# 5 Diversity Oriented Synthesis: A Challenge for Synthetic Chemists

A. Bender, S. Fergus, W. R. J. D. Galloway, F. G. Glansdorp, D. M. Marsden, R. L. Nicholson, R. J. Spandl, G. L. Thomas, E. E. Wyatt, R. C. Glen, D. R. Spring

*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK.*

e-mail: drspring@ch.cam.ac.uk
Tel.: +44 1223 336498
Fax: +44 1223 336362

This article covers the diversity oriented synthesis (DOS) of small molecules in order to generate a collection of pure compounds that are attractive for lead generation in a phenotypic, high throughput screening approach useful for chemical genetics and drug discovery programmes. Nature synthesizes a rich structural diversity of small molecules, however, unfortunately, there are some disadvantages with using natural product sources for diverse small molecule discovery. Nevertheless we have a lot to learn from nature. The efficient chemical synthesis of structural diversity (and complexity) is the aim of DOS. Highlights of this article include a discussion of nature's and synthetic chemists' strategies to obtain structural diversity, and an analysis of molecular descriptors used to classify compounds. The assessment of how successful one diversity oriented synthesis is versus another is subjective, therefore we use freely available software (www.cheminformatics.org/diversity) to assess structural diversity in any combinatorial synthesis.

## 5.1 Introduction

Diversity oriented synthesis (DOS) aims to synthesize a collection of compounds that differ substantially in their molecular structure (Burke and Schreiber 2004;

Schreiber 2000; Spring 2003). This has application in aspects of chemical genetics and drug discovery.

Chemical genetics is the study of biological systems using small molecule ('chemical') intervention, instead of only genetic intervention (Schreiber 1998; Schreiber 2003; Spring 2005). Cell–permeable and selective small molecules can be used to perturb protein function rapidly, reversibly and conditionally with temporal and quantitative control in any biological system. Alternatively, biological tools can be used to study protein function such as gene knockouts/knockins, RNAi; but these tools act at the level if the gene, rather than protein and cannot be used in some situations (e.g. essential gene knockouts). Nevertheless they are general, fast, cheap and selective relatively. In order to exploit the advantages of the small molecule approach of chemical genetics advances must be made in finding selective small molecules to any protein quickly, cheaply and with adequate selectivity. But we should be encouraged that even after the billions of years of evolution, nature still uses small molecules for signalling, protection and other essential functions. In drug discovery programmes in major pharmaceutical companies there are teams of synthetic chemists whose roles involve adding new potential drug leads to the companies compound collection. These libraries usually contain upwards of half a million compounds. But what should all these compounds look like?

The first point to appreciate is that chemical space is astronomic (Figure 1). Chemical space is synonymous with multi-dimensional molecular descriptor space, where descriptors are characteristics of the compounds such as molecular weight. In the context of this article, chemical space is defined as the total molecular descriptor space that encompasses all organic compounds with a molecular weight less that 2000 daltons, i.e. most natural products and synthetic drugs. This chemical space is enormous. It has been estimated that the possible number of real organic compounds that are possible with a molecular weight less than 500 daltons is over $10^{60}$ (Bohacek et al. 1996). To put this in context, the number of atoms on earth is approximately $10^{51}$, so there are not enough atoms in the universe to explore all of chemical space, let alone the time it would take to

make everything! Therefore, we cannot make everything, so we have to be selective.

## Figure 1-Chemical Space.

The second point to appreciate is that biology survives with a surprisingly small number of small molecules, and for that matter a surprisingly small number of proteins. Simple life forms can function with a few hundred small molecules. Such life forms have genomes encoding less than a thousand proteins. The human proteome has been predicted to be around a quarter of a million proteins (O'Donovan et al. 2001). This is tiny in comparison to the number of proteins that are theoretically possible. The average size of natural proteins is 300 residues, and with the 20 proteinogenic amino acids this gives a staggering $10^{390}$ possibilities ($20^{300}$). Nature cannot have explored all these possibilities and therefore, we can take heart that we can find a small molecule probe for a biological question, or a drug, without having to make everything! This is due to that fact that there is more than one answer to any (biological) question (Figure 2). I am sceptical about biologically relevant chemical space being miniscule, as I would predict that the majority of the $10^{60}$ possible ‚drug-like' small molecules possible would have some biological activity, albeit often unwanted and unexploitable.

