Statistical design and application to combinatorial chemistry

Sally Rose

Combinatorial chemistry has matured from its early production-belt image of synthesizing the maximum number of compounds in the minimum amount of time. It is now a discovery technology driven by good experimental design, with a focus on producing high numbers of quality drug-like compounds. Complex statistical and optimization algorithms are routinely applied to ensure that the library is designed with the best chance of achieving its goals, while also facilitating efficient synthesis.

Sally Rose
BioFocus
Sittingbourne Research Centre
Sittingbourne
UK ME9 8AZ
tel: +44 1795 412300
fax: +44 1795 471123
e-mail: srose@biofocus.co.uk

▼ Combinatorial chemistry is generally defined as the synthesis of libraries of compounds containing all possible combinations of reagents. Thus, for a library synthesized from three sets of reagents (also termed monomers or building blocks), R^1, R^2 and R^3, with n, m and p numbers of different reagents, respectively, the combinatorial library is of size $n \times m \times p$. If $n = m = p = 10$, the library will contain 1000 (n^3) compounds, and each reagent will be represented 100 (n^2) times. This approach leads to efficient synthesis.

Significant advances in combinatorial library design have occurred in parallel with the development of high-throughput chemistry systems. However, several groups have now realized that, although synthetically efficient, a library designed using these principles has much redundancy of information and relatively poor diversity [1–3]. Therefore 'non-combinatorial' designs are becoming more popular.

In the early days of combinatorial chemistry, many groups were seduced by the concept of synthesizing large libraries of single compounds or mixtures. The libraries were often of poor quality and tended to have non-ideal physicochemical property profiles containing many high molecular weight and hydrophobic compounds, which have poor solubility and bioavailability characteristics. Libraries of mixtures had the added complication that deconvolution, fractionation or tag decoding were a prerequisite in identifying the active component. The numbers of compounds could be significantly increased by combining facile chemistry with automation. Although there were some success stories, these libraries frequently generated many false positives in HTS, which required significant effort to follow-up, and often led up blind alleys. The vast overhead in the data storage requirements that accompanied the combinatorial revolution, and the interpretation of a large volume of poor data, also caused problems.

More recently, the focus has moved towards a more rational, medicinal chemistry approach to parallel synthesis, resulting in higher quality, purified, single-compound libraries with drug-like [4–8] physicochemical properties. More emphasis is placed on design for a biological target, rather than simply maximizing diversity [9,10]. The chemistry can be more challenging and smaller libraries might be produced. However, higher hit rates against the target are achieved compared with the early combinatorial libraries, leading to an overall increase in efficiency.

Efficient library design is best achieved through dialogue between a combinatorial chemist and computational chemist. A clear goal is needed in terms of what the library aims to achieve. Commonly, this is either: (1) diversity, to supplement a corporate collection; or (2) biologically focussed, to discover or optimize hits against a specific target or family of targets.

Synthetic feasibility, the physicochemical property profile, molecular diversity, library design methodology and the requirement for high-throughput docking or pharmacophore matching, all need to be considered in the production of an appropriate library. Here, statistical approaches to diversity and library

1359-6446/02/$ – see front matter PII: S1359-6446(01)02093-1

design methodology for libraries of single compounds (as opposed to libraries of mixtures) will be reviewed.

Molecular diversity

Understanding molecular diversity is essential to optimizing library design for a particular task. If the assumption is made that a diverse library is more likely to provide hits in HTS [whereas lead optimization focussed around a library of similar (non-diverse) compounds is statistically more likely to produce active compounds than highly diverse compounds], then the library can be designed with the appropriate goals in mind. Quantifying 'relevant' diversity requires decisions to be made on the type of property, structural or pharmacophoric descriptors to use, which intermolecular distance measure and library diversity function is most appropriate, and whether reagent, product or reagent-biased product designs are to be used [11].

Descriptors

Molecules can be variously described by a vast number of descriptors related to structural features or molecular properties. A set of descriptors relevant to the goals of the library must be chosen. Livingstone offers a recent review of molecular properties [12], and van de Waterbeemd *et al.* review properties related to drug absorption and pharmacokinetics [13], which are increasingly being seen as important factors to consider in a library design exercise.

