A Practical Method for Targeted Library Design Balancing Lead-like Properties with Diversity

Michael J. Stocks,*[a] Gareth R. H. Wilden,[a] Garry Pairaudeau,[b] Matthew W. D. Perry,[a] John Steele,[b] and Jeffrey P. Stonehouse[a]

A practical and pragmatic method is demonstrated that aligns lead-like properties with compound diversity for the picking of compounds to synthesise from large virtual libraries. Methods are highlighted for decreasing synthetic attrition through the prior filtration of reagents sets grouped by reaction type. Also disclosed are protocols that use a combination of predicted physicochemical parameters and potential toxicological liabilities to enable the synthesis of lead-like compounds with a low potential risk of exhibiting toxicity or undesirable physicochemical properties. Lastly, a compound-picking process for a 2D compound matrix is demonstrated that maximises the diversity coverage whilst minimising synthetic effort. Thus a very highly optimised process is shown that delivers premium sample quality where lead-likeness and novelty are aligned to afford the best possible enhancement for the corporate compound collection.

Introduction

High-throughput screening (HTS)[1] is still seen by many in the pharmaceutical industry as the optimal way to discover new chemical hits for both new and previously explored pharmaceutical targets. A fundamental concern in this lead-generation approach is the high initial hit rates that are often observed, resulting from many compounds that act through undesired mechanisms. These include compounds that are either nonselective,[2] reactive,[3] or that interfere with the assay technology through either poor solubility[4] or aggregation.[5] These compounds can often mask the true actives of the screen and should therefore be removed from the screening collection.[6] The addition of new, high-value compounds that do not possess such liabilities should be a high priority for pharmaceutical companies involved in lead generation using HTS techniques.[7]

Observations from the analysis of early HTS campaigns resulted in the concept of lead-like compounds.[8] Current reviews[9] as well as our own in-house analysis have shown that many early, unfiltered libraries were biased towards compounds of high molecular weight, with many examples being synthesised that lie outside Lipinski's criteria for oral absorption,[10] and that contain functional groups with metabolic and chemical instability issues. A recent analysis of marketed oral drugs[11] has shown that the majority of oral drugs are within very well-defined ranges of molecular weight (340–360 Da on average) and lipophilicity (Clog$P$ 2.3–2.7).[12] Intuitively, lead-like libraries should therefore have their own well-defined physical properties, as it is generally accepted that during the lead-optimisation phase of a project, the molecular weight of the initial lead is seldom decreased (Figure 1). As a consequence, hard cutoffs based on molecular weight (450 Da, with 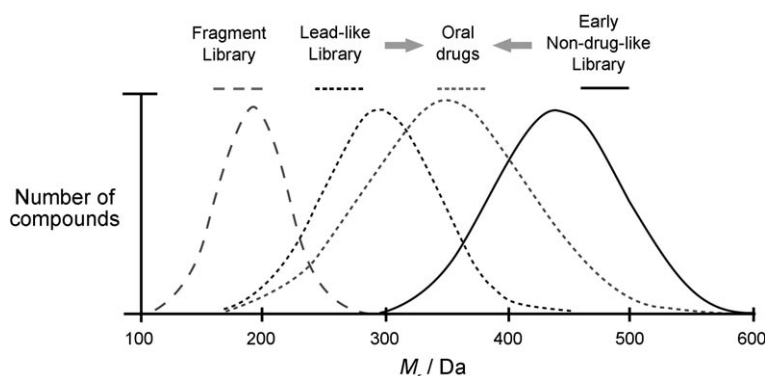90% of the library being < 400 Da) and calculated lipophilicity (Clog$P$ < 4 and > −1) have been applied to our lead-generation compound libraries.

Figure 1. The evolution of the lead-like library based on molecular weight consideration. Also shown is a molecular weight profile of a fragment library as well as the profile of early non-drug-like libraries, the optimisation of which to fit the molecular weight profile of oral drugs remains challenging.

