Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries†

Stephen D. Pickett,* Iain M. McLay, and David E. Clark

Rhône-Poulenc Rorer, Dagenham Research Centre, Rainham Road South, Dagenham, Essex, United Kingdom RM10 7XS

Received July 29, 1999

In this paper we address several issues in the design of lead optimization libraries. Multipharmacophore descriptors were first developed in the context of designing diverse compound libraries. One reason for favoring such descriptors is the importance of the pharmacophore hypothesis in understanding the interaction of a compound with a protein target. Allied to this is the proposal that sampling over all potential pharmacophores leads to diversity in a biologically relevant space. We present results in support of this argument and also demonstrate that such methods are applicable to the design of focused libraries where the aim is to design the library toward a known lead or leads. This portability is important because it means that the same descriptors can be used for diverse library design, screening set selection, and focused library design, giving a consistent approach. We also examine the question of designing libraries with improved pharmacokinetic properties and show that it is possible to derive simple and rapidly computable descriptors applicable to the prediction of drug transport properties. Furthermore, these can be applied in the context of library design, although it may be necessary to synthesize libraries in a noncombinatorial manner to obtain the best results. To address this problem, we describe a Monte Carlo search procedure that allows the selection of a near-combinatorial subset in which all library members satisfy the design criteria. We present an example from our own work that illustrates how consideration of calculated log *P*, molecular weight, and polar surface area in the design of a combinatorial library can lead to compounds with improved absorption characteristics as determined by experimental Caco-2 measurements.

1. INTRODUCTION

Combinatorial chemistry is now firmly established as a powerful tool available to the medicinal chemist in the pursuit of new drug candidates.[1] Combinatorial methods provide a way to generate very large numbers of compounds in a relatively short period of time (compared to traditional synthesis of singles). However, this very aspect of combinatorial chemistry in itself presents a problem. A balance needs to be struck between making everything possible and the constraints of economics, logistics, and time. In other words, there is a need to select the products to be synthesized from the vast pool of those that could possibly be made. This is normally done in such a manner as to maintain the combinatorial nature of the library, e.g., for a two-dimensional library, selecting *M* reagents at R1 and *N* at R2 to give $M \times N$ products. There have been several reports of methods for performing this selection to ensure "diversity" of either reagents or (better) final products.[2−4] The major differences between the reported approaches are whether the selection is performed by considering the diversity of the reagents or the final products and how "diversity" is defined and measured. For more details, the reader is directed to several recent reviews.[5−8]

Once an initial hit or lead compound has been identified through general screening, the compound is optimized, the purpose being to increase the robustness of the potential drug in humans. The optimization process is multiparametric, involving simultaneous optimization of biological properties (in vitro and in vivo potency), physicochemical properties (e.g., log *P*, pK_a), pharmaceutic properties (e.g., solubility, crystallinity), and pharmacokinetic properties (absorption, metabolism, distribution, and elimination). The effort expended on each of these areas will change as the project progresses, but it is important to consider all aspects as early as possible in the drug discovery process.

In this paper, we present methods applicable to the design of lead optimization libraries. In particular, we focus on two key areas. First, given one or more lead compounds, how can we select follow-up compounds for screening and design follow-up libraries to take advantage of this (limited) knowledge? In this context, and with our methodology, the latter is a constrained case of the former. This enables us to build upon our experiences with diverse library design[4,9−11] and make use of the same descriptors. Second, how can libraries be designed to improve the likelihood that the compounds will be fit for purpose (e.g., compounds for oral administration need to be absorbed in the intestine and CNS drugs need to penetrate the blood−brain barrier)? In what follows, the next section gives a general introduction to the issues and our approaches to them. Section 3 details the methodology we have used. Section 4 gives two examples: one of selecting a follow-up screening set and the other of focused library design. Section 5 discusses various approaches to addressing the problem of "drug-likeness" and includes a successful application of the methods to a real

† Paper presented at the Fifth International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June 6−10, 1999.

* Corresponding author. Email: stephen.pickett@rp-rorer.co.uk. Tel: +44 181 919 3353. Fax: +44 181 919 2029.

