# Methods for Library Design and Optimisation

D.V.S. Green* and S.D. Pickett

*Department of Cheminformatics, GlaxoSmithKline, Stevenage, UK*

**Abstract:** The introduction of combinatorial chemistry groups into pharmaceutical companies provoked a desire for efficient and effective methods for library design and optimisation. This, in turn, has resulted in a large number of scientific publications, detailing a variety of approaches to the problem. This review attempts to describe the major works in the literature, to set them in context both chronologically and scientifically, and to identify the outstanding challenges that must be addressed, if this area of research is to maintain the rapid progress seen hitherto.

## 1. INTRODUCTION

The past decade has seen the pharmaceutical industry invest heavily in automation, most notably in screening, compound handling and chemical synthesis. The medicinal chemist has access to an impressive, and ever increasing range of chemical reactions, which can be carried out in parallel on robotic systems [1]. For those wishing to build large collections of compounds for high throughput screening, or those lead optimisation projects that are able to utilise these automated reactions, the associated increase in throughput offers the opportunity to reduce the cycle time from lead identification through to candidate selection. However, although the efficiency of a chemistry team equipped with such automation, measured in terms of number of compounds produced, may be increased by, say, a hundred fold, this by no means guarantees the effective exploration of the accessible chemical space. For example, should a lead series be conveniently synthesised by a three component reaction, and each reactant could be drawn from a pool of 200 (a rather conservative example, should the required functional groups be from common reactive species such as amines, aldehydes or acids), the number of possible products is 8 million, which is approximately the same number of compounds as registered in Chemical Abstracts! It is therefore, clear that some selection processes must occur to enable an efficient navigation through these possible products, now universally referred to as "Virtual" compounds. These processes have evolved from simple reagent selection procedures, through to the state of the art methods, which utilise sophisticated multi-dimensional optimsation algorithms [2]. Before these algorithms are explored in detail, it is useful to define some terms that will be used throughout this text. A Virtual Library is a set of compounds that could be synthesised from the application of a particular reaction scheme to an appropriate set of reactants. The process of Library Design, or Optimisation, is the application of a process that attempts to identify the optimal set of reactants given the design objectives. These objectives will involve considerations such as activity prediction, ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties, and the size of library to be made, cost of reagents and synthetic efficiency (including combinatorial constraints).

*Address correspondence to this author at the Department of Cheminformatics, GlaxoSmithKline, Stevenage, UK; E-mail: darren.vs.green@gsk.com

## 2. PRODUCTS VERSES REACTANT BASED DESIGN

The most facile Library Design method is an extrapolation of traditional medicinal chemistry, whereby the chemist uses their experience and knowledge of the SAR to choose the next molecule(s) to make, or in this case, the set of reactants to use. Indeed, this was the method of choice for the pioneers of automated chemistry. However, experience has shown [3] that this approach was only successful for the most practical of the design objectives set out above, namely cost of reagents and synthetic efficiency, at the expense of the biological objectives, with the result that many of the combinatorial libraries synthesised contained no biologically active compounds, and many suffered from high molecular weight and lipophilicity. There are several reasons for this, but from a design consideration the most telling is the desire to exploit a reaction to yield large numbers of products. This gives rise to the use of multi-component reaction schemes that, when coupled with a desire to use a "diverse" set of reactants, lead to the large molecular weight products. For those groups seeking to synthesis large numbers of diverse structures, this design methodology was also flawed, partly due to groups choosing to work around large chemical cores, such as benzodiazepines (1) [4], and partly because reactant lists were not selected with product diversity in mind. That, product based diversity designs are instrinisically superior to reactant based designs, was convincingly demonstrated by Gillet *et al.* [5]. The results of this work indicated that the most diverse set of compounds are chosen from a virtual library by use of unconstrained product selection, followed by the application of a combinatorial constraint to the product selection (to satisfy synthetic efficiency objectives), with the reactant based selections consistently rated as the poorest method. However, product based design has several obstacles, not shared by the relatively simple reactant based method. Particularly when using computational measures of diversity, it is imperative that the Virtual Library is comprised of reactants that a chemist would be happy to use, because many undesirable reactants (for example toxic, expensive or unreactive moieties) would be selected by a diversity measure because they are "different". In response to this, tools were developed, which allowed the application of

1389-5575/04 $45.00+.00 

filters to lists of molecules [6,7]. The other major overhead of product based design is the combinatorial explosion. In our previous example, a chemist using reactant based design would only have to consider 200+200+200=600 reactants. In a product based design, the 8 million products must have their structures enumerated, and a selection method applied, which will involve the application of an algorithm and often take quite a time. This is due to the combinatorics of the selection procedure (the majority of examples will assume the use of full combinatorial synthesis). For example, there are $10^{26}$ different ways to choose 10x10 amines and acids from a virtual library of 100x100 reactants [8]! For all but the most simple of library designs, the combinatorics of the problem remove the possibility of a systematic evaluation of the solutions, and therefore all modern methodologies rely on a stochastic method (such as a Genetic Algorithm, Simulated Annealing or Monte Carlo algorithm) to find solutions that meet the design objectives.

## 3. LIBRARY DESIGN PRE-REQUISITES

Before a detailed examination of the available algorithms for library optimisation, there are some common pre-requisites, which can be obtained in a variety of ways. Firstly, a reaction scheme and associated reactants are needed. The majority of algorithms will use molecular descriptors based on the product structure, such as molecular weight, calculated logP, a predicted affinity for a protein from a docking algorithm and so on. Thus, a means to enumerate the product structures is needed. The two most prevalent methods (see Fig. **1**) are fragment marking (sometimes referred to as "clipping", as found in commercial products such as Central Library [9], CombiLibMaker [10] and the ACCORD toolkit [11]) and reaction-based enumeration (as implemented with the Daylight toolkit [7], and Afferent [12]). The enumeration may be accessed in three ways: full enumeration, where the whole set of product structures is enumerated up front; dynamic enumeration, where the product structures are enumerated as required; and implicit enumeration, where product properties are obtained without product structure enumeration. Implicit enumeration has been developed to cope with very large libraries of product structures (> 1 million products). One approach [13] uses a neural network to predict product properties on the basis of the properties of reactants. A facile and intuitive example of this would be to predict logP (the standard algorithms, such as clogP [14], predict by the addition of values for molecular fragments found in the product structure), but several more complex properties are claimed to have been modelled with success. The second methodology [15] aims to reproduce the product properties exactly, without enumeration of the structures. This method works only for properties that can be described as additive with respect to structure: molecular weight, logP, "Lipinski" properties [16]. Both methods report impressive speeds of property calculation: the Lipinski properties were calculated with the CLUMBER program for a million member benzodiazepine library in 96 seconds (10,460 products/sec).