Commonly, property descriptors are sub-divided into 2D (and 1D) or 3D, indicating which type of structural representation of a molecule is required for its calculation. Properties that can be calculated from a 2D structural representation include topological indices, ClogP, molecular weight, atom counts, and number of hydrogen bond donors and acceptors. These properties are readily calculable for large databases and virtual libraries. 2D to 3D structure converters, such as Corina (Accelrys, Cambridge, UK) and Concord (Tripos, St Louis, MO, USA), enable rapid generation of a low energy 3D conformation from a 2D structure, which can be used for calculation of 3D properties. 3D properties include energies of the highest-occupied molecular orbital (HOMO) and lowest-unoccupied molecular orbital (LUMO), dipole moment and vector, and shape properties. Caution should be applied when including 3D properties as molecular descriptors in a library design exercise, as most are highly dependent on the conformation (and often orientation) of the molecules, and can thus be misleading for large diverse libraries, unless strict rules on conformation and orientation can be defined and adhered to.

Labute [14] has described a set of 32 descriptors that he has named '2.5D', which he found to be generally applicable to structure–activity relationships and that can be applied to diversity studies. They are derived from the contribution to the van der Waals molecular surface area made by each atom and their relative polarizability, hydrophobicity and electrostatic contributions. The properties are calculated from a 2D structure (connection table), but implicitly contain information on the 3D structure.

Another relatively recent extension to modelling diversity is through Pearlman and Smith's BCUT (Burden Chemical Abstract Services, University of Texas, Houston, TX, USA) descriptors [15,16]. These complex descriptors encode a large amount of chemical information in a low-dimensional space. The descriptors are eigenvalues derived from matrices containing information on the diagonal elements related to physicochemical properties, and off-diagonal elements related to topology and connectivity.

Structural descriptors can be generated from pre-defined fragment lists, such as MACCS (Molecule Access System) (or ISIS) keys (MDL Information Systems, San Leandro, CA, USA) or BCI fragments (Barnard Chemical Information, Sheffield, UK), or by using 2D fingerprints such as those generated by Unity (Tripos) and by Daylight Software (Daylight, Mission Viejo, CA, USA). These descriptors comprise a bit string of 0s and 1s, indicating the presence or absence of structural features. Both approaches have been widely used to define diversity, and an analysis by Brown and Martin [17] suggests they are both valid for mapping chemical space onto biological activity space, although MACCS keys were marginally favoured in this study.

Like MACCS keys and 2D fingerprints, 3D pharmacophore descriptors comprise bit strings identifying the presence or absence of specific 3- or 4-point pharmacophores in a molecule [3,16]. A pharmacophore is defined as a particular set of three or four features (selected from an aromatic centre, hydrophobic group, H-bond donor, H-bond acceptor, acid group, basic group or N^+), which are separated by pre-defined distance ranges. An analysis by Pötter and Matter [18] suggested that 2D structural descriptors were preferred to 3D pharmacophore triplets for modelling diversity in large libraries.

Recently, Schneider *et al.* [19] proposed the use of a 2D pharmacophore autocorrelation vector that removes the requirement to generate 3D molecular structures. Atoms are assigned generalized types (as described before, but excluding aromatic centres), and a vector is constructed that summarizes the information on the shortest routes (number of bonds) between each of the possible generalized atom pairs.

Distance measures

Once the diversity descriptors have been selected, decisions as to which intermolecular distance measure to use are required [20]. Distance measures are used to facilitate