10.1021/ci990261w CCC: $19.00
Published on Web 12/07/1999

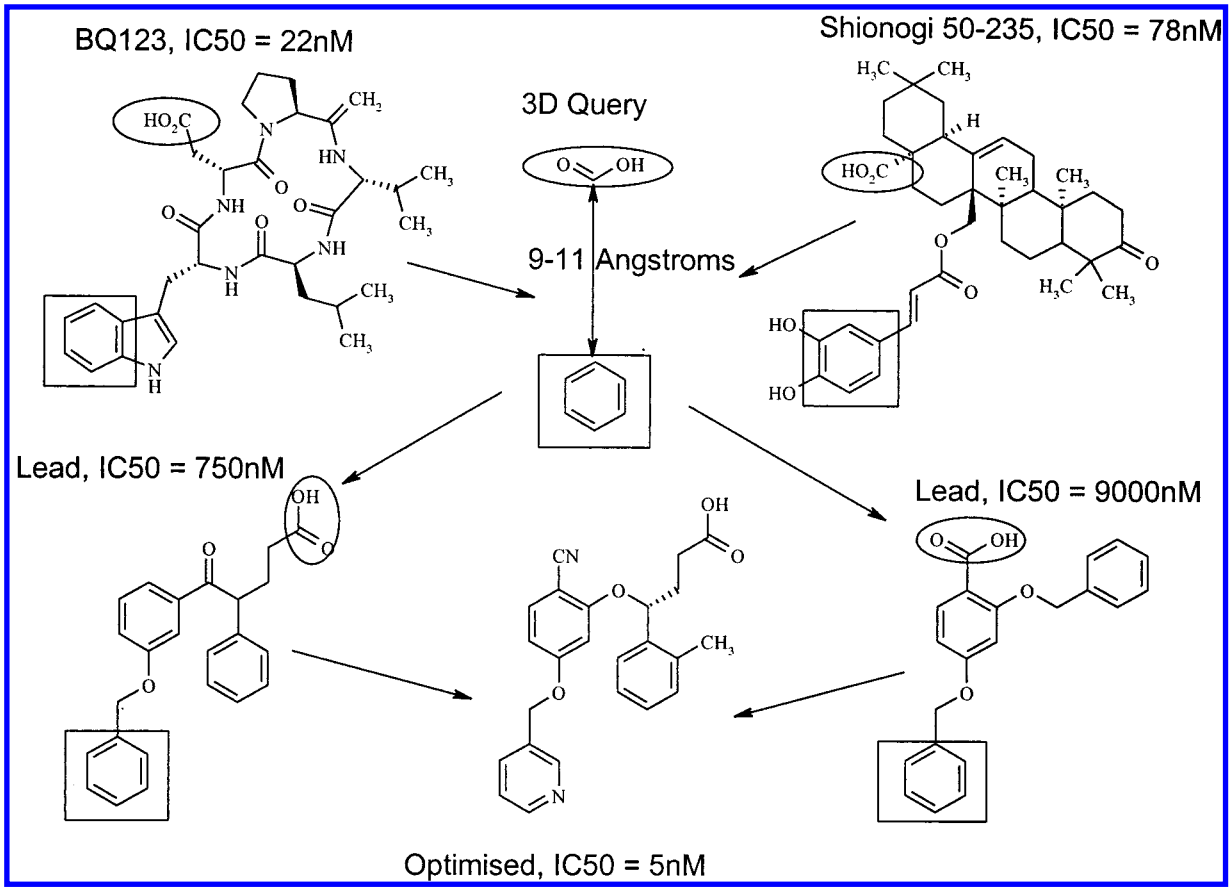

Figure 1. Elucidation of a pharmacophore for endothelin A antagonism and its use in the discovery of potent and selective compounds. BQ123 and Shionogi were flexibly overlayed, and a two-point pharmacophore was identified. Searching of the corporate database and subsequent screening of the hits identified the two lead compounds shown, which were optimized to give a compound with an IC50 of 5nM.[17]

design problem. Section 6 contains a general discussion and conclusions.

2. LEAD OPTIMIZATION LIBRARY DESIGN

Follow-up Screening Sets and Focused Library Design. Lead compounds can be obtained from a variety of sources, for instance, the corporate collection, combinatorial libraries, competitor patents, or natural products. Two-dimensional (2-D) similarity/substructure searching[12−14] provides a rapid way of identifying close analogues within a compound collection for follow-up screening. Three dimensional (3-D) searching[15] is a powerful tool for locating compounds that are structurally distinct from the lead but share a similar 3-D arrangement of key features necessary, although not sufficient, for biological action, i.e., a pharmacophore. There are several published successes of such an approach,[16−18] and there are clear advantages over 2-D methods. For instance, the pharmacophore represents the *three-dimensional* relationship between key features of the ligand−receptor interaction such as hydrogen bonding, hydrophobicity, charge−charge interactions, etc. Methods have also been developed for considering conformational flexibility during the search, and this can increase the hit rate of a search significantly.[19−22] The difficulty lies in being able to define the pharmacophore precisely enough to be useful. For example, in our own work on endothelin antagonists,[17] the two-point pharmacophore from BQ123 was too broad to be generally useful and only when the Shionogi compound was included could a sufficiently precise pharmacophore be defined (Figure 1).

The development of the pharmacophore key as a descriptor for diverse library design was driven in part by the successes arising from 3-D searching.[9] The nature of the descriptor means that it is applicable not just to diverse library design. Once the pharmacophore key has been calculated (see Methods), it can be used in the same way as any other binary key. Thus, the key for a lead compound(s) can be compared to the precalculated keys for a compound collection and the compounds ranked on the basis of their pharmacophoric similarity to the lead(s) with no need to define a particular pharmacophore of interest. Section 4 exemplifies this approach, using the pharmacophore key from a conformationally flexible RGD (Arg-Gly-Asp) tripeptide to generate a screening set enriched with known fibrinogen antagonists. If the compound collection represents a virtual library of products, then it should be possible to select the reagents in such a way as to ensure that the products on average have a high degree of similarity with the lead(s). A genetic algorithm (GA) provides one possible method for enabling a combinatorial selection of products to be made.

Several other approaches to focused library design have been published. For example, Sheridan and Kearsley used similarity with a probe calculated from topological descriptors to score tripeptoids generated from a large collection of possible residues using a GA.[23] However, the final selection was not a true combinatorial library, and the user is required to select fragments from their frequency of occurrence in the GA output. A similar approach has been reported by Cho et al.[24] The HARPick procedure developed at RPR[4,25]

provides an alternative strategy based upon pharmacophore keys. The design can be focused toward (or away from) the pharmacophore distribution of a compound collection as a whole rather than just considering pairwise molecular similarities.

Designing in "Drug-likeness". To define the term "drug-likeness", we need to consider the properties that a molecule must possess before it can be called a drug. Several recent papers have aimed to tackle this problem by considering compounds from the World Drug Index[26] (WDI) as "drugs" and a set of compounds from commercial catalogues, e.g., the Available Chemicals Directory[27] (ACD) or the SPRESI database[28] as "nondrugs". Methods such as neural networks,[29,30] recursive partitioning,[31] or genetic algorithms[32] have then been used to derive rules or descriptor weights that classify the compounds as drugs or nondrugs. However, these approaches suffer from several limitations. For instance, Sadowski and Kubinyi[30] note that the ACD contains "a significant number of undetected drugs" even after substantial filtering, and the interpretation of a trained neural network to derive rules is not possible. Nevertheless, the results are encouraging, and such approaches provide a means for selecting suitable compounds from supplier databases or company collections. For example, in subsequent work, the approach of Gillet et al.[32] has been used in the selection of a corporate screening set.[33] For lead optimization, however, the needs are somewhat different and these methods do not supply the necessary detail to allow a medicinal chemist to modify the properties of a molecule in a rational way. What actually confers "drug-likeness" on a molecule is a very complex issue covering a wide range of different aspects: physical form, formulation, absorption, metabolism, toxicity, plasma protein binding, etc. We have decided, therefore, to consider each of these areas independently to derive models that are predictive and interpretable so that the medicinal chemist is given clear guidelines on how to tackle a particular problem.

In this paper, we shall focus on one particular aspect, that of predicting absorption following oral administration. One of the most well-known approaches to absorption prediction is that developed by Lipinski et al.,[34] in which some medicinal chemistry rules-of-thumb were formalized into the widely applied "Rule-of-5". The rules were derived from an analysis of 2245 drugs from the WDI[26] believed to have entered phase II trials on the basis of their having been assigned either a USAN (United States adopted name) or INN (international nonproprietary name). The Rule-of-5 states that if a compound satisfies any *two* of the following rules, it is likely to exhibit poor intestinal absorption:

(1) Molecular weight greater than 500 Da.

(2) Number of hydrogen bond donors greater than five (a donor being any O−H or N−H group).

(3) Number of hydrogen bond acceptors greater than 10 (an acceptor being any O or N including those in donor groups).

(4) Calculated log *P* greater than 5.0 (if ClogP[35] is used) or greater than 4.15 (if MlogP[36] is used).