Results and Discussion

Over the past several years, AstraZeneca has been involved in a company collection enhancement (CCE) programme with the aim to increase both the size, and more importantly, the quali-

[a] Dr. M. J. Stocks, Dr. G. R. H. Wilden, Dr. M. W. D. Perry, Dr. J. P. Stonehouse
Department of Medicinal Chemistry, AstraZeneca R&D Charnwood
Bakewell Road, Loughborough, Leics, LE11 5RH (UK)
Fax: (+ 44) 1509 645571
E-mail: mike.stocks@astrazeneca.com

[b] Dr. G. Pairaudeau, Dr. J. Steele
Department of Medicinal Chemistry, AstraZeneca R&D Mölndal
Pepparedsleden 1, 43183 Mölndal (Sweden)

 WILEY InterScience® DISCOVER SOMETHING GREAT

1

These are not the final page numbers! ↗↗

ty of the corporate compound collection to serve as a bank for HTS. Key to this process has been the establishment of definitions for our lead-like profile and the development of robust library design tools to enable an optimal selection of compounds for synthesis from large virtual libraries. The selected targets are chosen to exclude unwanted functionality and structural features whilst maintaining core and lead-like physical properties for all molecules synthesised.

There are many reports that deal with the design and creation of entire compound collections,[13] but that is beyond the scope of this article. Throughout the remainder of this paper, our concepts and ideas are illustrated with an actual example from our recent library work to show the progress of a single library for the collection through all the stages, from initial idea generation to commitment to library synthesis. Paramount to our compound diversity thinking is the design and synthesis of novel drug-like scaffolds as well as the establishment of new multi-component reactions to enable library synthesis of lead-like compounds from readily available reagents.

Analysis of ideas for novelty

The incorporation of novel chemical structures into the corporate compound bank is one of the main driving forces behind this initiative. Primary sources of this novelty come from a combination of identification of new commercial building blocks, ideas presented in the literature, and chemists' intuition.[14]

There have been many recent examples of the incorporation of spirocyclic bis-amine ring systems into novel pharmaceutical compounds.[15] The benefits of these relatively rigid, low-molecular-weight scaffolds include the orientation of substituents attached at the two nitrogen atoms in a well-defined three-dimensional manner. A virtual enumeration for all possible spirocyclic ring systems was employed in which the total number of heavy atoms (limited to carbon and two nitrogen atoms) was chosen to be >7 but <11, with a related imposed constraint that the minimum ring size was to be four atoms and the maximum ring size was to be six atoms. This resulted in 21 possible spirocyclic ring systems (Figure 2) after the removal of potentially unstable aminals.

A search was run in which the frequency of these substructures was examined in published patents, limiting the search to exemplified compounds (Figure 3). Interestingly, the spirocyclic bis-amines that are well exemplified within patents are those that show availability from commercial suppliers; however, the potentially more interesting examples are most probably less represented due to their perceived synthetic complexity and their lack of "on the shelf" availability.

From the analysis it became apparent that the spirocyclic azetidines were the least represented class of compounds, and so we focused our synthetic effort within this area. We chose compound 20 as our key target based on a combination of its potential to be a piperazine mimic,[16] its achirality, and relative scarcity in published examples.

Our first synthesis of the spirocyclic bis-azetidine ring system 20 has already been described.[17] For library synthesis, we
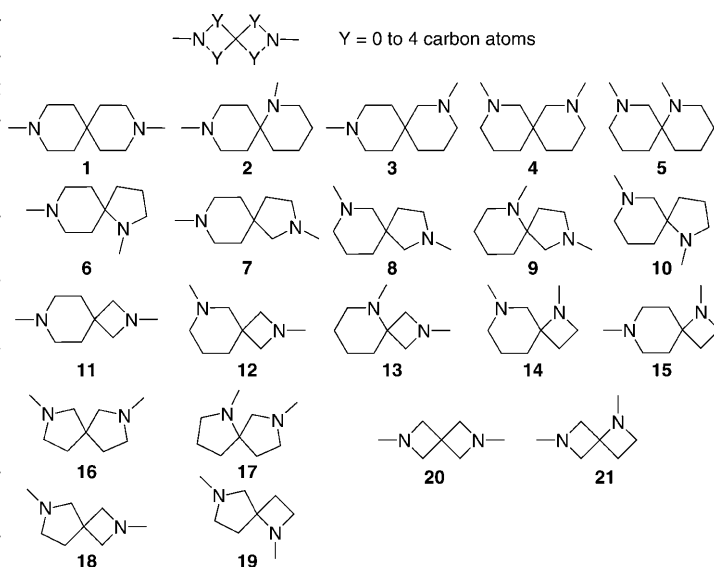

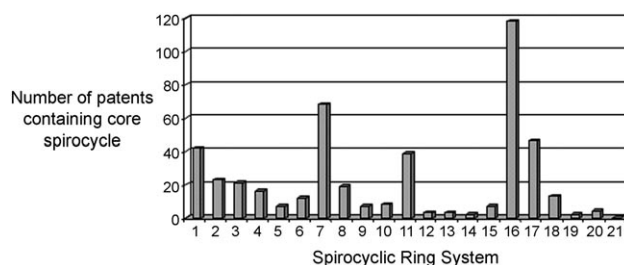**Figure 2.** Spirocyclic diamines considered as potential bifunctional scaffolds.

