Statistical Molecular Design of Building Blocks for Combinatorial Chemistry

Anna Linusson,*,† Johan Gottfries,‡ Fredrik Lindgren,§ and Svante Wold†

Research Group for Chemometrics, Umeå University, S-901 87 Umeå, Sweden, AstraZeneca R&D, S-431 83 Mölndal, Sweden, and Umetrics Office, Amiralsgatan 20, S-211 55 Malmö, Sweden

Received July 19, 1999

The reduction of the size of a combinatorial library can be made in two ways, either base the selection on the building blocks (BB's) or base it on the full set of virtually constructed products. In this paper we have investigated the effects of applying statistical designs to BB sets compared to selections based on the final products. The two sets of BB's and the virtually constructed library were described by structural parameters, and the correlation between the two characterizations was investigated. Three different selection approaches were used both for the BB sets and for the products. In the first two the selection algorithms were applied directly to the data sets (D-optimal design and space-filling design), while for the third a cluster analysis preceded the selection (cluster-based design). The selections were compared using visual inspection, the Tanimoto coefficient, the Euclidean distance, the condition number, and the determinant of the resulting data matrix. No difference in efficiency was found between selections made in the BB space and in the product space. However, it is of critical importance to investigate the BB space carefully and to select an appropriate number of BB's to result in an adequate diversity. An example from the pharmaceutical industry is then presented, where selection via BB's was made using a cluster-based design.

Introduction

In the pharmaceutical industry the search for new leads and the following structural optimization is of critical importance. Traditionally, screening of natural products together with pharmacological and physio/physiopathological knowledge has been the major source for lead discovery. However, during the last 10 years combinatorial chemistry, along with high-throughput screening (HTS), has developed as a new route to drug discovery.

In combinatorial chemistry a large library of compounds is generated (synthesized), as combinations of building blocks (BB's). These molecules are tested for biological activity, by HTS, with the scope to identify a novel molecule that binds to a receptor, an enzyme, or any other molecular target of interest.[1−3] However, the number of compounds that can be synthesized and tested by means of combinatorial chemistry is limited, and a selection seems necessary. The selection can be made randomly or in a systematic way, e.g., by statistical design.[4−8] The use of statistical designs results in a well-balanced data set suitable for mathematical modeling, i.e., quantitative structure−activity relationship (QSAR). When the relationships between the chemical structures and the biological activity are investigated in this manner, the results are reliable, interpretable, and thus more useful for the next step in drug development. The optimal selection should comprise high diversity (great variety) and should have full coverage of the property space but no redundancy. However, this optimum is difficult to find since the guarantee of high diversity and full coverage often leads to redundancy.

In combinatorial chemistry the reduction of compounds to synthesize can be performed in two different ways, either to make the selection of the BB's or to virtually construct the final products and use them for the selection. Gillet et al. have claimed that more diverse libraries result from applying dissimilarity-based compound selection on the product level rather than at the BB level, using the Daylight fingerprint representation of the molecules.[9] We here propose that this is not a general outcome and that BB-based selections can provide as good diversity and coverage as product-based selections. Applying the design to the BB space rather than to the full library has several advantages: (i) fewer compounds need to be characterized, allowing more advanced methods to be used that yield more reliable results, (ii) the number of different BB's will be reduced, which facilitates the combinatorial process, and (iii) the interpretation and use of QSAR will be more direct, since the output will give information directly about desired properties of the BB's.

To address this question, we further investigate in this paper the consequences of making the selection in the BB space rather than in the product space and what kind of selection method that is preferable. In a first example, virtual BB libraries of 20 carboxylic acids and 20 amines were constructed. The correlation between the characterization of the BB's and the resulting products was investigated. Three selection methods were tested: in the first two, selection algorithms are used directly on the data set, while in the third, a cluster analysis precedes the selection algorithm. For comparison the three methods were applied both to the two BB

* Corresponding author. Tel: +46-(0)90-786 69 62. Fax: +46-(0)-90-13 88 85. E-mail: Anna.Linusson@chem.umu.se.
† Umeå University.
‡ AstraZeneca R&D.
§ Umetrics Office.

10.1021/jm991118x CCC: $19.00 © 2000 American Chemical Society
Published on Web 03/08/2000

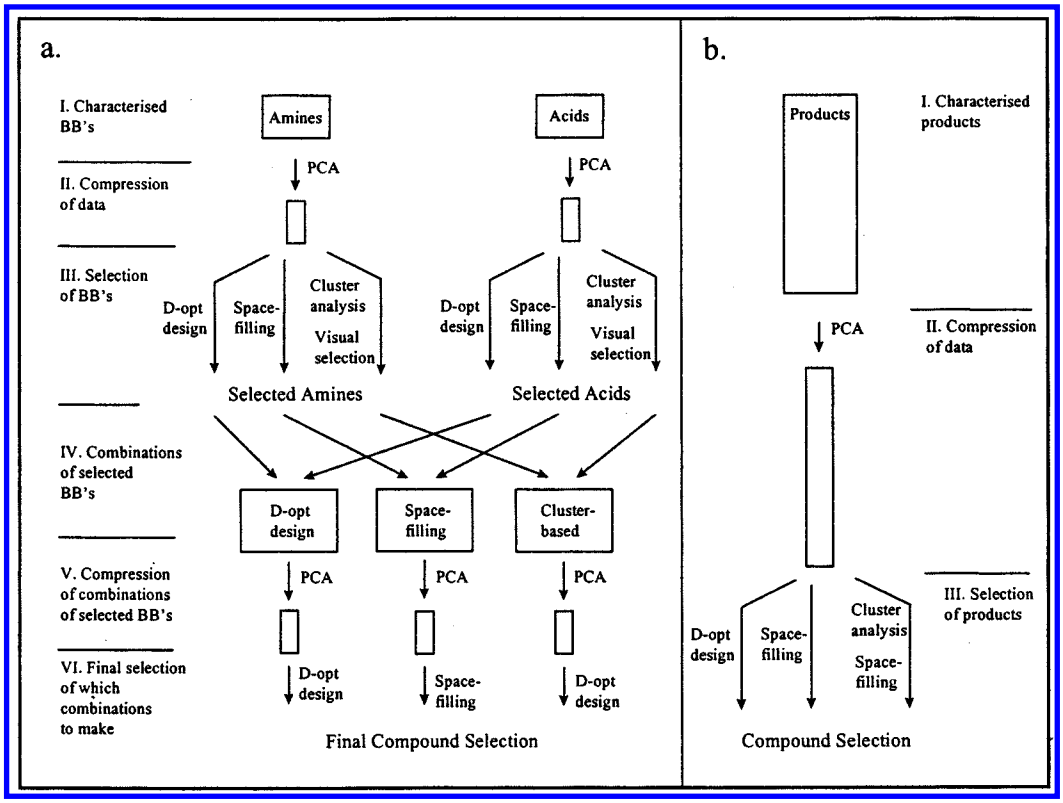

Figure 1. Flowchart over the selection procedure in the BB space (a) and in the product space (b).

sets and to the full product library of 400 amides (Figure 1). The results of the two approaches (BB design and product library design) were quantitatively evaluated and compared regarding diversity and coverage.