The rules are very quick to calculate and were designed to give few false negatives (i.e., absorbed compounds misclassified as being not absorbed). This approach can provide a very useful initial screening of a virtual library or compound collection to filter out compounds predicted to have absorption problems but is not very discriminating when used in isolation and can generate many false positives. Recently, Palm et al. have reported a strong correlation between the polar surface area (PSA) and fractional absorption in humans of 20 carefully chosen drugs.[37] Their methodology involves calculating the "dynamic" polar surface area from an ensemble of low-energy conformers and as such is quite time-consuming. However, we have shown how using just a single low-energy conformer for PSA calculations can yield an equally good correlation for this data set.[38] This single conformer PSA is very rapid to calculate and thus provides a good partner to the Rule-of-5 in virtual library analysis. PSA also provides a good descriptor for modeling blood−brain barrier penetration.[39−41] In this paper, we show how application of these methods as design criteria can produce libraries with improved absorption properties using a recent example from our own work. However, it will be seen from this example that, when applying such methods, it is not (always) possible to preserve the combinatorial efficiency of the library synthesis. Thus, we also present a stochastic design procedure that not only ensures that all selected compounds satisfy our criteria but also that the resultant library is as close to a full combinatorial library as possible, thereby ensuring synthetic efficiency.

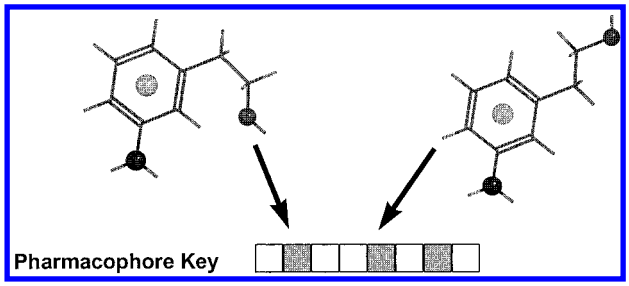

Figure 2. Calculation of pharmacophore keys. Reproduced with permission from ref 8.

3. METHODS

Pharmacophore Descriptors. The basic idea behind the calculation of pharmacophore descriptors is shown in Figure 2, and details of our methodology have been given elsewhere.[9−11,42] For a defined conformation of a molecule, triangles (tetrahedra) are formed from all combinations of three (four) pharmacophoric points and are recorded in a bit string. Each bit represents a particular combination of pharmacophore points (donor−acceptor−aromatic, donor−basic−aromatic, etc.) and distances. We use the ChemDiverse module of Chem-X[43] to calculate the descriptors. Atom typing is performed using an in-house customization[9,42] of the standard Chem-X parametrization. A systematic search procedure is used for the conformational analysis and is controlled by in-house scripts.[11,42] Chem-X offers the possibility of seven distinct pharmacophore types: hydrogen bond donor, hydrogen bond acceptor, aromatic center, hydrophobic point, acid, base and Nplus. These types are user-definable, and, in particular, Nplus can be redefined to identify groups such as hydroxyl (i.e., groups that can act both as donors and acceptors[11]) or to define a specific atom or region of the molecule as a reference point, which is useful in diversity-related tasks.[6,42] In the work described here, we have used either 3- or 4-point pharmacophores defined from

Table 1. Distance Ranges Used for Pharmacophore Key Calculations

minimum distance (Å)	maximum distance (Å)	minimum distance (Å)	maximum distance (Å)
0	2.0	5.8	7.9
2.0	2.5	7.9	10.6
2.5	3.2	10.6	14.3
3.2	4.3	14.3	19.5
4.3	5.8	19.5	>19.5

just the first six types with hydroxyls, for example, being included in the definitions of donors and acceptors.

Combining 15 distance ranges and 4-point pharmacophores, it is possible to arrive at several hundred million geometrically valid pharmacophores. For the 4-point pharmacophores, we use 10 distance ranges, resulting in a key containing 24.4×10^6 bits. The distance ranges used in this work are shown in Table 1. A ChemLib routine has been coded to write individual pharmacophore keys to disk for subsequent postprocessing and gives considerable saving in terms of disk space. A standard Chem-X 4-point key takes about 3 MB. The ChemLib routine utilizes the internal storage architecture of Chem-X to write out the bit strings for matched geometries (combinations of 4-centers) only. The exact size depends on the numbers and types of pharmacophores set, but a 4-point key with 5000 pharmacophores occupies about 40 KB. 3-D structures are generated from SMILES[44] strings using CONCORD.[45,46]

Multipharmacophore Searching. The pharmacophore bit strings can be used in exactly the same way as standard 2-D keys. Thus, it is possible to compare the keys by counting the number of set bits in common or calculating similarity coefficients. A program was written in C to compare the pharmacophore keys of one or more probe molecules to individual molecule pharmacophore keys, outputting the number of set bits (pharmacophores) in common. A second program uses this information to rank the compounds on the basis of their similarity to the probe(s) using a number of measures including the number of set bits in common and the Tanimoto coefficient.

Rule-of-5. The parameters required for assessing the Rule-of-5 are described above. Programs have been written using the Daylight Toolkit[47] to calculate the number of donors and acceptors according to the Lipinski definitions.[33] ClogP[34] is used to estimate log *P*.

Polar Surface Area. Polar surface areas were calculated using the method described by Clark.[38] The molecule is encoded as its neutral species in SMILES[44] format with appropriate stereochemical designations where the stereochemistry is known. CONCORD[45,46] is used to convert the SMILES representation into an approximate 3-D structure. This conformation is then energy-minimized using the maximin2 minimizer in SYBYL.[48] Minimization is terminated after either 1000 iterations or when a gradient of less than 0.05 kcal mol^{-1} $Å^{-1}$ is attained. The minimized conformation is passed to the MOLVOL program developed by Dodd and Theodorou.[49] MOLVOL computes the van der Waals molecular surface area for the conformation and outputs the contributions of the individual atoms to the surface area. The atomic radii used are very similar to those employed by Palm et al.[37] Finally, an in-house Fortran program sums the contributions of the polar atoms (N, O, and H attached to N or O) and outputs the PSA value. This procedure requires approximately 10 CPU seconds (SGI R10000) per molecule, most of which is spent in the optimization step.

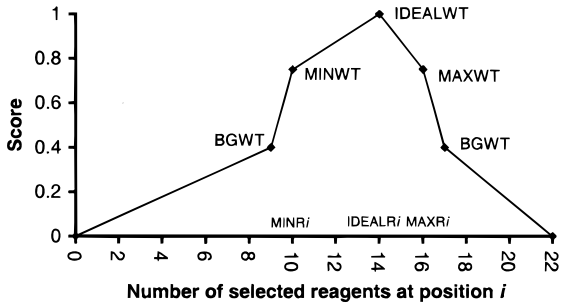

Figure 3. Scoring function for the Monte Carlo reagent selection program. MINR*i*, IDEALR*i*, and MAXR*i* are user-defined minimum, desired, and maximum number of reagents at position *i*. MINWT, IDEALWT, and MAXWT are the associated weights. BGWT is a background weight defined to guide solutions that are far from a good score toward a good solution.

