J Comput Aided Mol Des (2009) 23:513–525
DOI 10.1007/s10822-009-9264-5

Design of compound libraries for fragment screening

Niklas Blomberg · David A. Cosgrove · Peter W. Kenny · Karin Kolmodin

Received: 24 December 2008 / Accepted: 16 February 2009 / Published online: 13 March 2009

Abstract Approaches to the design of libraries for fragment screening are illustrated with reference to a 20 k generic fragment screening library and a 1.2 k generic NMR screening library. Tools and methods for library design that have been developed within AstraZeneca are described, including Foyfi fingerprints and the Flush program for neighborhood characterization. It will be shown how Flush and the BigPicker, which selects maximally diverse sets of compounds, are used to apply the Core and Layer method for library design. Approaches to partitioning libraries into cocktails are also described.

Keywords Fragment based · Fragment based drug discovery · Fragment based lead generation · Fragment screening · FBDD · FBLG · NMR screening · Screening library · Library design · Molecular complexity · Molecular similarity · Neighborhood · Fingerprint · Foyfi · Flush · Bigpicker · Filter · Leatherface · SMARTS · Solubility

Introduction

Over the last decade, Fragment-Based Drug Discovery (FBDD) has emerged [1–8] as a paradigm in the search for new medicines. A 2007 publication [5] detailed 'seven compounds undergoing or approved for clinical trials, whose origins lie in fragment-based screening'. Fragment-based methods provide a number of advantages over conventional screening approaches, the most important of which is that lead compounds are assembled from proven molecular recognition elements. Screening fragments typically provides access to a larger chemical space and enables this to be searched at a more controllable resolution. Fragment screening can be seen as a process by which fragments and hot spots [9, 10] on the protein surface reveal each other.

Although it has delivered encouraging results to date, the FBDD approach should not be seen as a panacea for the pharmaceutical industry's current ills. There is an undistinguished history in Pharma of over-selling useful technologies such as high throughput screening (HTS), combinatorial chemistry and virtual screening. Put another way, technology is a good servant but a poor master. Success in FBDD is more probable if structural information is available [3] and generating this will prove difficult for certain target classes for some time. Even when the target protein can be crystallized, obtaining structures for weakly bound complexes can still prove challenging.

The development of FBDD can be traced back to 1996, when Fesik and co-workers at Abbott Laboratories introduced Structure Activity Relationships by Nuclear Magnetic Resonance (SAR by NMR) [11] as an approach to lead discovery. Coincidently, the multiple solvent crystal structures [12] (MSCS) method, a precursor of X-ray crystallographic screening [7] had been described by Ringe and colleagues earlier that year. However, it can be argued that the concept of SBDD had appeared earlier in the computational chemistry literature when the multiple copy simultaneous search [13] (MCSS) and LUDI [14] methods were reported by Miranker and Karplus in 1991 and by

N. Blomberg
AstraZeneca R&D Mölndal, Pepparedsleden 1, 431 83 Mölndal, Sweden

D. A. Cosgrove · P. W. Kenny (✉)
AstraZeneca R&D Alderley Park, Macclesfield SK10 4TG, UK
e-mail: pwk.pub.2008@gmail.com

K. Kolmodin
AstraZeneca R&D Södertälje, 151 85 Södertälje, Sweden

Boehm in 1992. These approaches can be linked to the GRID method introduced by Goodford in 1985 [15]. Jencks described a theoretical basis for the benefits of linking fragments as early as 1981 [16].

The two pre-requisites for fragment screening are a robust assay system capable of quantifying weak binding (~mM) and a library of compounds with good aqueous solubility. In fragment screening the power of the assay is defined by the weakness of binding that can be measured reliably. If weak binding can be detected and quantified, compounds of lower molecular weight and complexity can be screened and smaller screening libraries can be used. The availability of a high quality assay can allow exploitation of a weakly binding compound that would not otherwise be a viable starting point for synthesis. This possibility suggests that leadlikeness [17] should be considered a function of both molecular structure and assay capability.

Solubility is an issue that needs to be addressed carefully when selecting compounds for screening at high concentration. It is not particularly difficult to identify compounds of low molecular weight that have good aqueous solubility. Low lipophilicity and the presence of ionizable groups both favor aqueous solubility although analysis showing that hits from NMR screens are more lipophilic than non-hits suggests that minimization of lipophilicity is a poor strategy for library design [4]. Approaches to managing the risk of selecting poorly soluble fragments have been described [18].

Selection of compounds for fragment screening has been described in a number of publications [6, 19–21]. Selection of compounds can be generic with respect to target or directed using protein structural information or known ligands. Generic compound selection is more appropriate when little is known about the target or when specialized formatting of the library is necessary, as is the case for tethering [22]. Other rationales for using generic screening libraries include plasticity of hot spots [23] and the current dearth of reliable methodology for prediction of binding affinity. Fragments are usually assayed as mixtures (cocktails) in NMR and X-ray crystallographic screening, which also favors the use of generic libraries. A typical fragment screening campaign will make use of both generic and directed compound selection as shown in Fig. 1 which summarizes the generalized work flow for fragment-based lead generation (FBLG). It is common to follow up hits by testing close analogs and the availability of these can be used as a criterion for selection of compounds in the initial screening library. Although fragment-based approaches are usually thought of as only relevant to lead discovery, availability of a suitable assay would enable SAR to be explored in a lead optimization framework using fragments (Fig. 2).

A number of criteria for selection of fragments for screening have been suggested. Typically these involve restriction of molecular weight, numbers of non-hydrogen atoms, lipophilicity, hydrogen bonding donors and acceptors, polar surface area and rotatable bonds. Researchers at Astex have proposed [24] the 'rule of 3' (Ro3) although they do not present supporting analysis and other groups appear to have arrived at useful endpoints by screening fragments that are not Ro3-compliant [25]. The molecular

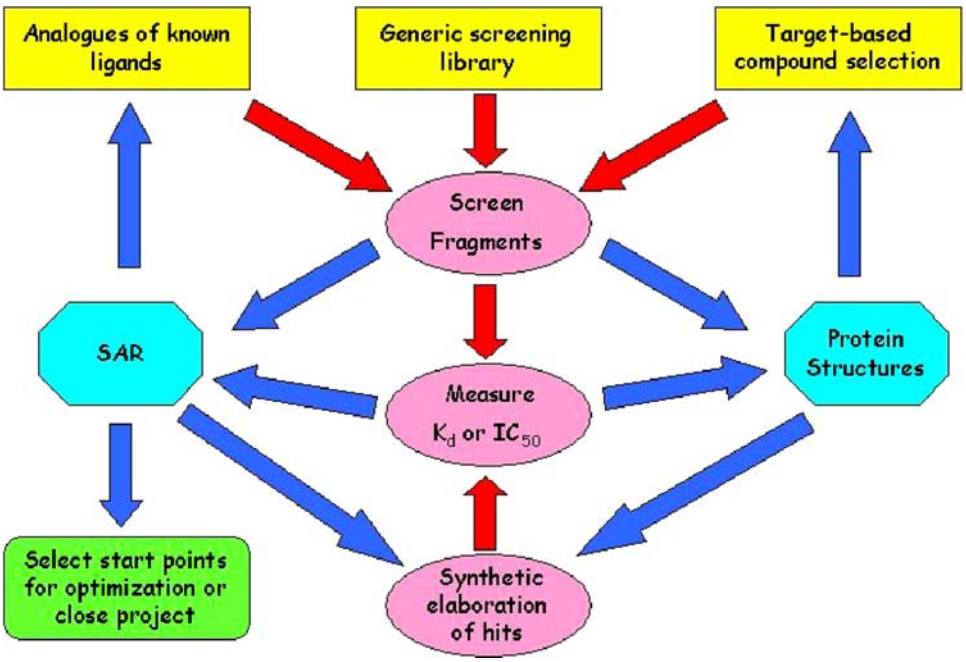

Fig. 1 Generalized workflow for fragment based lead generation with color-coded flows of compounds (*red*) and information (*blue*). The decision as to whether acceptable starting points for further optimization have been generated is based primarily on SAR although availability of high quality structural information will also be a factor