The results from these virtual libraries were then used in a real combinatorial chemistry problem (example 2). A peptide-like structure, with promising bioactivity, was used as a starting point for optimization of pharmcological activity.

Diversity, Coverage, and Redundancy

Diversity is a widely used but little defined term in the combinatorial chemistry field. Loosely diversity is easily understood as spread over a given space and equivalently lack of similarity within a set of compounds. Statistical design is developed to make a selection of cases or experiments (here molecules) as representative as possible given the varying factors, their ranges, and a model. Thereby the use of statistical design in a certain sense automatically maximizes the diversity of the selected set. The measures of adequacy of a design (determinant, condition number, etc.) hence are quantitative and direct measures of diversity.

Coverage is in a way the opposite of diversity and means, again loosely, how well a space is covered by a set. With a linear model, a statistical design tends to have rather poor coverage since only peripheral points are selected. Hence, to improve coverage, at least one center point should be included for linear designs, or alternatively, quadratic or cubic designs can be used. The latter two give substantially more design points than a linear model.

Redundancy is a term with negative association, indicating that unnecessary "much" has been done. In the present context redundancy tends to mean that an unnecessarily large library has been selected to solve a given problem. However, since a set of structural descriptors never is complete, and since its summarization by principal component analysis (see below) by necessity leads to an approximate set of coordinates, some redundancy may be recommendable. Thus, adding a randomly selected set of compounds to the designed set provides some additional security that important features have not been neglected. How large this additional "random" set should be is, of course, related to the actual circumstances.

Computational Methods

Characterization of Structures. In example 1, 20 carboxylic acids and 20 amines were used as the two sets of BB's. This initial selection was made with the objective of diversity. From these 20 × 20 BB's, 400 amides were computer generated. Both the BB's and the 400 amides were characterized by the same 37 descriptors (Table 1, Figure 1a: I, 1b: I) calculated by an "in-house" program developed at AstraZeneca R&D.[10] In example 2, the peptide-like structures were divided in four parts: one constant scaffold and three varied fragments. The latter comprised 12 basic, 5 specific, and 8 acidic fragments. From these parts, a virtual library containing the 480 possible molecules was generated. The 25 fragments and the 480 compounds were characterized by 13 semiempirical descriptors (Table 2) using Spartan (AM1).[11]

Data Analytical Methods: PLS and PCA. Two data analytical methods were used in both examples: partial least-squares projections to latent structures (PLS) and principal component analysis (PCA).

PLS is a regression method that calculates a projection which both takes the variance (in **X**) and the correlation between **X** and **Y** into consideration, i.e., finds latent structures in **X** that correlate with latent structures in **Y**.[12−14] PLS was used to investigate whether the characterization of the BB's contained sufficient information to describe the amides, i.e., to investigate the correlation between the reactants and the

Table 1. Structural Descriptors[10] Used for the Characterization of the Carboxylic Acids, Amines, and Amides

no.	name of descriptor
1	molecular weight
2	number of H-bond donors
3	number of H-bond acceptors
4	proportion of hydrogens
5	proportion of heteroatoms
6	proportion of halogens
7	proportion of fluorine
8	proportion of chlorine
9	proportion of iodine
10	proportion of carbon
11	proportion of sulfur
12	proportion of oxygen
13	proportion of nitrogen
14	number of rings
15	number of bonds
16	number of rotatable bonds
17	proportion of rotatable bonds
18	Balaban index
19	Centric index
20	Zagreb M1 index
21	Zagreb M2 index
22	Randic 0th index
23	Randic 1th index
24	sum of electrotopological values over all atoms
25	sum of electrotopological values over heteroatoms
26	sum of electrotopological values over halogens
27	sum of electrotopological values over carbons
28	Kier & Hall kappa 1 index
29	Kier & Hall kappa 2 index
30	Kier & Hall kappa 3 index
31	Petit John's R2 index
32	Petit John's D2 index
33	Petit John's i2 index
34	Harary number
35	Schultz index
36	total symmetry index
37	dyad symmetry index

Table 2. Semiempirical Structural Descriptors[11] Used for the Characterization of the Peptide-like Compounds and Their Fragments

no.	name of descriptor	abbreviation
1	atomic weight	AWeight
2	highest occupied molecular orbital	HOMO
3	lowest occupied molecular orbital	LUMO
4	electronegativity	Eneg
5	hardness	Hard
6	estimated polarizability	Epolar
7	molecular volume	MolVol
8	surface area	SurfA
9	ovality	Oval
10	log *P* (Ghose–Crippen)	LPGC
11	log *P* (Dixon)	LPDix
12	calculated solvation energy	AM1aq
13	lowest negative charge (Mulliken)	QMin

products. The BB descriptors were used as the **X**-matrix and the descriptors of the amides as the **Y**-matrix (Figure 2).

