adapted their DVS software to identify what they term a ***receptor-relevant subspace,*** where the BCUT metrics are selected to best group the active compounds within a population (in fact, it is possible to have several groupings of actives within the space) (see Section 2.2.1.1).

Comparing two populations by pharmacophore coverage, although straightforward, does ignore the contribution from individual compounds. This is important, in that two libraries could cover similar regions of pharmacophore space but individual compounds in the two libraries could be displaying different subsets of the total pharmacophores covered. This prompted Pickett et al. (70) to explore an alternative approach. In this case, a number of potential scaffolds were available and the aim was to find which of these would best complement previously synthesized libraries. Virtual libraries were generated using a predefined set of reactants and pharmacophore fingerprints were calculated for these and the previously synthesized libraries. By use of mea-

sures proposed by Turner et al. (69b), the virtual libraries were compared to the synthesized libraries at both a whole library level and an individual molecule level. From this analysis it was possible to select the scaffold that best complemented the previously synthesized libraries.

An alternative methodology based on the ring content of a database, using precalculated structure-based ***hashcodes*** has been proposed (110). The comparison of the hashcode tables can be used to compare two databases and the number of distinct ring-system combinations can be used as an indicator of database diversity. A method for diversity assessment called the ***saturation diversity*** approach, based on picking as many mutually dissimilar compounds as possible from a database was also proposed. The methods were used to compare a number of public databases and gave similar results.

4.6 Pharmacophore-Based Fingerprints

The examples of GPCR library design (described in Section 5.1.2) and protein-site design for Factor Xa (described in Section 5.3) illustrate the use and relevance of pharmacophore-based fingerprints in library design. A pharmacophoric bias has been a major component of many library designs (111), used in the context of focused or biased libraries. Their broad applicability is important, with the same descriptors being used for diverse library design, screening set selection, and focused library design. This provides a consistent approach that extends to protein-site based pharmacophores as discussed above. Their ability to determine the similarities and differences between structurally diverse molecules and sites is very powerful. An ensemble pharrnacophore data set measure is often used, which attempts to condense the individual molecule pharmacophore fingerprints into a single measure that describes the important features of the data set as a whole (36, 37, 78a,b).

McGregor et al. (112) have recently published a version of pharmacophore fingerprinting (the PharmPrint method) applied to QSAR and focused library design that uses a limited basis set of 10,549 three-point pharmacophores. They included the usual six pharmacophoric features, plus an additional definition of ***other*** for all remaining unassigned atoms. A subset of the MDDR database (13a) was used to define a reference set of bioactive molecules, separated into target classes (gene families). The discriminating power of several molecular descriptors was measured using the target class assignments for this set, and it was found that the pharmacophore fingerprint outperformed other descriptors.

4.7 Combined Pharmacophore Fingerprints and BCUTs

Library design using a simultaneous optimization of BCUT chemistry-space descriptors (11) and four-point pharmacophore fingerprints has been reported (32d, 37d). The authors investigated the feasibility and results in terms of complementarity of simultaneously optimizing two product-based descriptors for reactant selection from large virtual libraries. Diversity around a chosen chemistry was the goal of the studies reported, but the approach could equally be applied to optimize to a desired distribution of properties, say, from sets of biologically active compounds. A simulated annealing algorithm (97) was used to combine both components in a single optimization procedure. The choice was based on the ease of implementation and the ability to include multiple components in the objective (23d), an important goal in many recent designs, if only to modulate physicochemical properties to druglike ranges. In this example a small, fully enumerated virtual library of 86,140 amide compounds was constructed from carboxylic acids and primary amines present in the ACD (Available Chemicals Directory). The products of the optimized and random starting reactant sets were compared using average nearest-neighbor distances, and the Hopkins' statistic (113), which evaluates the degree of clustering in a data set, together with the four-point pharmacophore fingerprint diversity. The potential utility for very large virtual libraries, where precalculation of all the pharmacophore fingerprints would not be feasible, was illustrated by calculating four-point pharmacophore fingerprints for virtual library compounds on the fly. The fingerprints were calculated during the optimization procedure and stored in a compact encoded form, with

previously calculated fingerprints reused as needed (calculation times ~1–5 s with conformational sampling per structure on an SGI R10000 machine). Diversity was evaluated for the BCUT chemistry space using the ratio of filled to possible filled cells for the virtual library. Four-point pharmacophore diversity was evaluated by the number of unique pharmacophores and the total number of pharmacophores in the product subset, with the goal to optimize both the pharmacophoric uniqueness of each compound selected and the total number of pharmacophores exhibited. Encouraging results were obtained, with additional work necessary to develop a more general function.

4.8 Oriented-Substituent Pharmacophores

OSPPREYS is a pharmacophore diversity descriptor developed specifically for combinatorial library design by Martin and Hoeffel (43). Advantage is made of the common scaffold, so calculations are performed on the sets of substituents. This enables a more detailed pharmacophoric description of the library products than through calculations that could be practically performed on the enumerated products. By avoiding the problems of having to analyze many products with many conformations per product, and an explicit dependency on the scaffold, a higher spatial resolution could be obtained. The analysis of enumerated combinatorial libraries by pharmacophoric methods is generally limited to smaller virtual libraries, with three- or four-point pharmacophores, and limited conformational sampling, requiring new calculations for every library. The oriented-substituent pharmacophores (OSPs) were developed as a compromise approach between reactant and product-based methods to rectify these limitations. To recapture most of the orienting information that is lost in fragmenting the enumerated products into substituents, two additional points are added to each ordinary one-, two-, and three-point substituent pharmacophore, necessitating approximations through the ***combinatorial conformer*** and the ***template alignment*** assumptions. The OSPPREYS analysis does, however, account for up to nine-point pharmacophore similarity in the products of a library with three diversity sites. In addition, the consideration of relatively rigid substituents reduces the number of structures to analyze by up to 10^{10} compared to that of a full product-based analysis. This permits a thorough conformational sampling of very large virtual libraries that would be too slow on enumerated structures. A Euclidean property space for diversity analysis is possible because of the small number of pairwise substituent similarities, enabling options not possible by counting set bits in a library union fingerprint. The database of oriented substituent fingerprints is transferable between libraries, within the restrictions of the noted approximations. A major limitation in using OSPPREYS is that it can be applied only within a combinatorial library, and not between libraries. OSPPREYS is well suited to maximizing the diversity of scaffolds independently, and can be used to build a screening file based on such diversity.

4.9 Integration

The previous sections have outlined the basic methodology that has been developed in the areas of molecular diversity, similarity analysis, and library design. Traditionally, use of these methods was limited to a small number of exponents within a computational chemistry group because it involved bringing together a diverse set of software tools and data sources. Combinatorial and high throughput chemistry is now well integrated into the research process and there is a need for bench chemists to have access to such tools. The Cousin system developed at Upjohn has been in use since 1981 (114) and has recently evolved to the ChemList system. The system includes tools for the browsing of dissimilar compounds from substructure searching, useful in reactant selection, for example. Gobbi et al. (115) have described the development of the CICLOPS system in use at Novartis. The system provides functionality for designing and registering libraries and associated tasks such as accessing reactant availability. Tools are provided for filtering reactant lists and selecting a diverse subset of reactants if required. The system is PC-based and is built around the Daylight chemical information system (13b) and associated tool kits with custom Windows clients to control the process.

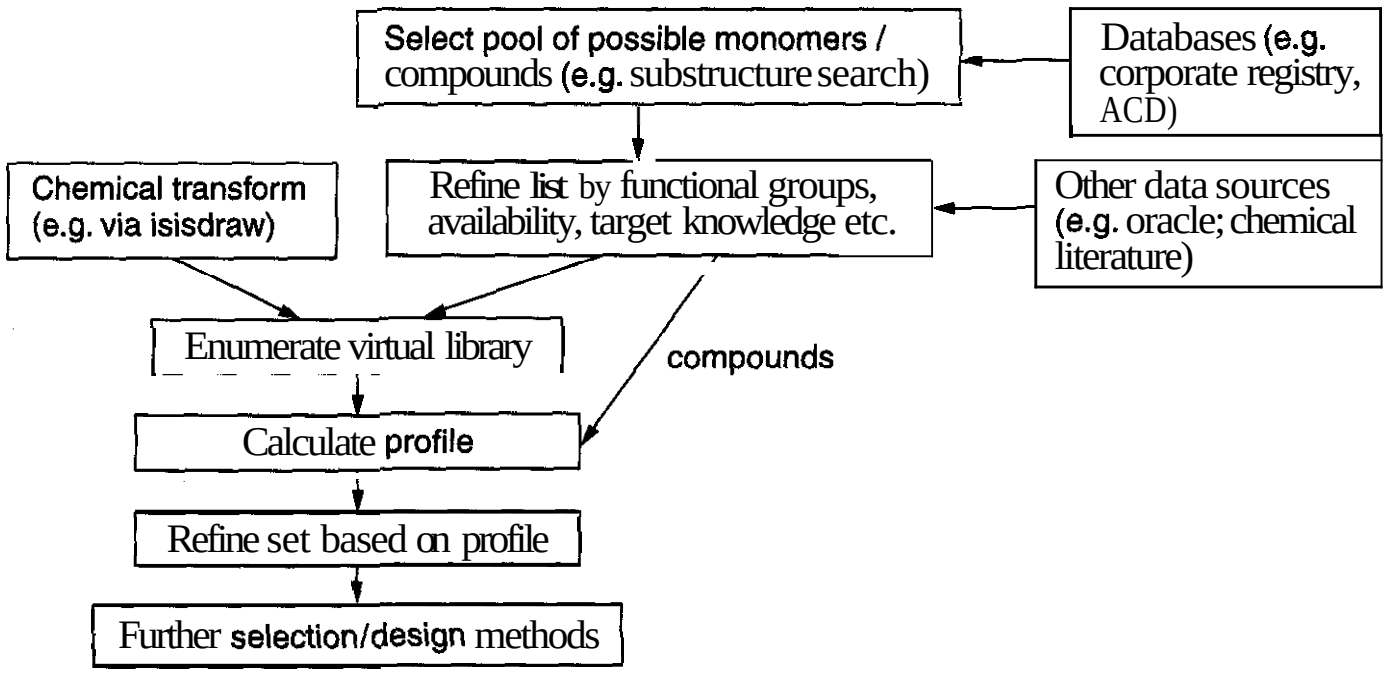


Figure 5.16. The workflow used within ADEPT (A Daylight Enumeration and Profiling Tool; GlaxoWellcome, UK) for compound selection and library design. [Reproduced from A. R. Leach and M. M. Hann, Drug Discovery *Today*, 5, 326–336 (2000), with permission of Elsevier Science.]

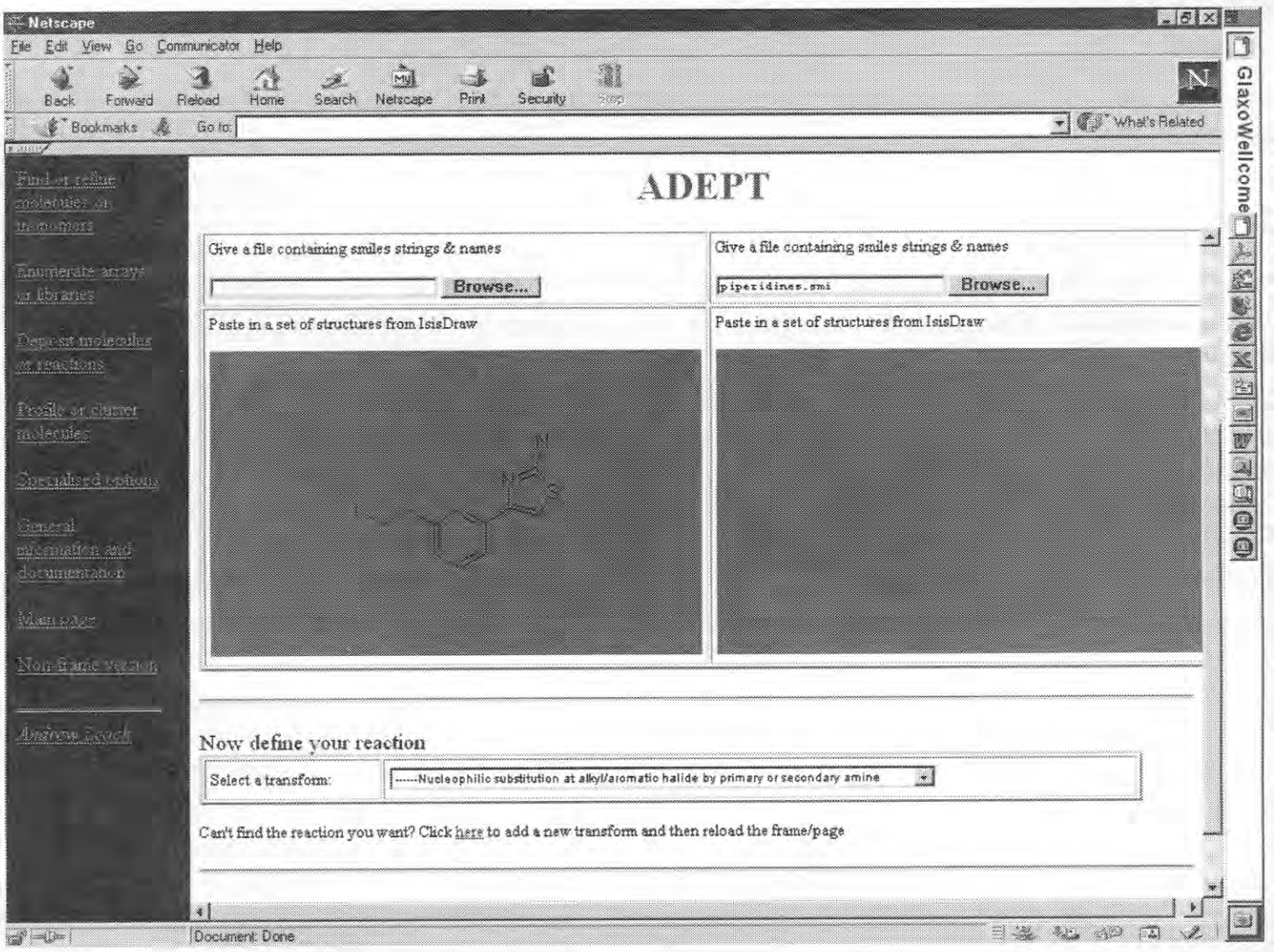
The ADEPT (A Daylight Enumeration and Profiling Tool) suite of programs developed at GlaxoWellcome (116) is a Web-based system providing access to a wide range of library design functionality, again based around the Daylight tool kit. Figure 5.16 provides an outline of the process workflow. Reactant lists are generated from searches in databases of in-house and commercially available monomers. A variety of filters can be applied to reduce the size of the lists. These include filters on molecular weight, rotatable bond count, and substructure filters to remove unwanted functionality. After library enumeration, various property histograms are calculated. This allows the user to further refine the reactant choice.