The second pre-requisite for any algorithm is the "scoring function", by which a proposed library can be compared against the design objectives. The simplest scoring functions involve similarity to a lead compound, and for this the most common measure is the Tanimoto index, applied to

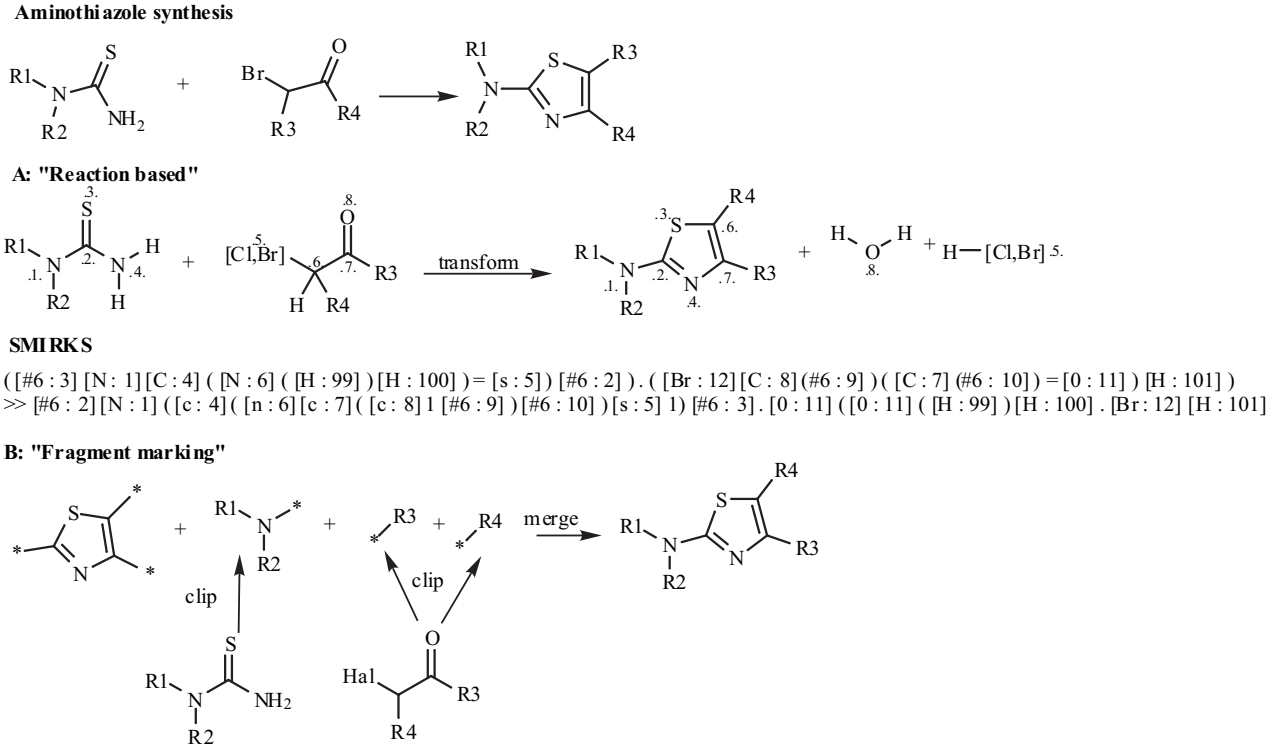


**Fig. (1).** Common methods for virtual library enumeration. In reaction-based methods, the reaction is encoded in a machine-readable format such as SMIRKS [18]. When this reaction is provided with two reactants, they will react if they posses the functional groups defined by the SMIRKS. In fragment marking, or “clipping”, the relationship between reactants and product must be specified beforehand, often by hand, and the enumeration component of the process is then a relatively trivial step.

structural fingerprints such as those used in chemical database systems [17,18]:

$I_{tanimoto}$ = (A and B)/(A or B)

The Tanimoto returns a value between 1 (exactly similar) and 0 (no similarity at all). There are a variety of alternative similarity measures, all of which have their own particular merits [19]. The opposite to similarity is Diversity, and some computational schemes take this literally, so that the diversity of a collection of structures can be computed using the Tanimoto similarity between every pair of the n molecules in the collection:

$I_{diversity} = (\Sigma(1- I_{tanimoto}))/n$

It is more common to describe diversity in terms of a metric that is bounded, which enables libraries of different sizes to be compared. 3D pharmacophore fingerprints [20], cell-based models [21,22,23 Schnur] and clustering [24] have all been proposed. Recently, the concept of information content has also been applied to library design [25]. This attempts to design libraries optimised to ask specific questions, and to allow an efficient deconvolution of screening results in order to answer those questions. For example, in a High Throughput Screen on a protein about which nothing but the peptide sequence is known, one might wish to ask "which 3D pharmacophores does this protein bind to?". Information-theoretic methods are an alternative to diversity methods in this situation.

Finally, perhaps the most familiar scoring functions are those derived for protein-ligand docking methods. As the emphasis of this review is on the design and optimisation algorithms required to address the particular problems of combinatorial library design, readers are referred to several published reviews [26,27,28,29].

Scoring functions can be conveniently classified as "static" or "dynamic" [30]. Static scores relate to properties of the product or reactant that do not change throughout the design process, for example if a product fails the Lipinksi Rule of 5, it always fails no matter what other products are present in the library selected. Dynamic scoring functions result from the consideration of the set of products in the library selected. For example, in a diversity analysis, a product may be considered redundant if it is too similar to several other products in the set, but this result will change if the product is included in a different selection. The scoring function may then be set a design goal, which is usually either to minimise (e.g. the size of the library to be synthesised, or cost of reactants), or maximise (e.g. maximise the diversity of the library, or the number of compounds made that are predicted to be active). Additionally, particularly when considering molecular properties, the design objective may involve a less well-defined goal, such as the fit of a property profile to a reference profile, or the minimisation of the number of products falling outside of a particular range. (Fig. **2**) illustrates the most common implementations of these objectives.

With a means to access the virtual library products and their properties, scoring function(s) and design objective(s), a library optimisation algorithm may be applied.

## 4. LIBRARY OPTIMISATION ALGORITHMS

There exist several reviews of the literature in this area [2,3,31,32], but a précis of the earlier methods will be presented, in order to place the most recent developments in proper context.