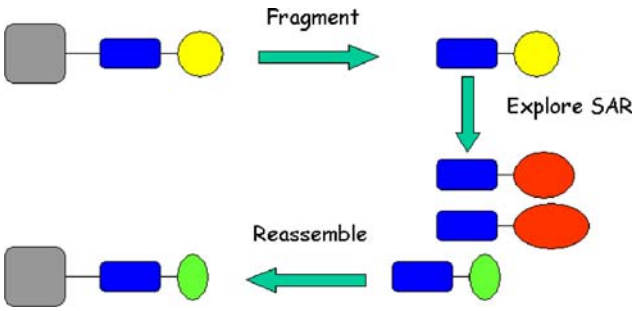

Fig. 2 Application of fragment-based methods in lead optimization. The lead compound is first fragmented to generate a starting point for synthesis that is more conveniently modified. A number of bioisoteric replacements for the functional group shown in *yellow* are evaluated and the best of these (*green*) can be then be incorporated into the more complex scaffold

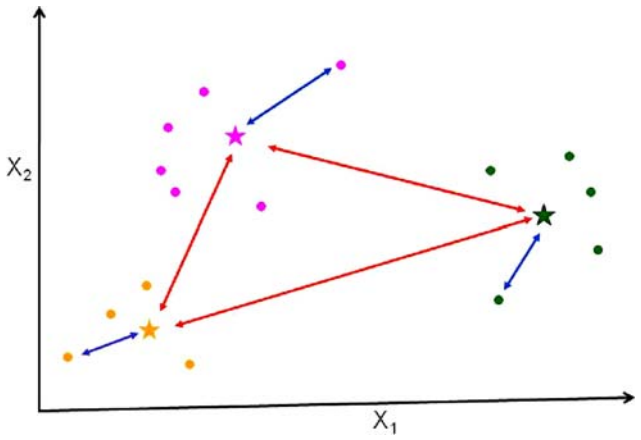

Fig. 3 Diversity and coverage illustrated by a 2D representation of chemical space defined by X_1 and X_2. Similar structures are close to each other and *stars* indicate structures used to sample this region of chemical space. The *color-coded arrows* represent the diversity (*red*) and coverage (*blue*) of the set of representative structures

complexity model [26] introduced by Hann and colleagues in 2001 provides an important theoretical framework for screening library design that is especially relevant to FBDD. Needle screening, described in 2000, represents an alternative approach to controlling molecular complexity although the term was not stated explicitly [27].

Screening library design is often viewed as a two-step process. First acceptable compounds are identified and then a representative subset of these is selected for use as a screening library, typically using some measure of molecular similarity. One objective of screening library design is to select a subset that achieves good coverage of the set of compounds that it samples. Effective coverage requires that most compounds in the full compound set have a close structural analog in the subset (Fig. 3). It should be stressed that diversity is usually a necessary, but not sufficient condition, for good coverage.

This article describes methods used for screening library design within AstraZeneca and illustrates these with reference to generic libraries that have been assembled for fragment-based screening. Screening precursors of the NMR library described in this work has led to identification of a number of exploitable starting points for further optimization [28–32].

Computational tools

Overview

A number of computational tools used in the design of fragment screening libraries were developed in house at Zeneca (1993–1999) for analysis of HTS output and reagent selection for design of compound libraries. Versions of Flush (1995), Flush Clustering (1995), BigPicker (1996), Filter (1995) and Leatherface (1995) were all in place at the time the SAR by NMR approach had been introduced and the availability of these computer programs shaped the fragment library design strategy. A common feature shared by a number of these tools is the use of fingerprint-based measures of molecular similarity and the relevant programs have been written in a sufficiently flexible manner to allow them to be used with any appropriately written fingerprint that the user deems to be relevant [33].

The Foyfi fingerprint

A molecular fingerprint is an abstraction of a molecule's structure, usually as a bitstring or Boolean array. A good introduction to the theory of and motivation for fingerprints is provided in the Daylight fingerprint theory manual [34]. Daylight Chemical Information Systems pioneered the use of hashed fingerprints, of which the Foyfi fingerprints used in this study are another example, originally to speed up the searching of databases for molecules containing particular substructural motifs. However, because similar molecules possess very similar fingerprints, and dissimilar ones have dissimilar fingerprints, they rapidly became used as a more general measure of molecular similarity [35]. To our knowledge, no one has published a detailed recipe of how to create a hash-based molecular fingerprint, so we describe the Foyfi algorithm here in some detail. We make no claims of novelty, but offer it so that others will not in future need to re-invent the wheel as we did. Our algorithm is not the only one nor necessarily the best one, but it produces fingerprints which give similarity figures that the chemists engaged in drug discovery at AstraZeneca are comfortable with.

The essence of a hashed fingerprint of this sort is the production of all paths in a molecule up to a certain length,

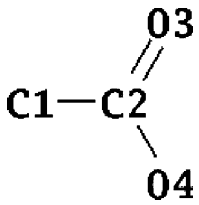

Fig. 4 Atom numbering of acetic acid molecule for Foyfi fingerprints

the conversion of each path into a large number, and the use of that number to set a small number of bits in a large bit string. A particular bit may be set by more than one path, but the aim is to produce a scheme where only a small number of different paths set exactly the same bits. A path in a molecule is the result of starting at one atom, moving to an atom connected to it, and then to an atom connected to that, and so on up to a given maximum size (7 atoms, for Foyfi). For example, in the acetic acid molecule (Fig. 4), the full path set, using the numbering scheme in the figure, is {1}, {2}, {3}, {4}, {1,2}, {2,3}, {2,4}, {1,2,3}, {1,2,4}. The conversion of these paths to a number then involves creating an array of integers consisting of the atomic number of each atom, the bond type (1 for a single bond, 2 for a double, 3 for a triple and 5 for an aromatic) of the bond between that atom and the next in the path and the transformation of this array into a single number using a hashing algorithm. We use the Super Fast Hash of Paul Hsieh [36] for this. The hashed number is then repeatedly divided by the number of bits in the fingerprint and the remainder used to set the corresponding bit in the fingerprint, until the hashed number is reduced to zero. Experience with this path-based fingerprint resulted in two additional refinements. As described, they could not distinguish between an indole and a quinoline. To address this, for rings of between 4 and 8 atoms, the paths were augmented with closed loops, where the first atom in the ring was also added at the end. For indole, this results in an additional path NCCCCN, and for quinoline NCCCCCN, which is sufficient to make their fingerprints different. A similar problem emerged with the fingerprints being unable to distinguish quinoline and isoquinoline. This was fixed by adding an augmented atom descriptor [37] where each atom and its immediate neighbors were encoded in a similar manner to the paths.