## Figure 2-Small molecule Challenge

If chemical space is huge and we cannot make everything, then firstly, what should we make, and secondly, how should we make it? The first question is discussed in the next section, and the second in sections 5.3 and 5.4 where nature's strategy is compared to strategies available to synthetic chemists.

## 5.2 What To Make?

Structural diversity is essential for lead generation in chemical genetics and drug discovery, as compounds that look the same structurally are likely to share similar physical and biological properties. The answer to the question „what to make?" is that it depends on what you want to use the compound for. If you are looking for an orally bioavailable drug with consideration of pharmacokinetics and the

therapeutic index between efficacy and toxicity, then several observations have been made as to molecular characteristics that are desirable, such as size, shape, allowed functional groups and solubility in water and organic solvents. Such ‚drug-like' compounds have been evaluated in different ways, the most famous of which is Lipinski's analysis of the World Drug Index that led to the ‚rule of five' (Lipinski et al. 1997). Each pharmaceutical company will have its own criteria for what to make. In the realm of chemical genetics there are many different situations where a small molecule may be required. If a small molecule were required for an *in vivo* animal model then ‚drug-like' characteristics would be sensible. If cell-based assays or *in vitro* assays are being used then a wider range of chemical space is exploitable than the restrictive chemical space defined by Lipinski's rules; nevertheless, selectivity is always required for high quality data. As regards allowed functional groups, we can be less prescriptive and even learn some lessons from nature. Nature makes an astonishing array of structural diversity in its secondary metabolites, and moreover they are often structurally complex too. Complex structures are likely to interact with biology more selectively than flat, simple molecules. Therefore, structural complexity is desirable because it is simple to kill cells unselectively, e.g. with bleach. Unfortunately, there are some disadvantages with using natural product extracts. Firstly, nature does not make secondary metabolites in a pure form for us to screen; therefore, the extracts are usually screened as mixtures of many compounds, leaving the problem of purifying and identifying the active component(s). Secondly, the natural product extract may come from a limited source, leaving a supply problem if the active compound is desired. Thirdly, the active natural product may be so complex structurally, such as vancomycin, that making analogues to optimize activity is a formidable synthetic challenge. Fourthly, chemistry space encompassed by natural products (and ‚drug-like' compounds) is unlikely to be the only region useful for discovering physical or biological properties of interest, and moreover, may not be the most productive region. These complications have led organic chemists to take the complementary approach of *synthesizing* structurally diverse and complex small molecules directly (Figure 3).

# Figure 3-DOS vs TOS

## 5.3 Nature's Strategies

The rich structural diversity and complexity of natural products have inspired all synthetic chemists. Many drugs in clinical use today are natural products or natural product derivatives. For example over the last twenty years, 5% of the 1031 new chemical entities approved as drugs were natural products, and another 23% were natural product derived (Newman et al. 2003). Natural products can be simple, such as serotonin and histamine, or complex structurally, such as vancomycin and taxol (Figure 4). They occupy a greater volume of chemical space relative to ,drug-like' compounds, but are still useful to the organisms that produce them at least. They tend to have less nitrogen, but more chiral centres and often have higher molecular masses (Clardy and Walsh 2004). Some natural products such as calicheamicin have highly reactive functional groups (ene diyne), yet are selective (Figure 4). Most of the rich diversity of secondary metabolites appear to come from organisms such as bacteria or plants (Clardy and Walsh 2004). But how do they make such as diverse range of compounds?

## Figure 4- Natural Product Structures

Biosynthetic routes to secondary metabolites are usually linear using simple building blocks usually from primary metabolism (such as amino acids for nonribosomal peptides, acyl-CoA thioesters for polyketides, isoprenyl diphosphates for terpenes). Unusual monomers are synthesized at the same time as the secondary metabolite, with the biosynthetic machinery being encoded in the same gene cluster. Once the monomer units have been added together in a linear fashion to give the scaffold, appendage diversification steps can be taken, for example oxidation (e.g. taxadiene to taxol, reticuline to morphine) or glycosidation (e.g. vancomycin). Nature has the advantage over present-day synthetic chemists in that it can use enzymes to conduct synthetic chemistry with usually complete chemo-, regio- and stereoselectivity. It should also be pointed out that we have identified only a small percentage of natural products to date. Improvements in culturing bacteria, combinatorial biosynthesis and secondary metabolite expression will undoubtedly lead to the discovery of new and exciting leads.