### 4.1 Frequency Based Methods

One of the earliest library optimisation procedures was published by Sheridan and Kearsley [33]. This utilised a fragment-based enumeration and a Genetic Algorithm, with a scoring function that attempted to maximise similarity to a

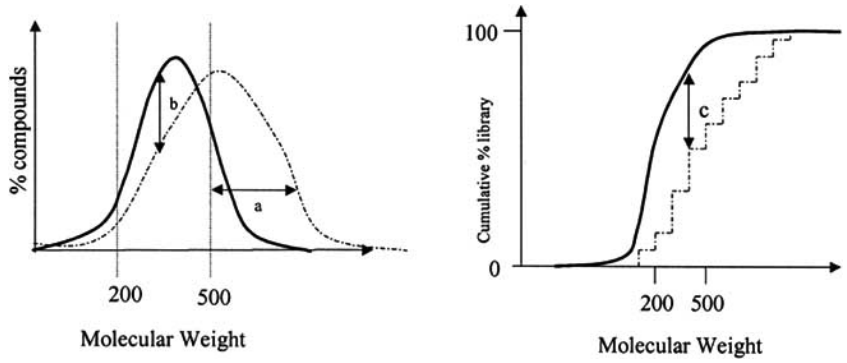


**Fig. (2).** Common methods for constraining the property profile of a library to that of a reference set. The graph shows a reference "drug like" set of compounds (bold line) which generally have molecular weight between 200 and 500. The dashed line represents a library with a significant shift towards large molecules (sadly, a frequent occurrence). The first method (left) is to assign a penalty to compounds which fall outside arbitrary limits, in this case set at below 200 and above 500 Daltons, the penalty being proprtional to the distance, a, of the compound from the desired limit. The penalty score would be the sum of penalties assigned over all the compounds in the library. The second method is to compute a measure of how close the library profile approximates the reference profile, by computing each distance b for all values of molecular weight. This can be accomplished by a Root Mean Square calculation. The third method (right) is to use a cumulative distribution, which enables the use of the $\chi^2$ test, for which the property has to be binned (the figure has molecular weight binned into 100s). The alternative Kolmogorov-Smirnov stastic [45] allows the use of binned or continuous distributions, and computes the difference between the distributions by summation of the distance c over all points, or bins, in the graph.

known lead molecule. Figure **3** illustrates how the algorithm achieves this. There was no combinatorial constraint applied, and thus the result was a collection of products that were of interest. In order to reduce these ideas to practice, a method for mapping a combinatorial set of reactants to the products was required. A simple procedure, now often referred to as MFA (Monomer Frequency Analysis [34] was applied to achieve this. For the top scoring product structures, the names of the reactants are extracted. The reactants are then ranked by the number of times they appear in the top scoring products, and the combinatorial library assembled from the top scoring reactants, with the assumption that a high proportion of the resulting products would appear in the top scoring list. In favourable circumstances, this methodology can be surprisingly effective, as demonstrated by Bravi *et al.* when describing a refinement to the algorithm, known as PLUMS [35]. A deficiency of MFA is that there is no consideration of the combinatorial relationship of the reactants, e.g. an acid A may have a high frequency of occurrence in the selected products through combination with a lot of different amines. If these particular amines are only ever selected when combined with acid A, then through the MFA method they will not have a high frequency of occurrence. The acid A would therefore, not be combined with the most favourable set of amines. PLUMS addresses this issue by working in reverse order to MFA, by the successive removal of poorly scoring reactants (the worst are reactants that do not appear in any selected product). The algorithm can be run until just one of each reactant remains. The output is a graph of library size versus the number of selected products contained in the library. This information can then be used to make an informed choice as to the size of library to be synthesised, and the proportion of "desired" products contained in the library. Table **1** compares the effectiveness of MFA and PLUMS for an example amide library.

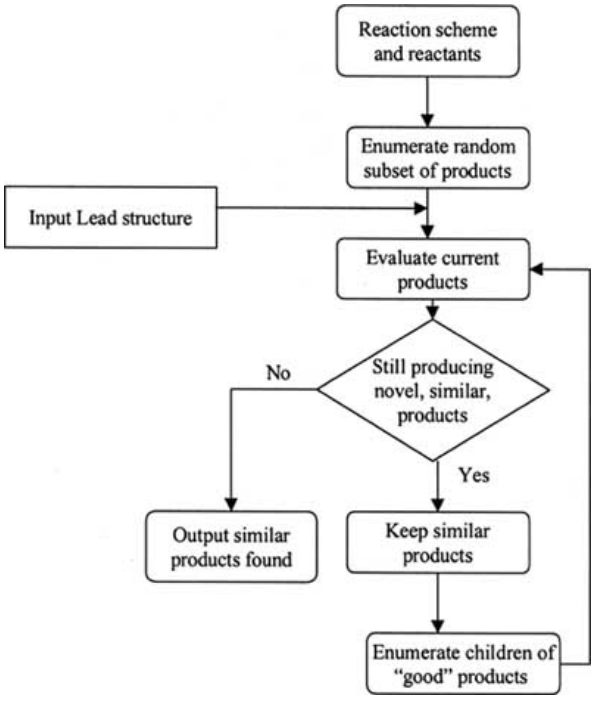


**Fig. (3).** A simple representation of the flow within evolutionary algorithms such as that developed by Sheridan and Kearsley [33]. Such algorithms typically require several hundred to a few thousand iterations to produce a reasonable set of suggestions for synthesis.

### 4.2 Stochastic Methods

Although freqency based methods are simple to develop and fast to apply, they are not applicable to some of the more complex designs demanded by the synthetic community. In particular, the emergence of 3D pharmacophore fingerprint methodologies [36] promised a more biologically relevant description of chemical diversity than had previously been accessible. These desires were to inspire the development of the first algorithm to successfully combine multiple design criteria, Harpick [20]. Harpick utilised Simulated Annealing to navigate the combinatorial design space, with molecules described by product properties- 3D pharmacophore fingerprints, physicochemical properties (Molecular Weight, clogP etc.) and a cost estimate drawn from the ACD database [37]. To distinguish good libraries from bad, Simulated Annealing requires a scoring function. This was implemented as a weighted sum.