Figure 3. Frequency within patents of various spirocyclic bis-amines.

required a differentially protected bis-azetidine in multi-gram quantities. We developed a large-scale synthesis by modification of our earlier procedure (Scheme 1).

The azetidine chloroester 1[18] underwent hydrogenolysis in the presence of di-tert-butyldicarbonate to give the Boc-protected azetidine 2. Selective reduction with sodium borohydride/calcium chloride[19] gave the alcohol 3, which was oxidised under Swern[20] conditions to the aldehyde 4. Imine
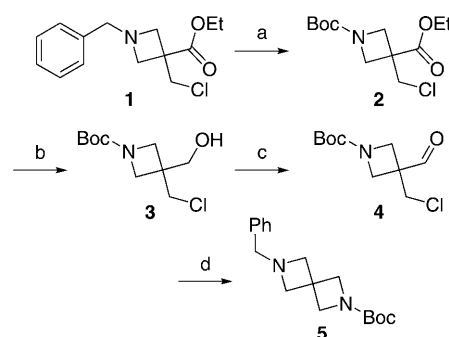

Scheme 1. Synthesis of the differentially protected bis-azetidine. Reagents and conditions: a) 10 % Pd/C, EtOAc, Boc₂O; b) CaCl₂, NaBH₄, EtOH, 20 °C; c) (COCl)₂, Et₃N, DMSO, CH₂Cl₂, −78 °C→RT; d) 1. PhCH₂NH₂, toluene/MeOH, 2. NaBH₄, 3. DMF(aq), 110 °C.

↖↖ These are not the final page numbers!

formation with benzylamine followed by reduction gave the amine, which was cyclised by heating to form the desired selectively diprotected compound **5**.

Reagent filtering

Having established the synthesis of the scaffold, a supply of reagents is required to decorate the scaffold to produce the library. While theoretically enumerating a library with all conceivable reagents and then selecting from the products is the ideal situation, this is impractical for production, where there are constraints on the number and availability of reagents. For this reason we use the AstraZeneca in-house reagent inventory and a database of available reagents from major commercial suppliers (ACD). We mine and process the databases in a regular and automated fashion.

In processing the reagents, the first criteria used are availability and price. Subsequent to this, the reagents are separated out by reagent class, and then further subdivided according to reaction type. In this way all reagents are filtered to remove undesirable chemical features as well as those that are incompatible to the proposed chemical transformation. To remove undesirable chemical features, physical property filters are applied, and the reagents are enumerated by a simple capping reaction (e.g. methylation or acetylation). The resulting structures are filtered through the corporate standard compound filters (which reject certain additional reactive functionalities, for example, aldehydes from amide coupling reactions) to remove unsuitable or toxicologically active groups, before the original reagent is regenerated (Scheme 2).
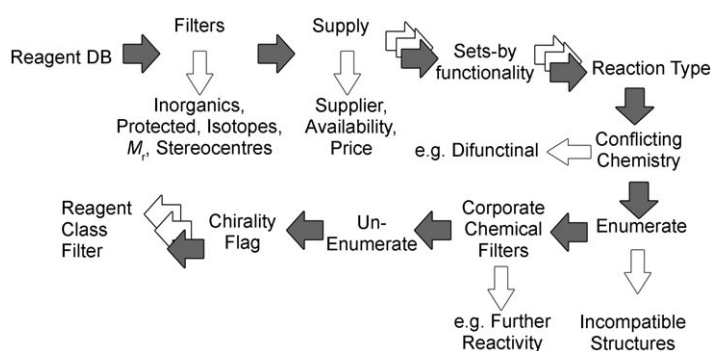
Scheme 2. Schematic representing reagent filtering by reaction type.

The output at this stage may be further filtered before saving, for example, to separate out primary and secondary amines or chiral reagents. Each of these files thus gives the chemist an excellent starting point from which to build chemical libraries, and removes much of the need to track back to reagent data once the final library has been produced (Table 1).

Table 1. Filtering the reagent database into core sets of reagents based on a selection of core reaction types.