Two or more structural descriptors often contain similar information, e.g., lipophilicity is correlated with electronic properties and molecular mass with the area and volume. Therefore, the true number of underlying principal properties is often much smaller that the original number of structural descriptors. PCA compresses a data set to its main "principal" structure, i.e., reexpresses the information in the 37 structural descriptors using a smaller number of new uncorrelated variables (Figure 1a: II, 1b: II). Calculated structural descriptors also contain noise with respect to the property that it should reflect; e.g., log *P* can be calculated in different ways with the purpose to express the water–octanol distribution and compared to that experimental value the descriptor will comprise noise. This noise, if random, will also be reduced by

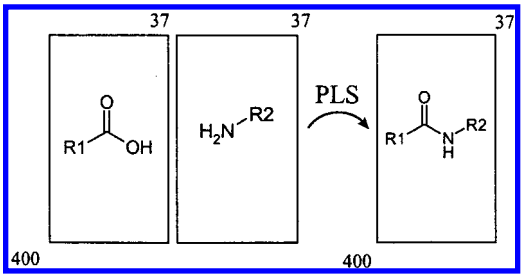

Figure 2. PLS was made, using the structural descriptors of the BB's as the **X**-block and the structural descriptors of the products as the **Y**-block, to investigate whether the characterization of the acids and the amines contained sufficient information about the amides.

PCA. PCA is expressed in terms of scores (**T**) and loadings (**P**):

$$\mathbf{X} = \mathbf{1}^*\bar{\mathbf{x}} + \mathbf{T}^*\mathbf{P}' + \mathbf{E}$$

where $\bar{\mathbf{x}}$ is a row vector of variable means and **E** is the residual matrix.[15] The scores are related to the objects while the loadings are related to the variables. A geometric view may be presented by score and loading plots. The optimal number of principal components was decided using the eigenvalue, the cross-validated R^2, and the interpretation of the loadings. The data set was scaled to zero mean and unit variance prior the PCA.

Statistical Molecular Design (SMD). The design variables used in the SMD's were unscaled scores resulting from separate PCA's of the BB/fragment sets and of the full library (Figure 1a: II, 1b: II). In example 1, the SMD's are applied both to the BB space and to the product space (Figure 1a: III, 1b: III) while in example 2 only applied to the BB space. The approach of example 2 was based on the results from example 1. Therefore, the selection method and the number of selected fragments were not varied and the selections were made solely in the BB space.

Three types of selection methods were investigated in example 1: (i) D-optimal design, (ii) space-filling design, and (iii) cluster-based design (Figure 1a: III, 1b: III). In the first two methods the selection algorithms were applied directly on the data set, while in the third a cluster analysis of the data preceded the selection.

(i) D-optimal design: D-optimal designs maximize *Det*(**X′X**) (D-optimal), where **X** is the coded model matrix with *n* rows (structures) and *p* terms (model variables) and *Det* denotes the determinant. Here, these designs are created to support a model between the structure of the compounds and the biological response, e.g., a quadratic model $\mathbf{y} = \mathbf{X}^*\mathbf{b} + \mathbf{e}$, where **y** is a column with *n* responses, **b** is a row vector containing the coefficients, and **e** is the residuals. For *k* factors, **X** has *k* columns for main terms, *k* columns for quadratic terms, and $k^*(k - 1)/2$ columns for interaction terms. By maximizing the volume of **X′X** (the determinant) the error variance of the coefficients **b** will be minimized.[16] The number of BB's to select is dependent both on the model and on the number of design variables. We have investigated models consisting of linear terms, linear and two-way interaction terms, linear and quadratic terms, and linear, two-way interaction, and quadratic terms. Note that when a model has interaction terms ($\mathbf{x}_j\mathbf{x}_k$) and/or square terms ($\mathbf{x}_k^2$), the **X** space has more dimensions than just the number of factors *k*. Consequently, the *Det*(**X′X**) will be a volume in a larger space, and the resulting design in the original nonexpanded space will also include interior points. We have also investigated different number of design variables, i.e., different numbers of PCA scores. The molecule closest to the calculated centroid was included in all designs, as an additional center point. The D-optimal designs were generated by the MODDE software.[17]

(ii) Space-filling design: Space-filling designs spread out the design points so an even coverage of the design space will be achieved. The space-filling design used here maximizes the

minimum Euclidean distance between the nearest neighbors of the selected BB's.[18] These designs are not explicitly based on a model nor is the number of selected BB's directly dependent on the number of design variables. When few objects (compounds) are selected, space-filling and D-optimal designs (with a center point) will give similar results.

(iii) Cluster-based design: The structural space of the BB's is investigated for groupings and singletons (cluster analysis) before the selection.[19,20] The selection is then made so that every structural feature is represented. Prior to the selection of compounds from the product space (example 1) fuzzy clustering[21,22] was made on the 400 amides. A space-filling design was then applied for each of the clusters.

Final Compound Selection. All combinations of the selected BB's are not necessary.[6,23] A smaller subset can be selected from the resulting products from the first SMD applied to the BB space (Figure 1a: IV−VI). This second design often gives more than one subset with satisfying properties. Hence, this final compound selection gives the opportunity of choosing the subset based on synthetic probability, experimental conditions, or intuition for specific molecules. This is a big advantage for those who are about to actually synthesize the molecules. It is also possible to exclude molecules that are impossible/very difficult to synthesize, before this second design is applied. Moreover, when it comes to the biological testing this subset of fewer compounds will, in respect of time and expenses, give the opportunity of using broader and more reliable test methods.

Evaluation. The objective was to select smaller subsets of compounds that covered the product space and not to lose much in diversity compared to the full library and the product library design. Since diversity is difficult to capture in one single number, more than one evaluation criterion has been used. In example 1, this was investigated by (i) the coverage of the score plots, (ii) the Tanimoto coefficient, (iii) the Euclidean distance, and (iv) the condition number and the determinant for each subset of molecules.

(i) The coverage of the score plots: The coverage of the selected subsets (by each approach and statistical design) in reference score plots was investigated. The score plots of the full library (400 amides, 5 components) were used as a reference. The coverage of the selected BB's in all combinations of the scores was evaluated, denoted as *Ql*. For satisfying coverage of each score plot (t_1 vs t_2, t_1 vs t_3, t_1 vs t_4,..., t_4 vs t_5, in total 10 score plots), 4 points were given (one for each quadrant), with a maximum value of 40. If one quadrant had poor representation 0.5 was assigned, and if there were no BB's in a quadrant 0 points were given. This criterion is crucial for the diversity and needs to be high in order to have a satisfying representation in the score plots. It is presented as percent of the maximum value.