The Flush program

Flush is a program that enables the examination of the neighborhood [38] of each of a set of molecules by reference to a second set. It calculates the fingerprint-based Tanimoto distance (1.0—Tanimoto similarity) between the first (probe) set and a second (target) set and presents these distances in various ways. The two output formats used most commonly are 'counts' and 'neighbors'. In the former, for each probe molecule, the numbers of target compounds within a Tanimoto distance of 0.1, 0.2, 0.3… 1.0 are recorded in a table. The counts are cumulative, such that all molecules within the 0.1 Tanimoto shell are also counted in the 0.2 shell and so on. This table gives a compact representation of the similarities between the molecules in the probe and target sets. The 'neighbors' output format requires a threshold value from the user. For each probe molecule, all the target molecules within the threshold distance are listed in order of ascending distance. It is possible to place each molecule in one of two classes (Active or Inactive, reflecting the initial purpose of the program, to assist in the analysis of HTS results) and the output is annotated accordingly. Each probe and target molecule name has its class appended in 'neighbors' output. In 'counts' output, the target molecule count is split between Active and Inactive molecules and the ratio of the two also given, thus giving 3 times as many columns in the output.

The active/inactive ratio of an active quantifies of the likelihood of finding other active compounds in the neighborhood of the compound of interest. A high ratio increases confidence that the activity observed for the hit and its neighbors are genuine. Identifying these patterns of activity is one of the objectives of analysis of HTS output. However, the active/inactive ratio is a relatively crude characterization of the neighborhood of an active compound because it takes no account of how likely the actives are to be found in the neighborhood by chance alone. For example, a neighborhood containing 2 active and 3 inactive compounds is more probable (and therefore less interesting) for a screen with a hit rate of 1% than one with 4 active and 6 inactive compounds, even though 40% of the compounds are active in each case. The probability p_i of finding i actives in a group of n compounds randomly selected from the output of a screen with hit rate h is given by:

$$p_i = \binom{n}{i} h^i (1-h)^{n-i} \quad (1)$$

Letting x equal the number of hits observed in a neighborhood of size n, then a score, S_{neighb} for the neighborhood can be calculated from the probability that at least x actives will be found by chance alone in a group of n randomly selected compounds:

$$S_{\text{neighb}} = -\log \sum_{i=x}^{i=n} p_i \quad (2)$$

Using the previous example of a screen with a 1% hit rate, values of S_{neighb} can be calculated for neighborhoods of 5 compounds with 2 hits ($S_{\text{neighb}} = 3.01$) and 10 compounds with 4 hits ($S_{\text{neighb}} = 5.70$). Note that the active compound that defines the neighborhood is included

in the analysis, which allows singletons to be brought onto the same scale as active compounds with larger neighborhoods. This approach to scoring neighborhoods has been used in analysis of virtual screening output [39]. Cumulative binomial probabilities have been used in an analogous manner to score the performance of chemical series in HTS [40].

Flush clustering

The clustering program flush_clus is an implementation of the sphere-exclusion algorithm of Taylor [41], which has also been reported independently by Butina [42]. The algorithm requires the specification of a threshold Tanimoto distance that defines the boundary of a cluster; all molecules in the cluster have a Tanimoto distance from the cluster centre, or seed, that is no higher than this threshold. The clustering proceeds as follows:

1. For each molecule in the set, find all its neighbors within the threshold.
2. Take the molecule with the largest neighbor list. This molecule (the 'seed' of the cluster) and its neighbors becomes the next cluster. In the case of a tie, the seed with the largest initial neighbor list (before the clustering removed some neighbors) is preferred. If there is still a tie, the seed with the name lower in lexicographical order is preferred, a somewhat arbitrary decision but one taken so that the algorithm remains fully deterministic.
3. Remove all the cluster members from the neighbor lists of all other molecules.
4. Return to step 2 until all molecules are within a cluster.

This algorithm has several advantages over other clustering methods. It is fully deterministic since there is no random element at any stage. All molecules have a predetermined similarity with the cluster centre although, owing to the non-Euclidean nature of the Tanimoto similarity measure, it is not possible to say anything about the similarities of other pairs of molecules within a cluster. The run time is dependent only on the square of the number of molecules. The algorithm is very easily and efficiently parallelized, since the majority of the time is spent calculating the neighbor lists, which can be done independently of each other. Parallelizing the algorithm has another significant benefit, which was in fact the principal driver for the creation of a parallel-processing version. The neighbor lists can become very demanding of memory when clustering a large molecule set (of the order of hundreds of thousands of molecules). In the mid-1990s when the program was written this was a major consideration, since computer memory was much smaller than is common today (64 MB was considered large, for example). By spreading the calculation across a number of computers (using the PVM parallel programming environment) it was possible to use the memory of all of them to hold the neighbor lists making it possible to cluster much larger sets. The increase in run-time that resulted from parallelization was a significant but secondary advantage.

One consequence of the algorithm is the production of 'false singleton clusters.' The final clusters in the output are invariably singleton clusters, where the only member is the seed. Some of these will be true singletons, i.e. molecules lacking neighbors within the clustering threshold, but others (the false singletons) will be singletons by virtue of the fact that their neighbors were placed in other larger clusters in a previous iteration of the algorithm. The flush_clus program offers the opportunity of performing a final sweep through the clusters using a larger similarity threshold and placing the singleton molecules within the cluster for which it has the greatest similarity with the seed, so long as this is within the threshold. Whether the sweep is performed or not, for each cluster seed the number of neighbors it had initially within the threshold is output so as to distinguish the true singletons from the others.

The BigPicker program

BigPicker is a program that selects a subset of molecules from a larger set in a way that attempts to maximize the diversity of the subset. It is a Monte Carlo (MC) approach in which the score of the subset is assessed as a single number, the shortest Tanimoto distance between any two fingerprints in the subset, and it attempts to maximize this score. It is thus an example of a Maximin algorithm. The initial subset is selected at random, and the pair of fingerprints closest to each other identified. One step of the MC algorithm is as follows: One fingerprint of the closest pair is selected at random, and swapped with a fingerprint randomly selected from the remainder of the larger set. The score is then re-calculated, taking note of the fact that the Tanimoto distances need only be calculated for the new fingerprint to the ones remaining in the subset—there is no need to re-calculate the full distance matrix. If the score has improved, the swap is accepted automatically. If it has declined, the Boltzmann factor p is calculated:

$$p = \exp(-T\Delta S) \tag{3}$$

where ΔS is the difference between the old score and the new and T, the effective temperature, is a scaling number. If p is less than a number randomly selected between 0 and 1 the swap is accepted and, if more, it is rejected. The MC steps are repeated until some pre-determined number have been carried out, or until another pre-determined number of attempted swaps have been rejected, i.e. it has not been possible to improve the score of the subset. At points

during the running of the program, T is adjusted to attempt to keep the acceptance/rejection ratio at approximately 0.1. Generally a small number of runs are carried out from different random start points, and the highest-scoring one returned.