A product-based library design algorithm, PLUMS (117), has been developed to ensure that combinatorial constraints are satisfied in the design. The algorithm successively removes the monomer that adds least value to the library as governed by two terms, the effectiveness (number of molecules meeting user-defined criteria such as property ranges, fit to pharmacophore or dock to protein site) and efficiency (ratio of effectiveness to library size). The algorithm is sufficiently fast to work within the Web-based environment of ADEPT. Figure 5.17 shows screen shots from ADEPT, illustrating how a library can be specified and the resulting product histograms. A similar system has been implemented at Vertex (118a). A key component of this system is the REOS filtering tool (118b), which applies filters on molecular weight, lipophilicity, unwanted substructures, rotatable bond counts, and so forth to remove "obviously bad" compounds.

4.10 Structure-Based Library Design

Structure-based library design uses 3D structures of the biological targets to direct the design and selection of templates/scaffolds and of reactants that will produce compounds that can fit into the target and thus are likely to bind and have biological activity. The experimental structural information *can* be derived by a structural biology approach, using X-ray crystallography or NMR spectroscopy. Computational models can be built and used (e.g., homology modeling techniques for closely related proteins), but an experimental structure is always preferred. A structural biology approach can also be used to identify molecules or fragments thereof that bind to a target. For example, NMR screening (3) can be used to identify potential scaffolds or reactants for a combinatorial library that bind to a target site and is able to detect very low affinity binding (in the millimolar range, compared to the low micromolar range from biological screening); this *can* be done without the need to determine the 3D structure of the target.

(a)



(b)

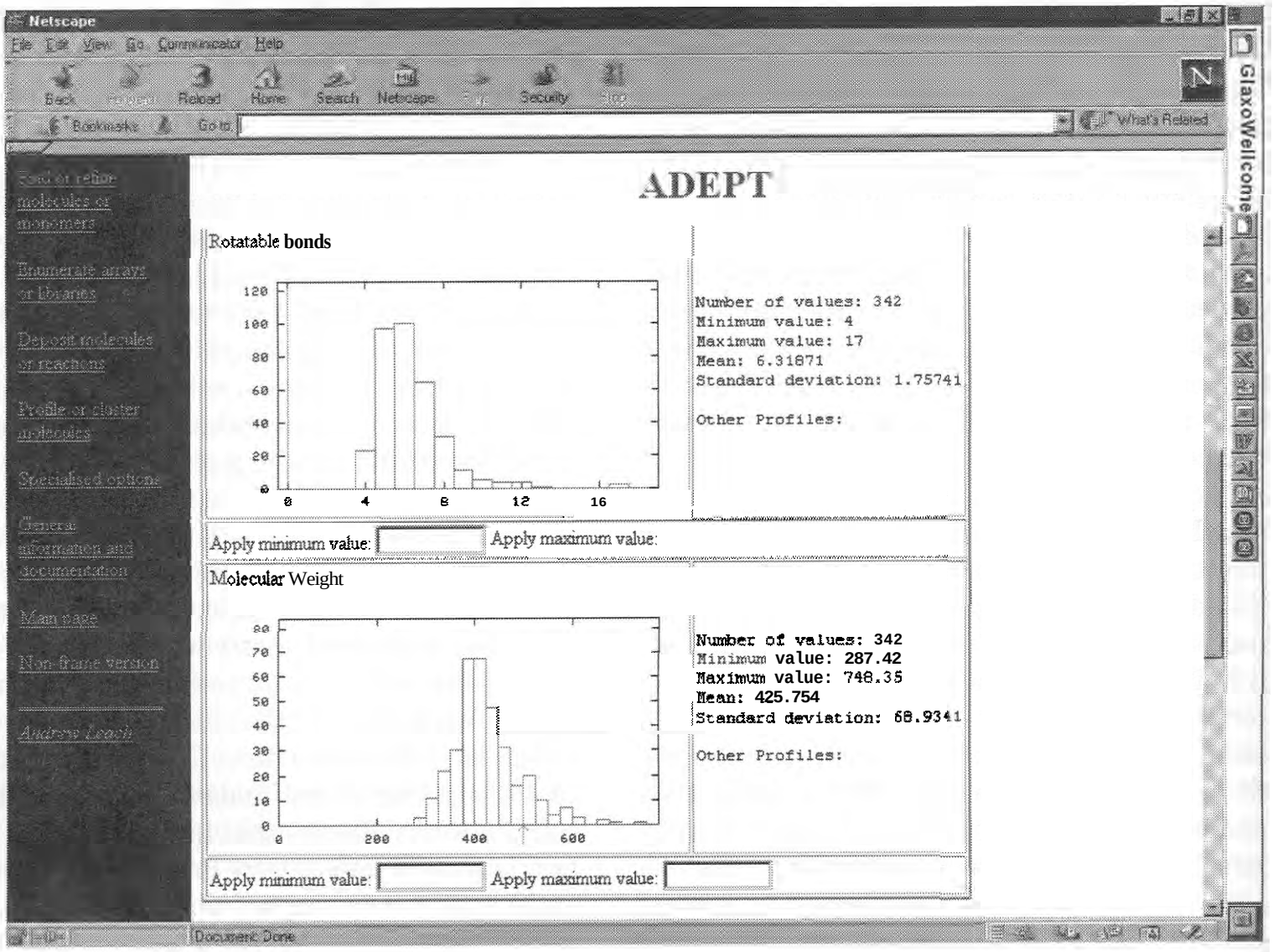


Figure 5.17. Screen-shots from ADEPT. (a) A simple two-component library composed of an aminothiazole template and a series of piperidines specified with ADEPT. (b) Histograms of rotatable bonds and molecular weight for the enumerated virtual library, aiding the medicinal chemist in the design of the library. [Reproduced from A. R. Leach and M. M. Hann, *Drug Discovery Today*, 5, 326–336 (2000), with permission of Elsevier Science.]

Structure-based drug design (SBDD) is the topic of another chapter, and key issues such as the scoring functions for the ligand-receptor interaction are not discussed further here. The ability to combine SBDD with combinatorial chemistry enables a focused design approach that can explore a range of ideas, reducing the dependency on SBDD limitations (structural information, scoring, conformational sampling, etc.). The ability to obtain the X-ray or NMR structure of new potent molecules complexed with their targets can also be critical for the next iteration, in that computational structure-based design methods may be unable to predict alternative and new binding modes, especially because the protein site is normally kept rigid and unpredicted conformational changes can take place during the binding process. A review by Stahl (119) discusses the technology that directly uses receptor three-dimensional structures, discussing relevant topics such as scoring functions, receptor-ligand docking, and practical applications. Bohm and Stahl (120) have reviewed structure-based library design in terms of molecular modeling merging with combinatorial chemistry.

The synergy between combinatorial chemistry and *de novo* design has been discussed by Leach et al. (121). They present an approach wherein a template (corresponding to the *central core* of a combinatorial library) is positioned within an acyclic carbon chain whose length and bond orders are systematically varied. The conformational space of each resulting structure (core plus chain) is explored, to determine whether it is able to link together two or more strongly interacting functional groups or pharmacophores located within a protein binding site. In a second phase, 2D queries are derived from the molecular skeletons and used to identify possible reactants from a database that would enable the all-carbon linking chains to be replaced by more synthetically feasible groups.

Sheridan et al. (122) have published on designing targeted libraries with genetic algorithms, extending earlier work, to use the GA with 3D scoring methods and showing that the approach of assembling libraries from fragments in high scoring molecules is a reasonable one. Example applications to two situations are described: (*1*) where the 2D structure of some actives (diverse angiotensin II antagonists) is known, with the goal to design a library that best resembles the actives; and (*2*) to simulate the situation where an active site (stromelysin-1 in this case) is available and the requirement is to design a library of structures likely to bind to it.

Tondi (123) discusses several examples in which structure-based drug design and combinatorial library synthesis have worked successfully together in a complementary way. These include the discovery of:

- Potent nonpeptide inhibitors of cathepsin D (124), which uses CombiBUILD (125), a derivative of the DOCK (126a,b) approach, with this structure-based selection approach yielding seven times as many hits as a diversity-based procedure.
- Thrombin inhibitors (127), where Böhm et al. used LUDI to dock and score computationally available primary amines and then score the virtual library generated from benzaldehydes with the top-scoring hit.
- Novel inhibitors of matrix metalloproteinases (128): Rockwell et al. (128a) used a combinatorial library at the beginning of the work to suggest leads suitable for further optimization that required a conformational change at the binding site, and a structure of the complex to enable iterative optimization; Szardenings et al. (128b) used SBDD to design the starting scaffold, with synthesis guiding the introduction of diversity.
- Thymidylate synthase inhibitors (129), using DOCK to identify the starting lead.

The CombiDOCK program (126c), based on DOCK, enables the evaluation of very large virtual libraries by using structure-based combinatorial docking. Multiple docked orientations of the scaffold are used to evaluate reactants separately at each of the substitution positions. The total docking score for each product is rapidly estimated by summing the contributions from reactants at each position (which are attached as in the final product to the docked scaffold, which may be a computationally convenient anchor fragment formed during the reaction rather than a syntheti-

cally used chemical). Further checks are made for the highest scoring structures (e.g., for steric interactions between reactants at the different substitution positions). This approximation produces an enormous speed-up over docking all the individual compounds, which, from a time perspective, rapidly becomes prohibitive for large combinatorial libraries. From the scores it is possible to select combinations of reactants that produce compounds complementary to the protein binding site. Combinatorial restraints can be applied as required to obtain the most efficient use of reactants and robotics, with an evaluation of any reduction in the inclusion of higher scoring compounds.

Different strategies for combining diversity and structure-based design in site-focused libraries and the DOCK-based CombiBUILD algorithm are discussed in a review (125), as an example of how lead compounds can be rapidly identified by combining diversity with structure-based design in site-focused libraries.

Lamb et al. (130) have published on the design, docking, and evaluation of multiple libraries against a family of targets, using a similar divide-and-conquer algorithm for side chain selection that enables the exploration of large lists of reactant substituents with linear rather than combinatorial time dependency. The method consists of three main stages: (1) docking the scaffold, (2) selecting the best substituents at each site of diversity, and (3) comparing the resultant structures within and between the libraries. The scaffold docking procedure, in conjunction with a novel vector-based orientation filter, was shown to be effective for several protease targets, reproducing experimental binding modes.

The application of the powerful combination of SBDD and combinatorial chemistry is not limited to lead discovery or the optimization of potency, but also to the optimization of the selectivity (using knowledge of the structures of related targets) and pharmacokineticl druglike properties of a molecule. For example, the structure of a ligand-receptor complex can clearly indicate areas where chemical modifications could be made to modulate these other properties, without directly affecting binding/potency. Models/structures of ligands with the cytochrome P450 metabolizing enzymes are also now becoming available.