(ii) The Tanimoto coefficient: The Tanimoto coefficient (T) is given by

$$T_{i,j} = \sum_{k=1}^{p} x(i,k) \times x(j,k) / [\sum_{k=1}^{p} x(i,k)^2 + \sum_{k=1}^{p} x(j,k)^2 - \sum_{k=1}^{p} x(i,k) \times x(j,k)]$$

where i and j are two compounds (amides) from a descriptor matrix $\mathbf{X}$ with p variables (the scores from the PCA). The Tanimoto coefficient is an example of an association coefficient, which also include the cosine coefficient and the Dice coefficient. These coefficients differ only in the way they normalize the dot products of the two vectors.[24] Commonly these coefficients are used for binary descriptors, but here we have applied the Tanimoto coefficient to continuous variables. The coefficients for the selected subsets were calculated as an average using the Tanimoto coefficients between all molecules, denoted as T_{abs} (the absolute values are used) and T_{norm} (the lowest possible value, −1/3, is added and normalized). Also the average using the Tanimoto coefficient for every molecule to its nearest neighbor (using both T_{abs} and T_{norm}), denoted as T_{near}[7] was calculated. This gives a measure of the similarity within the different subsets and should therefore be as small as possible.

(iii) The Euclidean distance: For comparison, the average of the Euclidean distances between all objects was calculated (E_{mean}) and also the average using the Euclidean distance between every molecule and its nearest neighbor (E_{near}). The diversity measurements (both the Tanimoto coefficients and the Euclidean distances) were made by unscaled PCA scores from the full library (the 400 amides). The E_{mean} and E_{near} should be as large as possible for good coverage and high diversity.

(iv) The condition number and the determinant: The condition number for each model matrix (including main, two-way interaction, and quadratic terms) was also calculated. The scores from the PCA made on the 400 amides were used in coded form, i.e., transformed to the range of −1 to 1. The condition number is the highest singular value divided by the lowest, and it shows the roundness of a selected group of points. This is directly related to lack of correlation between the model variables in the data set. For orthogonal designs such as factorial designs, where every factor is varied independent of each other, the condition number is 1. When the correlation increases between the design variables of the selected molecules, the condition number will increase. For an optimal selection the condition number should be as low as possible. The determinant (denoted *Det*) for each subset was calculated as log(*Det*($\mathbf{X}'$*$\mathbf{X}$)), where $\mathbf{X}$ is the same matrix as the one used for the calculation of the condition number. To receive stable reliable models the determinant should be as large as possible.

In example 2, the score space of the full library (the 480 compounds) was used to evaluate the final product library derived from the BB-based selection procedure.

Results and Discussion

Selection in the BB Space vs Selection in the Product Space. The first step of the investigation was to make sure that the characterization of the BB's/fragments contained sufficient information of the final products, since this is a prerequisite for the approach. For both examples 1 and 2, the PLS gave a good correlation between the characterization of the BB's and the final products, respectively. Figure 3 shows how every descriptor variable (the product space) is described by PLS using the BB's compared to the results from the PCA made solely on the products. This comparison is fair, since it is the underlying properties of the molecules that are of interest not the actual descriptors. In example 1, 80% of the variation in the BB's describes 86% of the variation in the amides (10 components); for example 2, 98% of the variation in the BB's describes 92% of the variation in the products (7 components). This shows that both a one-dimensional and a semiempirical characterization of the BB's appeared to contain appropriate information about the final products, which shows that selections can be made in the BB space rather than in the product space.

Statistical Molecular Designs and Number of Selected BB's. In example 1, the PCA of the amides gave five principal components, which were used as design variables for the product selection (Figure 1b: II). Twenty-five amides were selected with good coverage using a D-optimal design (main, two-way interaction, and quadratic terms) which is viewed in Figure 4a. Among these 25 selected amides 12 different carboxylic acids and 10 different amines were present (Figure 5). Also with the space-filling design 25 amides were selected with appropriate coverage. The fuzzy

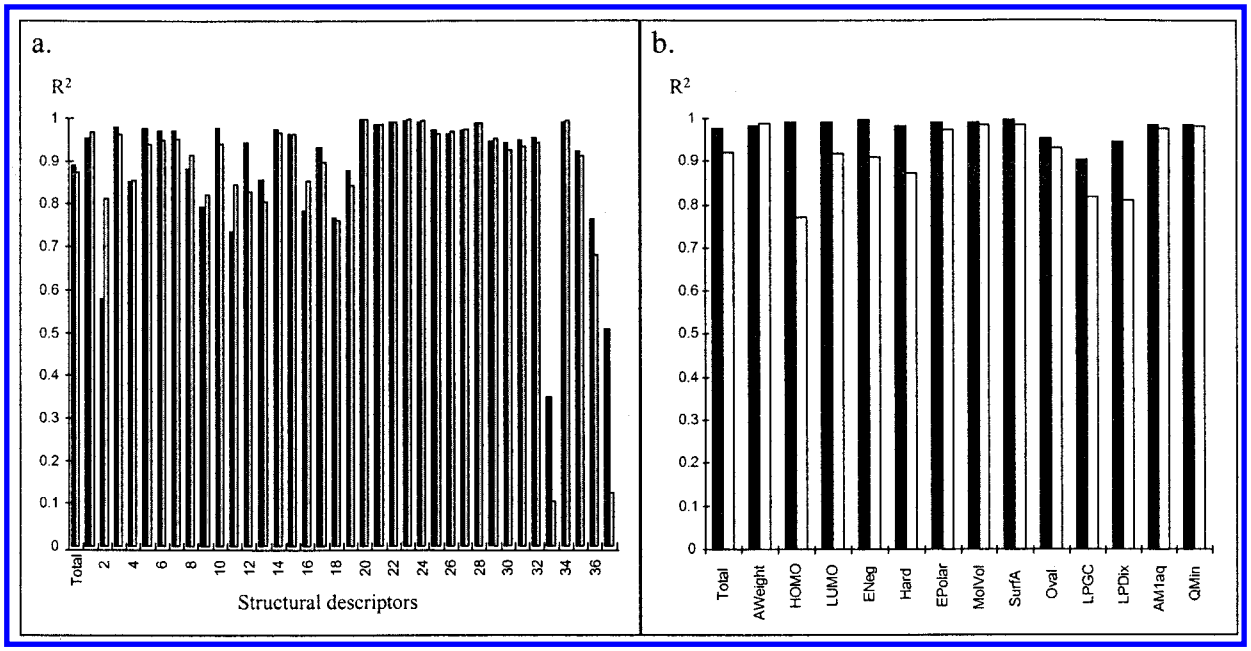

Figure 3. Explained variation of the structural descriptors of the products, when a PCA was made on the products (the black bars) compared to when a PLS was made as explained in Figure 2 (the gray bars): (a) the 400 amides characterized by 37 one-dimensional descriptors; (b) the 480 peptide-like structures characterized by 13 semiempirical descriptors.