**Table 1. Comparison of the Monomer Frequency Analysis (MFA), PLUMS and VOLGA (A Product-Based Method Similar to the Galoped Program [39]) Methods for Library Optimisation [35]**

| Method | Number of Desirable Compounds Synthesised |
|---|---|
| MFA | 39 |
| PLUMS | 69 |
| VOLGA | 69 |
| DMFA | 69 |

A virtual library of 10,000 compounds (100 acids and 100 amines) was analysed, and 409 "desirable" compounds selected. The challenge for the algorithms was to synthesise as many of these desirable compounds as possible using a 10x10 array. Dynamic Monomer Frequency Analysis (DMFA) [35], a modified monomer selection procedure, was developed as a result of these comparisons, and is included for completeness.

The Diversity property scores for each library were based on the number of 3-point 3D pharmacophore patterns found, whilst the property scores were based on the fit to a required property profile (see Fig. **2**). The use of 3D pharmacophores for the design of diversity libraries was further extended by the COMPSEL and DIVSEL methods [38].

At the time that these fledgling methods were developed, there was widespread utilisation of split-mix, solid phase synthetic chemistry methodologies, with the resultant libraries screened as pools. A common methodology for the deconvolution of active compounds from the pool was Mass Spectrometry. It was thus advantageous to minimise the number of products in the library, which have the same molecular weight. However, this was to be accomplished whilst maximising the diversity of the entire library. Around this time, scientists were much influenced by the seminal study of Gillet *et al*. [5] which quantified the superiority of product based designs over reactant selection when designing diverse libraries. The GALOPED program of Brown and Martin [39] allowed an elegant amalgamation of the various design criteria by encoding combinatorial libraries such that

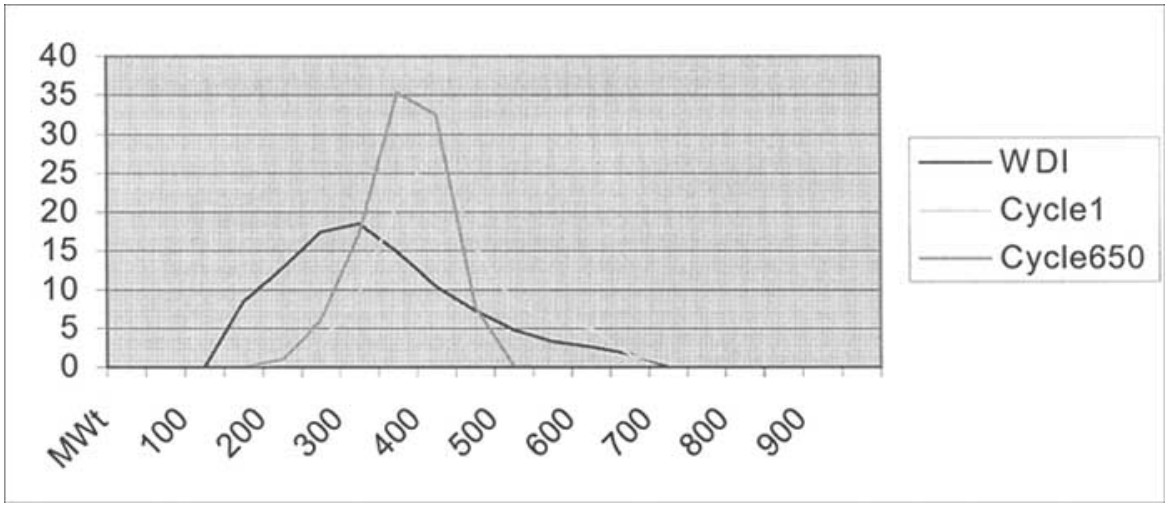


**Fig. (4).** The Molecular Weight profile of a library before and after optimisation with the SELECT program. The Kolmogorov-Smirnov statistic has a tendency to force a distribution towards the middle of the property range, at the expense of the extremes.

they could be evolved by a Genetic Algorithm. Although similar to the work of Sheridan and Kearsley, this work introduced some modifications necessary for application in support of real-life combinatorial chemistry groups, in particular the constraint that the solution must be combinatorial. As with Harpick, a weighted mean scoring function was applied to combine Diversity, size of library and the molecular weight redundancy.

In order to perform their product versus reactant diversity experiments [5], Gillet *et al*. had also developed a Genetic Algorithm based combinatorial selection method. This period also saw the publication of the Rule of 5 [Lipinski Rof5], and the realisation by the scientific community that the quest for diversity had been less than successful at generating molecules, which could be optimised into drugs [40, 41]. As a result, the SELECT program [42] provided a flexible framework for the optimisation of multiple design criteria. Diversity was measured by the fast cosine-similarity method [43], or the slower near-neighbours approach [44], and any property distribution could be defined and used to derive the optimisation, using the Kolmogorov-Smirnov [45] or $\chi^2$ statistic applied to the cumulative distribution. Neither of these statistics allow a completely satisfactory measure of the similarity between distributions. Figure **4** shows a characteristic property profile obtained after optimisation with the Kolmogorov-Smirnov method.

Researchers at 3D-pharmaceuticals have explored numerous methods by which to improve the efficiency and effectiveness of combinatorial library design [46-52]. In particular, the group has optimised the use of Simulated Annealing methods in library optimisation. For example, standard Simulated Annealing protocols enable the solution space to be explored by small transitions from one solution to another. Whether or not the algorithm should step to a particular solution or not is usually governed by the scoring function, E, of the solution. Lower (better) scores are always accepted, whilst the adoption of a higher (poorer) scoring solution is governed by the Metropolis acceptance criterion:

The variation of the "temperature", T, of the system allows the search to sample larger or fewer proportions of higher scoring solutions, thus allowing a broader or narrower search, respectively. The constant $K_B$ is reflective of the cost of jumping from one solution to another. For multi-objective optimisation, although the group adopts a weighted-sum scoring function, the cost function cannot be known a priori. This has been mitigated by use of an adaptive procedure, by which $K_B$ is allowed to vary during the optimisation.

The work of Waldman and co-workers at Accelrys has concentrated on the use of Monte-Carlo prodecures by which to optimise a library [53]. The Cerius-2 software [54] allows the use of a wide variety of scoring functions, which may be optimised through the use of penalty functions and constraints (Fig. **2**). Large virtual libraries are managed through the use of the CLUMBER programme [15]. They have also prescribed a set of principles with which to judge diversity scoring functions. Very few of those published to date are able to achieve these sensible criteria, which gives an indication of the difficulty of research in this area.

There are several other methodologies published [55,56], which adopt a subset of the methodologies outlined above, and reflect the current dominance of product-based, stochastic optimisation procedures which, utilise a weighted-sum scoring function.