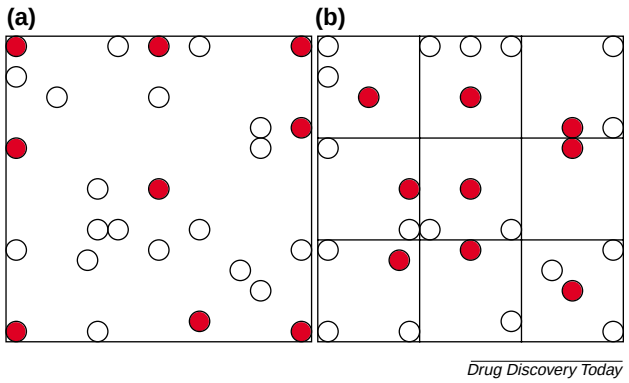


Figure 1. Schematic 2D representation of the different sampling of diversity space by **(a)** Euclidean distance to maximize diversity, and **(b)** a cell-based selection method, which selects the compound nearest the cell centre. Selected compounds are shown in red.

sampling of diversity space and to compare libraries. Euclidean distance is the most common distance measure used for properties, though the cosine distance measure has been recommended by Turner *et al.* [21] as being less faster to calculate, with a time order, O, related to the number of compounds, N [i.e. O(N), rather than O(N^2) as for Euclidean distance]. Gorse *et al.* [20] propose a City Block or Squared Euclidean measure, which also gives a time dependence of order O(N). The Tanimoto distance is widely used for measuring the similarity of compounds when they are described by structural descriptors, such as 2D fingerprints and MACCS keys.

Some distance-based methods can result in the selection of a large number of outliers (i.e. compounds with extreme descriptor values). This might not be ideal. Therefore, methods such as OptiSim [22] have been developed to enable the user to define a balance between diversity and how representative a selection is of the total set.

Diversity space can also be sampled by dividing the property space into cells and selecting a representative compound from each cell, commonly the compound nearest the cell centre. This is only suitable when considering small numbers of descriptors (i.e. a low dimensional space) because the number of cells rises exponentially with each additional descriptor; that is, if each descriptor, d, is divided into p cells, then the number of cells equals p^d. Using too many descriptors results in very sparse cell occupancy.

A comparison of the selection obtained using an Euclidean (distance-based) and cell-based approach is given in Fig. 1. The Euclidean approach gives better diversity in this example, but cell-based methods generally tend to give more even sampling of space. Cell-based measures can be used to compare libraries by studying their relative cell-occupancies.

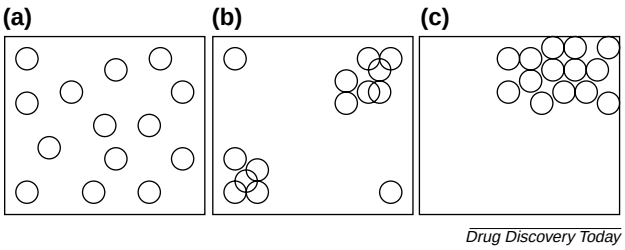


Figure 2. A schematic representation of library diversity. **(a)** A library with members that sample the diversity space evenly (the ideal situation). **(b)** A library that covers the same diversity space but the compounds cluster. **(c)** A library with even sampling of space, but only limited diversity.

Library diversity functions

Diversity functions aim to quantify the diversity of the combinatorial library. Figure 2 compares the diversity for three example libraries. A recent paper by Waldman *et al.* [23] discusses diversity functions and their relative limitations in describing coverage of property space by combinatorial libraries. The following types of general algorithms have been applied:

- Intermolecular distances; for example, the sum of the pairwise intermolecular dis-similarities (distances) over all molecules [21,24] or average nearest neighbour distance [25].
- Cell-based methods; for example, counting the number of cells occupied in a library or looking for even coverage of the cells [23].
- Waldman *et al.* [23] have recently described novel methods based on the minimum-spanning tree.

Currently, the cell-based methods are preferred to distance measures. The minimum-spanning tree approach appears to be marginally superior to the cell-based methods, but is computationally expensive.

Combinatorial library design methodology

Final molecule or reagent diversity?

The chemist is also faced with the decision of whether to design the library using either:

(1) Reagent-based design; that is, selecting a set of suitable reagents for each site of variation in the library.
(2) Reagent-biased product design; that is, using an optimization algorithm to select a subset of reagents that give the best properties in the final compound set without exhaustive use of different reagents.
(3) Product-based design; that is, creating a virtual library from a set of relevant reagents and selecting a set of final products to synthesize from this.