5 EXAMPLE APPROACHES

5.1 General Target Class-Focused Approaches

5.1.1 Defining the Chemical/Biological Space. The design of target class (gene family) libraries or compound subsets requires the definition of a biologically relevant chemical space. This "biological" space can then be used for the design and selection of biased/focused libraries and compound subsets. Many approaches can be taken, adapting the use of a wide variety of similarity/diversity descriptors (discussed in Section 2.1) to the identification of properties associated with a particular target class or subset thereof. The goal is to identify a feature or set of features that, ideally, is specific, but more generally "enriched" for the target(s) of interest. A common approach is to identify chemical substructures that are characteristic for the target class, and use these for the design. The simplest approach is to include such substructures in the library, but the co-occurrence of other features is often needed, and the quantification of this provides an enhanced design. An example of this combined approach is discussed in the next section, using the pharmacophore fingerprints expressed relative to "privileged" substructures. This provides a convenient cell-based partitioning approach. Alternatively, it is possible to identify properties that are enriched for a particular target class, without reference to any particular substructures: 1D (e.g., physicochemical), 2D (e.g., ISIS keys, BCUTs), and 3D (e.g., pharmacophore fingerprints) properties can all be used. BCUTs have been used within a target (to identify a receptor-relevant subspace, in which actives cluster), to differentiate within a target class (e.g., ion channel openers vs. blockers) and for general target class analysis. BCUT chemical space provides a way to quantify the "diversity" of certain properties within actives for a target class, as well as to identify any particular combination of properties that actives share. BCUTs have

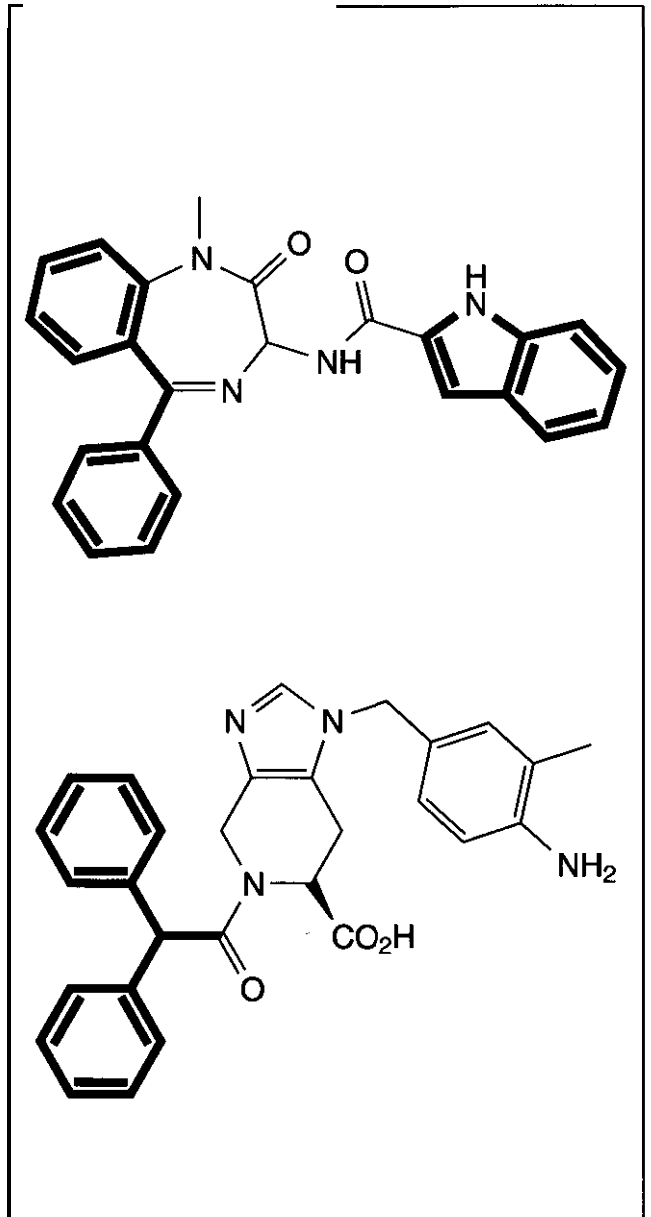
been used to select representative subsets from libraries biased to a target class (59a).

5.1.2 7-Transmembrane C-Protein-Coupled Receptors. Examples of a product-based combinatorial library design that use four-point pharmacophore fingerprints in a "relative" diversity mode have been described for the design of combinatorial libraries that contain "privileged" substructures focused on 7-TM GPCRs. These are a large family of very important biological targets lacking high resolution experimental 3D structures of the human targets; therefore most design has focused around the ligands. The occurrence of common "privileged" substructures for 7-TM GPCRs, often spanning several targets, provides a useful focused design method. Some example structures are shown in Fig. 5.18.

A useful modification was made to the standard pharmacophore descriptor evaluation (37, 80) by forcing one of the points in the pharmacophoric description to be a privileged substructure. This provides a novel quantification of all the 3D pharmacophoric shapes, and thus important 3D information relevant to the biological activity of the ligands, relative to the substructure. This builds on the ability of 3D pharmacophoric descriptors to separate chemical structure from ligand binding properties, and enables a fingerprint to be generated that describes the possible pharmacophoric shapes from the viewpoint of that special point/substructure (see Fig. 5.19). A relative or internally referenced measure of diversity is thus created, enabling new design and analysis methods (see Section 2.2.5). The goal of the published method was to design novel structures, accessible through combinatorial chemistry, that have one or more privileged substructure reactants/cores, and are enriched in the relative 3D pharmacophoric shapes of known ligands. The method identifies patterns with other key features that need to be present with the privileged substructure, such as acids and bases. The optimization can also include an enrichment in pharmacophoric shapes containing the privileged substructure that are not in existing structures, enabling the exploration of new 3D pharmacophoric diversity focused around a feature known to be important for biological activity.

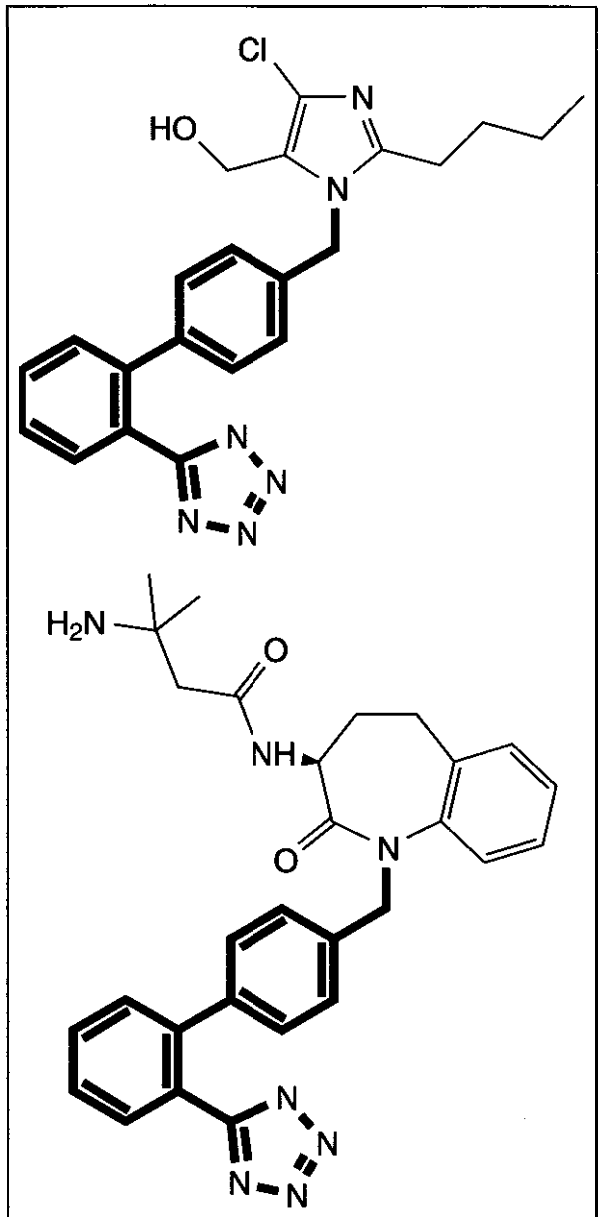
The Ugi reaction (131), a four-component condensation reaction, was chosen and more than 100,000 compounds were synthesized. Privileged substructures such as biphenyl tetrazole were used, for example, at the amine position (see Fig. 5.20). Other GPCR privileged substructures such as diphenyl methane, biphenyl tetrazole, and indole were used to focus the pharmacophore descriptors (see Figs. 5.18 and 5.21). GPCR ligands reported to be active at receptors with peptidic endogenous ligands were identified from the MDDR (13a). These compounds were used to provide the reference data for the design by calculating the union pharmacophore fingerprint of compounds containing the privileged substructure (see Fig. 5.22 for an example structure). A virtual combinatorial library was then created, and for a particular reactant position, the privileged pharmacophore fingerprints were calculated for each candidate reactant over all the products that would be generated if it were used in the library. Either previously selected or a representative set of reactants were used for the other three components to generate the virtual Ugi products.

The combinatorial library was then designed by comparing for each reactant the fingerprint generated from the resultant products with the fingerprint for the known drug ligands (MDDR-fingerprint). Reactants were selected by identifying, on a position-by-position basis, reactants that gave products that matched the greatest number of these MDDR-exhibited privileged pharmacophores. The design goal was recalculated after the selection of each reactant, by removing the pharmacophores matched by the products generated by that reactant from the target list. Subsequent reactants were thus picked based on their ability to match the remaining pharmacophores. The approach used was to select the first reactant as the one that would give library compounds with the most number of privileged pharmacophores in common with the drug set. The process was continued until no more reactants could be found that contributed a nontrivial number of new privileged pharmacophores. Optimization methods such as the HARPick approach (described in Section 4.3) could be used to enable other properties, such as flexibility and physicochemical



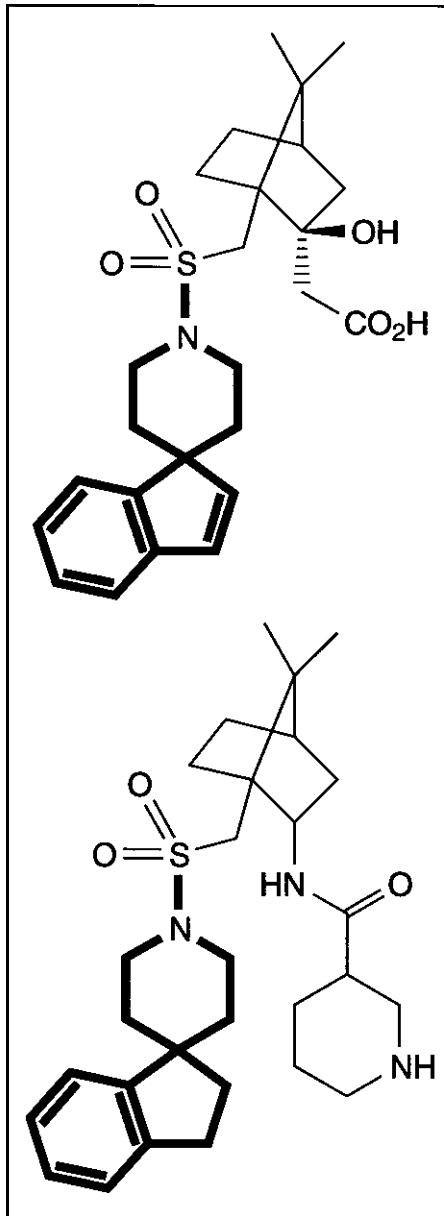
MK 239
CCK-A antagonist

PD 132,177
AT-2 ligand



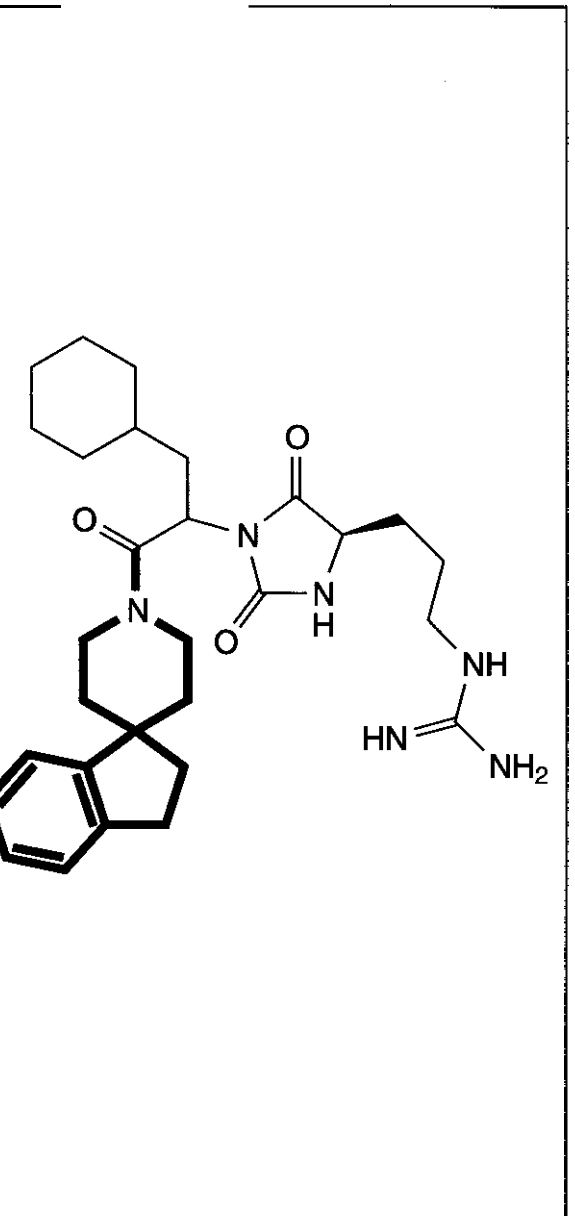
Losartan
AT-1 antagonist

MK-0751
GH secretagogue agonist



L-366,509
oxytocin antagonist

L-368,112
GH secretagogue agonist



Novartis
C5a agonist

Figure 5.18. Examples of 7-TM GPCR "privileged" substructures highlighted on known active compounds. Note that these often occur together with another characteristic feature (e.g., basic or acidic group).