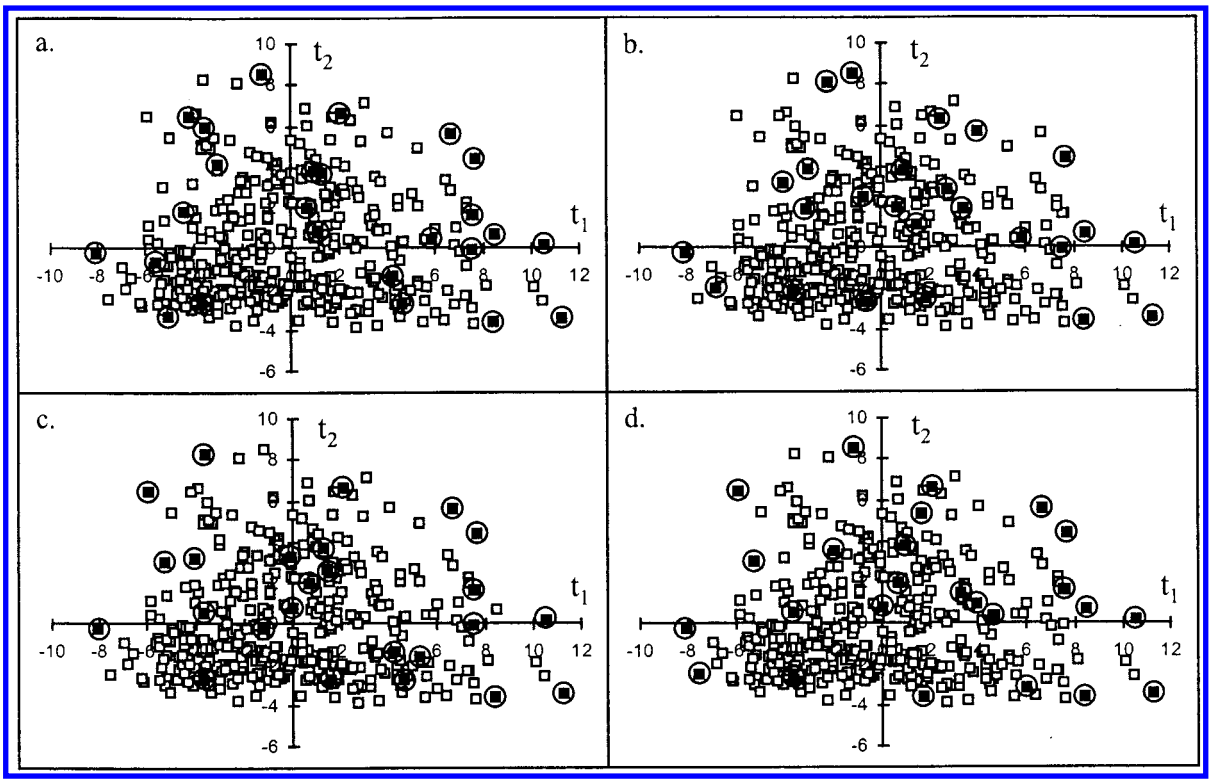

Figure 4. Coverage of the selected amides (black squares) in the total product space (white squares) was used to evaluate the results from the statistical designs (*Ql*): (a) 25 amides selected by a D-optimal design from the total product space; (b) 25 amides resulted from a second design applied to 49 amides (7 carboxylic acids and 7 amines) selected in the BB space by a cluster-based design; (c) 25 amides resulted from a second design applied to 81 amides (9 carboxylic acids and 9 amines) selected in the BB space by a D-optimal design; (d) 25 amides resulted from a second design applied to 81 amides (9 carboxylic acids and 9 amines) selected in the BB space by a space-filling design.

clustering of the amides resulted in eight groups: those with high nitrogen content, high oxygen content, polyaromatic, long-chained aliphatic, fluorine content, small rings, heteroaromatic, and iodide/chlorine content. A space-filling design was applied for each of the eight groups using the local scores (five PCA scores) for each group. Three compounds were selected from each, except for the "small rings" group, which contained the largest number of molecules, from which four were selected.

For the two BB sets, the amines and the carboxylic acids, the PCA resulted, in both cases, in five principal components (Figure 1a: II), which were used as design variables. For the D-optimal designs the number of PCA scores was varied along with the level of the D-optimal