Since the work reported in this paper was carried out, we have significantly speeded up the above procedure by omitting the optimization step and using the SAVOL3[50] program to calculate the polar surface area from the CONCORD-generated conformation. The procedure is very fast, allowing the processing of 10 or more structures per second, but the output PSA values still give an excellent correlation ($r^2 = 0.94$) with the human fractional absorption data presented by Palm et al.[37]

Monte Carlo Algorithm for Library Design. As shown by the example below, applying Rule-of-5 and PSA criteria to a virtual library can lead to synthetically inefficient (noncombinatorial) designs. A Monte Carlo procedure was therefore developed to choose a subset from the virtual library such that the Rule-of-5 and PSA criteria are satisfied in each product while ensuring that the combination of reagents is as near to the full combinatorial selection as possible. The Monte Carlo search is carried out in reagent space but evaluates the score for each potential product. The method is exemplified for two reagent positions but is extendable to more.

Assume a virtual library of size **N** from a combination of **A** × **B** reagents. The user specifies the following input files: acceptable virtual library products and the reagents, the score file, and, optionally, a list of reagents that must be included. The score file specifies for each reagent position, *i*, the minimum allowable, maximum allowable, and ideal number of reagents to be selected at that position, MINR*i*, MAXR*i*, IDEALR*i*, and the associated weights for the scoring function, MINWT, MAXWT, IDEALWT, and BGWT. BGWT is the background weight for a number of selected reagents MINR*i* − 1 or MAXR*i* + 1 and is useful to guide poorly scoring solutions toward a good solution. The reagents are coded by a bit string of length **A** + **B** with 1 marking the inclusion of the reagent in the library. To initialize the procedure, a random set of bits are set on: **a** from **A** and **b** from **B**. The full combinatorial sublibrary **a** × **b** is generated and scored. The scoring function is a series of simple linear functions for each position, *i*, as shown in Figure 3, where the *x*-axis indicates the number of reagents currently selected at position *i*. The user defines MINWT, MAXWT, IDEALWT, and BGWT in the score file. In

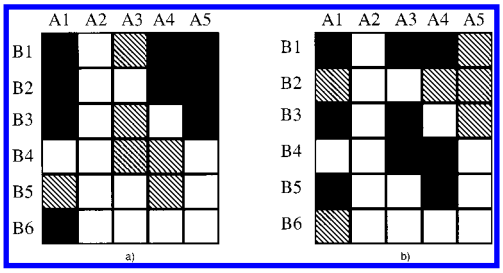

Figure 4. Selecting nine products from a 5 × 6 virtual library. The shaded squares represent products that satisfy some user-defined criteria (e.g., PSA). The darker shading represents a possible selection of nine products satisfying the criteria. (A) Nine products are selected from 3 × 4 reagents, but reagent B6 is used only once. (B) Nine products are selected from 3 × 4 reagents with each reagent used at least twice.

addition, the number of acceptable products in the sublibrary is counted, NACC. A penalty function is included to penalize a solution where NACC is less than a user-defined minimum. Initial trials with this scoring function showed that an additional constraint is also required and is best understood by reference to Figure 4. Let us assume that we require a library of nine products from a virtual library of 5 × 6 products. The shaded squares represent the products that satisfy some user-defined criteria. It can be seen that no 3 × 3 subset of R1 and R2 exists. However, several 3 × 4 subsets with nine good products exist. In the solution indicated by the heavier shading in Figure 4a, one of the R2 reagents is used only once (B6) and the R1 groups are used between two and four times. This is not synthetically efficient and is of particular concern when much effort is required to prepare a reagent or attach it to a resin. The problem is exacerbated when considering libraries of several hundred products. From a synthetic efficiency point-of-view, the solution indicated by the heavier shading in Figure 4b is preferred where each R1 is used three times and each R2 at least twice. Thus, an additional constraint was added to the scoring function at the request of the chemists to ensure that each reagent position is used a user-defined minimum number of times in the final solution. Once the solution has been scored, the usual Metropolis criterion is used to decide whether to select the solution. The next potential solution is then generated by randomly mutating one of the **A** + **B** bits from 0 to 1 or vice versa. Acceptable solutions can be obtained in a matter of minutes on a SGI R10000 processor.

4. MULTIPHARMACOPHORE SEARCHING—FOCUSED LIBRARY DESIGN

Reagent Profiling. A commonly encountered situation in lead discovery is an initial screening library that shows a preference for a particular group at one position. For a follow-up library, a medicinal chemist often requests a focused set of reagents to be selected that are similar to this substituent. Take Figure 5a as an example, where we wish to select reagents similar to the structure shown. Some of the results from a 2-D search for acids taken from the ACD are shown in Figure 5b. Daylight fingerprints have been used as the similarity measure. This set is conservative (they are all benzoic acids) and structurally homogeneous. Figure 5c shows examples of compounds selected using pharmacophore similarity with the lead. Three-point pharmacophore keys were calculated for the reagents, and they were ranked

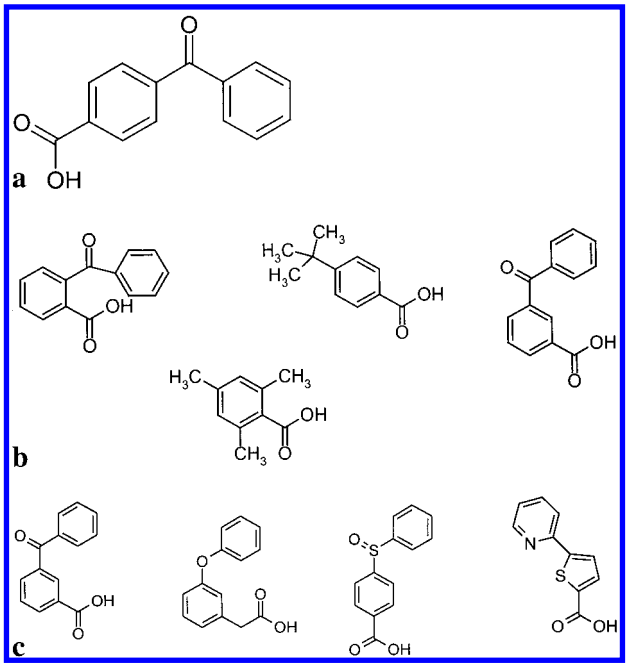

Figure 5. Reagent selection using 2-D and 3-D similarity searching methods. (a): Target reagent. (b) Some reagents selected by 2-D similarity searching. (c) Some reagents selected by multipharmacophore searching.

on pharmacophore overlap with the lead. The set is more structurally diverse than the 2-D set. Potentially important pharmacophores from the lead such as the carbonyl acceptor are picked up in a variety of ways, e.g., the pyridine nitrogen and the sulfoxide oxygen. Phenyl acetic as well as benzoic acids are present. There is much additional information to be gained by combining these approaches. The 2-D approach will detect close analogues for very local exploration, while the multipharmacophore similarity results in a larger structural diversity but still remains focused on the potentially important pharmacophores within the lead. We have used this combination of approaches with some success in lead optimization projects, including one instance where the 2-D search resulted in only one analogue of the substituent of interest. In this case, the 3-D multipharmacophore approach was crucial to the design.