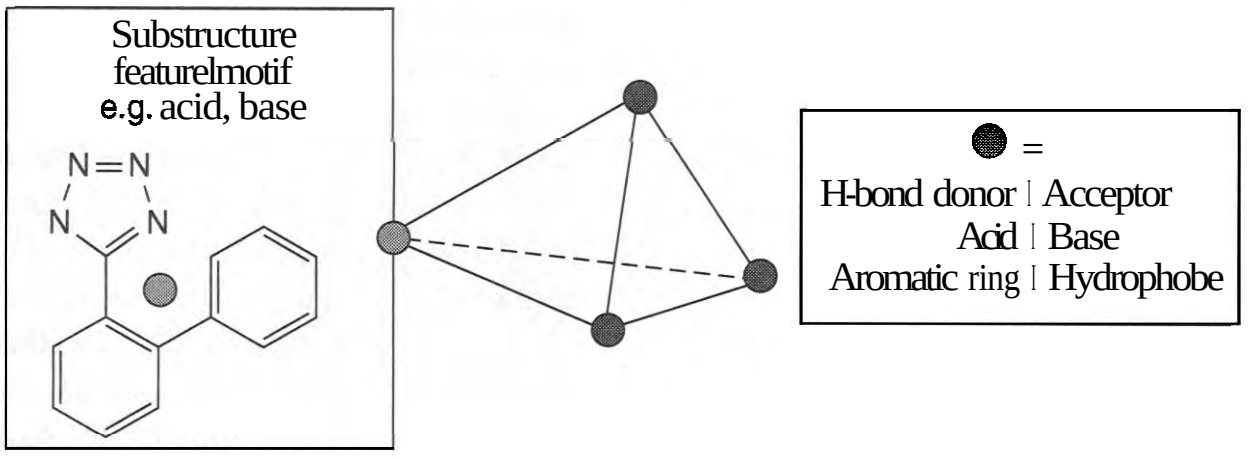


Figure 5.19. Example of a "privileged" four-point pharmacophore. Here biphenyl tetrazole, a substructure seen in a number of GPCR inhibitors, is specifically defined as a pharmacophore feature, using a centroid dummy atom. Only pharmacophores that include this type are included in the fingerprint, thus providing a relative measure of diversity/similarity with respect to the privileged feature.

properties, to be optimized also. The total number of pharmacophores (this time without reference to the privileged substructures) can also be monitored and optimized. Example results from one of the Ugi library optimizations are shown in Fig. **5.23.**

This design illustrates an advantage of a partitioning (cell-based) approach. The pharmacophore fingerprint can be used to monitor progress, to quantify how much of the desired goal has been accomplished, and to evaluate whether a given chemistry can yield further compounds that match the design criteria and/or explore new pharmacophoric space.

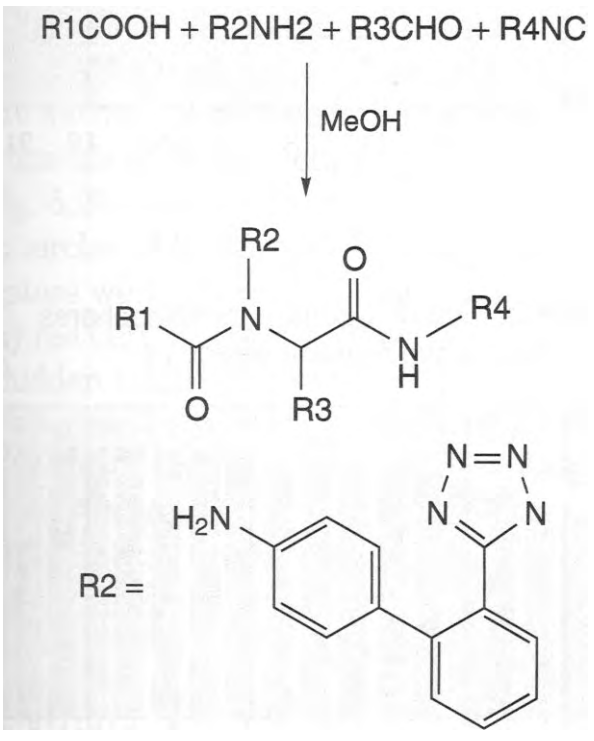


Figure 5.20. Example of the Ugi chemistry with biphenyl tetrazole incorporated as a "privileged" group at the amine position.

The example here used only a binary fingerprint, but even more powerful results can be obtained when a count for each potential pharmacophore is included. The authors showed that for these designed Ugi libraries the same Ugi chemistry could indeed yield significant new diversity for multiple 14,000 compound libraries, but that after three libraries diminishing returns were obtained. They used the understandable nature of the pharmacophore descriptor by analyzing the remaining MDDR-pharmacophore fingerprint to show that most of the remaining pharmacophores to be matched contained acids and/or bases. A modified chemistry approach was therefore developed using protected acids (t-butyl esters) and bases (BOC protected) in the Ugi reaction. The unmatched cells in the MDDR-fingerprint can be related back to the compounds that

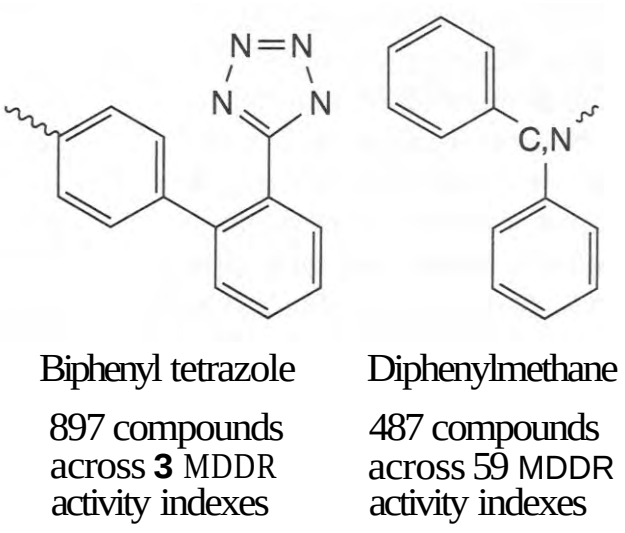


Figure 5.21. Examples of 7-TM GPCR "privileged" motifs found in the MDDR database.

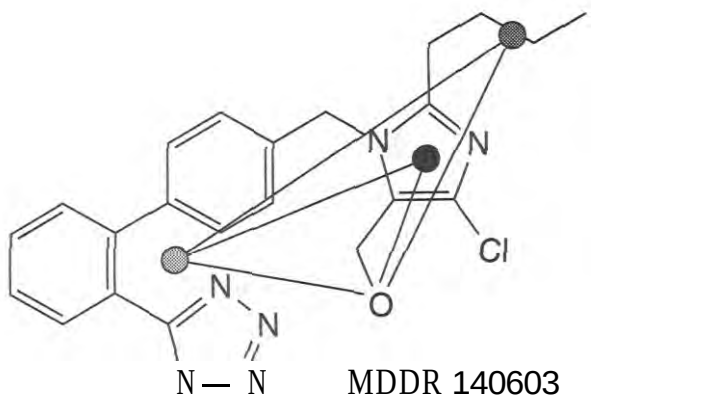


Figure 5.22. Example of pharmacophore feature assignments involving the biphenyl tetrazole "privileged" substructure and the total four-point potential pharmacophores calculated for a GPCR antagonist. Note that just the subset (40%) of the total pharmacophores that contained the "privileged" substructure was used for the library design.

generated them, enabling a truly iterative design of further libraries. In Fig. 5.24 the increasing size of the pharmacophore fingerprint from four consecutive Ugi libraries is illustrated, together with the distribution of pharmacophoric features in MDDR pharmacophores that had not been matched.

Another example for GPCR library design is the use of BCUT metrics as the basis for target class-focused approaches to accelerated drug discovery. A particularly interesting example is work done by Wang and Saunders (32e,i) at Neurocrine Bioscience in their effort to discover novel nonpeptidic ligands for a particular member (GPCR-1) of the GPCR-PA+ family of receptors activated by peptides carrying an obligatory positive charge. They and their colleagues performed a thorough search of the literature and identified a few hundred ligands of the various members of the GPCR-PA+ family. Knowing that it is usually not useful to follow up hits or leads showing very poor affinity, they eliminated ligands with less

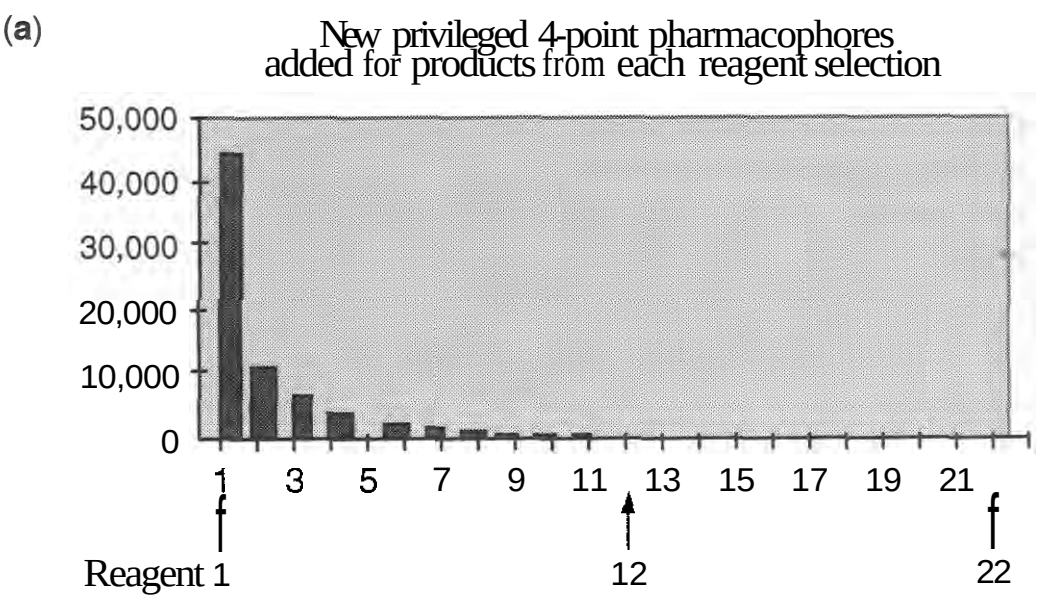


Figure 5.23. (a and b) Contributions per acid reactant of pharmacophores for optimization in the Ugi reaction (with biphenyl tetrazole as the "privileged" motif at the amine position). The order shown is the final selected order of reactants, based on obtaining the maximum number of new privileged pharmacophores per additional reactant. Histogram a shows the number of new pharmacophores added by each new selected reactant in the "privileged" pharmacophoric space defined by known GPCR compounds containing the biphenyl tetrazole; shown in histogram b is the matching increase in the total number of pharmacophores for the library for each new selected reactant.