Although product-based selections are generally accepted to maximize diversity [24], this is not the only important

criterion, and Linusson *et al*. [26] found reagent-based selections gave good diversity and were often more practical. Product-based designs are incompatible with the synthesis of all-combination libraries. Reagent-biased product designs represent a balance between diversity and practicality. Some examples of the different approaches are outlined next.

Reagent diversity

Martin and Wong [27] have developed software ('Tailor') to select reagents to provide a specific property profile in the final library. The reagents are allocated to a series of overlapping bins that describe their characteristics, such as hetero-aromatic, rigid low molecular weight, or fragments found in the top-200 selling drugs. The chemist can then define the bins he wishes to consider and the number of reagents required from each bin. A D-optimal design using a parallel Fedorov search algorithm is used to select the final set of reagents to maximize reagent diversity within the specified constraints. The method was shown to perform significantly better than a random selection for maximizing diversity in a subset of 17, 32 or 51 aldehydes selected from a set of 908 candidates. A further example of reagent-based diversity design can be found in Ref. [26].

Reagent-biased product design

Several groups are using genetic algorithms, simulated annealing or random sampling techniques to select an optimal subset of reagents to use in combinatorial library synthesis for a diverse or targeted library [25,28–32]. One of the first such approaches was HARPick from Good and Lewis [33]. The methods attempt to balance synthetic practicality with good design principles by restricting the number of different reagents. The optimization is commonly targeted towards one or a combination of several of the following properties:

- drug-likeness;
- cost;
- final compound diversity or similarity to a known active analogue; or
- maximizing the scoring function for a docking study or pharmacophore fit.

Zheng *et al*. (GlaxoSmithKline, King of Prussia, PA, USA) have developed a combinatorial design program, PICCOLO [29], which can be used to select a subset of reagents that enable optimization of a wide variety of factors, including:

- similarity to known leads;
- reagent diversity and coverage of property space;
- product novelty with respect to the corporate collection;
- Lipinski properties;
- liabilities against P450 enzymes;
- aqueous solubility;
- molecular flexibility;
- mass spectrum redundancy; and
- reagent price.

The optimization uses simulated annealing to minimize a total penalty score for the combinatorial library.

Bravi *et al*. [34] (GlaxoSmithKline, Stevenage, UK) have developed PLUMS, which reduces the size of a large virtual combinatorial library by iteratively removing reagents. PLUMS is used to target a library synthesis such that the library best conforms to a set of constraints, such as a quantitative structure–activity relationship (QSAR) or pharmacophore hypothesis. Products that satisfy the constraints are termed 'hits'. The worst reagent is removed at each iteration to improve the efficiency and effectiveness of the library until the user-defined (or 'optimal') library size is obtained. The most 'effective' library is defined as the smallest library containing all the hits, whereas the most 'efficient' library is defined as the largest sub-library containing only hits and no unfavourable products.

PICCOLO therefore enables a variety of diverse factors relevant to drug discovery to be taken into account during the design process, whereas PLUMS is focussed on designing a library to test a specific user-defined hypothesis to explain activity.

Product-based designs

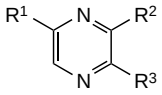
Product-based designs or 'cherry-picking' are not compatible with the practicalities of synthesis of large combinatorial libraries, but might be appropriate for small non-combinatorial libraries and lead optimization studies. Generally, such designs would be used when information on QSAR, pharmacophore or docking studies can be used to select individual compounds that are predicted to be active.

Non-combinatorial library design methodology

Full combinatorial designs are the most efficient way to facilitate rapid synthesis of large numbers of compounds. However, such designs can contain much redundancy of information and several groups are developing algorithms and software for non-combinatorial designs.