## 5.4 Synthetic Chemist's Strategies

Synthetic chemists have the advantage over nature with respect to a wider selection of building blocks and chemical reactions (nature does not seem to have discovered alkene metathesis-at least not via Ru, Mo or W catalysis). A collection of compounds with the highest level of structural diversity will consist of molecules that have incorporated different building blocks, stereochemistries, functional groups and molecular frameworks (Spring 2003). Consider a coupling reaction that involves a substrate, a building block (or more than one building block for multicomponent coupling reactions), and a reagent to give the product. In simple terms, strategies to generate structural diversity would involve varying the building block [(i) appendage decoration], reagent [(ii) constitutional isomer generation, (iii) stereoisomer generation, (iv) divergent reaction pathways] or substrate [(v) divergent folding pathways] (Scheme 1). The most successful syntheses of structural diversity incorporate multiple strategies.

**Scheme 1- Diversity Generation Strategies**

*Appendage decoration* is the most straightforward diversity-generating processes and a central feature in combinatorial chemistry particularly to improve the biological activity of a drug lead; it involves the use of coupling reactions to attach different building blocks to a common molecular framework (cf. nature's strategy). Many examples are available from the literature of this approach to combinatorial synthesis. If only appendage decoration is used in the library synthesis then all the products will have the same molecular frameworks, which is ideal if a focussed library is required. Nevertheless, if a very diverse range of building blocks is used, then although the scaffold is the same, the overall structural diversity can be very high. In order to generate an even greater degree of structural diversity in the molecular scaffold, other strategies need to be incorporated into the synthesis too.

*Constitutional isomer generation* involves using chemoselective and/or regioselective reactions to synthesize different product isomers. *Stereoisomer generation* involves using reactions that proceed with diastereoselectivity and/or enantioselectivity.

*Divergent reaction pathways* are a very efficient way of generating structural diversity, particularly, diverse molecular frameworks and functional groups. Skeletal diversity is generated by using different reagents to change a common substrate into a collection of products having varied molecular skeletons.

*Divergent folding pathways* utilizes substrates with different appendages that pre-encode skeletal information into a collection of products having different molecular skeletons using common reaction conditions. Most DOS libraries use several strategies to generate structural diversity. For example, Oguri and Schreiber have elegantly demonstrated that six structurally diverse indole alkaloid-like frameworks can be generated by shifting the relevant functionality around three points on a starting scaffold (Scheme 2). A rhodium-catalysed tandem cyclization-cycloaddition reaction was used to efficiently generate distinct frameworks (**1** and **2**) with complete diastereocontrol (Oguri and Schreiber 2005).

Scheme  2- Oguri...

But how do you assess the *degree* of structural diversity that is created? Intuition? It is clear that a less subjective method of assessment is required to assess diversity.

## 5.5 Assessment Of Molecular Diversity

In order to assess the molecular diversity of a collection of compounds on a large scale it is necessary to use computer algorithms that, generally speaking, consist of two operations. Firstly, the structures are put into chemical descriptor space using molecular descriptors; and secondly, diversity in chemical descriptor space is calculated (Xue and Bajorath 2000). The calculation of molecular descriptors creates an abstract representation of the molecule (Bender and Glen 2004; Brown and Martin 1996). The representations of molecules can be classified according to their dimensionality (Willett et al. 1998):

(i) One-dimensional (1-D) where bulk properties such as volume, molecular weight and log P (Downs et al. 1994).

(ii) Two-dimensional descriptors (2-D) are derived from the connectivity table of a molecular structure (Estrada and Uriarte 2001).

(iii) Three-dimensional descriptors (3-D) use geometrical information from points in 3D space.