Screening Set Selection/Product-Based Design. Screening set selection can proceed in an analogous fashion to the reagent profiling. The pharmacophore key of the lead compound(s) is calculated and used as a probe against the keys of the proposed virtual library or available screening compounds. To show how this may work in practice, we take the example of selecting a screening set of fibrinogen receptor antagonists. Fibrinogen binds to its receptor via the RGD motif.[51] Solution NMR and crystal structures of the fibronectin type III domain containing this motif[52−54] show it to be located in a disordered loop region, suggesting a degree of flexibility. A number of compounds have been identified as fibrinogen receptor antagonists covering a wide range of structural classes;[55] see Figure 6. Given the flexibility of the RGD loop, it would be difficult to identify a reasonable single pharmacophore from a knowledge of the crystal structure alone.

The 4-point pharmacophore keys for a 100 000 compound subset of the RPR collection were calculated and stored on

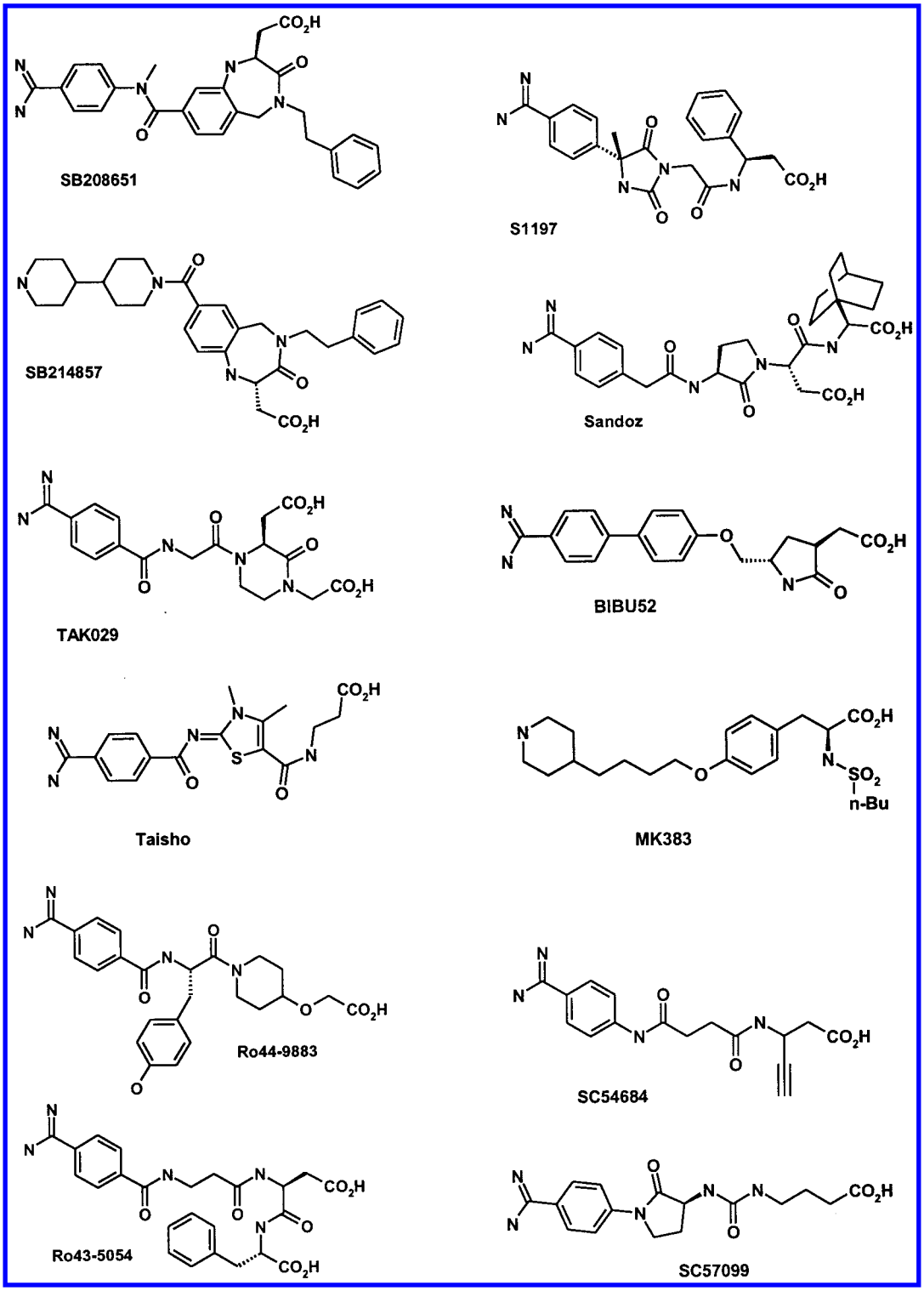

Figure 6. Structurally diverse fibrinogen receptor antagonists.

disk as described in Methods. The 4-point pharmacophore keys of the RGD tripeptide, N and C-capped with amide groups, and for each of the compounds in Figure 6 were calculated with a full systematic conformational analysis. The RGD key was then used as a probe against the combined (RPR + Fibrinogen) set. Results are presented in Table 2, compounds having been ranked by pharmacophore overlap with RGD. The following information is presented: the number of hits, i.e., compounds with at least one pharmacophore in common with RGD, the highest and lowest rankings of known antagonists from Figure 6, and how many of these are found in the top ranked 50, 100, ..., 1000 compounds. Screening just 1% of the database would have found 7 out of 12 known actives in the data set with all 12 found in the top 3%. Table 2 also gives the results obtained using each of the known antagonists as a probe. While the results are variable, in the best case (BIBU52), all other antagonists are found in the top 1000 ranked compounds.