Everett *et al*. [1] (Pfizer, Sandwich, UK) have developed an integrated in-house IT library system that aids in the automation of all aspects of synthesis, purification, design and registration of non-combinatorial libraries. It consists of two systems, LiCRA and S1D. LiCRA is used to create, register and help automate library synthesis. S1D selects reagents to maximize diversity, but such that each reagent is used an equal number of times and the design falls within predefined limits for specified physicochemical properties, such as ClogP and molecular weight. The system

(a)



(b)

No. of reagents	PAD™ library size	Combinatorial library size
n	n^2	n^3
10	100	1000
30	900	27,000

(c)

Library size	No. of different reagents at each site		Repeat occurrences of each reagent in library	
	PAD™	Combinatorial	PAD™	Combinatorial
~1000	32	10	32	100

Drug Discovery Today

Figure 3. A comparison of the Predictive Array Design™ (PAD™) and combinatorial designs for a library with three sites of variation (R^1, R^2 and R^3) and equal numbers (*n*) of reagents at each site. The PAD design facilitates the synthesis of smaller libraries with considerably more variation at each substitution site than the combinatorial design. **(a)** A scaffold with three sites of variation. **(b)** Comparison of PAD™ and combinatorial library size for 10 and 30 different reagents at R^1, R^2 and R^3. **(c)** Comparison of the number of different reagents and repeat occurrences of reagents in a PAD™ and combinatorial library of ~1000 compounds.

can also be used to enable the chemist to cherry-pick required compounds from the combinatorial virtual design and ensure relevant coverage of drug-like physicochemical property space.

Pickett *et al.* [3] have developed a system that uses a Monte Carlo algorithm to select sets of reagents to optimize the number of acceptable products in the final non-combinatorial library (based on user-defined criteria), while ensuring each reagent is used at least a pre-defined minimum number of times for synthetic efficiency. The system uses pharmacophore keys and physicochemical properties to define the acceptability of the final product and can be used for diversity or similarity designs.

Wood and Rose [2] have proposed a sampling system based on the experimental design methodology, Latin Squares, to select n^2 compounds from an n^3 all-combinations library such that all possible pairwise combinations of reagents are present in the synthesized library. The method, Predictive Array Design™ (PAD™), has recently been enhanced to include a simulated annealing optimization process to ensure relevant coverage of physicochemical property space [35]. A comparison of PAD™ with the full combinatorial approach is shown in Fig. 3.

These non-combinatorial designs aim to balance synthetic practicality with good design principles in a similar way to reagent-biased product designs. All the non-combinatorial designs enable a greater number of reagents to be used in the library than any full combinatorial design for an equivalent sized final library. They therefore enable greater diversity. PAD™ is fundamentally different from the other two approaches and from reagent-biased product design in that it does not use an optimization algorithm to drive reagent selection. Reagent selection can be made by the chemists using whichever tool(s) they desire. The novelty comes from the way the reagents are paired together using classical experimental design to ensure all pairwise combinations of reagents are present in a library with three or more sites of variation.

Concluding remarks

In the 1970s and 1980s, QSAR analysts and statisticians preached the value of using experimental design when synthesizing small sets of compounds to ensure good coverage of physicochemical property space and facilitate subsequent QSAR analysis on the activity data. However, in practice, medicinal chemists generally went their own way and decided which compounds to synthesize themselves. The advent of combinatorial chemistry and the enormous numbers of compounds that could potentially be synthesized has led to a resurgence of interest in experimental design methods, and this has developed into the fields of molecular diversity and library design. Chemists are now regularly asking computational chemists for assistance in experimental design. Although commercial diversity and design software is available, many companies have chosen to develop in-house systems tailored to their own requirements.

Design methods were initially focussed on full-combinatorial designs to maximize the number of compounds that could be produced in a short time while aiming to produce relatively diverse libraries with drug-like properties. However, combinatorial chemistry has now matured and non-combinatorial designs are being developed that place even more emphasis on design than on maximizing output. These new designs are being taken up by medicinal and combinatorial chemists keen to achieve relatively high-throughput rates without losing focus on the biological target.