Since binding of a ligand to a target is an event in space, the geometry of the ligand in relation to that of the binding pocket is critical. Therefore, is it still advisable to use a 2-D method over a 3-D method in certain situations? Molecules are not rigid entities, they are conformationally flexible, especially if many single bonds are present in a molecule, this leads to a 'curse of dimensionality' when dealing with 3-D information. In addition, since the active (binding) conformation of a structure is usually unknown, most of the possible conformations cannot be excluded. Dealing with the complete conformational ensemble results in an increase in noise, since virtually every spatial arrangement can be assigned to the ligand. 2-D methods on the other hand do not explicitly capture shape; shape is implicitly contained in the connectivity table. Therefore the information required is greatly reduced, eliminating noise. This leads to a much faster generation of results while usually retaining their validity. Atom environment descriptors are employed as a molecular representation (Bender et al. 2004), as shown in Figure 5. For diversity assessment, we can calculate the average number of atom environments per molecule. The absolute number of features necessarily increases if non-identical structures are added, but here we are interested in a diversity measure relative to the size of the library. This software is freely available via a web interface at www.cheminformatics.org/diversity.

**Figure 5- Atom environment descriptors**

To test this computational assessment of structural diversity a range of combinatorial libraries was chosen from the literature, and an 'ideal diverse library' consisting of 40 diverse natural products. The diversity values of each library are shown in Table 1.

# Table 1. Diversity Values

The diverse libraries generally result in a higher value of diversity than the focussed libraries; however, certain limitations require highlighting when evaluating the diversity of a collection of compounds. The diversity value is dependent on the number of compounds in the collection; therefore, very small libraries (library members < 10) give illogical results that should be utilized with caution. Also, since the programme compares compounds using two factors: (i) the hybridization of the atoms and (ii) the variation of atoms, a focussed library using a common scaffold with varying appendages that contain a wide variety of elements and different degrees of hybridization will give a higher value than perhaps expected. This programme is a useful tool in assessing the diversity of a collection of compounds; however, it should be employed with due care upon understanding some of its limitations as outlined above.

## 5.6 Technology Aspects

If chemical genetics is going to become more accessible then the synthesis and screening of diverse compound collections needs to be done in a much smaller, faster and cheaper way. These considerations are also attractive to the drug discovery industry where profit margins are being squeezed. Synthesis using microwaves has accelerated compound production to a degree, but really order of magnitude step changes are required to make chemical genetics more competitive relative to biological techniques. Microarray and microfluidics technologies have the potential to make such a step change.

## 5.7 Conclusion

The diversity oriented synthesis of small molecules is a challenge to synthetic chemists, requiring new strategies to generate appendage and skeletal diversity. Progress has been made recently and we have assessed the structural diversity achieved by using a free computer programme (www.cheminformatics.org/diversity) that utilizes fragment-based molecular descriptors to quantify the structural diversity of collections of small molecules. If DOS is to be more useful generally the process of selective small molecule discovery to modulate the function of a given protein will need to be more efficient and economical.

# References

Bender A, Mussa HY, Glen RC (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. J Chem Inf Comput Sci 44:1708-1718

Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2:3204-3218

Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modelling perspective. Med Res Rev 16:3-50

Brown RD, Martin YC (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J Chem Inform Comput Sci 36:572-584

Burke MD, Schreiber SL (2004) A planning strategy for diversity-oriented synthesis. Angew Chem Int Ed 43:46-58

Clardy J, Walsh C (2004) Lessons from natural molecules. Nature 432:829-837

Downs GM, Willett P, Fisanick W (1994) Similarity searching and clustering of chemical-structure databases using molecular property data. J Chem Inf Comput Sci 34:1094-1102

Estrada E, Uriarte E (2001) Recent advances on the role of topological indices in drug discovery research. Curr Med Chem 8:1573-1588

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3-25

Newman DJ (2003) Natural products as sources of new drugs over the period 1981-2002. J Nat Prod 66:1022-1037

O'Donovan C, Apweiler R, Bairoch A (2001) The human proteomics initiative (HPI). Trends Biotechnol 19:178–181

Oguri H, Schreiber SL (2005) Skeletal diversity via a folding pathway: synthesis of indole alkaloid-like skeletons. Org Lett 7:47-50

Schreiber SL (1998) Chemical genetics resulting from a passion for synthetic organic chemistry. Bioorg Med Chem 6:1127-1152

Schreiber SL (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. Science 287:1964-1969

Schreiber SL (2003) The small molecule approach to biology. Chem Eng News 81:51-61

Spring DR (2003) Diversity-oriented synthesis; a challenge for synthetic chemists. Org Biomol Chem 1:3867-3870

Willett P, Bernard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983-996

Xue L, Bajorath J (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry and virtual screening. Combinatorial Chemistry & High Throughput Screening 3:363-372

Legends:

Figure 1. Chemical Space.