A key feature of BigPicker is the ability to pre-select some of the subset (the so-called 'keep' molecules) and thus perform a diversity-based selection about these molecules. Distances between two 'keep' molecules are excluded from the score calculation, but distances between a 'keep' molecule and another in the subset are included. Obviously, if the minimum interdistance includes a 'keep' molecule, it cannot be swapped out, so the other one is automatically selected for swapping. It is also possible to label some or all the molecules as being in a particular class, and specify how many of each class should appear in the subset. Thus if selecting a set of amines, one might want to specify that one quarter of the subset comprises primary amines, one quarter secondary, one quarter tertiary and the remainder from any of the classes. The score is calculated as normal, but the swapping is done in such a manner as to preserve the required makeup of the subset so if a primary amine is selected to swap out of the subset, another one is selected to replace it.

The Filter program

The substructural matching capability required for screening library design extends beyond simply determining whether or not a substructure is present in a molecule. In particular it is important to be able to count occurrences of substructural elements such as non-hydrogen atoms and anionic groups. The Filter program determines whether a structure satisfies a series of substructural requirements defined in SMARTS (SMiles ARbitary Target Specification) notation [43]. Each requirement must be satisfied and performance gains can be achieved by ensuring that the most easily tested and most restrictive requirements are tested first.

Specification of substructural requirements is best illustrated by examples. Filter can be used to search for the occurrence of a particular substructural pattern in a database of structures. For example, the following specifies a 3/5-amino pyrazole in a manner that allows matching of both tautomers:

```
AminoPyrazole [NH2]c1[nD2][nD2]cc1      1
```

The first field is a name to be associated with the SMARTS string in the second field and the third field specifies that exactly one instance of the substructure must be present in the structure for it to be acceptable. In screening library design, it is common to specify an acceptable range in the number of instances of a substructure that is acceptable. For example a requirement for 10 through 20 non-hydrogen atoms can be specified as:

```
NonHydrogen      *      10 – 20
```

It is also possible to define atom types that can be used to specify more complex atom types. In this case an asterisk is used as the third field to indicate that it is not necessary to count the number of instances of the substructural target as illustrated by the following definition of non-fluorine halogens.

```
HevHalo      [Cl,Br,I]   *
```

Library design

General aspects of fragment library design

The design and characteristics of two fragment screening libraries are described in this article. The 20 k generic fragment screening library (GFSL05) was assembled to provide easy access to compounds for FBLG projects and is generic with respect to both target and screening technology. The 1.2 k generic NMR library (APGNMR07) was designed primarily for protein-detect NMR screening although it is intended to be generic with respect to target. GFSL05 was built to take advantage of the speed, throughput and flexibility with which solution samples can be dispensed. Use of solution samples also makes more efficient use of material since dead volumes associated with the dissolution process can lead to significant quantities of material being discarded when concentrated stock solutions are used only for a single screen. The rationale for assembling APGNMR07 was similar and reinforced by the need for more specialized formatting of this library. Both libraries were assembled from compounds that were available in house or from external vendors.

The Core and Layer (CaL) approach (Fig. 5) used to select compounds for these libraries can be summarized as sampling from a sequence of pools of structures. It was the ability to define a 'keep' set for BigPicker that led to the development of CaL as an approach to library design and it is no exaggeration to state that tactical capability shaped strategy in this case. Normally, the pools are selected using progressively less restrictive criteria so the order of a pool in the sequence determines its size and the perceived attractiveness of the compounds in it, although this does not always have to be the case. Pool *i* is sampled to maximize the molecular diversity of the compounds selected with respect to both each other and the compounds already selected from the *i*-1 pools. The compounds selected from pools 1 to *i*-1 can be seen as providing a 'core' to shape sampling from pool *i*, which is termed a 'layer'.

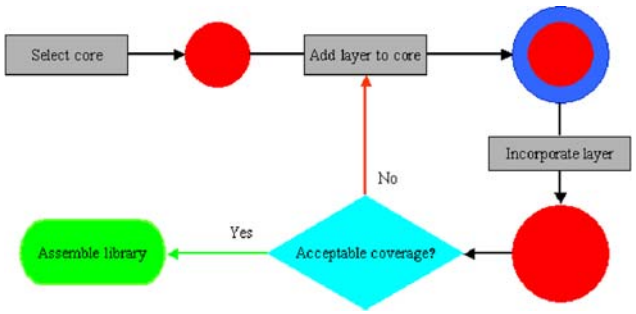

Fig. 5 Core and layer library design biases compound selection away from previously sampled regions of chemical space

Both Flush and BigPicker can be used to apply CaL and to some extent the programs function in a complementary manner when used for this purpose. The link between BigPicker and CaL is particularly clear because the core structures are designated as a 'keep' set. The structures selected by BigPicker to be dissimilar to the 'keep' set, comprise the layers. Once a layer has been selected, the structures in it are added to the keep set in preparation for selection of the next layer. One characteristic of BigPicker relevant to library design is that its selection algorithm favors structures that lack close neighbors. This issue can be addressed by ensuring that selection is only made from structures with an acceptable number of neighbors.

Flush can be combined with Flush_clus to provide a practical implementation of CaL. Flush is used to identify structures that are dissimilar to the core and the program allows considerable flexibility is how this can be done. For example, acceptable structures can be specified as having less than 3 neighbors in the core set with Tanimoto coefficient greater than 70%. This similarity threshold is typically increased as more layers are added. Flush_clus is

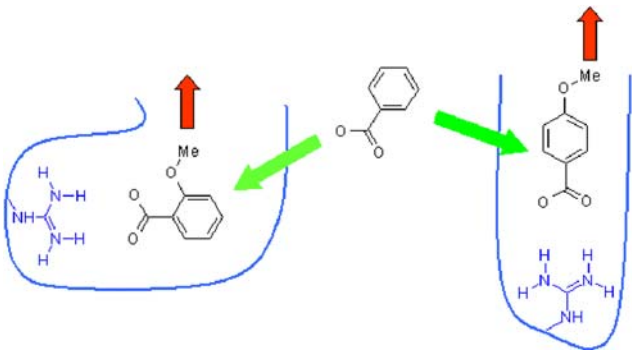

Fig. 6 Extent of substitution as measure of molecular complexity. Each of the isomeric methoxybenzoic acids binds to one site and not to the other. Provided that the assay is capable of detecting the weaker binding of benzoic acid, it will deliver a fragment hit against both targets and analog screening will lead to the identification (*green arrows*) of the appropriate methoxybenzoic acid, which can be further elaborated (*red arrows*)

then used to sample from the set of structures that are adequately dissimilar to the core structures.