This example shows that the multipharmacophore descriptors contain important information relating to the biological activity. It is straightforward to extend this approach to combinatorial library design by considering it as a con-

Table 2: Results of Multipharmacophore Similarity Searching for Fibrinogen Receptor Antagonists (See Text for Explanation)

probe	hits	top	lowest	N50	N100	N150	N200	N500	N1000
RGD	23 802	6	2982	2	4	4	4	6	8
BIBU52	37 724	1	822	4	4	5	5	6	11
MK383	57 756	2	11178	2	2	2	2	5	5
RO43-5054	73 288	1	4190	2	2	2	2	5	7
RO44-9883	64 126	1	3238	2	2	3	3	5	6
S1197	63 960	1	2422	3	3	3	3	6	7
SANDOZ	66 543	1	3696	2	2	3	3	5	6
SB208651	47 853	1	5654	4	5	5	5	6	6
SB214857	48 127	7	18016	2	3	4	4	5	6
SC54684	11 220	1	1999	6	7	7	8	8	10
SC57099	8 154	1	2257	7	7	7	7	9	10
TAISHO	27 044	1	1492	6	7	7	8	9	9
TAK029	38 641	1	2216	4	5	5	5	6	9

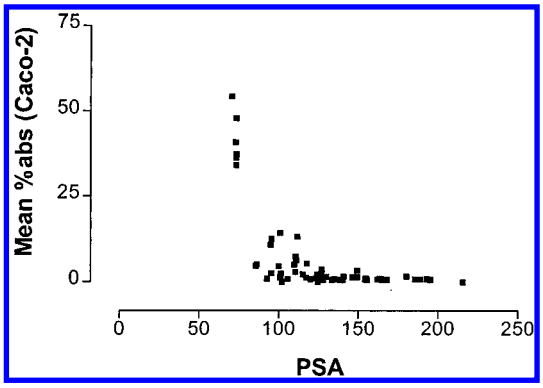

Figure 7. In-house experimental Caco-2 absorption data plotted against PSA (in Å^2). The percent absorbed figure is calculated by simply taking the peak area for the acceptor chamber and dividing it by the value for the donor chamber. The surface area of the inserts is 0.33 cm^2 (Costar HTS 24 well plate). The value shown is the mean taken from two experiments each run over a 3 h time period.

strained application of the screening set selection methodology. We have implemented a GA-based combinatorial selection procedure that optimizes the average pharmacophore similarity of the selected library subset with a lead or leads by considering the individual product contributions.[56]

5. DESIGNING IN DRUG-LIKENESS

In Section 4, the discussion focused on what may be termed chemical diversity or similarity, and we have shown that the multipharmacophore descriptors are also relevant for biological similarity. However, lead optimization involves the discovery of compounds that are not only active but also possess the correct pharmacokinetic properties. Transport properties are particularly important for compounds that are to be taken orally and blood−brain barrier penetration is either to be desired or avoided depending upon the therapeutic target. The Rule-of-5[34] provides one filter to assess a compound's potential for absorption. The work of Palm et al.[37] using polar surface area (PSA) as a descriptor was mentioned earlier. One of the findings from this group's work was that a compound possessing a PSA >140 Å^2 is likely to show less than 10% fractional absorption in humans. Our work using single-conformation-based PSA values has borne out this finding with a larger literature data set,[38] as have in-house data from a Caco-2 absorption assay (see Figure 7).

Blood−brain barrier penetration is another consideration of importance in some drug discovery projects. PSA has shown itself to be a useful descriptor in this context too.[39−41]

Table 3: Comparison of Mean Absolute Errors of Prediction of Different log *BB* Prediction Methods Using Three Test Sets from the Literature[a]

test set	number of compounds	mean absolute errors of prediction				
		Abraham[58]	Lombardo[59]	Norinder[60]	Luco[61]	RPR[41]
1[58]	7	0.30	N. A.	N. A.	0.42	0.37
2[59]	5	N. A.[a]	0.41	0.52	0.29	0.24
3[61]	25	N. A.	N. A.	N. A.	0.43	0.50

[a] N.A. indicates that no results were presented for this test set.

Equation 1[41] was derived from the Abraham data set[57] using the PSA calculation procedure mentioned earlier and the ClogP program:[35]

$$\log BB = -0.0148(\pm 0.001)\mathrm{PSA} + 0.152(\pm 0.036)C\log P + 0.139(\pm 0.073)$$
$$n = 55,\ r = 0.887,\ s = 0.354,\ F = 95.8 \quad (1)$$

where n is the number of compounds, r is the correlation coefficient, s is the standard error, and F is the Fisher value, a measure of the statistical significance of the equation. The standard errors of the correlation coefficients are given in parentheses.

The predictive capability of eq 1 has been tested on three test sets taken from the literature. Table 3 compares the performance of eq 1 with a number of other methods for predicting BBB penetration.[58−61] As can be seen, eq 1 performs more-or-less equally well in terms of prediction compared to these other methods and, in our view, has several advantages over them. First, it is very rapid to calculate compared to methods such as those of Lombardo et al.[59] and Norinder et al.[60] that require semiempirical or ab initio molecular orbital calculations combined with conformational analysis. Second, it is fully automated, requiring no human intervention in contrast to the method of Abraham et al.,[57,58] which currently requires manual dissection of the compound of interest and summation of descriptor values for the resulting fragments. (We note that recent work has sought to automate this procedure.[62]) Finally, the quantities that comprise eq 1 are easily interpretable; it is thus straightforward for a medicinal chemist to conceive alterations to the molecular structure in order to alter the values of PSA and/or *C* log *P* and thereby change the predicted log *BB*. This contrasts with the approach of Luco[61] wherein many topological descriptors are input to a partial least-squares procedure. It is not always straightforward to use the principal components output by such a method directly in the process of design.

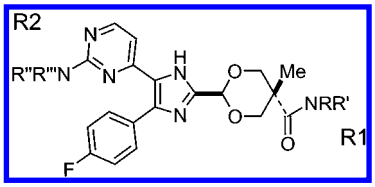

Figure 8. Generic structure of designed library.

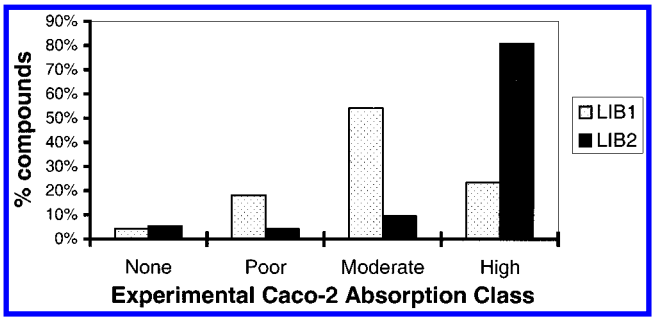

Figure 9. Comparison of experimental Caco-2 absorption data for a fully combinatorial library (LIB1) and a library designed to have good absorption properties (LIB2) (see text).