References

1 Everett, J. *et al.* (2001) The application of non-combinatorial chemistry to lead discovery. *Drug Discov. Today* 6, 779–785

2 Wood, J. and Rose, V.S., [BioFocus] (1999) Method of designing chemical substances. PCT WO 99/26901

3 Pickett, S.D. *et al.* (2000) Enhancing hit-to-lead properties of lead optimization libraries. *J. Chem. Inf. Comput. Sci.* 40, 263–272

4 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* 23, 3–25

5 Sadowski, J. and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* 41, 3325–3329

6 Ajay *et al.* (1998) Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.* 41, 3314–3324

7 Clarke, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of 'drug-likeness'. *Drug Discov. Today* 5, 49–58

8 Sadowski, J. (2000) Optimisation of the drug-likeness of chemical libraries. *Perspect. Drug Discov. Des.* 20, 17–28

9 Ajay *et al.* (1999) Designing libraries with CNS activity. *J. Med. Chem.* 42, 4942–4951

10 Jacoby, E. (2001) A novel chemogenomics knowledge-based ligand design strategy – application to G protein-coupled receptors. *Quant. Struct.-Act. Relatsh.* 20, 115–123

11 Gorse, D. and Lahana, R. (2000) Functional diversity of compound libraries. *Curr. Opin. Chem. Biol.* 4, 287–294

12 Livingstone, D.J. (2000) The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* 40, 195–209

13 van de Waterbeemd, H. *et al.* (2001) Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* 44, 1313–1333

14 Labute, P. (2000) A widely applicable set of descriptors. *J. Mol. Graph. Model.* 18, 464–477

15 Pearlman, R.S. and Smith, K.M. (1998) Novel software tools for chemical diversity. *Perspect. Drug. Discov. Des.* 9, 339–353

16 Mason, J. and Beno, B.R. (2000) Library design using BCUT chemistry space descriptors and multiple-four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity. *J. Mol. Graph. Model.* 18, 438–445

17 Brown, R.D. and Martin, Y.C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 36, 572–584

18 Potter, T. and Matter, H. (1998) Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* 41, 478–488

19 Schneider, G. *et al.* (1999) 'Scaffold-hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* 38, 2894–2896

20 Gorse, D. *et al.* (1999) Molecular diversity and its analysis. *Drug Discov. Today* 4, 257–264

21 Turner, D.B. *et al.* (1997) Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* 37, 18–22

22 Clark, R.D. (1997) OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* 37, 1181–1188

23 Waldman, M. *et al.* (2000) Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graph. Model.* 18, 412–426

24 Gillett, V.J. *et al.* (1997) The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Comput. Sci.* 37, 731–740

25 Gillett, V.J. and Nicolotti, O. (2000) Evaluation of reactant and product-based approaches to the design of combinatorial libraries. *Perspect. Drug. Discov. Des.* 20, 265–286

26 Linusson, A. *et al.* (2000) Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.* 43, 1320–1328

27 Martin, E. and Wong, A. (2000) Sensitivity analysis and other improvements to tailored combinatorial library design. *J. Chem. Inf. Comput. Sci.* 40, 215–220

28 Sheridan, R.P. *et al.* (2000) Designing targeted libraries with genetic algorithms. *J. Mol. Graph. Model.* 18, 320–334

29 Zheng, W. *et al.* (2000) PICCOLO: a tool for combinatorial library design via multicriterion optimization. *Pac. Symp. Biocomput.* 588–599

30 Rassokhin, D.N. and Agrafiotis, D.K. (2000) Kolmogorov–Smirnov statistic and its application in library design. *J. Mol. Graph. Model.* 18, 368–382

31 Beroza, P. *et al.* (2000) Applications of random sampling to virtual screening of combinatorial libraries. *J. Mol. Graph. Model.* 18, 335–342

32 Brown, R.D. *et al.* (2000) Combinatorial library design for diversity, cost efficiency and drug-like character. *J. Mol. Graph. Model.* 18, 427–437

33 Good, A.C. and Lewis, R.A. (1997) A new methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. *J. Med. Chem.* 40, 3926–3936

34 Bravi, G. *et al.* (2000) PLUMS: a program for the rapid optimization of focused libraries. *J. Chem. Inf. Comput. Sci.* 40, 1441–1448

35 Lipkin, M.J. *et al.* Predictive Array Design™. A method for sampling combinatorial chemistry library space. *SAR and QSAR in Environ. Res.* (in press)