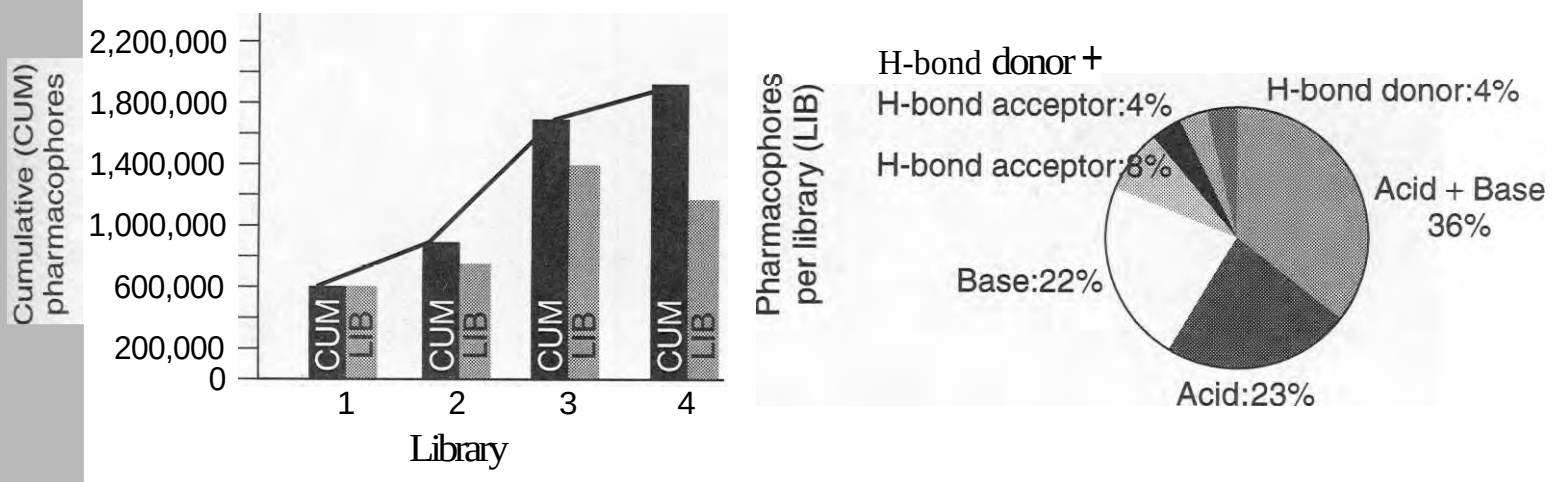


Figure 5.24. On the left is shown the cumulative (black) total number of four-point pharmacophores from consecutive 14,000 sets of Ugi libraries designed for 7-TM GPCR targets, together with the total number of pharmacophores in each library (in gray). Note the diminishing yield of new pharmacophores with later libraries, indicating that a change in strategy is needed. On the right are shown the features present in the resultant unrepresented pharmacophores (i.e., found in 7-TM GPCR biphenyl tetrazole-containing compounds in MDDR but not in synthesized libraries), indicating a strategy change to include more acids and bases together with the biphenyl tetrazole.

than 100 μM affinity for the corresponding receptor. Significantly, they also eliminated ligands with better than 1 μM affinity for the corresponding receptor. This very unusual step was taken in an effort to convince their colleagues that the method they intended to use was not reliant on knowing the answer ahead of time. This left 187 ligands with affinities mostly between 10 and 70 μM for various members of the GPCR-PA+ family of receptors. Using these compounds, they perceived a three-dimensional BCUT subspace within their corporate chemistry space that clusters the ligands of individual members of the GPCR-PA+ family and appears to be appropriate for this target class. The positions of all 187 ligands in the 3D chemistry space shown in Fig. 5.25 were originally indicated by open cyan circles. All ligands of some but not all receptors were then color-coded as indicated. Many red GPCR-2 and yellow GPCR-4 ligands are hidden under the green GPCR-1 ligands. The gray oval provides a crude indication of the region of chemistry space of interest for GPCR-PA+ receptors.

Figure 5.26 indicates the positions of roughly 2000 Neurocrine compounds selected from 14 different combinatorial libraries based on 14 different and proprietary scaffolds. Rather than selecting compounds only near the known ligands of GPCR-1, their receptor of interest, Wang and Saunders also selected compounds spanning the entire GPCR-PA+ receptors. This was done to further convince their colleagues, as explained below. All 2000 compounds were screened for activity against the GPCR-1 receptor. Those testing positive were retested in a secondary, functional assay. All but two compounds having better than 100 nM affinity for the GPCR-1 receptor are colored blue and/or are located within the blue oval. All but one compound having better than 10 nM affinity for the GPCR-1 receptor are colored red and/or are located within the red oval. All compounds with better than 2 nM affinity are colored green and are located within the two small green ovals within the larger green oval, consistent with the two crude clusters of GPCR-1 ligands seen in Fig. 5.25. The fact that these two small ovals each contain products from several different libraries (scaffolds) suggests the possible existence of two binding modes for this receptor. It is also significant to note that, although the authors intentionally synthesized compounds within the entire region of interest for GPCR-PA+ receptors, the only compounds showing significant affinity for the GPCR-1 receptor were located close to the known GPCR-1 ligands (compare with Fig. 5.25), thus supporting the use of BCUT coordinates (on receptor-relevant axes) as a valid approach to virtual high throughput screening. The tight clustering of GPCR-PA+ ligands in both figures clearly suggests that BCUT metrics represent, albeit in a relatively

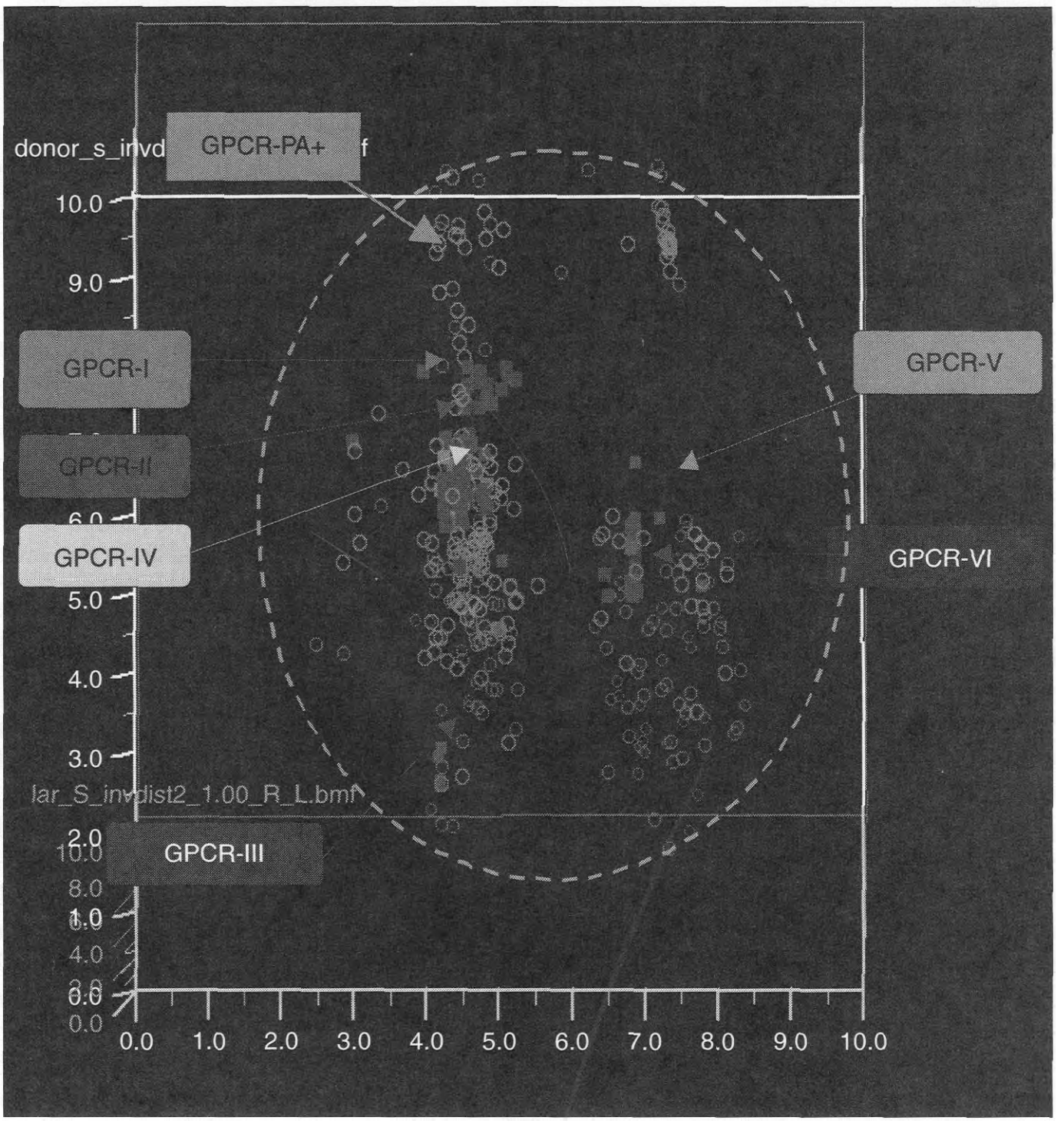


Figure 5.25. The 3D subspace most receptor relevant for members of the GPCR-PA+ family of receptors. Points indicate coordinates of 187 published ligands of various GPCR-PA+ receptors. Some have been color-coded by receptor for illustrative purposes. See Refs. 32e,i and text for further details. See color insert.

crude fashion, the same sort of information as would be represented in a description of the pharmacophore for the receptor of interest.

5.2 Property-Biased Design

The use of pharmacophoric descriptors in enhancing the hit-to-lead properties of lead optimization libraries has been described (76). Pharmacophore fingerprints, based on the Chem-X/ChemDiverse multiple pharmacophore descriptors, were used and several issues in the design of lead optimization libraries were addressed. The applicability of pharmacophoric methods to the design of focused libraries was demonstrated in this case, where the aim was to design the library toward a known lead or leads. The authors also investigated the design of libraries with improved pharmacokinetic properties. Simple and rapidly computable descriptors applicable to the prediction of drug transport properties were used, and the results illustrate a common problem: to obtain the best results it may be necessary to synthesize libraries in a noncombinatorial manner. A Monte Carlo search procedure was devised to enable the selection of a

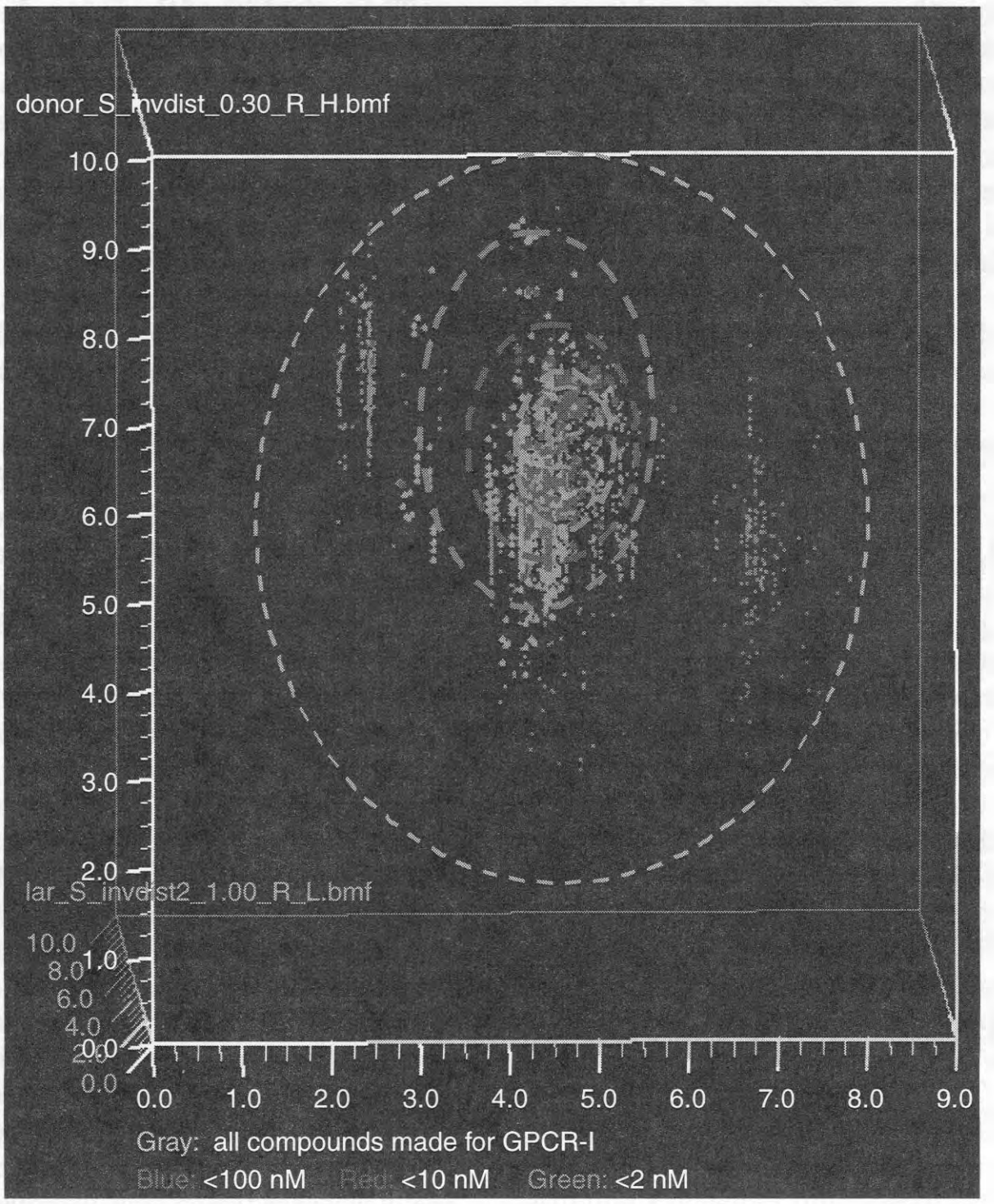


Figure 5.26. The same **3D subspace** as in Fig. 5.25, rotated slightly to provide a better viewing perspective. Points indicate coordinates of about 2000 combinatorial products selected from 14 different libraries. Color-coding indicates affinity for the GPCR-1 receptor. See Refs. 32e,i and text for further details. See color insert.

near-combinatorial subset in which all library members satisfy the design criteria. By including calculated log P, molecular weight, and polar surface area in the design of a combinatorial library, it was shown that the compounds with improved absorption characteristics (as determined by experimental Caco-2 measurements) could be obtained.