This section concludes with an example from some recent work where we have applied such ideas in the design of a lead optimization library. Figure 8 shows the core structure of a two-component library.[63] A first library in this series (LIB1) was synthesized in a fully combinatorial manner with one of us (I.M.M.) selecting reagents considering some of the properties above but in a nonoptimal and nonautomated manner. It was largely the difficulties in performing this selection that led us to develop the Monte Carlo search algorithm described in Methods for a second library (LIB2). The virtual library for LIB2 contained 1485 compounds, and these were filtered using the Rule-of-5 and PSA to give 770 compounds. The initial goal had been to synthesize a library of 400 compounds in a 20 × 20 array. Unfortunately, no such subset of the 770 compounds exists. The selection of a near-optimal subset is a nontrivial exercise by hand, and so the Monte Carlo algorithm described in Methods was applied. Chemistry considerations (manuscript in preparation) limited the desired number of R2 reagents to 20. This was further complicated by the desire to include a small number of R2 groups known to be active from a previous library. More flexibility was allowed in the choice of R1. The set of constraints can be summarized as follows: only include products that meet the Rule-of-5 and PSA design criteria, $N(\mathrm{R1}) \geq 20$, $N(\mathrm{R2}) = 20$, select at least 400 compounds, use each R2 at least 20 times (see Methods for a fuller explanation of this constraint), and include certain R2. Upon application of the Monte Carlo search algorithm, the preferred solution selected 24 R1 groups and 20 R2 (all used at least 20 times), giving a total of 449 products satisfying the design criteria.

The designed library was synthesized, and Figure 9 compares the experimental Caco-2 absorption profiles of libraries LIB1 and the fully designed library, LIB2. The classification into low (<2% absorbed), medium (2−20% absorbed), and high (>20% absorbed) absorption is applied in the same manner across all measurements by the group performing the tests and hence is not specific to this case. It is clear that LIB2 has a much improved profile compared to LIB1. However, just as importantly, this improvement in the absorption characteristics was not at the expense of biological activity. For LIB1, 60% of the compounds were more active than a target compound, while for LIB2, the corresponding figure was 85%.

6. CONCLUSIONS

In this paper, we have addressed several issues in the design of lead optimization libraries. Multipharmacophore descriptors were first developed in the context of designing diverse compound libraries. One reason for favoring such descriptors is the importance of the pharmacophore hypothesis in understanding the interaction of a compound with a protein target. Linked to this is the proposal that sampling over all potential pharmacophores leads to diversity in a biologically relevant space.[9] The results in section 4 provide some support for this argument. However, in the context of this paper, the use of the RGD pharmacophore key to identify fibrinogen antagonists shows that such methods are also applicable to the design of focused libraries where the aim is to design the library toward a known lead or leads. This is important because it means that the same descriptors can be used for diverse library design, screening set selection, and focused library design, giving a consistent approach. In addition, recent reports have shown that such descriptors are also applicable in the context of site-directed design.[10,42]

In section 5, we addressed the question of designing libraries with improved pharmacokinetic properties. It is possible to derive simple and rapidly computable descriptors applicable to the prediction of drug transport properties. Furthermore, these can be applied in the context of library design, although it may be necessary to synthesize libraries in a noncombinatorial manner to obtain the best results. This lack of 100% synthetic efficiency is potentially an issue even for smaller lead optimization libraries, and so we have developed a Monte Carlo search procedure that allows the selection of a near-combinatorial subset in which all library members satisfy the design criteria while being as synthetically efficient as possible. Application of these methods to a real-life example led to the synthesis of compounds with improved absorption properties as measured by experiments using Caco-2 monolayers. Such an approach should be equally applicable in a structure-based design project where a docking algorithm could be used to filter the virtual library. In conclusion, consideration of calculated log *P*, molecular weight, and polar surface area in the design of libraries can give compounds with improved absorption characteristics and provides a valuable tool to the medicinal chemist in lead optimization.

ACKNOWLEDGMENT

We thank Clive Brealey, Nicky Wilsher, Martin Barrett, and Gary Wilkinson for the Caco-2 absorption data and Richard Lewis (now Eli Lilly, U.K.) and Paul Bamborough (now GlaxoWellcome, U.K.) for helpful discussions in the early stages of this work.

REFERENCES AND NOTES

(1) Dolle, R. E. Comprehensive Survey of Chemical Libraries Yielding Enzyme Inhibitors, Receptor Agonists and Antagonists, and Other Biologically Active Agents: 1992 through 1997. *Mol. Diversity* **1998**, *3*, 199−233.