### 4.3 Methods that Incorporate Reactant Selection

The success of product-based design methods has somewhat obscured the benefits of reactant-based selection, in particular the facts that product-based designs can be difficult for non-experts to use, and that reactant-based methods map directly to synthetic processes. Therefore, some groups have sought to keep to a reactant-based design, whilst mitigating for the worst excesses of these methods.

One of the earliest publications in Library Design methods was the work of Martin and co-workers at Chiron [57]. This involved the use of D-optimal design to select diverse subsets of reactants, in the assumption that this would lead to diverse products. The implementation was very flexible, and allowed much intuition to be imparted by

the chemist. Reactants were placed into bins, depending on which properties were to be optimised. From these bins, the user decided how many reactants should be selected. For example, in order to mimic a property distribution of drug like compounds, more compounds would be selected from the middle bins than from the extreme. The use of multiple design criteria was allowed by a sequential selection process, which could be used to ensure that, say, high scoring reactants in a docking procedure were added to the design first, and then others were selected in order to achieve the required property distribution. However, the use of a sequential, and hence order dependent, selection process is a major disadvantage, as the many trade-offs between reactants (e.g. one that scores well in a docking method but which yields products with high clogP) cannot be explored or optimised.

The reactant-biased product-based (RBPB) method of Pearlman and Smith [58] tries to combine the benefits of reactant-based and product based designs. A large number of constraints can be incorporated into the design to allow control of library size, numbers of reagents at each position and, importantly for "real-life" application, ensuring that the proposed library fits with the appropriate format (plate layout) of the automation to be applied in the synthesis. The methodology has been implemented in the DiverseSolutions (DVS) suite of programs that includes the novel BCUT [21] descriptors for defining chemical space and efficient cell-spaced diversity algorithms. All potential sub-libraries satisfying the user-defined constraints are considered. As the library design evolves, a library score is calculated that is a function of the product scores for the candidate libraries. Reactant scores are a function of two terms: a term based on the product scores for products containing the reactant and previously selected reagents at the other positions, and a term for the potential products containing the reactant. The reactant scores are transitioned from the possible to the actual as the design proceeds. Product scores can be based on similarity measures for focussed designs or (cell-based) diversity measures for diverse designs. The algorithm also makes it possible to quickly evaluate the effect of changing a reactant or selecting a replacement, as is often necessary in the iterative environment of real world library design. Even if great care has been taken in selecting the initial reactant pool, a selected reactant may not be available at the time of synthesis or may not validate in trial reactions.

A hybrid of MFA and product-based designs has been described by Graham *et al*. [59]. The paper places great emphasis on the practical use of library design, in that combinatorial constraints and plate layouts are examined. Briefly, the virtual library is clustered. For each substituent position R1, R2, R3 etc. the reactants are ranked by the number of clusters that their product structures appear in. This is a rough measure of the diversity imparted by the use of that reactant (although has the same defects as the original MFA as described above). The most diverse reactant is selected, the clusters associated with the chosen reactant are removed from the list, and the next reactant selected. With these reactant lists in hand, a smaller virtual library is constructed from them. A subset of products, one from each cluster in the library, is selected from them. This will be a non-combinatorial list. This set of structures, represented by a matrix (see Fig. **5**), is rearranged by swapping rows and columns to ensure top left of the matrix is as dense as possible i.e. contains the reactants which, when combined combinatorially, give the most products in the desired list of compounds.

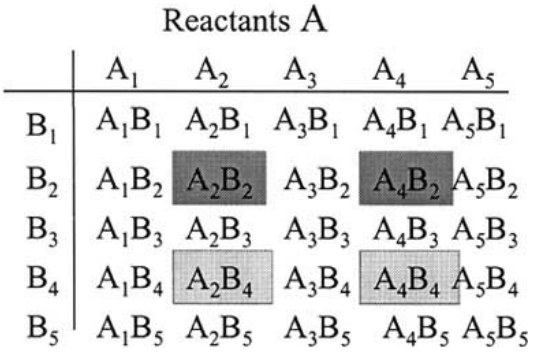


**Fig. (5).** Matrix representation of combinatorial lib.

A true hybrid approach between reactant- and product-based designs is the OptDesign methods [30]. For each reactant R1, R2, R3 etc, the reactant is assigned a class (e.g. "<$10/g" or "available in house"). Each class can be weighted to adjust the proportion of representation in the library. The design procedure is iterative, much like the original Chiron method [56], and seeks to build the library through successive blocks (plates) of combinatorial products. Reactant level filters can be applied- for example, clogP or molecular weight ("Static" filters), reactant similarity (reject reactants similar to those already used) and product similarity (reject products similar to those already in the selected set).

### 4.4 Multi-Objective Methods

All of the above optimisation methods suffer from the same problem. That is a restrictive formulation of overall scoring function. The stochastic, product-based methods use a weighted-sum fitness function to merge all the criteria together, whilst the reactant-based methods are difficult to extend to more than two objectives, as they are order dependent, and therefore unable to explore the trade off between different solutions. Indeed, it has been recognised as a major problem [30]. The weighted-sum fitness function is a problem because there is no way, a priori, of determining the relative weights of say diversity and fit to a molecular weight profile. Instead, several runs are necessary with different weighting schemes. As the number of objectives (terms in the function) increases the balance between the weight terms becomes ever more complex. The very fact of assigning weights in the first place can also have an impact on the solutions evaluated by the search algorithm.

An attempt to mitigate this has been made through the application of sensitivity analysis to the D-optimal design method (see section 4.3)[60]. A further bin is established, which is filled with all reactants. For each bin of interest, four different designs are attempted, by increasing and decreasing the allocated quota by one and then by two. The effect on, say, the library diversity, of these allocation changes can be monitored. Repetition across each pair of design objectives can therefore, give a guide as to which criteria are correlated, and this can be taken into consideration by the chemist.