The use of computational methods such as reactant clustering and library profiling to maximize reactant diversity and optimize pharmacokinetic parameters has been described (132), with four-point pharmacophore fingerprint analysis used to quantify the added diversity gained by using two independent synthetic routes.

5.3 Site-Based Pharmacophores

Pharmacophore fingerprints generated from complementary site points can be used to direct combinatorial library design and to investigate selectivity. An example of the pharmacophore fingerprinting method for selectivity studies has been validated (37a,b) in studies of three closely related serine proteases: thrombin, Factor Xa, and trypsin. Site points were

positioned in the active site of each protein using the results of GRID (42) analyses (see Fig. 5.5), and receptor-based four-point pharmacophore fingerprints were generated. Fingerprints were also generated using full conformational flexibility for some highly selective and potent thrombin and Factor Xa inhibitors. Receptor-based similarity was investigated as a function of common potential three- and four-point pharmacophores for each ligand/receptor pair. The results indicated that the use of just the common potential four-point pharmacophores could give information pertaining to relative enzyme selectivity; when three-point pharmacopliores were used, however, poor resolution of enzyme selectivity was observed. The thrombin inhibitor thus exhibited greater similarity with the complementary four-point pharmacophore fingerprints of the thrombin active sites than with the potential pharmacophore keys generated from the other enzymes; a similar result was found for the Factor Xa inhibitors with the Factor Xa site.

Clearly, the inclusion of the shape of the binding site should improve the resolution, and the DiR (Design in Receptor) approach (133) refines the process, requiring that the pharmacophoric match fits the shape of the target site (i.e., is sterically compatible with the site). This clearly provides much additional information at the expense of greatly increased calculation time. Within the DiR approach, two-, three-, and four-point potential site pharmacophores can be used. This provides interesting new library design possibilities, in that it is possible to evaluate which ligands are able to fit in the site by matching at least one set of pharmacophoric features, and to quantify which pharmacophore hypotheses are matched. A subset of ligands can then be designed that match as many different pharmacophoric hypotheses as possible, and the biological screening of the resultant compounds can determine which bind best. Alternatively, pharmacophore constraints can be applied to a shape-driven searching approach, and Good et al. (34) have shown the effectiveness of this with the DOCK virtual screening/docking approach, in which the addition of pharmacophore constraints improved both the enrichment and speed of the process.

The active site of the Factor Xa serine protease (134) has been used for combinatorial library design (37c,d) using the DiR approach. GRID analyses using probes for hydrogen bond donors, acceptors, bases, acids, and hydrophobes resulted in 23 complementary site points being added (see Fig. 5.5). The shape of the active site was defined using 162 protein atoms. To ensure that a relevant area of the binding site was being explored (based on knowledge of X-ray protein-ligand complexes), site pharmacophores were forced to contain a hydrophobe or aromatic ring centroid point from both the S1 and S4 regions of the binding pocket. By using this focused approach, a "diversity" of matched site pharmacophores was obtained, representing a sampling of "reasonable" binding modes related to those experimentally observed and, thus, presumably having a higher probability of giving rise to biological activity. This focused approach reduced the total number of site pharmacophores from 5393 to 775 [using the seven distance ranges setting (37a) and considering all distances in the 1–15 Å range]. The approach was validated by the identification of feasible binding models (37c), similar to that experimentally observed for a known Factor Xa inhibitor. The Ugi four-component condensation reaction (131)(see Fig. 5.20) was used for the study and is capable of producing very large numbers of different structures from commercially available reactants. An example of the power of the method was given, whereby products were selected semimanually from a small virtual library of 432 products (37c,d). Products were constructed from the four reactant sets: carboxylic acids ($R_1 \times 3$), amines ($R_2 \times 2$), aldehydes ($R_3 \times 3$), and isonitriles ($R_4 \times 24$). The pharmacophore-based site analysis showed the optimum positions of substitution and chain length for benzamidine-containing fragments (targeted to the aspartate-containing S1 pocket) and the optimum lengths of other hydrophobic reactants (targeted to the S4 pocket) to produce compounds that would sample the maximum number of binding modes. In this case the groups were always forced to be in the S1 and S4 pockets to maintain "reasonable" binding modes, although this restriction could be excluded to probe even further potential binding modes. Thus,

as the identity of the matched site pharmacophore(s) was known for each compound, target site-based diversity of binding modes could be explored in the design process. An optimized selection of reactants was possible, and the value to the design of reactants with different chain lengths could be evaluated.

6 CONCLUSIONS AND FUTURE DIRECTIONS

Similarity and diversity metrics have been successfully used for a variety of tasks, including virtual screening, subset selection, and combinatorial library design. Databases of virtual compounds (e.g., from validated combinatorial chemistry protocols and reactants) can be used for both virtual screening and library design (virtual screening on virtual libraries with additional combinatorial constraints). The ability to exploit rapidly large virtual libraries of compounds that could be made by validated combinatorial chemistry protocols provides very powerful virtual screening and library design approaches. Future directions for library design will involve the application of such approaches in a fully integrated fashion (e.g., the ADEPT tool described in Section 4.10) and further enhancements to the constraints necessary to achieve druglike compounds (e.g., 80% compliance to the Rule of 5, predictive models for metabolism- and toxicity-related issues). Where the goal is lead generation (e.g., to enrich the compound screening file for high throughput screening), a focus will be on target classes (gene families) of interest, and the generation of compounds with leadlike properties, such as a lower molecular weight. The move away from combinatorial libraries to sparse arrays and noncombinatorial (cherry-picked) libraries (90) will continue, enabling more effective designs with control of associated properties. However, as more property constraints are applied to the library designs for leadlike/druglike properties, the need to include positive design elements to ensure good biological activity is emphasized. The goal for drug discovery is thus to identify targets and to generate compounds that are at the intersection of chemical, biological, and druglike property (e.g., absorption, toxicophores) space. Different targets and different expected routes of administration will require different constraints, and an element of diversity (with constraints toward a drug-occupied chemical space) will remain important, to enable the most effective use of combinatorial library chemistry and to discover new leads for both established and new targets.

REFERENCES

1. M. Johnson and G. Maggiora, ***Concepts and Applications of Molecular Diversity,*** John Wiley & Sons, New York, 1990.
2. (a) P. Willett, J. M. Barnard, and G. M. Downs, ***J. Chem. Inf. Comput. Sci.,*** 38,983 (1998); (b) P. Willett, ***Curr. Opin. Biotechnol., 11,*** 85 (2000); (c) P. Willett, Ed., Perspectives in Drug Discovery Design, Vols. 718, Kluwer Academic, Dordrecht/Norwell, MA, 1997; (d) J. S. Mason and M. A. Hermsmeier, ***Curr. Opin. Chem. Biol.,*** 3,342 (1999).
3. (a) J. M. Moore, ***Curr. Opin. Biotechnol., 10,*** 54 (1999); (b) P. J. Hajduk, T. Gerfin, J.-M. Boehlen, M. Haeberli, D. Marek, and S. W. Fesik, ***J. Med. Chem.,*** 42,2315 (1999); (c) C. A. Lepre, ***Drug Discovery Today, 6,*** 133 (2001); (d) J. Fejzo, C. A. Lepre, J. W. Peng, G. W. Bemis, Ajay, M. A. Murcko, and J. M. Moore, ***Chem. Biol.,*** 6, 755 (1999).
4. J. S. Mason and S. D. Pickett, ***Perspect. Drug Discov. Des., 7/8,*** 85 (1997).
5. (a) R. D. Brown, ***Perspect. Drug Discov. Des., 7/8,*** 31 (1997); (b) Y. C. Martin, R. D. Brown, and M. G. Bures in E. M. Gordon and J. F. Kerwin, Jr., Eds., ***Combinatorial Chemistry Molecular Diversity Drug Discovery,*** Wiley-Liss, New York, 1998, pp. 369385; (c) M. G. Bures and Y. C. Martin, ***Curr. Opin. Chem. Biol., 2,*** 376 (1998); (d) Y. C. Martin, M. G. Bures, and R. D. Brown, ***Pharm. Pharmacol. Commun.,*** 4, 147 (1998).
6. (a) D. K. Agrafiotis in P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner, Eds., ***The Encyclopedia of Computational Chemistry,*** Vol. *1,* John Wiley & Sons, Chichester, UK, 1998, pp. 742–761; (b) D. K. Agrafiotis and V. S. Lobanov, ***J. Chem. Inf. Comput. Sci.,*** 39, 51 (1999); (c) D. K. Agrafiotis, J. C. Myslik, and F. R. Salemme, ***Annu. Rep. Comb. Chem. Mol. Diversity,*** 2, 71 (1999); (d) D. K. Agrafiotis, V. S. Lobanov, D. N. Rassokhin, and S. Izrailev,

Methods Princ. Med. Chem., 10 (Virtual Screening for Bioactive Molecules),265 (2000).

7. *R. A. Lewis, S. D. Pickett, and D. E. Clark in K. B. Lipkowitz and D. B. Boyd, Eds., Reviews in Computational Chemistry, Vol. 16, Wiley-VCH, John Wiley & Sons, New York, 2000, pp. 1–51.*
8. *D. C. Spellmeyer and P. D. J. Grootenhuis, Annu. Rep. Med. Chem., 34, 287 (1999).*
9. *H. Matter and M. Rarey in G. Jung, Ed., Combinatorial Chemistry, Wiley-VCH Verlag GmbH, Weinheim, Germany, 1999, pp. 409–439.*
10. *Y. C. Martin, J. Comb. Chem., 3,231* (2001).
11. *(a) R. S. Pearlman, Network Sci., 2,* (6/7) (1996), *available at: http://www.netsci.org/Science/Combichem/feature08.html;* (b) *R. S. Pearlman and K. M. Smith, Perspect. Drug Discov. Des., 9,* 339/355 *(1998);* (c) *R. S. Pearlman and K. M. Smith, Drugs Future, 23,885* (1998); (d) *R. S. Pearlman and K. M. Smith, J. Chem. Inf. Comput. Sci.,* **39**, 28 (1999).
12. *J. S. Mason, A. C. Good, and E. J. Martin, Curr. Pharm. Des., 7, 567* (2001).
13. *(a) MDL Information Systems* Inc., San Leandro, *CA, URL: http://www.mdli.com;* (b) *C. A. James, D. Weininger, and J. Delaney, Daylight Theory Manual, version 4.72, Daylight Chemical Information Systems, Inc., URL:* http://www.daylight.com/dayhtml/doc/theory/theory.toc.html; (c) UNITY/SLN *manual available from Tripos, Inc., 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, URL:* http://*www.tripos.com.*
14. *(a) R. D. Brown and Y. C. Martin, J. Chem. Inf. Comput. Sci.,* **37**, 1 (1997); (b) *R. D. Brown and Y. C. Martin, J. Chem. Inf. Comput. Sci., 36, 572 (1996).*
15. *D. R. Flower, J. Chem. Inf. Comput. Sci., 38, 379 (1998).*
16. *M.* Randić, *J. Am. Chem. Soc., 97,6609 (1975).*
17. *(a) L. B. Kier,* L. *H. Hall, W. J. Murray, and M.* Randić, *J. Pharm. Sci., 64, 1971 (1975);* (b) *L. B. Kier, L. H. Hall, and W. J. Murray, J. Pharm. Sci., 64,1974* (1975).
18. *L. H. Hall and L. B. Kier, J. Mol. Graph. Modell., 20, 4 (2001).*
19. *M.* Randić, *J. Mol. Graph. Modell., 20, 19* (2001).
20. *(a) L. H. Hall and L. B. Kier in K. B. Lipkowitz and D. B. Boyd, Eds., Reviews in Computational Chemistry, Vol. 2, VCH Publishers, New York, 1991, pp. 367–422;* (b) *MOLCONN-Z, Hall Associates Consulting, 2 Davis Street, Quincy, MA, available from* eduSoft *L.C.* http://www.eslc.vaviotech.com.
21. *(a) R. A. Lewis, J. S. Mason, and I. M.* McLay, *J. Chem. Inf. Comput. Sci., 37,599* (1997).
22. *D. J. Cummins, C. W. Andrews, J. A. Bentley, and M. J. Cory, J. Chem. Inf. Comput. Sci., 36, 750 (1996).*
23. *(a) W. G. Glen, W. J. Dunn* III, *and D. R. Scott, Tetrahedron Comput. Methodol., 2,349 (1989);* (b) *C. Cheng, G.* Maggiora, *M. Lajiness, and M. J. Johnson, J. Chem. Inf. Comput. Sci., 36, 909 (1996);* (c) *M. Hassan, J. P.* Bielawski, *J. C.* Hempel, *and M. Waldman, Mol. Div., 2, 64 (1996);* (d) *D. K.* Agrafiotis, *J. Chem. Inf. Comput. Sci., 37,841 (1997).*
24. *R. E. Cahart, D. H. Smith, and R.* Venkataraghavan, *J. Chem. Inf. Comput. Sci.,* **25**, 64 *(1985).*
25. *R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, J. Chem. Inf. Comput. Sci., 27, 82 (1987).*
26. *(a) S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D.* Andose, *R. T. Mosley, and R. P. Sheridan, J. Chem. Inf. Comput. Sci., 36, 118* (1996); (b) *R. P. Sheridan, M. D. Miller, D. J. Underwood, and S. K. Kearsley, J. Chem. Inf. Comput. Sci., 36,128 (1996);* (c) *R. P. Sheridan, J. Chem. Inf. Comput. Sci., 40, 1456 (2000).*
27. *(a) G. Schneider, W. Neidhart, T. Giller, and G.* Schmid, *Angew. Chem. Znt. Ed. Engl., 38,2894 (1999);* (b) *G. Schneider,* O. Clement-Chomienne, *L. Hilfiger, P. Schneider, S. Kirsch,* H.-J. *Bohm, and W. Neidhart, Angew. Chem. Int. Ed. Engl., 39,4130 (2000).*
28. *D. Gorse and R. Lahana, Curr. Opin. Chem. Biol., 4,287 (2000).*
29. *C. A.* Lipinski, *F. Lombardo, B. W. Dominy, and P. J. Feeney, Adv. Drug Deliv. Rev., 23, 3 (1997).*
30. *F. R. Burden, J. Chem. Inf. Comput. Sci., 29, 225* (1989).
31. DiverseSolutions *was developed by R.* S. Pearlman *and K. M. Smith at the University of Texas, Austin, and is distributed by* Tripos, *Inc., St. Louis, MO.*
32. *(a) H. Gao, J. Chem. Inf. Comput. Sci., 41,402 (2001);* (b) *D.* Stanton, *J. Chem. Inf. Comput. Sci.,* **39**, 11 *(1999);* (c) *D. Schnur, J. Chem. Inf. Comput. Sci.,* **39**, 36 *(1999);* (c) *D. Schnur and P. Venkatarangan in A. K. Ghose and V. N. Viswanadhan, Eds., Combinatorial Library Design and Evaluation, Marcel Dekker, New York, 2001, pp. 473–501;* (d) *B. R.* Beno *and J. S. Mason, Drug Discovery Today, 6, 251*