Reactant Type	Reaction Class	Subsets Created
Alkyl Amines	Reductive Amination Amide Coupling Sulfonamide Formation Alkylation SnAr Displacement	Primary/Secondary/All
Aryl Amines	Sulfonamide Formation Amide Formation Buchwald Chemistry	Primary/Secondary/All
Acids	Acyl Sulfonamides Amide Formation	
Sulfonyl Chlorides	Sulfonamide Formation	
Alkylating Agents	Alkylation Epoxide Ring Opening	Reactive Heterocycles Alkylating Agents
Alcohols and Thiols	Alkylation SnAr Displacement	Alcohols/Thiols
Aldehydes	Reductive Amination	
Ketones	Reductive Amination	Chiral Cyclic/Acyclic Symmetrical

Compound filtering

Once a complete library is enumerated,[21] another extensive array of filters are applied to the resulting structures in two separate processes. The first of these takes the set of compounds that would never be produced, and annotates and discards these from the process. The second stage analyses the remaining compounds for those most desirable to make and presents that selection to the chemist. In a normal reaction, there is also a pool of compounds available that, while perfectly acceptable to the chemist, are not the ideal selected set. These can be mined later, should changes to the reagent set become necessary.

The first stage of this process is to apply a series of filters to the compounds based on calculated and predicted physical properties. Compounds are passed through all the filters in parallel, and the final results are combined. While it may be computationally more expedient to run the quicker filters first, removing most compounds before the more computationally expensive runs are conducted, this technique presents the end user with a list of all of the failings of a particular compound, and should it later become necessary to re-include some compounds into the design from the filters, then all the data have been calculated with which to make that decision. The exact nature of the filters run can also be fine-tuned as part of the process to adapt to changes in design. The first stage of this process is to generate the canonical tautomer of the compound through the company's in-house standardisation rules. After this, the filters applied include simple physical property filters; a molecular weight of 450 Da is the maximum tolerated,

These are not the final page numbers! ↗↗

along with a calculated log *P* range of −1.0 to 4.0. Compounds are subjected to an exact match search against the corporate compound collection, and any matching compounds from the virtual library are flagged and removed. A cluster analysis (using our in-house-developed fingerprint tools, with a Tanimoto radius of 0.2) is carried out on all structures in the current screening collection, and virtual compounds with any near neighbours to those already represented are removed from the library at this stage. If a library is to be targeted at a specific bioactive molecule, then the Tanimoto clustering can be relaxed to increase the number of close analogues to be synthesised within the defined cluster. The most complex calculations carried out on the compounds are profiling for predicted solubility, genotoxicity (against an AMES dataset), suspected hERG activity, and reactive metabolite formation liability (Figure 4). If a compound is predicted to have liabilities in
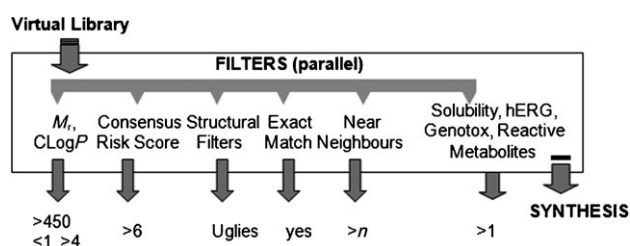
Figure 4. Flowchart that illustrates the process for evaluating and picking compounds. The number of near neighbours (*n*) can be varied to reflect library type (i.e. diversity or hit exploitation).

more than one of the categories of solubility, hERG activity, or genotoxicity, then it is removed from further consideration.