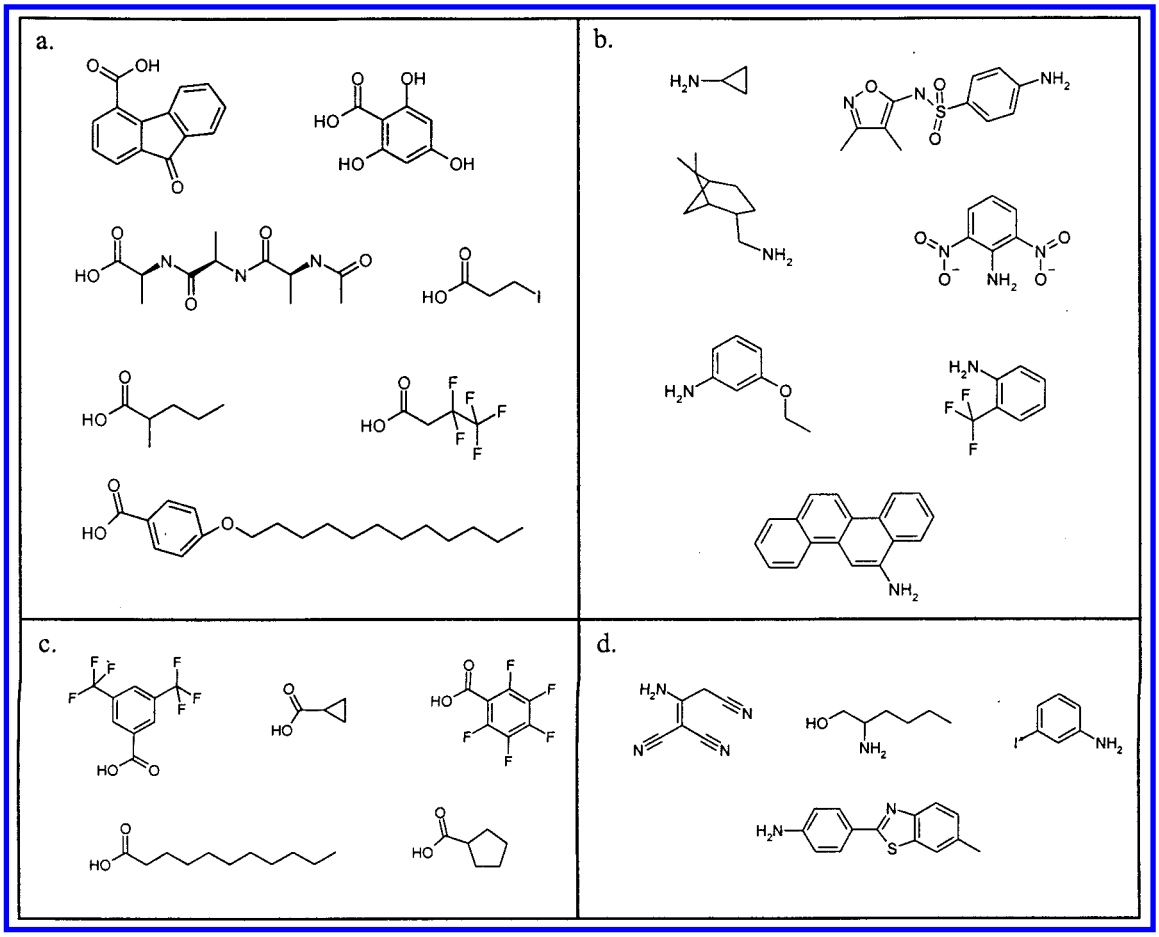

Figure 5. BB's that were selected by cluster-based design in the BB space: (a) the carboxylic acids and (b) the amines. Also shown, the additional reactants that were selected by the D-optimal design in the total product space: (c) the carboxylic acids and (d) the amines. The *m*-ethoxyaniline was not selected by the design in the product space.

designs applied (see Table 3, Statistical method). Five to twelve BB's were selected, where five is the statistically minimum number of BB's since five PCA scores are used. In the selection using the space-filling design all five components were utilized independent of the number of selected BB's. The cluster analyses of the BB sets showed that the carboxylic acids contained three main groups (long-chained aliphatic, fluorine content, "normal") and three singletons (iodide content, polyaromatic, high nitrogen/carbonyl content) and the amines contained three main groups (polyaromatic/heteroaromatic, chlorine/fluorine/iodide content, small rings) and one singleton (two nitro groups). BB's were selected from each group together with the singletons resulting in 7 acids and 7 amines (Figure 5a,b).

A second selection was made independently of the first SMD in those cases where more than 25 compounds were the result of the combined BB's. This was achieved by combining the selected BB's, e.g., 9 amines and 9 acids yielding 81 amides (Figure 1a: IV). From these a selection of 25 amides was made (Figure 1a: VI), using local PCA scores from the 81 compounds (Figure 1a: V). In those cases where the BB's were selected by D-optimal design and cluster-based design, a D-optimal design (including main, two-way interaction, and quadratic terms) was used for this second selection, while where the space-filling designs had been used for the BB selection it was also used for the second selection (Figure 1a: VI). The results of the evaluation of all selections can be seen in Table 3.

Deciding the number of selected BB's is a trade off between how many that can be afforded in the combinatorial process and what concomitant loss of the diversity is acceptable. For instance, if the diversity demand is very high, 11 amines and 12 acids can be selected from the reactant pools without losing any diversity at all (Table 3, no. 7). By setting the diversity demand to $Ql > 90\%$ and the other criteria (the Tanimoto coefficient, the Euclidean distance, the condition number and the determinant) not to differ much from the product selection, the cluster-based design would fulfill these requests by using 7 amines and 7 acids (Table 3, no. 13, Figure 4b). The D-optimal design required 9 amines and 9 acids to accomplish this (Table 3, no. 6, Figure 4c), which corresponds to a model based upon four principal components, using linear and quadratic terms. The space-filling designs were not able to satisfy these demands with 9 amines and 9 acids (Table 3, no. 11, Figure 4d). If the diversity demands were lowered to $Ql > 85\%$, the space-filling design (9×9, Table 3, no 11) and the D-optimal design (7×7, Table 3, no. 5) were also acceptable.

Another interesting and important result is that selecting 5 amines and 5 amides will not give sufficient diversity. If the number of selected BB increases to 7 the diversity resembles the diversity obtained by selection from the product space. This is a crucial point: when the selection is made in the reactant space it is important that a sufficient number of BB's is selected. The result obtained by Gillet et al.,[9] indicating that it

Table 3. Evaluation of the Selected Subsets*

no.	statistical method	selection	*Ql*	E_{near}	E_{mean}
1	D-optimal design	25	95.8	4.93	10.52
2	D-optimal (main terms, t_1-t_3)	5*5 = 25	71.2	2.4	8.14
3	D-optimal (main + interaction terms, t_1-t_2)	5*5 = 25	73.8	4.11	9.36
4	D-optimal (main terms, t_1-t_5)	7*7 = 49 => 25	80.0	4.09	9.84
5	D-optimal (main + quadratic terms, t_1-t_3)	7*7 = 49 => 25	87.5	4.66	10.26
6	D-optimal (main + quadratic terms, t_1-t_4)	9*9 = 81 => 25	92.5	4.99	10.17
7	D-optimal (main + quadratic terms, t_1-t_5)	11*12 = 132 => 25	96.2	5.11	10.29
8	space-filling design	25	92.5	5.56	10.39
9	space-filling design	5*5 = 25	82.5	3.55	9.06
10	space-filling design	7*7 = 49 => 25	83.8	5.16	10.25
11	space-filling design	9*9 = 81 => 25	86.2	5.54	10.59
12	cluster-based design	25	91.2	3.83	8.91
13	cluster-based design	7*7 = 49 => 25	93.8	4.46	9.91