(2001); (e) X.-C. Wang and J. Saunders, Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, August 26–30, 2001 (2001), MEDI-012; (f) E. L. Stewart, P. J. Brown, J. A. Bentley, and T. M. Willson, Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, August 2640, 2001 (2001), COMP-182; (g) Y. Gao and V. Goodfellow, Abstracts of Papers, 221st ACS National Meeting, San Diego, CA, 2001 (2001) MEDI-235; (h) X.-C. Wang and J. Saunders, Abstracts of Papers, 221st ACS National Meeting, San Diego, CA, 2001 (2001), MEDI-207; (i) J. Saunders, Proceedings of the IBC Conference on Drug Discovery by Design, Boston, MA, November 5–8,2001; (j) B. Pirard and S. D. Pickett, J. Chem. *Inf.* Comput. Sci., 40,1431 (2000).

33. (a) A. C. Good and J. S. Mason, Reviews in Computational Chemistry, Vol. 7, VCH, New York, 1995, pp. 67–127; (b) G. W. A. Milne, M. C. Nicklaus, and S. Wang, SAR QSAR *Environ.* Res., **9**, 23 (1998); (c) W. A. Warr and P. Willett, Design of Bioactive Molecules, American Chemical Society, Washington, DC, 1998, pp. 73–95.
34. A. C. Good, J. S. Mason, and S. D. Pickett, Methods *Princ.* Med. Chem., 10 (Virtual Screening for Bioactive Molecules), 131 (2000).
35. S. D. Pickett, J. S. Mason, and I. M. McLay, *J.* Chem. *Inf.* Comput. Sci., 36,1214 (1996).
36. E. K. Davies in I. M. Chaiken and K. D. Janda, Eds., Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery, American Chemical Society, Washington, DC, 1996, pp. 309–316.
37. (a) J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. R. Hulme, and R. F. Labaudiniere, J. Med. Chem., 42,3251 (1999); (b) J. S. Mason and D. L. Cheney, Proc. *Pac.* Symp. *Biocomput.*, 4, 456 (1999); (c) J. S. Mason, and D. L. Cheney, Proc. *Pac.* Symp. Biocomput., **5**, 576 (2000); (d) J. S. Mason and B. R. Beno, J. Mol. Graph. *Modell.*, 18, 438 (2000).
. (a) M. J. McGregor and S. M. Muskal, *J.* Chem. *Inf.* Comput. Sci., 39, 569 (1999); (b) M. J. McGregor, and S. M. Muskal, J. Chem. *Inf.* Comput. Sci., 40, 117 (2000).
. E. K. Bradley, P. Beroza, J. E. Penzotti, P. D. J. Grootenhuis, D. Spellmeyer, and J. L. Miller, *J.* Med. Chem., 43,2770 (2000).
0. D. Horvath in A. K. Ghose and V. N. Viswanadhan, Eds., Combinatorial Library Design and Evaluation, Marcel Dekker, New York, 2001, pp. 429–472.
. (a) R. Poulain, D. Horvath, B. Bonnet, C. Eckhoff, B. Chapelain, M.-C. Bodinier, and B. Déprez, J. Med. Chem., 44,3378 (2001); (b) R. Poulain, D. Horvath, B. Bonnet, C. Eckhoff, B. Chapelain, M.-C. Bodinier, and B. Déprez, J. Med. Chem., 44, 3391 (2001).
42. (a) P. J. Goodford, J. Med. Chem., 28, 849 (1985); (b) D. N. A. Bobbyer, P. J. Goodford, and P. M. McWhinnie, J. Med. Chem., 32,1083 (1989); (c) The GRID program is developed and distributed by Molecular Discovery Ltd.
43. E. J. Martin and T. J. Hoeffel, *J.* Mol. Graph. *Modell.*, 18,383 (2000).
44. A. R. Leach, D. V. S. Green, M. M. Hann, D. B. Judd, and A. C. Good, J. Chem. *Inf.* Comput. Sci., 40, 1262 (2000).
45. (a) K. A. Andrews and R. D. Cramer, J. Med. Chem., 43,1723 (2000); (b) R. D. Cramer, M. A. Poss, M. A. Hermsmeier, T. J. Caulfield, M. C. Kowala, and M. T. Valentine, J. Med. Chem., 42,3919 (1999); (c) R. D. Cramer, R. D. Clark, D. E. Patterson, and A. M. Ferguson, J. Med. Chem., 39,3060 (1996).
46. (a) H. Matter and T. Potter, J. Chem. *Inf. Comput.* Sci., 39, 1211 (1999); (b) V. J. Van Geerestein, H. Hamersma, and S. P. Van Helden in H. Van de Waterbeemd, B. Testa, and G. Folkers, Eds., Computer-Assisted *Lead* Finding and Optimization, Verlag Helvetica Chimica Acta, Basel, Switzerland, 1997, pp. 159–178.
47. M. Hahn, J. Chem. *Inf.* Comput. Sci., 37, 80 (1997).
48. O. F. Giiner, M. Waldman, R. Hoffmann, and J.-H. Kim in O. F. Giiner, Ed., Pharmacophore Perception, Development and Use in Drug Design, International University Line, La Jolla, CA, 2000, p. 213.
49. R. Carbo, L. Leyda, and M. Arnau, Znt. J. Quantum Chem., 17,1185 (1980).
50. A. C. Good, E. E. Hodgkin, and W. G. Richards, *J.* Chem. *Inf.* Comput. Sci., 32,188 (1992).
51. (a) D. J. Wild and P. Willett, *J.* Chem. *Inf. Comput.* Sci., 36, 159 (1996); (b) D. A. Thorner, D. J. Wild, P. Willett, and P. M. Wright, J. Chem. *Inf.* Comput. Sci., 36,900 (1996).
52. A. Schuffenhauer, V. Gillet, and P. Willett, J. Chem. *Inf.* Comput. Sci., 40, 295 (2000).
53. R. D. Cramer III, D. E. Patterson, and J. D. Bunce, J. *Am.* Chem. *Soc.*, 110, 5959 (1988).
54. G. Cruciani, P. Crivori, P.-A. Carrupt, and B. Testa, THEOCHEM, 603, 17 (2000).
55. (a) W. Guba and G. Cruciani in K. Guberrtofte and F. S. Jorgensen, Eds., Molecular Modeling and Prediction of Bioreactivity, Plenum, New York, 2000, pp. 89–95; (b) P. Crivori, G. Cru-

ciani, P.-A. Carrupt, *and B. Testa, J. Med. Chem., 43,2204 (2000).*

56. *M. Pastor, G. Cruciani, I. McLay, S. Pickett, and S. Clementi, J. Med. Chem., 43, 3233 (2000).*
57. *E. J. Martin, J. M. Blaney, M. A.* Saini, *D. C. Spellmeyer, A. K. Wong, and W. H. Moos, J. Med. Chem., 38, 1431 (1995).*
58. *C. A. James, D. Weininger, and J. Delaney, Fingerprints-Screening and Similarity, Daylight Theory Manual* v4.72, *Daylight Chemical Information Systems, Inc., URL: http://www.* daylight.com/dayhtml/doc/theory/theory.toc. *html.*
59. (a) *P. R.* Menard, *J. S. Mason, I. Morize, and S. Bauerschmidt, J. Chem. Inf. Comput. Sci., 38, 1204 (1998);* (b) *P. R.* Menard, *R. A. Lewis, and J. S. Mason, J. Chem. Inf. Comput. Sci., 38,497 (1998).*
60. *P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, UK, 1987.*
61. *R. A.* Jarvis *and E. A. Patrick, ZEEE Trans. Cornput., C-22,1025 (1973).*
62. *(a) P. Willett, V. Winterman, and D.* Bawden, *J. Chem. Inf. Comput. Sci., 26,109 (1986);* (b) *J. B.* Dunbar, *Perspect. Drug Discou. Des.,* 7/8, *51 (1997).*
63. *T. N.* Doman, *J. M.* Cibulskis, *M. J.* Cibulskis, *P. D.* McCray, *and D. P. Spangler, J. Chem. Inf. Comput. Sci., 36, 1195 (1996).*
64. *(a) J. M.* Barnard *and G. M. Downs, J. Chem. Inf. Comput. Sci., 37,141* (1997); (b) *J. M.* Barnard, *and G. M. Downs, Perspect. Drug Discov. Des.,* 7/8, *13* (1997).
65. *G. M. Downs, P. Willett, and W. Fisanick, J. Chem. Inf. Comput. Sci., 34, 1094 (1994).*
66. *(a) M. Lajiness, Perspect. Drug Discov. Des.,* 7/8, 55 *(1997);* (b) *V. J. Gillet and P. Willett in A. K. Ghose and V. N. Viswanadhan, Eds., Combinatorial Library Design and Evaluation, Marcel Dekker, New York, 2001, pp.* 379–*398;* (c) *R. D. Clark, J. Chem. Inf. Comput. Sci., 37, 1181* (1997); (d) *T. Potter and H. Matter, J. Med. Chem., 41,478* (1998).
67. *B. D. Hudson, R. M. Hyde, E.* Rahr, *and J. Wood, Quant. Struct.-Act. Relat., 15, 285 (1996).*
68. *R. E.* Higgs, *K. G. Bemis, I. A. Watson, and J. H. Wikel, J. Chem. Inf. Comput. Sci., 37,861 (1997).*
69. *(a) J. D. Holliday,* S. S. Ranade, *and P. Willett, Quant. Struct.-Act. Relat., 14, 501 (1995);* (b) *D. B. Turner, S. M.* Tyrrell, *and P. Willett, J. Chem. Inf. Comput. Sci.,* 37, 18 *(1997).*
70. *S.* D. *Pickett, C. Luttmann, V. Guerin, A. Laoui, and E. James, J. Chem. Inf. Comput. Sci., 38, 144 (1998).*
71. *J. Mount, J. Ruppert, W. Welch, and A.* Jain, *J. Med. Chem., 42, 60 (1999).*
72. *M. Snarey, N. K. Terrett, P. Willett, and D. J.* Wilton, *J. Mol. Graph., 15, 372 (1997).*
73. *M. Waldman, H. Li, and M.* Hasan, *J. Mol. Graph. Modell., 18,412 (2000).*
74. *M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller, and N.* Ramsden, *J. Chem. Inf. Comput. Sci., 39, 897 (1999).*
75. *(a) D. E. Clark and S. D. Pickett, Drug Discovery Today, 5, 49* (2000); (b) *P. J. Eddershaw, A. P. Beresford, and M. K. Bayliss, Drug Discovery Today,* 5, 409–414 (2000); (c) *H. van de Waterbeemd, D. A. Smith, K. Beaumont, and D. K. Walker, J. Med. Chem., 44, 1313 (2001).*
76. *S. D. Pickett, D. E. Clark, and I. M. McLay, J. Chem. Inf. Comput. Sci., 40,263 (2000).*
77. *D. E. Clark, J. Pharm. Sci., 88,807 (1999).*
78. *(a) A. C. Good and R. A. Lewis, J. Med. Chem., 40, 3926 (1997);* (b) *R. A. Lewis, A. C. Good, and S. D. Pickett in H. Van de Waterbeemd, B. Testa, and G. Folkers, Eds., Computer-Assisted Lead Finding and Optimization, Verlag Helvetica* Chimica Acta, Basel, *Switzerland, 1997, pp. 135–156;* (c) *V. J. Gillet, P. Willet, J. Bradshaw, and D. V. S. Green, J. Chem. Inf. Comput. Sci., 39, 169* (1999).
79. *G. Grassy, A. Yasri, R. Lahana, J. Woo, S.* Iyer, *M. Kaczorek, R. Folc'h, and R. Buelow, Nat. Biotechnol., 16, 748 (1998).*
80. *J. S. Mason in P. M. Dean and R. A. Lewis, Eds., Molecular Diversity Drug Design,* Kluwer Academic, *Dordrecht, Netherlands, 1999, pp. 67–91.*
81. *(a) A. C. Good and I. D. Kuntz, J. Comput.-Aided Mol. Des., 9, 373 (1995);* (b) *M. J.* Ashton, *M. Jaye, and J. S. Mason, Drug Discovery Today, 1, 71 (1996).*
82. *J. H. Van Drie and R. A.* Nugent, *SAR QSAR Environ. Res., 9, 1 (1998).*
83. *A. C. Good, Internet J. Chem., 3* (2000), http://www.ijc.com/article/2000v3/9/.
84. *D. E. Patterson, R. D. Cramer, A. M.* Ferguson, *R. D. Clark, and L. E. Weinberger, J. Med. Chem., 39,3049 (1996).*
85. *A. C. Good, J. S. Mason, D. V. S. Green, and A. R. Leach in A. K. Ghose and V. N.* Viswanadhan, *Eds., Combinatorial Library Design and Evaluation, Marcel Dekker, New York, 2001, pp. 399–428.*