Alongside this filtering method, a ranking process is also carried out. While a hard cutoff value in a physical property descriptor is necessary, it is, by definition, very arbitrary in nature. Where there are several steps, compounds with values just over the threshold for the previous criterion are treated equally with compounds just under the threshold for the following criteria—and this may not always lead to a good property distribution. For this reason, a sliding scale is applied to all numerical descriptors computed for the compounds, including polar surface area and counts of donors and acceptors that are not used in the first-stage filtering process. The relative weight of each descriptor can also be varied, and a summation of these values is taken as a "consensus risk" for each compound in the library. In this analysis, the properties of the virtual compounds are calculated, and a risk score is attached to each parameter. For example, if a compound has up to three hydrogen bond donors, then we see little risk within the compound (risk score = 0); however, if it has more than five hydrogen bond donors, then we deem this parameter as being potentially high risk (risk score = 3). The consensus risk score of each compound is then calculated taking into account all the calculated physical parameters shown (Figure 5). In this way, no compounds are synthesised which have a consensus risk score of > 6, and as a consequence, we believe that we maintain a very
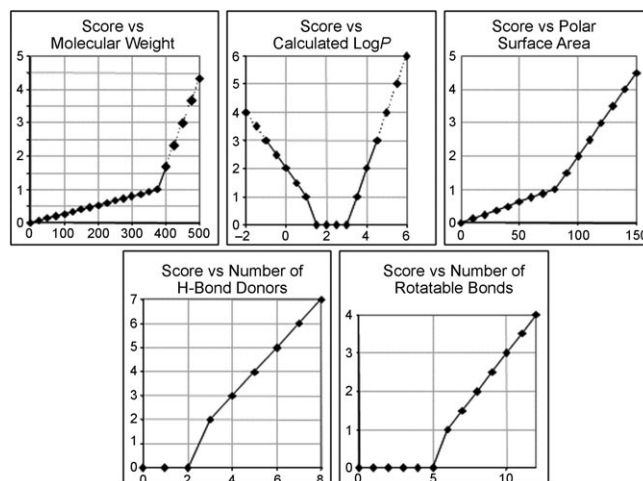
Figure 5. Consensus risk analysis for virtual compounds within the library. Consensus risk score is the sum of the individual elements scored according to the above criteria.

high level in the quality of the compounds selected for synthesis that may not have been present with hard cutoff values alone.

Should a library be targeting a compound with known toxicology issues, such as cytochromes P450, then additional filters can be added to the process at this stage. Compounds that pass all of these filters are saved into the pool from which the actual library can be picked. A visual report is given to the chemist at this stage, profiling the distribution of all of the calculated properties both inside the library and for those compounds that failed, along with the reasons for failure. This information has proved invaluable in selecting the next round of building blocks to complete the library design.

For the compound library based around the spirocyclic *bis*-azetidine (Scheme 3), the N−H azetidine **6** could be allowed to
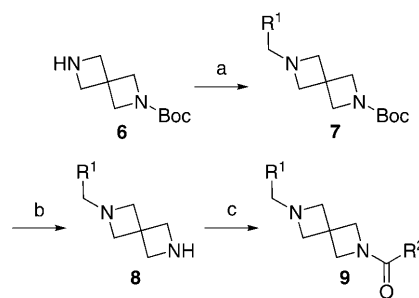
Scheme 3. Synthetic proposal for library based on the *bis*-azetidine scaffold. *Reagents and conditions*: a) R^1CHO, $NaBH(OCOCH_3)_3$, NMP; b) TFA, CH_2Cl_2, RT; c) R^2CO_2H, HATU, NMP, Et_3N.

react with a range of aldehydes to afford *N*-alkylated azetidines **7**. Deprotection followed by acylation with an array of carboxylic acids gave a library of *bis*-azetidines **9**.

From the AstraZeneca Charnwood in-house reagent stock, there were 3681 carboxylic acids and 768 aldehydes available. These were further filtered by the methods described to 592

carboxylic acids and 204 aldehydes—reagents in which we would have a high degree of confidence for the delivery of final compounds. Further refinement of these sets to include only achiral reagents gave us final enumeration sets of 499 carboxylic acids and 201 aldehydes (a virtual library of 100 299 compounds). A total of 64 652 compounds from the virtual library passed through the initial enumeration, based on $M_r <$ 450 Da and Clog$P < 4.5$ cutoffs. To increase compound quality, these properties were further constrained to only synthesise compounds with $M_r < 400$ and Clog$P < 4$ (Figure 6). This resulted in 12 837 possible compounds worthy of synthesis by passing all filters.
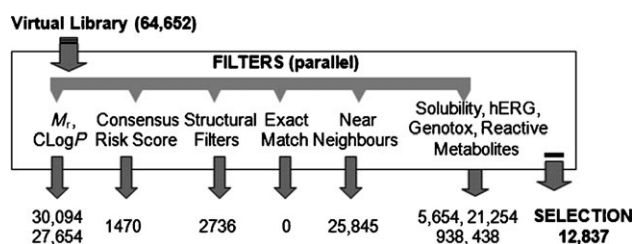

Figure 6. Evaluation of the virtual library, showing 12 837 compounds passing all the generic filters and hard M_r and Clog$P$ cutoffs. The numbers refer to the number of compounds failing each of the filters in parallel.