(2) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem*. **1995**, *38*, 1431−1436.
(3) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(4) Good, A.C.; Lewis, R. A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick. *J. Med. Chem*. **1997**, *40*, 3926−3936.
(5) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in Diversity Profiling and Combinatorial Series Design. *Mol. Diversity* **1998**, *4*, 1−22.
(6) Mason, J. S.; Hermsmeier, M. A. Diversity Assessment. *Curr. Opin. Chem. Biol*. **1999**, *3*, 342−349.
(7) Drewry, D. H.; Young, S. S. Approaches to the Design of Combinatorial Libraries. *Chemom. Intell. Lab. Syst.* **1999**, *48*, 1−20.
(8) Lewis, R. A.; Pickett, S. D.; Clark, D. E. Computer-Aided Molecular Diversity Analysis and Combinatorial Library Design. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, in press; Vol. 16.
(9) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1214−1223.
(10) Mason, J. S.; Pickett, S. D. Partition-based Selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.
(11) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB—Strategies for the Design and Comparison of Combinatorial Libraries using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.
(12) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 644−649.
(13) Fisanick, W.; Lipkus, A. H.; Rusinko, A. Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 130−140.
(14) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 7, pp 1−66.
(15) Good, A. C.; Mason, J. S. Three-Dimensional Structure Database Searches. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 7, pp 67−118.
(16) Wang, S.; Milne, G. W. A.; Yan, X.; Posey, I.; Nicklaus, M. C.; Graham, L.; Rice, W. G. Discovery of Novel, Non-Peptide HIV-1 Protease Inhibitors by Pharmacophore Searching. *J. Med. Chem.* **1996**, *39*, 2047−2054.
(17) Astles, P. C.; Brown, T. J.; Handscombe, C. M.; Harper, M. F.; Harris, N. V.; Lewis, R. A.; Lockey, P. M.; McCarthy, C.; McLay, I. M.; Porter, B.; Roach, A. G.; Smith, C.; Walsh, R. J. A. Selective Endothelin A Receptor Ligands. 1. Discovery and Structure−Activity of 2,4-Disubstituted Benzoic Acid Derivatives. *Eur. J. Med. Chem*. **1997**, *32*, 409−423.
(18) Marriott, D. P.; Dougall, I. G.; Meghani, P.; Liu, Y. J.; Flower, D. R. Lead Generation Using Pharmacophore Mapping and Three-Dimensional Database Searching: Application to Muscarinic M3 Receptor Antagonists. *J. Med. Chem*. **1999**, *42*, 3210−3216.
(19) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.
(20) Moock, T. E.; Henry, D. R., Ozkabak, A. G.; Alamgir, M. Conformational Searching in ISIS/3D Databases. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 184−189.
(21) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.
(22) Clark, D. E.; Jones, G.; Willett, P.; Kenny, P. W.; Glen, R. C. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 197−206.
(23) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm To Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 310−320.
(24) Cho, S. J.; Zheng, W.; Tropsha, A. Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Peptide Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259−268.
(25) Lewis, R. A.; Good, A. C.; Pickett, S. D. Quantification of Molecular Similarity and Its Application to Combinatorial Chemistry. In *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, 1997; pp 135−156.
(26) World Drug Index; Derwent Information (http://www.derwent.com).
(27) Available Chemicals Directory (ACD); Molecular Design Limited, San Leandro, CA 94577.
(28) The SPRESI database is distributed by Daylight Chemical Information Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691.
(29) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish Between Drug-like and Non-Drug-like Molecules? *J. Med. Chem*. **1998**, *41*, 3314−3324.
(30) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem*. **1998**, *41*, 3325−3329.
(31) Wagener, M.; van Geerestein, V. J. Analysing Large Datasets with Decision Trees: Discriminating between Potential Drugs and Nondrugs. Abstracts of the 217th American Chemical Society Meeting, March 21−25, 1999, Anaheim, CA; American Chemical Society: Washington, DC, 1999; COMP-25.
(32) Gillet, V. J.; Willett, P.; Bradshaw J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(33) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.
(34) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(35) ClogP. Daylight Chemical Information Software, version 4.51; Daylight Chemical Information Inc.: 27401 Los Altos, Suite #370, Mission Viejo, CA 92691.
(36) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(37) Palm, K.; Stenberg, P.; Luthmann, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res*. **1997**, *14*, 568−571.
(38) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J. Pharm. Sci*. **1999**, *88*, 807−814.
(39) Kansy, M.; van de Waterbeemd, H. Hydrogen-Bonding Capacity and Brain Penetration. *Chimia* **1992**, *46*, 299−303.
(40) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Chretien, J. R.; Raevsky, O. A. Estimation of Blood−Brain Barrier Crossing of Drugs Using Molecular Size and Shape, and H-bonding Descriptors. *J. Drug Targeting* **1998**, *6*, 151−165.
(41) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and its Application to the Prediction of Transport Phenomena. 2. Prediction of Blood−Brain Barrier Penetration. *J. Pharm. Sci*. **1999**, *88*, 815−821.
(42) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem*. **1999**, *42*, 3251−3264.
(43) ChemDiverse. Oxford Molecular Group plc, The Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, United Kingdom.
(44) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(45) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*, 1−7.
(46) CONCORD v4.02; Balducci, R.; McGarity, C. M.; Rusinko, A., III; Skell, J.; Smith, K.; Pearlman, R. S (University of Texas at Austin). Distributed by Tripos, Inc.: 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144.
(47) SMILES Toolkit. Daylight Chemical Information Software, version 4.51; Daylight Chemical Information Inc.: 27401 Los Altos, Suite #370, Mission Viejo, CA 92691.
(48) SYBYL 6.4.2; Tripos, Inc.: 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144.
(49) Dodd, L. R.; Theodorou, D. N. Analytical Treatment of the Volume and Surface Area of Molecules Formed by an Arbitrary Collection of Unequal Spheres Intersected by Planes. *Mol. Phys.* **1991**, *72*, 1313−1345.
(50) Pearlman, R. S.; Skell, J. M.; Deanda, F. SAVOL3: Algorithms for Atomic Contributions to Molecular Surface Areas and Volumes; available from Prof. R. S. Pearlman, Laboratory for Molecular Graphics and Theoretical Modeling, College of Pharmacy, University of Texas, Austin, TX 78712 (Email: pearlman@vax.phr.utexas.edu).
(51) Eldred, C. D.; Judkins, B. D. Fibrinogen Receptor Antagonists: Design and Clinical Applications. *Prog. Med. Chem*. **1999**, *36*, 29−90.

(52) Main, A. L.; Harvey, T. S.; Baron, M.; Boyd, J.; Campbell, I. D. The Three-Dimensional Structure of the Tenth Type III Module of Fibronectin: An Insight into RGD-mediated Interactions. *Cell* **1992**, *71*, 671−678.

(53) Dickinson, C. D.; Veerapandian, B.; Dai, X. P.; Hamlin, R. C.; Nguyen, H. X.; Ruoslahti, E.; Ely, K. R. Crystal Structure of the Tenth Type III Cell Adhesion Module of Human Fibronectin. *J. Mol. Biol.* **1994**, *236*, 1079−1092.

(54) Leahy, D. J.; Aukhil, I.; Erickson, H. P. 2.0 Å Crystal Structure of a Four-domain Segment of Human Fibronectin Encompassing the RGD Loop and Synergy Region. *Cell* **1996**, *84*, 155−164.

(55) Mousa, S. A.; Cheresh, D. A. Recent Advances in Cell Adhesion Molecules and Extracellular Matrix Proteins: Potential Clinical Implications. *Drug Discovery Today* **1997**, *2*, 187−199.

(56) Pickett, S. D.; Clark, D. E.; Lewis, R. A. Multi-pharmacophore Descriptors for 3D Similarity Searching and Design. Abstract of the 215th ACS National Meeting, 29 March−2 April, 1998, Dallas, TX; American Chemical Society; Washington, DC, 1998; COMP 009.

(57) Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen Bonding. 33. Factors that Influence the Distribution of Solutes Between Blood and Brain. *J. Pharm. Sci.* **1994**, *83*, 1257−1268.

(58) Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen Bonding. Part 36. Determination of Blood Brain Distribution Using Octanol−Water Partition Coefficients. *Drug Des. Discovery* **1995**, *13*, 123−131.

(59) Lombardo, F.; Blake, J. F.; Curatolo, W. J. Computation of Brain-Blood Partitioning of Organic Solutes via Free Energy Calculations. *J. Med. Chem.* **1996**, *39*, 4750−4755.

(60) Norinder, U.; Sjöberg, P.; Österberg, T. Theoretical Calculation and Prediction of Brain-Blood Partitioning of Organic Solutes Using MolSurf Parametrization and PLS Statistics. *J. Pharm. Sci.* **1998**, *87*, 952−959.

(61) Luco, J. M. Prediction of the Brain−Blood Distribution of a Large Set of Drugs from Structurally Derived Descriptors Using Partial Least-Squares (PLS) Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396−404.

(62) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Free Energy Relation Descriptors using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 835−845.

(63) Bamborough, P. L.; Collis, A. J.; Halley, F.; Lewis, R. A.; Lythgoe, D. J.; McKenna, J. M.; McLay, I. M.; Porter, B.; Ratcliffe, A. J.; Wallace, P. A.; et al. Preparation of Imidazolyl-cyclic Acetals as TNF-alpha Inhibitors; PCT Int. Appl. (1998) WO9856788.

CI990261W