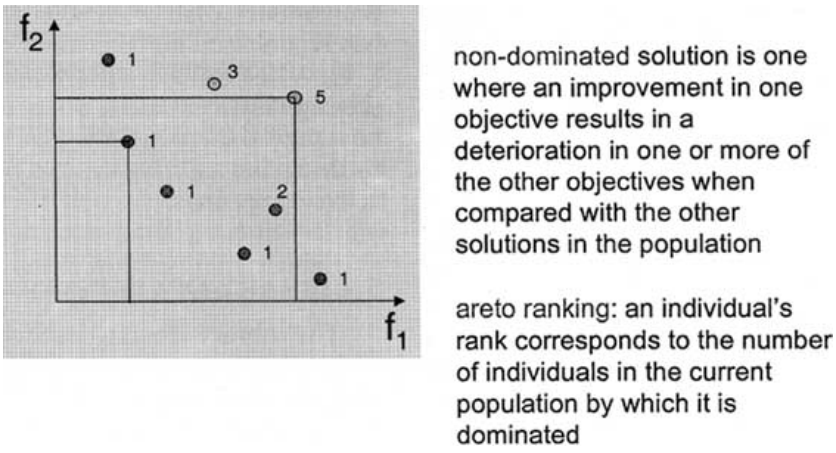


**Fig. (6).** An illustration of Pareto Ranking.

A more radical solution to this problem, which recognises the competition between objectives, is embodied in a sophisticated enhancement of the SELECT methodology, MoSELECT, which employs a Multi-Objective Genetic Algorithm (MOGA) [61]. The method allows a completely flexible environment in which to design libraries with multiple design criteria, without the need to decide how these should be combined to define a scoring function. This is often assigned arbitrarily in previous design methods, or ignored because the methods are unable to incorporate multiple objectives. MoSELECT uses the concept of Pareto Optimality (see Fig. **6**) to rank solutions according to their scores against all the objectives. By using this method, a family of non-dominated solutions is evolved, all of which have some advantage over the others in at least one of the design criteria, and are therefore of interest to the chemist. These solutions can be interrogated by the chemist, and other criteria such as the celebrated "chemist's eye" used to select the most appropriate library to make. This can be thought of as a shift from design (whereby the optimisation algorithm attempts to find the best solution available) towards decision support, where the algorithm is used to present a series of solutions to the scientist, all of which represent ways of achieving the stated design objectives. Advantages of this method over other stochastic approaches include the ability to simultaneously optimise many objectives without having to choose a weighting scheme *a priori*. By not curtailing the search space by use of property constraints and weights, the genetic algorithm is more efficient at searching the chemical space, and may find solutions which are often made inaccessible to other search procedures (see Fig. **7**). Proponents of the reactant-based approach often criticise product-based methods because of the cumbersome link back to synthetic procedures. MOGA alleviates these difficulties, as reactant based criteria, such as cost [62] may be included alongside product based objectives, without having to consider how they should be weighted. In addition, recent development of MoSelectII

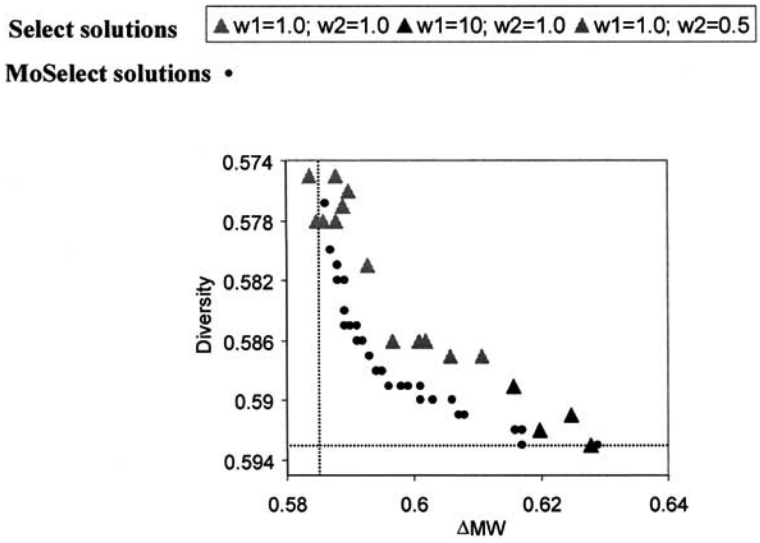


**Fig. (7).** MoSELECT vs SELECT. The SELECT program was run multiple time to design a combinatorial library based on structural Diversity and the fit of the molecular weight profile of the library to that of the WDI. The weights for the objective were varied, and the results shown by the coloured triangles. The solutions found by a single run of the MoSELECT algorithm are able to span the combined set of single solutions produced by multiple SELECT optimisations, without the user having to decide which weights to apply.

[63] illustrates how this methodology may be used to explore other synthetic practicalities, such as library size and configuration, against the theoretical criteria of diversity and molecular properties (see Fig. **8**).

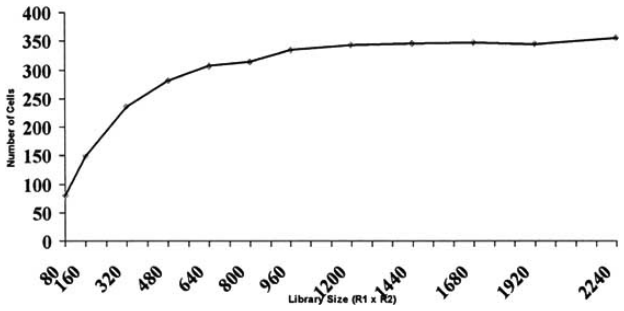


**Fig. (8).** Library size vs Diversity. An illustration of the MoSELECTII program, which is able to treat the library size as an objective, as well as a traditional objective such as diversity. This type of trade-off plot, showing how library diversity at first increases as the library size increases, and then reaches a plateau, can be generated automatically.

### 4.5 DOCKING METHODS

A full review of docking methods and scoring functions is beyond the scope of this article and the reader is referred to several excellent recent reviews of this topic [26,27,28]. At the simplest level a docking algorithm can be used to provide a filter for library design, where all enumerated structures are docked and either the docking score itself or some binary scheme (docks or not) based on a score cut-off, used as one of the design objectives. However, such an approach may be prohibitively time constrained even for moderately sized virtual libraries. Hence several groups have taken advantage of the combinatorial nature of the problem to speed up the calculations. For example, the CombiBuild approach [64, 65], based on the popular docking program DOCK, provides an excellent example of the potential of such methods. The scaffold is predocked (or the position taken from a crystal structure) and each substituent position evaluated independently. Probability maps for each position are used to reduce the influence of steric overlaps between positions. The program was used to design a non-peptide library targeted against Cathepsin D, with impressive results when compared to random or diverse reagent selections. In the CombiDOCK approach, multiple positions of the scaffold are generated and used to evaluate each substituent position independently. Product scores are generated by summing over the constituent reagents with a further check on the higher scoring structures to remove bad intramolecular contacts for example. [66]. Lamb *et al*. [67] expanded on this "divide and conquer" approach to allow the evaluation of multiple libraries against several protein targets. The docking programme FlexX [68] uses an incremental build up procedure for docking. The programme FlexX$^C$ [69] extends this approach to combinatorial libraries by docking the core first and efficiently sampling at each reagent position. This can result in a 30-fold improvement in docking time compared to docking each molecule independently.