Whilst the spirocyclic *bis*-azetidine may well function as a piperazine mimic, a potential liability of this substitution is the observed increase in basicity between the piperazine and *bis*-azetidine analogues (Figure 7). This increase may well lead to
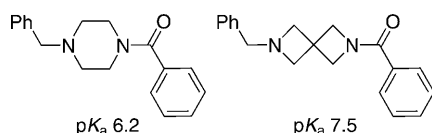

Figure 7. Measured pK_a values of represented examples of the piperazine and *bis*-azetidine analogues.

an increase in hERG activity relative to the corresponding piperazine analogues, and indeed ~33% of the virtual library was predicted to have a high probability of hERG activity. Whilst this predicted toxicological issue was a major reason for attrition in the filtering process, removal of these compounds can only enhance the overall quality of the eventual library committed to synthesis.

The next stage of the process is to determine the minimum number of compounds required to give maximum coverage of possible diversity space for the library.

Selection of compounds from the library

Diversity is probably the most overused and misunderstood tool in the gamut of computational chemistry. While "diversity" methodology is applied to compound selection, it is important to remember that nothing can be diverse in itself, but needs a

point of reference against which to be considered. A far better term to use in this context is that of coverage. The purpose of this chemical initiative is to create compounds to enrich the compound bank and ultimately provide a greater chance of conducting a successful HTS campaign against an emerging target, and therefore this must be the yardstick against which the composition of each library must be judged. As the original design will have been biased towards a target or chemistry of interest, and all overlap with the existing compound bank has already been removed, we need only care about selection from inside the library itself. For the greatest chance of finding and prosecuting an early-stage HTS hit, a cluster of ~10–15 compounds has been shown to have the best probability of success.[22] Paradoxically, it has also been shown that nearest-neighbour compounds are often inactive. There is a clear need to balance these two criteria in any library selection. The use of an in-house fingerprint to generate these clusters has led to a replication of this study using internal data. A factor that quickly became apparent is that if a whole-molecule pharmacophore-based descriptor is used (instead of an atomistic/connectivity-based descriptor), then the clusters that form in structural space and pharmacophore space bear little resemblance to one another. As a result of this work, the method chosen to sample compounds was to annotate the compounds in the library according to the clusters of both these descriptors. Singleton clusters are given extra leeway and, where possible, are assigned to the nearest cluster within a suitable extended Tanimoto radius of their location (Figure 8).
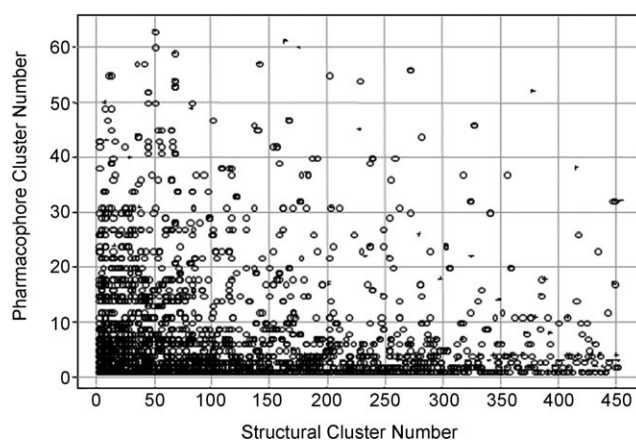

Figure 8. Cluster analysis of 12 837 compounds, showing excessive population in certain clusters. Compounds depicted as "+" are cluster centres in either metric. The pharmacophoric cluster number relates to those compounds clustered using the pharmacophore-based descriptor, whereas the structural cluster number relates to those clustered using the structural fingerprint. The total number of virtual compounds selected is the combination from these two methods, thus we fully cover both pharmacophoric and structural fingerprints within the design.

A selection from the library should cover as much of the cluster space in both pharmacophoric and structural space. The first round of the selection should identify and select all of the largest cluster centres for the synthesis set. Reagents are then selected (and thus compounds) and added to the selec-