Primary selection of compounds for each pool is performed by imposing substructural requirements (e.g., number of non-hydrogen atoms; number of rotatable bonds) defined using SMARTS notation before applying filters based on sample properties (e.g., availability of solid: $\geq$100 mg). Our view of molecular complexity in the context of library design is conceptually closer to needle screening [27] than the Hann model [26] in that we restrict extent of substitution in order to control complexity (Fig. 6). SMARTS notation is especially appropriate for determining extent of substitution because it allows terminal atoms and ring/chain nature of bonds to be specified.

Availability of analogs is another key determinant of the suitability of a compound for inclusion in a screening library. Analogs of a probe structure can be identified automatically by identifying structures for which the molecular similarity with respect to the probe exceeds a user-defined threshold. Molecular similarity can be thought of as defining a neighborhood around the probe structure and the analogs found in there are often referred to as neighbors [38]. We have found Tanimoto coefficient calculated using Foyfi fingerprints to be a particularly useful measure of molecular similarity in library design. Within a chemical series, structures with the most neighbors tend to be the most prototypical and least elaborated. Furthermore the availability of a number of close structural analogs of a compound is an indication that the compound is synthetically accessible and the structural template can be readily elaborated synthetically. The cluster sampling method of Flush identifies the structures with the most neighbors as the 'cluster seeds'.

It is usually necessary to elaborate fragment hits to achieve the low-micromolar potencies that are associated with leads generated by conventional high throughput screening so synthetic accessibility is an issue that needs to be considered in selection of fragments. It might be thought that selecting compounds with synthetic handles such as hydroxyl groups and carboxylic acids would lead to hits that can be readily elaborated. However, synthetic elaboration of a carboxylic acid, for example amide formation by coupling with an amine, changes its characteristics to the extent that the binding of the fragment may no longer be predictive of binding of the derivatives. In their article on library design, Schuffenhauer et al. note 'the carboxylic group as a potential linking group is present in a masked form as a methyl ester' [19] and we have found it useful to focus on 'prototypical reaction products'. These are low molecular complexity compounds that incorporate substructures that can be constructed using reliable synthetic procedures. Examples include amides, sulfonamides, ethers, amines (especially when synthesized by reductive amination) and heterocycles linked by nitrogen, oxygen or

sulfur. Within the prototypical reaction product framework, the amide can be described as presenting the carboxylic acid and amines as reagents to the target protein.

Aqueous solubility is the key physical property in fragment screening since compounds are typically assayed at high concentration in order that weak binding can be detected. Solubility is favored by low lipophilicity and the presence of ionizable groups in the molecule although, as noted in the introduction, excessively polar fragments are less likely to bind to screening targets. We have used measurements from a high-throughput solubility assay to model the risk of selecting poorly soluble fragments as a function of ClogP [18] and to check the solubility of more lipophilic compounds before including these in the library. Once the initial analysis has been carried out, solubility was only measured for neutral compounds with ClogP of at least 2.2 and the measured values were required to exceed the upper limit of 100 mM for the assay. The Leatherface [44] molecular editor was used with an in house ionization and tautomer model to set protonation states of molecules.

Solubility in dimethylsulfoxide (DMSO) is also an issue because high concentration stock solutions are used. These were checked routinely for evidence of precipitation after delivery to primary liquid store (PLS) as part of the process of assembling GFS05 in soluble form. Precipitate was observed for 29 out of 525 (5.5%) samples registered as adducts but for only 89 of the 4,440 (2.0%) non-adducts. Most samples registered as adducts are salt forms (e.g., sodium salt; amine hydrochloride) so this observation was interpreted as evidence that salt forms were more likely to lead to precipitation from DMSO.

Shape matching [45] provides a measure of molecular similarity that is orthogonal to Tanimoto coefficient calculated from Foyfi fingerprints. CaL is compatible with any meaningful measure of molecular similarity and a recently developed method [46] for molecular shape indexing was also used to select compounds. The molecular shape indexing uses a set of reference shapes that are selected to be representative of structures in a multi-conformer database of structures. A candidate library of compounds can be mapped onto the reference shapes to determine which of the latter are under-represented in the library. These reference shapes are independent of chemistry and are potentially transferable between compound collections. Molecular shape indexing was used as a secondary selection tool after a large proportion of library compounds had already been selected. One attraction of combining Foyfi-based similarity with molecular shape indexing in this manner is that it highlights situations in which small structural modifications lead to significant changes in conformational preferences.

Generic NMR screening library

The original generic NMR screening library (1 k) was built using BigPicker to apply the Core and Layer method with a maximum non-hydrogen atom count of 19. An explicit cutoff for ClogP was not used in the design of this library although the property was carefully monitored. The core of the library was built from fragments that consisted of polar warheads (e.g., carboxylic acid, hydroxyl, primary amide, primary amine, heterocyclic nitrogen) linked (0–2 atoms) to benzene rings with at most one other substituent (methyl or chloro). The first layer consisted of compounds selected by restricting extent of substitution. Since BigPicker was used to apply CaL, an additional requirement for the availability in house of least five neighbors at a Tanimoto coefficient of 0.80 was also enforced. Substructural and neighborhood restrictions were relaxed for selection of compounds in the outer layers.

The original generic NMR library served as a basis for the APGNMR07 library and the 1:1:4 cation:anion:neutral ratio of the earlier library was retained. In some cases it proved possible to acquire material to replenish depleted samples and when this was not possible, the Flush program was used to identify close analogs of these compounds. Solubility and purity information that had not been available previously was used to de-select compounds from the original library. The acceptable portion of the original NMR library and close analogs of sample-depleted compounds was used as the core for APGNMR07. CaL was applied by first using Flush to select structures that were adequately dissimilar to the core compounds, followed by selection of a representative set of these using Flush Clustering. Neighborhood restriction, used with BigPicker, is not necessary for Flush Clustering because this procedure selects structures with the largest numbers of neighbors. Molecular shape indexing [46] was used to select 54 compounds for inclusion in APGNMR07 and these were among the last fragments to be selected.

X-ray screening library

A library of 570 compounds was selected from the original NMR library for X-ray crystallographic screening. The need for good aqueous solubility was seen as particularly critical for this screening methodology and only those compounds with molecular weight below 200 Da and ClogP less than 2 were included in the X-ray library. It is important to note that this work pre-dated the availability of shape fingerprinting methodology [46] and the measured solubility data [18] that were used in the design of GFSL05 and APGNMR07.

Generic fragment screening library

GFSL05 was built in order to provide a readily exploitable source of compounds for fragment screening projects with diverse requirements. It was anticipated that many users would select target-directed subsets from the library. Emphasis was placed on achieving good chemical diversity and molecular recognition criteria were taken account of in fragment selection. For example the substructural filters used to select much of the library required the presence of at least one ionizable group or a strong hydrogen bond donor or acceptor. Care was taken to ensure that diverse types of ionizable group were adequately represented and the oxy-anions included sulfonates, phosphonates, acidic phenols and a number of hydroxyheterocycles [47] in addition to the more commonly encountered carboxylic acids. A number of nitrogen-based anions were also represented, including tetrazoles [48], acylsulfonamides, heterocyclic sulfonamides [49] and imides [50]. Another molecular recognition feature that was explicitly targeted was a hydrogen bond donor adjacent to an acceptor as found in pyrazole and lactams. The ionization state distribution of GFSL05 was not controlled to the same extent as for APGNMR07.