Several authors have adapted established *de novo* design programmes to take account of the combinatorial nature of the problem and also to get around one of the issues of such programmes, that of synthetic tractability. Thus, Johnson has reported on an adaptation of the programme SPROUT [70], VLSPROUT, specifically designed for efficiently scoring virtual libraries. Bohm *et al*. [71] have demonstrated how the LUDI programme [72] can be used in this context by designing a library against Thrombin, discovering several active compounds.

## 5. LIBRARY OPTIMISATION IN PRACTICE

The library design methods described above are in widespread use within the pharmaceutical industry and a number of successful applications are emerging. Pickett *et al*. [73,74] used a Monte Carlo search algorithm to optimise both the combinatorial efficiency and bioavailability of a library targeted against p38 MAP kinase. The design objectives incorporated descriptors relevant for drug absorption (polar surface area, Lipinski's rule of 5 [75]) and the designed library showed a significant improvement (as measured by Caco-2 permeability) over a previously synthesised library. A number of potent and orally bioavailable p38 MAP kinase inhibitors were identified suitable for further biological investigation.

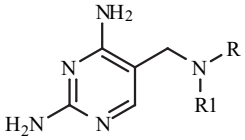


**Fig. (9).** The 2,4-diaminopyrimidine substructure which formed the basis of the virtual library for the Structure based library design work of Wyss *et al* [76]. The R groups derived from a set of secondary amine reagents.

A successful application of structure based library design has been reported by Wyss *et al*. [76]. A library of compounds was targeted against dihydrofolate reductase (DHFR) using a combination of structure-based and diversity-based selection approaches. The program FlexX [68] was used to dock 9884 library products with the constraint of a fixed 2,4-diaminopyrimidine core as defined by a crystal structure (Fig. **9**). 252 from 300 of the top ranked products were synthesised. In addition, a random selection of 150 compounds for which no dockings were found were combined with the lowest 150 scored solutions to give additional 300 compounds, of which 269 were synthesised. 500 compounds were also selected from the full product space by a diversity approach, as defined by the 3D properties of the reagents. The structure-based library gave a 21% hit-rate compared to the diverse library (3%), and the lowest scored solutions (1%).

## 6. CONCLUSION

This paper has catalogued the journey of Combinatorial Library Design from the initial philosophical debates (monomer vs product), through scientific implementation (stochastic algorithms, scoring functions) to decision support (MOGA). It is not an exaggeration to claim that combinatorial design, in the strictest sense, is essentially a solved problem. It is true that the implementations of these algorithms need to become more efficient, in order to cope

with the enumeration, search and evaluation of very large virtual libraries. Indeed, the size of these virtual libraries will continue to grow, as more commercial building blocks are made available, and further studies increase the scope of chemical reactions. However, the greatest problem facing practitioners is the accuracy of available scoring functions for docking, similarity, diversity and ADMET prediction. Until these are improved to the point that educated guesswork is no longer a competitor, the enormous potential of the elegant algorithms expounded in this text cannot be fulfilled. And only then may we truthfully claim to practice Library Optimisation.