C. D. Eldred and B. D. Judkins, *Prog. Med. Chem.*, **36**, 29 *(1999).*

(a) *V. J. Gillet, P. Willett, and J. Bradshaw, J. Chem. Inf. Comput. Sci., 38,165 (1998);* (b) *E. A.* Jamois, *M. Hassan, and M. Waldman, J. Chem. Inf. Comput. Sci., 40, 63 (2000);* (c) *R. S.* Pearlman *and K. M. Smith, Symposium on Combinatorial Chemistry, Abstracts of Papers, 217th ACS National Meeting, Anaheim,* CA, *March 1999, MEDI-012.*

(a) *J.* Sadowski *in A. K. Ghose and V. N.* Viswanadhan, *Eds., Combinatorial Library Design and Evaluation, Marcel Dekker, New York, 2001, pp. 291–300;* (b) *A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski, J. Comb. Chem., 1, 55 (1999);* (c) *Ajay, W. P. Walters, and M. A. Murcko, J. Med. Chem., 41, 3314 (1998);* (d) *J.* Sadowski *and H.* Kubinyi, *J. Med. Chem., 41,3325 (1998).*

T. Mitchell and G. A. Showell, *Curr. Opin. Drug Discov. Dev., 4, 314 (2001).*

J. R. Everett, M. Gardner, F. Pullen, *G. F. Smith, M. Snarey, and N. Terrett, Drug Discovery Today, 6, 779* (2001).

(a) *J. H. Van Drie and M. S. Lajiness, Drug Discovery Today, 3,274 (1998);* (b) *R. A. Lewis in P. M. Dean and R. A. Lewis, Eds., Molecular Diversity Drug Design,* Kluwer *Academic,* Dordrecht, *Netherlands, 1999, pp. 221–248 and other chapters therein.*

G. Jones, P. Willett, R. Glen, A. Leach, and R. Taylor in A. Parill *and M. Reddy, Eds., Rational Drug Design, ACS Symposium Series 719, American Chemical Society, Washington, DC, 1999.*

R. T. Koehler and H. O. *Villar, J. Comput. Chem., 21, 1145 (2000).*

R. D. Brown, M. Hassan, and M. Waldman, J. Mol. Graph. Modell., 18, 427 (2000).

(a) *S. J.* Teague, *A. M. Davis, P. D.* Leeson, *and T. Oprea, Angew. Chem. Znt. Ed. Engl., 38, 3743 (1999);* (b) *T. Oprea, et al., J. Chem. Inf. Comput. Sci., 41, 1308 (2001).*

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science, 220, 671 (1983).

(a) *L. Weber, Drug Discovery Today, 3, 379 (1998);* (b) *W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, The Art of Scientific Computing, Cambridge University Press, New York, 1984, pp. 444–455;* (c) *D. K. Agrafiotis, J. Chem. Inf. Comput. Sci., 37, 576* (1997); (d) *W. Zheng, S. J. Cho, C. L. Waller, and A. Tropsha, J. Chem. Znf. Comput. Sci., 39, 738* (1999).

D. E. Clark, Ed., Evolutionary Algorithms in Molecular Design, Wiley-VCH, New York, 2000.

R. P. Sheridan *and S. K. Kearsley, J. Chem. Inf. Comput. Sci., 35, 310 (1995).*

J. Singh, M. A. Ator, *E. P. Taeger, M. P. Allen, D. A. Whipple, J. E. Soloweji, S. Chowdhray, and A. M. Treasurywala, J. Am. Chem. Soc., 118, 1669 (1996).*

R. D. Brown and Y. C. Martin, J. Med. Chem., 40,2304 (1997).

(a) *C. M. Fonseca and P. J. Fleming in S. Forrest, Ed., Genetic Algorithms: Proceedings of the Fifth International Conference, Morgan Kaufmann, San Mateo, CA, 1993, pp. 416–423;* (b) *C. M. Fonseca and P. J. Fleming in K. De Jong, Ed., Evolutionary Computation, Vol. 3, The Massachusetts Institute of Technology, Cambridge, MA, 1995, pp. 1–16;* (c) *V. J. Gillet, W. Khatib, P. Willett, P. Fleming, and D.V. S. Green, J. Chem. Inf. Comput. Sci.,* **42**, 375–385 *(2002).*

A. Polinsky, R. D. Feinstein, S. Shi, and A. Kuki *in I. M. Chaiken and K. D. Janda, Eds., Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery, American Chemical Society, Washington, DC, 1996, pp. 219–232.*

D. S. Thorpe, A. E. Chan, A. *Binnie, L. C. Chen, A. Robinson, J. Spoonamore, D.* Rodwell, *S. Wade, S. Wilson, M. Ackerman-Berrier, H. Yeoman, S. Walle, Q. Wu, and K. F. Wertman, Biochem. Biophys. Res. Commun., 266, 62 (1999).*

(a) *G. M. Downs and J. M.* Barnard, *J. Chem. Inf. Comput. Sci.,* **37**, *59 (1997);* (b) *J. M.* Barnard, *G. M. Downs, A. von Scholley-Pfab, and R. D. Brown, J. Mol. Graph. Modell., 18, 452 (2000);* (c) *G. M. Downs and J. M.* Barnard, *2nd Joint Sheffield Conference on Chemoinformatics: Computational Tools for Lead Discovery, April 9–11, 2001. Presentation available at:* **http://cisrg.shef.ac.uk/shef2001/talks/downs.** *ppt;* (d) Accelerys, San Diego, *CA (previously MSI, Molecular Simulations Inc.),* http:// *www.accelrys.com.*

D. K. Agrafiotis and V. S. Lobanov, 2nd Joint Sheffield Conference on Chemoinformatics: Computational Tools for Lead Discovery, April 9–11, 2001. Abstract available at: http://cisrg. shef.ac.uk/shef2001/abstracts.htm.

N. E. Shemetulskis, J. B. Dunbar, *Jr., B. W.* Dunbar, *D. W. Moreland, and C. Humblet, J. Cornput.-Aided Mol. Des., 9,407 (1995).*

108. J. H. Voigt, B. Bienfait, S. Wang, and M. C. Nicklaus, J. Chem. *Inf.* Comput. Sci., 41, 702 (2001).

109. M. J. McGregor and P.V. Pallai, J. Chem. *Inf.* Comput. Sci., 37, 443 (1997).

110. (a) R. B. Nilakantan, N. Bauman, K. S. Haraki, and R. J. Venkataraghavan, J. Chem. *Inf.* Comput. Sci., **30, 65** (1990); (b) R. B. Nilakantan, N. Bauman, and K. S. Haraki, J. *Comput.*-Aided Mol. Des., 11,447 (1997).

111. E. J. Martin and R. E. Critchlow, J. Comb. Chem., **1, 32** (1999).

112. (a) M. J. McGregor and S. M. Muskal, J. Chem. *Inf.* Comput. Sci., 39, 569 (1999); (b) M. J. McGregor and S. M. Muskal, J. Chem. *Inf.* Comput. *Sci.*, 40, 117 (2000).

113. B. A. Hopkins, *Ann.* Bot., 18,213 (1954).

114. (a) T. R. Hagadone and M. S. Lajiness, Tetrahedron Comput. Methodol., 1, 219 (1988); (b) T. R. Hagadone, J. Chem. *Inf.* Comput. Sci., 32,515 (1992).

115. A. Gobbi, D. Poppinger, and B. Rohde, *Perspect.* Drug Discov. Des., **7/8,** 131 (1997).

116. A. R. Leach, J. Bradshaw, D.V. S. Green, and M. M. Hann, J. Chem. *Inf.* Comput. Sci., 39, 1161 (1999).

117. G. Bravi, D.V. S. Green, M. M. Hann, and A. R. Leach, J. Chem. *Inf.* Comput. Sci., 40, 1441 (2000).

118. (a) W. P. Walters, Presented at the Daylight User Group Meeting, MUG'99.1999. Available online at http://www.daylight.com/meetings/mug99/Walters/index.html; (b) W. P. Walters, M. T. Stahl, and M. A. Murcko, Drug Discovery Today, 3,160 (1998).

119. M. Stahl, Methods *Princ.* Med. Chem., 10 (Virtual Screening for Bioactive Molecules), 229 (2000).

120. H. J. Bohm and M. Stahl, Curr. Opin. Chem. Biol., 4, 283 (2000).

121. A. R. Leach, R. A. Bryce, and A. J. Robinson, J. Mol. Graph. *Modell.*, 18, 358 (2000).

122. R. P. Sheridan, S. G. SanFeliciano, and S. K. Kearsley, J. Mol. Graph. *Modell.*, 18, 320 (2000).

123. D. Tondi and M. P. Costi in A. K. Ghose and V. N. Viswanadhan, Eds., Combinatorial Library Design and Evaluation, Marcel Dekker, New York, 2001, pp. 563–603.

124. E. K. Kick, D. C. Roe, A. G. Skillman, G. Lin, T. J. A. Ewing, Y. Sun, I. Kuntz, and J. A. Ellman, Chem. Biol., 4,297 (1997).

125. D. C. Roe in P. M. Dean and R. A. Lewis, Eds., Molecular Diversity Drug Design, Kluwer Academic, Dordrecht, Netherlands, 1999, pp. 141–173.

126. (a) I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, J. Mol. Biol., 161, 269 (1982); DOCK is developed and distributed by the Kuntz Group, Dept. of Pharmaceutical Chemistry, 512 Parnassus, University of California, San Francisco, CA 94143–0446, URL: http://www.cmpharm.ucsf.edu/kuntz; (b) T. J. A. Ewing and I. D. Kuntz, J. Comput. Chem., 18, 1175 (1997); (c) Y. Sun, T. J. A. Ewing, A. G. Skillman, and I. D. Kuntz, J. Cornput.-Aided Mol. Des., 12,597 (1998).

127. (a) S. F. Brady, et al., J. Med. Chem., 41, 401 (1998); (b) H. J. Bohm, D. W. Banner, and L. Weber, J. Cornput.-Aided Mol. Des., 13, 51 (1999).

128. (a) A. Rockwell, M. Melden, R. A. Copeland, K. Hardman, C. P. Decicco, and W. F. DeGrado, J. *Am.* Chem. *Soc.*, 118,10337 (1996); (b) A. K. Szardenings, D. Harris, S. Lam, L. Shi, D. Tien, Y. Wang, D.V. Patel, M. Navre, and D. A. Campbell, J. Med. Chem., 41,2194 (1998).

129. D. Tondi, U. Slomiczynska, M. P. Costi, D. M. Watterson, S. Ghelli, and B. K. Shoichet, Chem. Biol., 6, 319 (1999).

130. M. L. Lamb, K. W. Burdick, S. Toba, M. M. Young, A. G. Skillman, X. Zou, J. R. Arnold, and I. D. Kuntz, Proteins Struct. *Funct.* Genet., 42,296 (2001).

131. I. Ugi and C. Steinbruckner, Chem. Ber., 94, 734 (1961).

132. T. F. Herpin, G. C. Morton, A. K. Dunn, C. Fillon, P. R. Menard, S. Y. Tang, J. M. Salvino, and R. F. Labaudiniere, Mol. Diversity, 4, 221 (2000).

133. C. M. Murray and S. J. Cato, J. Chem. *Inf.* Comput. Sci., **39, 46** (1999).

134. A. Tulinsky, K. Padmanbhan, K. P. Padmanbhan, C. H. Park, W. Bode, R. Huber, D. T. Blankenship, A. D. Cardin, and W. Kisiel, J. Mol. Biol., 232,947 (1993).

135. M. Rarey and M. Stahl, J. Cornput.-Aided Mol. Des., 15,497–520 (2001).