A number of fragment collections had been assembled in house before the decision to build GFSL05 was made. The compounds in these libraries were included in the core, provided that sufficient material was available and that no evidence of poor solubility or impurity had come to light. The two most important fragment collections to be incorporated into GFSL05 in this manner were the original NMR library and a collection assembled for screening against infection targets that has been described [4]. During the course of assembling GFSL05 other sources of compounds, including a kinase-directed set and the reagent collection from the AstraZeneca Charnwood site, became available and a number of these were included in the library within the CaL framework. GFSL05 is much larger than the NMR libraries and the order in which compounds were selected also reflected logistical constraints such as the need to source compounds, measure solubility and determine purity. Preparation of solution samples proved particularly resource-intensive because of the need to weigh samples.

Some information about aqueous solubility was available at the outset from a number of sources. Measured values of solubility were available in the in house databases for a number of low molecular weight compounds. The process of generating reference spectra for ligand-detect screening of the original NMR library had identified a number of poorly soluble fragments. Solubility had also been measured using nephelometry for a number of infection library compounds in a variety of assay buffers. Extensive use was made of the high throughput solubility assay in selecting compounds for inclusion in GFSL05. Measurements were made for neutral compounds with ClogP greater than 2.2, which were predicted to have an unacceptable risk of poor aqueous solubility [18].

CaL was applied using Flush and Flush Clustering as described for the APGNMR07 library. The substructural filters were less restrictive than those used to select compounds for APGNMR07 and the CaL mechanism was used to allow a controlled number of compounds with 21–25 non-hydrogen atoms to be selected. The ionization state distribution was not controlled to same extent as for APGNMR07. Molecular shape indexing [46] was also used in selection of compounds for GFSL05.

Cocktailing

X-ray screening library

It is common to screen fragments as mixtures when using NMR or X-ray crystallography for detection of binding since the throughputs that can be achieved with these techniques are typically lower than that of a biochemical assay [7]. Screening mixtures is often referred to as cocktailing and deconvolution is generally necessary once mixtures with hits have been identified. Techniques like X-ray crystallography and ligand-detect NMR that allow the binding of individual ligands to be observed can greatly reduce the amount of deconvolution required.

Once the difference electron density map provides evidence of binding, the identity of the bound fragment needs to be established. Normally, it would be necessary to perform a new round of soaking experiments with each of the ten compounds that were originally pooled in the cocktail. We developed a method for pooling the compounds by maximizing shape diversity in each cocktail to reduce the number of experiments in the second round of screening and simplify the hit identification process. An observed shape of electron density could then be more easily matched to a specific compound in the cocktail, since each compound would have a more or less unique molecular shape.

Molecular shape comparisons were performed using the Rapid Overlay of Chemical Structures (ROCS) program [51]. This program uses a representation of molecular shape based on atom-centered Gaussian functions [45] and generates a shape similarity between 0 (no match) and 1 (perfect match) for the aligned structures. A single 3D conformation was used since most compounds had no, or very few, rotatable bonds. Pair-wise shape comparisons were made for the compounds in the library, resulting in a symmetric 570 by 570 matrix with diagonal elements of

unity (self match) and shape scores between 0 and 1 for the off-diagonal elements. The objective was to partition the library into 57 groups of 10 compounds so that shape diversity was maximized in all groups.

A Monte Carlo protocol similar to that used by the BigPicker program was implemented for the optimization and two scoring functions were evaluated. The first was defined in terms of the maximum similarity between two members of a group, which is minimized to prevent two compounds in the same cocktail being too similar in shape. Letting s_{jk} represent the shape similarity of the j^{th} and k^{th} ($j \neq k$) compounds in each cocktail, this contribution of the i^{th} cocktail to this scoring function is defined by:

$$f_i = \max(s_{jk}) \quad (4)$$

The other scoring function was intended to optimize the overall diversity in the cocktails. Letting M equal the number of compounds in each cocktail, the contribution of the i^{th} cocktail to this scoring function is defined by:

$$g_i = \sum_{j=1}^{j=M-1} \sum_{k=j+1}^{k=M} s_{jk} \quad (5)$$

The scoring functions are computed by summing the contributions of cocktails.

The MC optimizations, where compounds were swapped between groups, were run for 1,000,000 steps. Swaps that resulted in a decrease in the value of the scoring function were automatically accepted. If the scoring function increased in value, swaps were accepted according to the Boltzmann criterion (Eq. 3) as for BigPicker. T was adjusted at 10,000 step intervals so as to give an acceptance ratio (number of accepted swaps/number of rejected swaps) for the unfavorable swaps between 0.1 and 0.3. Effective temperature fluctuated between 0.005 and 0.03 over the course of the MC optimization.

The Σg_i values for a number of cocktailing methods are compared in Table 1. The lowest value Σg_i value resulted from running the MC optimization with Σg_i as the scoring function. Using Σf_i as the scoring function led to a value of Σg_i of that was smaller than that resulting from random assignment of cocktails to mixtures. One particularly striking result is the low variation in the values of g_i for the different cocktails that results from MC optimization using Σg_i as the fitness function.

Table 1 Performance of different cocktailing methods

Method	Mean(g_i)[a]	SD(g_i)[b]	SE(g_i)[c]	Min(g_i)[d]	Max(g_i)[e]
MC (fitness: Σg_i)	35.79	0.03	0.004	35.74	35.83
MC (fitness: Σf_i)	36.20	0.87	0.115	34.41	39.04
'Reversed' MC (fitness: Σg_i)	38.97	2.03	0.269	26.96	44.18
Random	36.55	1.23	0.163	34.17	39.35

[a] Averaged over 57 cocktails

[b] Standard deviation

[c] Standard error

[d] Minimum value

[e] Maximum value

NMR screening library

The APGNMR07 library is stored as 200 mM stock solutions in d6-DMSO. The library is partitioned into 200 groups of 6 compounds, each of which has 1 anionic, 1 cationic and 4 neutral compounds. This allows the library to be screened as mixtures of 6 or 12 compounds while minimizing the risk of buffering problems that might result from too many acids or bases being combined in a single mixture.