Figure 2. Small Molecule Challenege.

Figure 3. Target oriented synthesis (TOS) versus diversity oriented synthesis (DOS). DOS concerns the efficient synthesis of structurally diverse (and complex) small molecules, *i.e.* where the molecules differ in their (i) attached groups, (ii) stereochemistry, (iii) functional groups and (iv) molecular frameworks. TOS aims to synthesize a single target. Synthetic pathways in DOS are branched and divergent and the planning strategy extends simple and similar compounds to more complex and diverse compounds. Retrosynthetic analysis concepts focus on the existence of a defined target structure. In DOS there is no single target structure and therefore retrosynthetic analysis cannot be used directly and a forward synthetic analysis algorithm is required. The three-dimensional grids of molecular descriptors illustrate the product(s) of the syntheses in chemical descriptor space.

Figure 4. Natural Product Structures.

Figure 5. Illustration of descriptor generation step, applied to an aromatic carbon atom. The distance ('layers') from the central atom is shown in brackets. Every heavy atom in the hydrogen-depleted structure of the molecule is assigned its Sybyl atom types. Sybyl atom types and are used to classify atoms according to the element type and hybridization state. An individual atom fingerprint is calculated for each heavy atom in the molecule capturing its local environment at a distance of n bonds. Frequencies of atom types at a given distance (n=0, 1, 2) are recorded.

Scheme 1. Diversity Generation Stratergies. Skeletal diversity can be generated by constitutional isomer and stereoisomer generation, divergent reaction pathways and divergent folding pathways.

Scheme 2. Strategies to give structurally diverse molecular frameworks by divergent folding pathways.
-
Table 1. Diversity value of nine collections of compounds. The diversity value is calculated on a scale from 0 to 100 incorporating the number of features per molecule. [a] To 3 significant figures. [b] Nearest integer value. [c] The 'ideal diverse library' consists of Acetic acid, Alliin, Ampicillin, Bee pheromone, Benzene, Bergenin, Beta carotene, Blebbistatin, Caffeine, Catechin, Cinnamic acid, Ciprofloxacin, Cocaine, Cortisone, Cyclosporin, Cysteine, D-glucose, Dpoamine, Erythromycin, Fluzanim, Fumiquinazoline G, Genistein isoflavonoid, Glucosamine, L-DOPA, Methane, Methanol, Morphine, Nandrolone, Omega-6 fatty acid, Phenylalanine, Quinine, Rapamycin, Serotonin, Streptomycin, Sucrose, Taxol, Testosterone, Vitamin A, Vitamin E and Vitamin K.

Figure 1

# *Small Molecule Challenge*

Where can we get millions of small molecules?

*The number of possible "drug-like" molecules has been calculated ($10^{62}$ to $10^{200}$) to be*

# *astronomic.*

RS Bohacek, et al. *Med. Res. Rev.* **1996**, *16*, 3.
MJ Owen *Biotech Advantage* **2002**, 6.

Figure 2

# *Small Molecule Challenge*

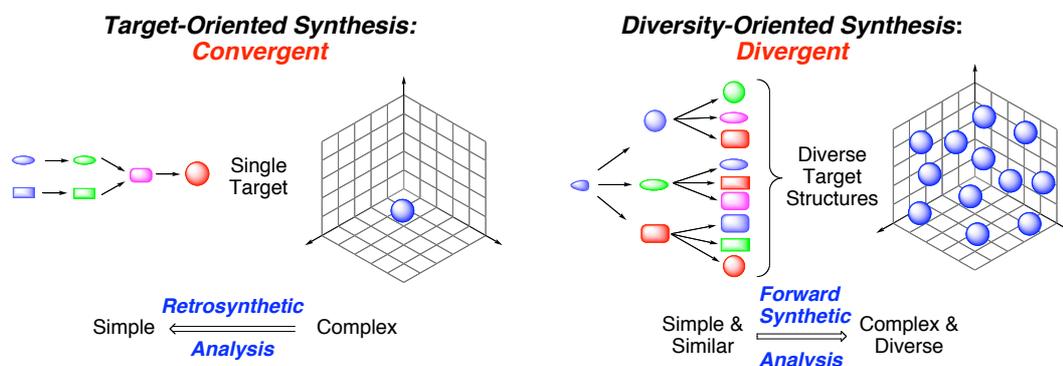Use existing chemistry techniques? **NO.**