no.	T_{norm}	T_{near}[a]	T_{abs}	T_{near}[b]	*Det*	cond no.
1	0.32	0.74	0.24	0.65	5.84	12.5
2	0.35	0.89	0.26	0.85	−8.93	210.3
3	0.34	0.76	0.23	0.68	−4.26	65.6
4	0.30	0.79	0.23	0.71	−0.77	29.9
5	0.32	0.75	0.22	0.67	2.74	23.9
6	0.30	0.71	0.21	0.61	4.21	15.8
7	0.31	0.70	0.22	0.61	5.66	13.0
8	0.31	0.71	0.23	0.61	2.38	90.4
9	0.30	0.80	0.23	0.73	−2.65	61.4
10	0.30	0.71	0.22	0.62	1.38	45.3
11	0.31	0.70	0.23	0.61	3.50	39.9
12	0.28	0.73	0.21	0.64	−2.34	85.7
13	0.32	0.76	0.23	0.68	3.31	24.2

Ql* = the coverage in the global score plots expressed as percent of the maximum value. E_{near} = the average Euclidean distance between every molecule and its nearest neighbor. E_{mean} = the average Euclidean distance of all pairs of BB's. T_{norm} = the normalized average Tanimoto coefficient. T_{abs} = the average Tanimoto coefficient using absolute values. T_{near} = the average Tanimoto coefficient of every molecule to its nearest neighbor, [a] using T_{norm} as similarity measure and [b] using T_{abs}. *Det* = log(*Det*(X′X**), where *Det* is the determinant and ***X*** the model matrix (including main, two-way interaction, and quadratic terms). cond no. = the condition number of the model matrix (including main, two-way interaction, and quadratic terms).

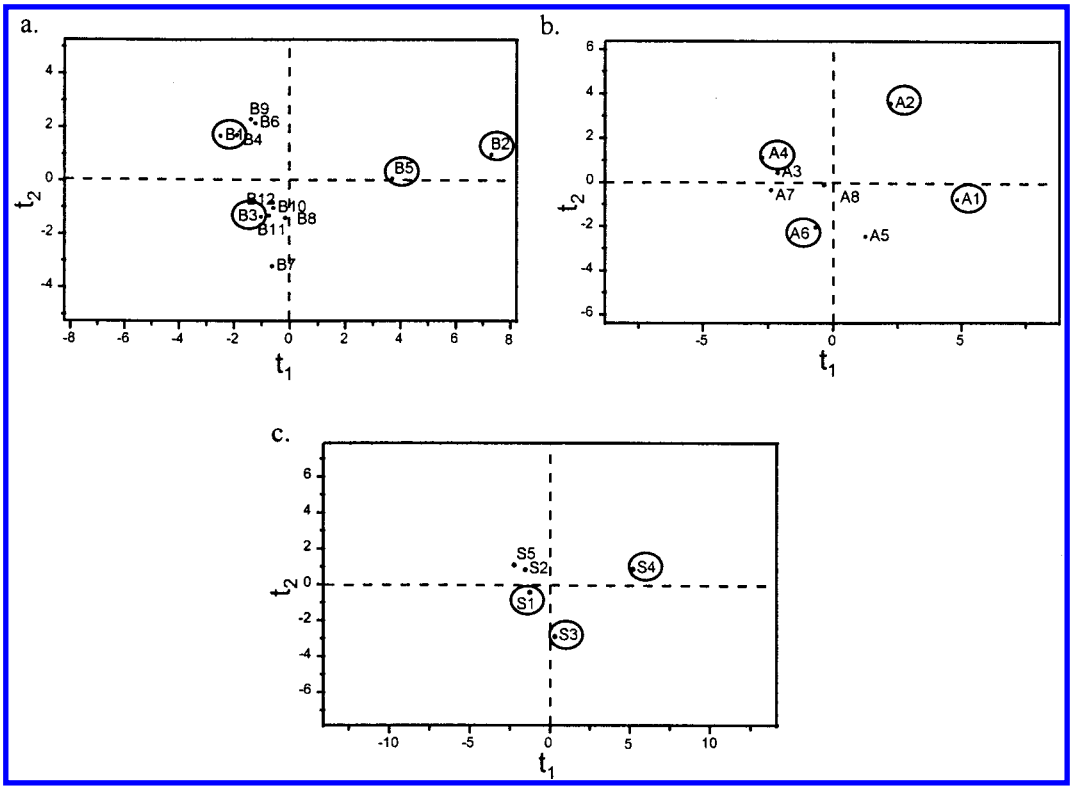

Figure 6. Fragments selected by cluster-based design, viewed in the score space: (a) the basic fragments, (b) the acidic fragments, and (c) the specific fragments.

is not possible to make selection in the reactant space without losing diversity, might be an effect of that to few BB's were selected, resulting in exclusion of unique structural features.

In summary, the selections in the product space are no. 1, 8, and 12. A comparison of the evaluation criteria for these three and the rest reveal that no. 6, 7, and 13 acquired almost the same result. This implies that

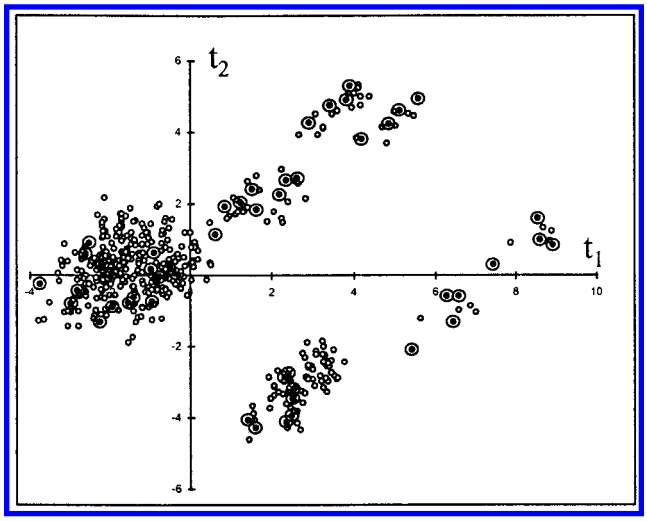

Figure 7. Distribution in the total product space ($n = 480$) of the 48 peptide-like structures that resulted from the fragment selection showed in Figure 6. From these 48 structures 22 were selected by a D-optimal design.

selections in BB's are equivalent to selections in product space.