Library characteristics

The ClogP distributions of the APGNMR07 and GFSL05 libraries are shown in Fig. 7. Both libraries contain compounds for which ClogP exceeds the Ro3 cut off of 3. Summary statistics for ClogP and non-hydrogen atom count are presented for the libraries in Table 2 and these are broken down according to whether compounds are ionized or neutral. The ionized compounds in GFSL05 are 0.4 log units more lipophilic than the neutral compounds, which is a consequence of the approach used to manage

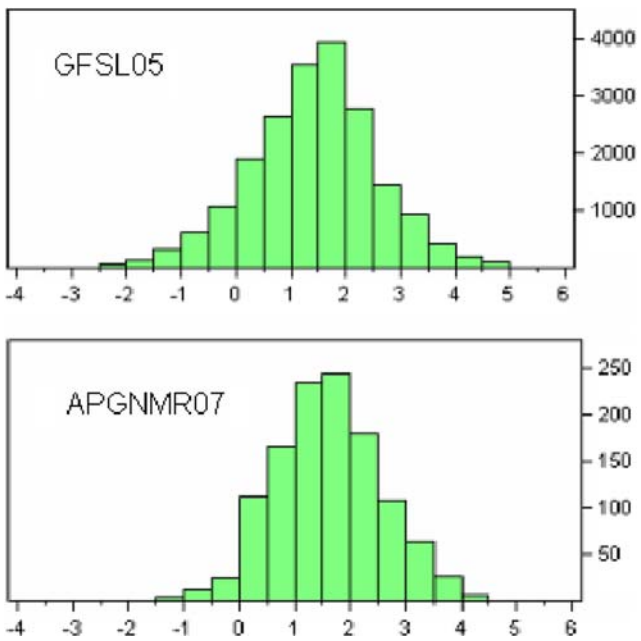

Fig. 7 Lipophilicity profiles for APGNMR07 and GFSL05 libraries

Table 2 Molecular size and lipophilicity summary statistics for GFSL05 and APGNMR07 screening libraries

Library	Property	Ionized[a]	N[b]	Mean	SD[c]	SE[d]
GFSL05	Non-H atoms	No	12,284	15.64	3.42	0.03
		Yes	8,058	16.04	3.58	0.04
	ClogP	No	12,284	1.248	1.056	0.010
		Yes	8,058	1.658	1.337	0.015
APGNMR07	Non-H atoms	No	800	13.66	2.18	0.08
		Yes	400	13.91	2.14	0.11
	ClogP	No	800	1.528	0.978	0.035
		Yes	400	1.718	1.006	0.050

[a] A compound was treated as ionized if one or more ionized form was generated for it by ionization and tautomer model

[b] Number of compounds

[c] Standard deviation

[d] Standard error

risk of poor solubility. Ionizable functionality is exploited to present the most lipophilic groups to screening targets. The compounds in APNMR07 are slightly more lipophilic than those in GFSL05 but have 2 fewer non-hydrogen atoms on average. The distribution of ionization types for GFSL05 is shown in Fig. 8. The proportion of compounds with anionic groups is somewhat larger than that of cationic compounds, which reflects the greater structural diversity of anionic groups.

Neighborhood analysis for the compounds in APGNMR07 and GFSL05 libraries is presented in Table 3. This shows that 70% of the compounds in GFSL05 have at least one neighbor, for which at least 10 mg is available in house as a single sample, with a Tanimoto coefficient of at least 0.85. This figure falls to 48% when the more demanding similarity criterion of Tanimoto coefficient $\geq$0.90 is used to define neighborhood. In contrast, 76% of the compounds in the much smaller APGNMR07 library have at least one neighbor available in house at this level of molecular similarity.

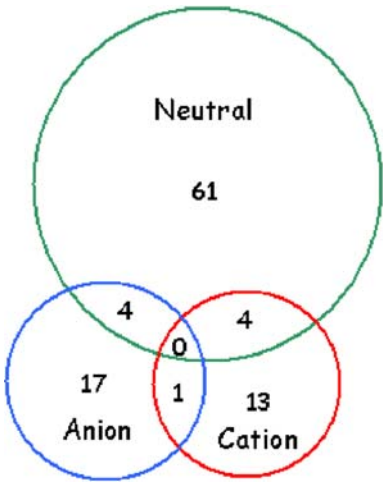

Fig. 8 Ionization state distribution for GFSL05. The ionization and tautomer model used to set protonation states generates multiple forms when the appropriate pKa value is thought to be close to the normal physiological pH of 7.4 as is the case for imidazoles and thiazolidinediones

Table 3 Neighborhood analysis for APGNMR07 and GFSL05 compound libraries

Library	Tanimoto	Comparison	0	1	2	3	4	≥5
APGNMR07	0.80	Available	3.0	4.9	6.1	6.7	4.3	75.1
APGNMR07	0.85	Available	9.1	13.0	11.8	9.3	7.3	49.6
APGNMR07	0.90	Available	23.6	21.6	15.0	9.9	8.1	21.9
APGNMR07	0.80	APGNMR07	79.2	15.1	4.8	0.8	0.0	0.2
APGNMR07	0.85	APGNMR07	87.2	10.3	2.1	0.3	0.1	0.1
APGNMR07	0.90	APGNMR07	92.8	6.5	0.5	0.2	0.0	0.1
GFSL05	0.80	Available	16.2	13.2	10.4	7.9	6.5	45.9
GFSL05	0.85	Available	30.1	18.9	11.9	7.5	6.0	25.4
GFSL05	0.90	Available	51.7	20.2	9.6	5.7	3.3	9.6
GFSL05	0.80	GFSL05	42.8	29.3	14.2	6.2	3.0	4.5
GFSL05	0.85	GFSL05	61.4	26.1	7.8	2.4	1.0	1.3
GFSL05	0.90	GFSL05	79.1	16.6	2.9	0.8	0.3	0.3

Acknowledgments It is a pleasure to acknowledge helpful and insightful discussions with Alex Breeze, Gill Burgess, Jeremy Burrows, Richard Button, Kevin Embrey, Rutger Folmer, Andrew Grant, James Haigh, Neil Hales, Jeff Morris, Paul Owen, Jens Petersen, Adam Shapiro, Ellen Simkiss, Steve St-Gallay, Dave Timms and Richard Ward.