- *Quality **and**, but not just, quantity counts.*

- *Structurally similar compounds have similar biological activities.*

- *There is more than one answer to every (biological) problem.*

*Need: **Structurally-Diverse***
*Small Molecule Collections*

Figure 3

# *Diversity-oriented synthesis*

Comparison of TOS and DOS



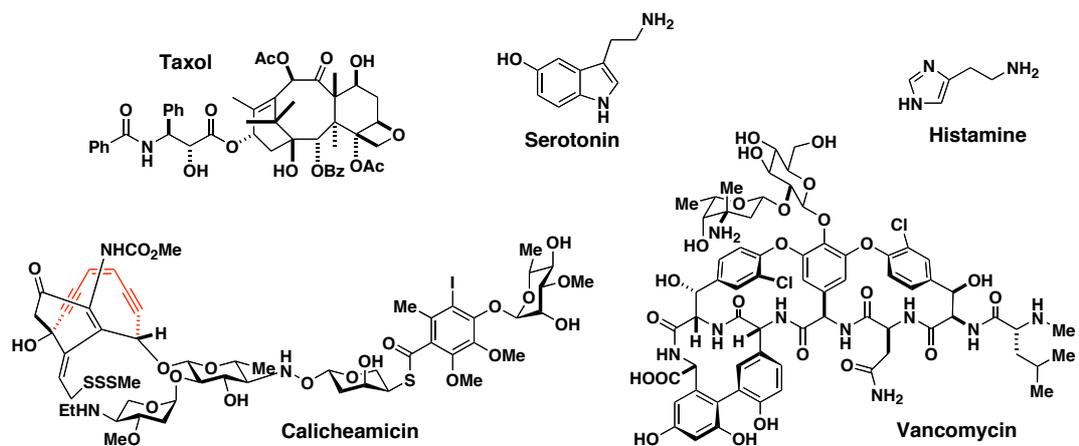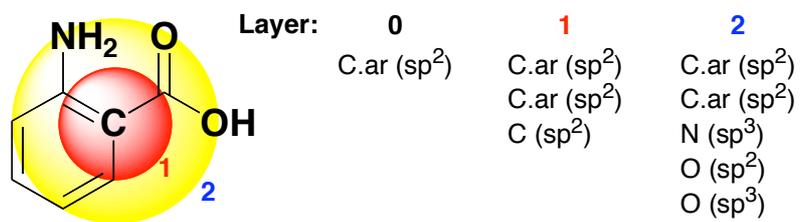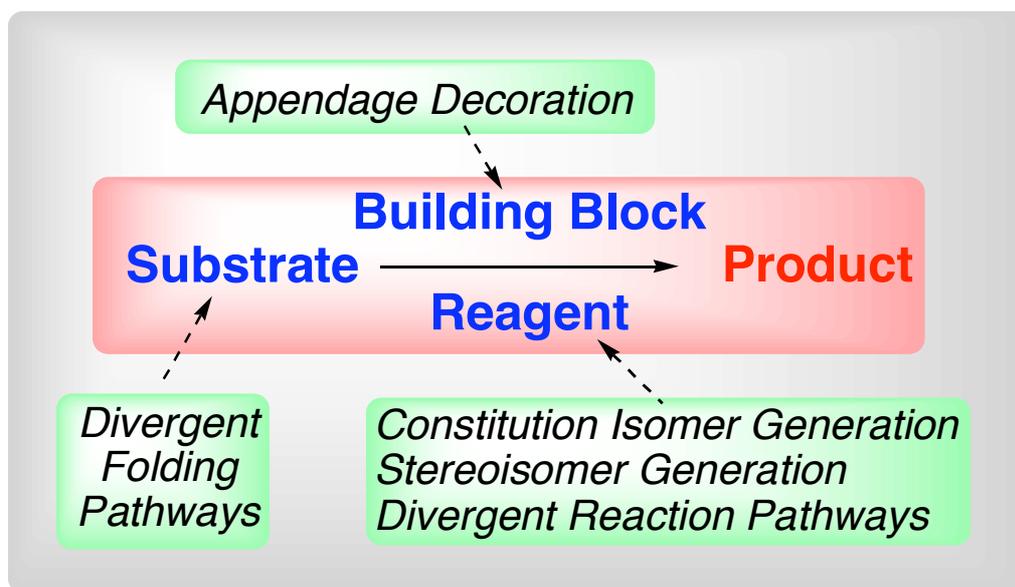*Org. Biomol. Chem.* **2003**, *1*, 3867-3870