## REFERENCES


[1] Dolle, R.E. *J. Comb. Chem*., **2002**, *4*, 369.
[2] Green, D.V.S *Prog. Med. Chem*., **2003**, *41*, 72.
[3] Leach, A. R.; Hann, M. M. *Drug Discovery Today*, **2000**, *5*, 326.
[4] DeWitt, S.H.; Kiely, J.S.; Stankovic, C.J.; Schroeder, M.C.; Cody, D.M.R.; Pavia, M.R. *Proc. Natl. Acad. Sci. USA,* **1993**, *90*, 6909.
[5] Gillet, V.J.; Willett, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 731.
[6] Leach, A.R.; Bradshaw, J.; Green, D.V.S.; Hann, M.M.; Delany, J.J. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 1161.
[7] Walters, W.P.; Murcko, M.A. in Ref. 28, **2000**, pp. 15-32.
[8] Agrafiotis, D.K. *IBM J. Res. Dev.*, **2001**, *45*, 545.
[9] MDL Central Library, available from MDL Inc., San Leandro, CA.
[10] CombiLibDBmaker available from Tripos Inc., St. Louis, MS.
[11] Accelrys ACCORD toolkit, available from Accelrys Inc., San Diego, CA.
[12] Afferent software, available from MDL Inc., San Leandro, CA.
[13] Agrafiotis, D.K.; Lobanov, V.S. *J. Comp. Chem*., **2001**, *22*, 1712.
[14] clogP available from BioByte Inc., Claremont, CA 91711.
[15] Barnard, J.M.; Downs, G.M.; von Scholley-Pfab, A.; Brown, R.D. *J. Mol. Graph. Model*., **2000**, *18*, 452.
[16] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. *Adv. Drug. Delivery Rev*., **1997**, *23*, 3.
[17] ISIS available from Molecular Design Limited, San Leandro, CA 94577.
[18] Weininger, D., Delany, J. Daylight Theory Manual, Daylight Chemical Information Systems Inc., Irvine, CA.
[19] Downs,G.M.; Willett, P. In *Reviews in Computational Chemistry;* Lipkowitz, K.B.; Boyd, D.B. Eds.; VCH Publishers, New York, **1995**; Vol. *7*, pp. 67-118.
[20] Good, A.C.; Lewis, R.A. *J. Med. Chem*., **1997**, *40*, 3926.
[21] Pearlman, R.S., Smith, K.M. *Perspect. Drug Discovery Des*., **1998**, *9, 10, 11*, 339-353.
[22] Lewis, R.A.; Mason, J.S.; McLay, I.M. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 599.
[23] Schnur, D. *Journal of Chemical Information and Computer Sciences*, **1999**, *39*, 36.
[24] Brown, R. D.; Martin, Y. C. *SAR and QSAR in Environmental Research*, **1998**, *8*, 23.
[25] Miller, J. L.; Bradley, E. K.; Teig, S. L *Journal of Chemical Information and Computer Sciences*, **2003**, *43*, 47.
[26] Good, A. *Curr. Opin. Drug Disc. Dev*., **2001**, *4*, 301.
[27] Walters, W.P.; Stahl, M.T.; Murcko, M.A. *Drug Discovery Today,* **1998**, *3*, 160.
[28] Böhm, H.J.; Schneider, G. *Virtual Screening for Bioactive Molecules*, Wiley-VCH; Weinheim, **2000**.
[29] Wang, R.; Lu, Y.; Wang, S. *Journal of Medicinal Chemistry*, **2003**, *46*, 2287.
[30] Clark, R. D.; Kar, J.; Akella, L.; Soltanshahi, F. *Journal of Chemical Information and Computer Sciences*, **2003**, *43*, 829.
[31] Drewry, D.H., Young, S.S. *Chemom. Intell. Lab. Syst.*, **1999**, *48*, 1-20.
[32] Schneider, G. *Current Medicinal Chemistry*, **2002**, *9*, 2095.
[33] Sheridan, R., Kearsley, S.K. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 310-320.
[34] Zheng, W., Cho, S.J., Tropsha, A. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 251-258.
[35] Bravi, G., Green, D.V. S., Hann, M.M., Leach, A.R. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1441-1448.
[36] Davies, E.K. in *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; American Chemical Society, Washington D.C., **1996**, pp. 309-316.
[37] Available Chemicals Directory, available from MDL Inc., San Leandro, CA.
[38] Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. *Journal of Chemical Information and Computer Sciences*, **1998**, *38*, 144.
[39] Brown, R.D.; Martin, Y.C. *J. Med. Chem*., **1997**, *40*, 2304.
[40] Hann, M.M.; Leach, A.R.; Harper, G. *J. Chem. Info. Comput. Sci*., **2001**, *41*, 856.
[41] Oprea, T.I.; Davis, A.M.; Teague, S.J.; Leeson, P.D. *J. Chem. Info. Comput. Sci*., **2001**, *41*, 1308.
[42] Gillet, V.J.; Willett, P.; Bradshaw, J.; Green, D.V.S. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 169.
[43] Holliday, J. D.; Ranade, S. S.; Willett, P. *Quantitative Structure-Activity Relationships*, **1995**, *14*, 501.
[44] Matter, H. *Journal of Medicinal Chemistry*, **1997**, *40*, 1219.
[45] Von Mises, R. *Mathematical Theory of Probability and Statistics*, Academic Press, New York, **1997**.
[46] Agrafiotis, D.K. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 841.
[47] Agrafiotis, D.K. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 576.
[48] Agrafiotis, D.K.; Lobanov, V.S. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 51.
[49] Agrafiotis, D.K.; Myslik, J.C.; Salemme, F.R. *Mol. Diversity*, **1999**, *4*, 1.
[50] Agrafiotis, D.K.; Lobanov, V.S., *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 1030.
[51] Lobanov, V.S.; Agrafiotis, D.K., *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 460.
[52] Rassokhin, D.N.; Agrafiotis, D.K. *J. Mol. Graph. Model*., **2000**, *18*, 370.
[53] Brown, R.D.; Hassan, M.; Waldman, M. *J. Mol. Graph. Model*., **2000**, *18*, 427.
[54] Cerius2 available from Accelrys Inc., San Diego, CA.
[55] Zheng, W.; Hung, S.T.; Saunders, J.T.; Seibel, G.L. *Pacific Symposium on Biocomputing, Honolulu, Jan 4-9*, **2000**, 588.
[56] Martin, E.J.; Critchlow, R.E. *J. Comb. Chem*., **1999**, *1*, 32.
[57] Martin, E.J.; Blaney, J.M.; Siani, M.A.; Spellmeyer, D.C.; Wong, A.K.; Moos, W.H. *J. Med. Chem.*, **1995**, *38*, 1431.
[58] Pearlman, R. S.; Smith, K. M. *Novel algorithms for the design of diverse and focussed combinatorial libraries*. Book of Abstracts, 217th ACS National Meeting, Anaheim, Calif., March 21-25, **1999**.
[59] Graham E.T., Jacober, S.P., Cardozo, M.G. *J. Chem. Inf. Comput. Sci*., **2001**, *41*, 1508.
[60] Martin, E.J.; Wong, A.K. *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 215.
[61] Gillet, V.J.; Khatib, W.; Willett, P., Fleming, P.J.; Green, D.V.S. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 375.
[62] Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, D. V. S. *J. Mol. Graph. Model.*, **2002**, *20*, 491.
[63] Wright, T.; Gillet, V. J.; Green, D. V. S.; Pickett, S. D. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 381.
[64] Kick, E.K.; Roe, D.C.; Skillman, A.G.; Liu, G.; Ewing, T.J.A.; Sun, Y.; Kuntz, I.D.; Ellman, J.A. *Chem. Biol*., **1997**, *4*, 297.
[65] Roe, D. C. *Molecular Diversity in Drug Design*, **1999**, 141.
[66] Sun, Y.; Ewing, T. J. A.; Skillman, A. G.; Kuntz, I. D. *Journal of Computer-Aided Molecular Design*, **1998**, *12*, 597.
[67] Lamb, M. L.; Burdick, K. W.; Toba, S.; Young, M. M.; Skillman, A. G.; Zou, X.; Arnold, J. R.; Kuntz, I. D., *Proteins: Struct., Funct., Genet*., **2001**, *42*, 296.
[68] Rarey, M.; Wefing, S.; Lengauer, T. *Journal of Computer-Aided Molecular Design*, **1996**, *10*, 41.
[69] Rarey, M.; Lengauer, T. *Perspectives in Drug Discovery and Design*, **2000**, *20*, 63.
[70] Gillet, V.J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A.P. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 207.
[71] Bohm, H-J; Banner, D. W.; Weber, L. *Journal of Computer-Aided Molecular Design*, **1999**, *13*, 51.
[72] Bohm H. J. *Journal of Computer-Aided Molecular Design*, **1992**, *6*, 593.
[73] Pickett, S.D.; McLay, I.M.; Clark, D.E. *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 263.

[74] McKenna, J. M.; Halley, F.; Souness, J. E.; McLay, I. M.; Pickett, S. D.; Collis, A. J.; Page, K.; Ahmed, I. *J. Med. Chem.*, **2002**, *45*, 2173.

.

[75] Clark, D. E. *Journal of Pharmaceutical Sciences*, **1999**, *88*, 807.

[76] Wyss, P.C.; Gerber, P.; Hartman, P.G.; Hubschwerlen, C.; Locher, H.; Marty, H-P.; Stahl, M. *J. Med. Chem*., **2003**, *46*, 2304.