tion to complete the best coverage of the clusters in the pool of compounds. Where multiple choices in each cluster are possible, then the risk score of the compounds is consulted, and the lowest-scoring compound (that with the least risk) is then selected. This process is repeated until the desired number of compounds has been selected. Use of this mechanism to select compounds not only ensures that the "diversity" space of a library is covered, but that the cluster membership of compounds for hit finding is also enhanced, and that the least risky compounds overall have been selected from the pool (Figure 9). Genetic algorithms[23] and Pareto optimisation have
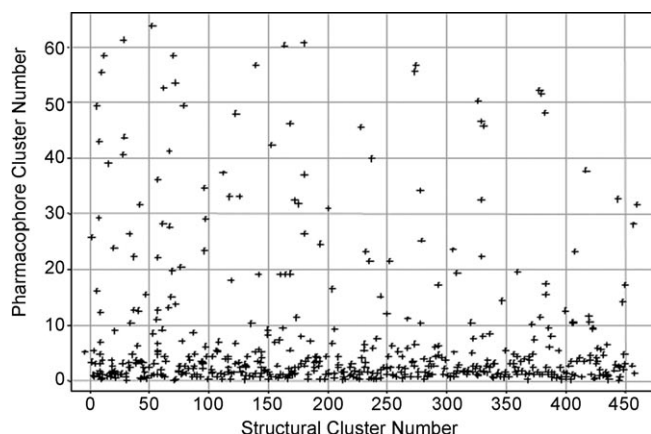
Figure 9. 518 Selected compounds from the library to show compound coverage in all clusters. Only the compounds from the cluster centres are chosen for synthesis.

been applied to this task, and perform well where multiple competing constraints in reagent space make the choice complex. In all other situations a python script performs the cluster-based selection.

To obtain maximal diversity coverage based on a combination of pharmacophore and structural fingerprint clustering (Tanimoto distance 0.2), the above analysis shows that we would need to synthesise only 4% (518 compounds) of the possible compounds from the virtual library (12837 compounds) that fulfilled all of our lead-like criteria based on a combination of hard physicochemical cutoffs ($M_r < 400$ Da, $\text{Clog} P < 4$) and filtering based on predicted toxicological and physicochemical liabilities.

Selection of the cluster centres may seem to give the best representation of the chemical space covered by the library, but there are several drawbacks to using this method. The first of these is that all of the clusters, including the singleton clusters, are sampled. Although this does cover the maximum of diversity space, it has been shown that singletons are not ideal for ensuring that compound classes are recognised in a high-throughput assay. For this, larger coverage of each cluster is needed. The second drawback is that a library of this size would need to be produced in a combinatorial fashion, and the reagents needed to synthesise this selection is $165 R^1 \times 272 R^2$—the resulting matrix is sparse indeed: 518 compounds from a possible 44880, a 1% reagent density from the square

case. A python script has been developed to perform a constrained selection from this matrix. This script takes the reagents used to construct the cluster centre of the largest cluster in each dimension as a starting point. Reagents are then added to the selection in turn according to which gives the largest enhancement of cluster and compound coverage at each stage, with the lowest-scoring products chosen in the event of a tie. The resulting selection is shown in Figure 10. In
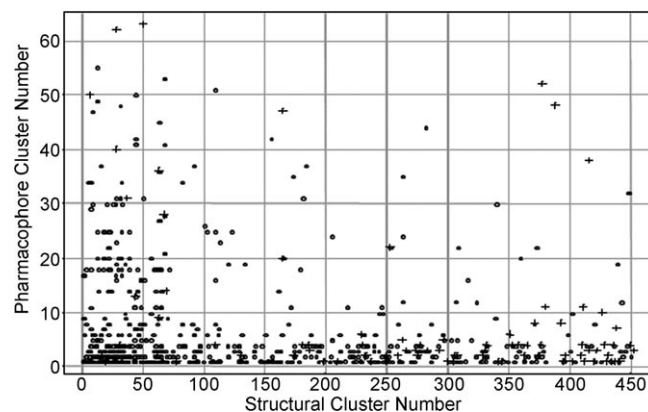
Figure 10. Final selection from the virtual library showing the final 1009 compounds selected for synthesis.

producing this result, constraints of 50 reagents in each dimension and ~1000 overall compounds were specified: the final selection is $50 R^1 \times 47 R^2$, giving 1009 compounds from a possible 2350 (a 43% reagent density). Although almost doubling the overall number of compounds needed to cover the library space, this results in a fivefold decrease (97 vs. 437) in the number of reagents needed to produce the final library. This library covers 62% of (non-singleton) structural cluster space, though it represents 84% of the compounds; it covers 81% of (non-singleton) pharmacophoric cluster space, while representing 99% of the compounds. Overall cluster coverage is 65%. In pragmatic terms, this is considered to be an excellent compromise in covering the space available to the library within a readily achievable synthetic set.