Figure 4



Taxol

Serotonin

Histamine

Calicheamicin

Vancomycin

Figure 5



| Layer: | 0 | 1 | 2 |
|---|---|---|---|
| | C.ar (sp$^2$) | C.ar (sp$^2$) | C.ar (sp$^2$) |
| | | C.ar (sp$^2$) | C.ar (sp$^2$) |
| | | C (sp$^2$) | N (sp$^3$) |
| | | | O (sp$^2$) |
| | | | O (sp$^3$) |

Scheme 1

Scheme 2.
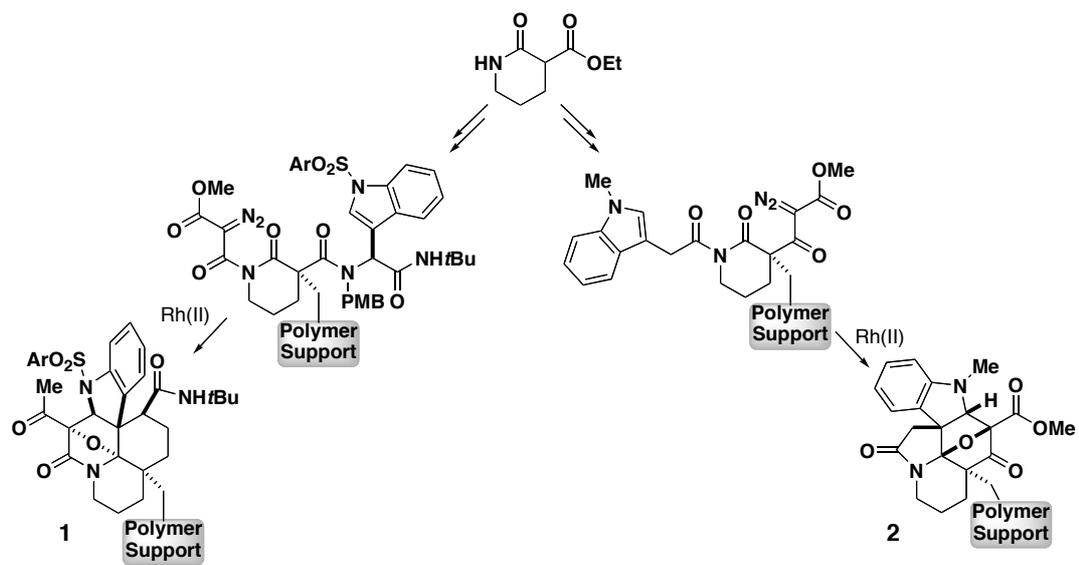
Table 1

| Library | Ref. | Number of Molecules | Number of Features | Features per Molecule[a] | Diversity Value[b] | |
|---|---|---|---|---|---|---|
| 1 | [11] | 168 | 38 | 0.226 | 2 | |
| 2 | [12] | 88 | 30 | 0.341 | 3 | |
| 3 | [13] | 49 | 50 | 1.02 | 10 | |
| 4 | [14] | 62 | 101 | 1.63 | 16 | *Increasing Molecular Diversity* |
| 5 | [18] | 24 | 56 | 2.33 | 23 | |
| 6 | [17] | 15 | 46 | 3.07 | 30 | |
| 7 | [16] | 18 | 72 | 4.00 | 39 | |
| 8 | [15] | 60 | 288 | 4.80 | 46 | |
| 9 | [19] | 10 | 55 | 5.50 | 53 | |
| Ideal[c] | - | 40 | 414 | 10.4 | 100 | |