References

1. Congreve M, Chessari G, Tisi D, Woodhead AJ (2008) J Med Chem 51:3661–3680. doi:10.1021/jm8000373
2. Hesterkamp T, Whittaker M (2008) Curr Opin Chem Biol 12:260–268. doi:10.1016/j.cbpa.2008.02.005
3. Hajduk PJ, Greer J (2007) Nat Rev Drug Discov 6:211–219. doi:10.1038/nrd2220
4. Albert JS, Blomberg N, Breeze AL, Brown AJH, Burrows JN, Edwards PD, Folmer RHA, Geschwindner S, Griffen EJ, Kenny PW, Nowak T, Olsson L-L, Sanganee H, Shapiro AB (2007) Curr Top Med Chem 7:1600–1629. doi:10.2174/156802607782341091
5. Jhoti H, Cleasby A, Verdonk M, Williams G (2007) Curr Opin Chem Biol 11:485–493. doi:10.1016/j.cbpa.2007.07.010
6. Barker J, Courtney S, Hesterkamp T, Ullmann D, Whittaker M (2005) Exp Opin Drug Discov 1:225–236. doi:10.1517/17460441.1.3.225
7. Hartshorn MJ, Murray CW, Cleasby A, Frederickson M, Tickle IJ, Jhoti H (2005) J Med Chem 48:403–413. doi:10.1021/jm0495778
8. Erlanson DA, McDowell RS, O'Brien T (2004) J Med Chem 47:3463–3482. doi:10.1021/jm040031v
9. DeLano WL (2002) Curr Opin Struct Biol 12:14–20. doi:10.1016/S0959-440X(02)00283-X
10. Bogan AA, Thorn KS (1998) J Mol Biol 280:1–9. doi:10.1006/jmbi.1998.1843
11. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Science 274:1531–1534. doi:10.1126/science.274.5292.1531
12. Allen KN, Bellamacina CR, Ding X, Jeffery CJ, Mattos C, Petsko GA, Ringe D (1996) J Phys Chem 100:2605–2611. doi:10.1021/jp952516o
13. Miranker A, Karplus M (1991) Prot Struct Funct Genet 11:29–34. doi:10.1002/prot.340110104
14. Boehm H-J (1992) J Comput Aided Mol Des 6:61–78. doi:10.1007/BF00124387
15. Goodford PJ (1985) J Med Chem 28:849–857. doi:10.1021/jm00145a002
16. Jencks WP (1981) Proc Natl Acad Sci USA 78:4046–4050. doi:10.1073/pnas.78.7.4046
17. Teague SJ, Davis AM, Leeson PD, Oprea T (1999) Angew Chem Int Ed 38:3743–3748. doi:10.1002/(SICI)1521-3773(19991216)38:24<3743::AID-ANIE3743>3.0.CO;2-U
18. Colclough N, Hunter A, Kenny PW, Kittlety RS, Lobedan L, Tam KY, Timms MA (2008) Bioorg Med Chem 16:6611–6616. doi:10.1016/j.bmc.2008.05.021
19. Schuffenhauer A, Ruedisser S, Marzinzik A, Jahnke W, Selzer P, Jacoby E (2005) Curr Top Med Chem 5:751–762. doi:10.2174/1568026054637700
20. Baurin N, Aboul-Ela F, Barril X, Davis B, Drysdale M, Dymock B, Finch H, Fromont C, Richardson C, Simmonite H, Hubbard RE (2004) J Chem Inf Comput Sci 44:2157–2166. doi:10.1021/ci049806z
21. Fejzo J, Lepre CA, Peng JW, Bemis GW, Ajay Murcko, MA MooreJM (1999) Chem Biol 6:755–769. doi:10.1016/S1074-5521(00)80022-8
22. Erlanson DA, Wells JA, Braisted AC (2004) Annu Rev Biophys Biomol Struct 33:199–223. doi:10.1146/annurev.biophys.33.110502.140409
23. Thanos CD, Randal M, Wells JA (2003) J Am Chem Soc 125:15280–15281. doi:10.1021/ja0382617
24. Congreve M, Carr R, Murray C, Jhoti H (2003) Drug Discov Today 8:876–877. doi:10.1016/S1359-6446(03)02831-9
25. Petros AM, Dinges J, Augeri DJ, Baumeister SA, Betebenner DA, Bures MG, Elmore SW, Hajduk PJ, Joseph MK, Landis SK, Nettesheim DG, Rosenberg SH, Shen W, Thomas S, Wang X, Zanze I, Zhang H, Fesik SW (2006) J Med Chem 49:656–663. doi:10.1021/jm0507532
26. Hann MM, Leach AR, Harper G (2001) J Chem Inf Comput Sci 41:856–864. doi:10.1021/ci000403i
27. Boehm H-J, Boehringer M, Bur D, Gmuender H, Huber W, Klaus W, Kostrewa D, Kuehne H, Luebbers T, Meunier-Keller N, Mueller F (2000) J Med Chem 43:2664–2674. doi:10.1021/jm000017s
28. Hohwy M, Spadola L, Lundquist B, Hawtin P, Dahmén J, Groth-Clausen I, Nilsson E, Persdotter S, Von Wachenfeldt K, Folmer RHA, Edman K (2008) J Med Chem 51:2178–2186. doi:10.1021/jm701509k
29. Geschwindner S, Olsson L-L, Albert JS, Deinum J, Edwards PD, De Beer T, Folmer RHA (2007) J Med Chem 50:5903–5911. doi:10.1021/jm070825k
30. Edwards PD, Albert J, Sylvester M, Aharony D, Andisik D, Campbell J, Chessari G, Congreve M, Folmer RHA, Geschwindner S, Koether G, Kolmodin K, Krumrine J, Mauger RC, Olsson L-L, Patel S, Spear N, Tian G (2007) J Med Chem 50:5912–5925. doi:10.1021/jm070829p
31. Black E, Breed J, Breeze AL, Embrey K, Garcia R, Gero TW, Godfrey L, Kenny PW, Morley AD, Minshull CA, Pannifer AD, Read J, Rees A, Russell DJ, Toader D, Tucker J (2005) Bioorg Med Chem Lett 15:2503–2507. doi:10.1016/j.bmcl.2005.03.068
32. Breeze AL, Green OM, Hull KG, Ni H, Hauck SI, Mullen GB, Hales NJ, Timms D (2005) Preparation of pyrroles as antibacterial agents. (2005) WO 2005026149
33. Kogej T, Engkvist O, Blomberg N, Muresan S (2006) J Chem Inf Model 46:1201–1213. doi:10.1021/ci0504723
34. Fingerprint theory manual, Daylight Chemical Information Systems Inc., Aliso Viejo, CA 92656, USA. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html, accessed 26th November 2008
35. Flower DR (1998) J Chem Inf Comput Sci 38:379–386. doi:10.1021/ci970437z
36. Hsieh P Hash codes. http://www.azillionmonkeys.com/qed/hash.html
37. Adamson GW, Lynch MF, Town WG (1971) J Chem Soc C 1971:3702–3706. doi:10.1039/j39710003702
38. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) J Med Chem 39:3049–3059. doi:10.1021/jm960290n
39. Krumrine JR, Maynard AT, Lerman CL (2005) J Med Chem 48:7477–7481. doi:10.1021/jm0501026
40. Nilakantan R, Nunn DS (2003) Drug Discov Today 8:668–672. doi:10.1016/S1359-6446(03)02793-4
41. Taylor R (1995) J Chem Inf Comput Sci 35:59–67. doi:10.1021/ci00023a009
42. Butina D (1999) J Chem Inf Comput Sci 39:747–750. doi:10.1021/ci9803381
43. SMARTS theory manual, Daylight Chemical Information Systems Inc., Aliso Viejo, CA 92656, USA. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed 19th December 2008
44. Kenny PW, Sadowski J (2005) Structure modification in chemical databases. Methods and Principles in Medicinal Chemistry, vol 23 (Chemoinformatics in Drug Discovery), pp 271–285. http://dx.doi.org/10.1002/3527603743.ch11
45. Grant JA, Pickup BT (1995) J Phys Chem 99:3503–3510. doi:10.1021/j100011a016
46. Haigh JA, Pickup BT, Grant JA, Nicholls A (2005) J Chem Inf Model 45:673–684. doi:10.1021/ci049651v

47. Matzen L, Engesgaard A, Ebert B, Didriksen M, Frølund B, Krogsgaard-Larsen P, Jaroszewski JW (1997) J Med Chem 40:520–527. doi:10.1021/jm9607212
48. Herr RJ (2002) Bioorg Med Chem 10:3379–3393. doi:10.1016/S0968-0896(02)00239-0
49. Bell PH, Roblin RO (1942) J Am Chem Soc 64:2905–2917. doi:10.1021/ja01264a055
50. Lipinski CA, Fiese EF, Korst RJ (1991) Quant Struct Act Relat 10:109–117. doi:10.1002/qsar.19910100205
51. ROCS OpenEye Scientific Software, Santa Fe, New Mexico, USA http://www.eyesopen.com/products/applications/rocs.html