In example 2, two principal components were used to describe each of the three different sets of fragments. Based on the above results (example 1) a cluster-based design was used in order to select a smaller number of fragments. Figure 6 shows the selection of fragments in the score space of the three sets of fragments. Since the fragment sets yielded fewer principal components (compared to example 1), fewer fragments could be selected without a loss in diversity: four basic fragments, three specific fragments, and four acidic fragments were selected, resulting in 48 compounds. The coverage in the product space is shown in Figure 7. A second selection, using D-optimal design, was applied to the selected and combined fragments ($n = 48$) resulted in 22 structures out of the virtual 480. By visual inspection these 22 still covered the reference space. These suggested 22 structures were subjected to synthesis and subsequent bioassessments.

Conclusion

We have shown here that it is possible to reduce the size of a combinatorial library by applying statistical designs to BB property spaces, without losing diversity compared to statistical design based on the final products.

The number of selected BB's influences the diversity of the final library. Selecting too few BB's has a negative effect on this diversity. Hence, it is of critical importance to investigate the properties of the BB sets prior to the designed selection, e.g., by cluster analysis and multivariate analysis of the clusters. Otherwise unique features might not be represented in the final library and the result will be a loss of diversity. It is therefore necessary to select an appropriate number of BB's to use in each designed sublibrary. However, these numbers need not be substantially higher than the statistically minimum number of BB's. This was shown in example 1, where raising the number of selected BB's from 5 to 7 resulted in an adequate diversity. In addition, the use of a second design to decide the final library yields an even smaller set to synthesize and submit for biological investigation. This will be of even greater importance as the BB sets increase in size. Additionally, a broad biological investigation (e.g., activity, bioavailability, degradation, toxicity, selectivity) of the compounds can be both expensive and time-consuming but nevertheless of major concern. By reducing the number of compounds to test, it is possible to put more expenses on the selected ones. The use of this second design also provides an opportunity of considering synthetic properties in the selection.

The advantages of basing the library selections on designed BB sets are substantial. The characterization of molecules becomes easier, saving both work time and CPU time. The QSARs based on the BB characterizations will give a direct feedback of which part of the molecule to modify in the next library to accomplish the required objective. The number of selected BB's is easy to control and to keep low in a simple and natural way.

In summary, a proper selection in the BB space based on rich statistical design results in a library containing as high diversity as a library based on product level selection. With the advantages of BB selection this method is therefore preferable.

Acknowledgment. Financial support from the Swedish Natural Science Research Council (NFR) and AstraZeneca R&D, Mölndal, is gratefully acknowledged. We are also grateful to Dr. Peter Pedaja at AstraZeneca R&D, Lund, for the calculations of the structural descriptors.

References

(1) Felder, E. R.; Poppinger, D. Combinatorial Compound Libraries for Enhanced Drug Discovery Approaches. *Adv. Drug Res.* **1997**, *30*, 111−199.
(2) Moos, W. H.; Pavia, M. R.; Ellington, A. D.; Kay, B. K. *Annu. Rep. Comb. Chem. Mol. Diversity 1* **1997**, 3−325.
(3) Szostac, J. Combinatorial Chemistry. *Chem. Rev.* **1997**, *97*, 347−510.
(4) 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.
(5) Young, S. S.; Farmen, M.; Rusinko, A., III. Random versus Rational. Which is Better for General Compound Screening? *Network Sci.* (electronic publication) **1996**, 2.
(6) Lundstedt, T.; Clementi, S.; Cruciani, G.; Pastor, M.; Kettaneh, N.; Andersson, P. M.; Linusson, A.; Sjöström, M.; Wold, S.; Nordén, B. Intelligent Combinatorial Libraries. In *Computer-Assisted Lead Finding and Optimization. Current Tools for Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, Wiley-VCH: Weinheim, 1997; pp 190−208.
(7) Pötter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.
(8) Drewry, D. H.; Young, S. S. Approaches to the Design of Combinatorial Libraries. *Chemom. Intell. Lab. Syst.* **1999**, *48*, 1−20.
(9) Gillet, V. J.; Willet, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generation Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(10) AstraZeneca R&D, S-431 83 Mölndal, Sweden.
(11) UNIX Spartan 4.0, Wavefunction, Inc., 18401 Von Karman Ave., Suite 370, Irvine, CA 92612.
(12) Wold, S.; Ruhe, A.; Wold, H.; Dunn III, W. J. The Collinearity Problem in Linear Regression. The Partial Least Squares Approach to Generalised Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.
(13) Manne, R. Analysis of Two Partial Least Squares Algorithms for Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 187−197.
(14) Höskuldsson, A. PLS Regression Methods. *J. Chemom.* **1988**, *2*, 211−228.
(15) Jackson, J. E. *A users guide to principal components;* Wiley: New York, 1991.

(16) Mitchell, T. J. An Algorithm for the Construction of "D-optimal" Experimental Design. *Technometrics* **1974**, *2*, 203–210.
(17) Modde 4.0 software, Umetri AB, Box 7960, S-907 19 Umeå, Sweden.
(18) Marengo, E.; Todeschini, R. A New Algorithm for Optimal Distance-based Experimental Design. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 37–44.
(19) Dunbar Jr., J. B. Cluster-based Selection. *Persp. Drug Discuss. Des.* **1997**, *7/8*, 51–63.
(20) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. Cluster-based Design in Environmental QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 383–390.
(21) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981.
(22) Linusson, A.; Wold, S.; Nordén, B. Fuzzy Clustering of 627 Alcohols, Guided by a Strategy for Cluster Analysis of Chemical Compounds for Combinatorial Chemistry. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 213–227.
(23) Andersson, P. M.; Linusson, A.; Wold, S.; Sjöström, M.; Lundstedt, T.; Nordén, B. Design of Small Libraries for Lead Exploration. In *Molecular Diversity in Drug Design;* Dean, P. M., Lewis, R. A., Eds.; Kluwer Academic Publishers: The Netherlands, 1999.
(24) Holliday, J. D.; Ranade, S. S.; Willet, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501–506.

JM991118X