Conclusions

A practical and pragmatic method has been illustrated that aligns lead-like properties with compound diversity for large virtual libraries. We have illustrated the processes established within AstraZeneca to select compounds for synthesis based on a combination of predicted physicochemical and toxicological liabilities. Having filtered reagents sets available based on reaction class has also greatly improved the efficiency of the process by pre-filtering those reagents with incompatible functional groups for a particular reaction class as well as those reagents that contain functional groups that derive non-lead-like products. Finally, from the example shown, a potential virtual library of 100299 compounds was sequentially filtered to remove compounds based on hard physicochemical cutoffs

↖↖ **These are not the final page numbers!**

Table 2. Pros (+) and cons (−) for each compound-picking method.		
Criterion	Picking Method 1 [165 R^1×272 R^2 → 518 compound library]	Picking Method 2 [50 R^1×47 R^2 → 1009 compound library]
Cost of synthesis	+ Lower cost for reagents and fewer purification steps required; − More reactants required	− Extra cost for synthesis, or expense at external company to purchase compounds; + Fewer reactants required
Number of compounds delivered	+ Increased diversity in fewer compounds	+ More close analogues to establish SAR
Number of compounds in collection	+ Fewer examples for HTS screening, so can populate HTS with more diverse compounds	+ Greater coverage for sub-setting HTS
Predicted success rate	− Difficult to validate due to sparse 2D matrix	+ Increased chemical validation owing to fewer reagents used
Ease of synthesis	− Cherry-picking required	+ More robust owing to fewer reagents
Suitability for one-step production	− More intermediates to synthesise	+ Fewer intermediates to synthesise
Suitability for two-step production	+ Fewer reactions required	− More reactions required

and predicted liabilities to afford a core set of 12 837 (12.8%) possible compounds for synthesis. Diversity analysis, based on a combination of pharmacophore and structural fingerprints afforded a final design size of 518 compounds (0.5% from original design, or 4% from the filtered design set) that covered complete diversity space for the entire virtual library. In practice, this approach leads to a very sparse 2D matrix that will have severe practical consequences when progressed into library production due to many reagents with multiple functional groups present. We have further shown, by constraining the 2D matrix with an in-house developed method, that we are able to select a more densely populated 2D matrix that delivers excellent coverage of both the pharmacophore and structural space for the whole virtual library.

It is important to consider the advantage of each of the two available picking processes that we have highlighted on a cost–benefit model (Table 2). Whilst there are pros and cons to each method, particularly dependent on the suitability of the chemistry for two-step productions or otherwise, we believe it is up to the chemist to determine, after careful consideration, which method would best suit their immediate plans for any given library.

The benefits of the approaches highlighted in this disclosure are: the high expected delivery of final products due to the initial reagent filtering; only compounds that fulfill AstraZeneca core lead-like criteria will be submitted for synthesis, and full coverage of pharmacophore and structural diversity will be accounted for in a relatively small sampling from the large virtual library. Thus, we have shown a very highly optimised process that delivers premium sample quality where lead-likeness and novelty are aligned, affording the best possible enhancement for the compound collection constrained within a pragmatically designed 2D matrix of library compounds. Finally, even though we have demonstrated our approaches within the constrains of a 2D chemical matrix, the processes have been readily adapted to cover both 1D and 3D compound libraries with additional multiple scaffolds.

Experimental Section

All reagents and solvents were obtained from commercial suppliers and used without further purification. All reactions were carried out under an inert N_2 atmosphere unless otherwise noted. Silica gel chromatography was performed using either glass columns packed with silica gel (200–400 mesh, Aldrich Chemical) or pre-packed silica gel cartridges (Biotage and ISCO). ^{1}H and ^{13}C NMR spectra were obtained from either a 300 MHz (^{1}H: 300 MHz, ^{13}C: 75 MHz) or 400 MHz (^{1}H: 400 MHz, ^{13}C: 100 MHz) Varian Unity Inova spectrometer. ^{1}H and ^{13}C NMR shifts (δ) are given in ppm and are measured relative to Me$_4$Si as standard. All final compounds were purified to >95% purity as determined by LC–MS analysis performed on an Agilent 1100 instrument; low-resolution MS data were collected on a Hewlett–Packard HP 6890 or HP 5973 MSD mass spectrometer using ESI modes (positive or negative).

1-*tert*-Butyl-3-ethyl-3-(chloromethyl)azetidine-1,3-dicarboxylate (2). *tert*-Butoxycarbonyl (Boc) anhydride (20.03 g, 85.41 mmol), ethyl-1-benzyl-3-(chloromethyl)azetidine-3-carboxylate (22 g, 81.34 mmol), and palladium hydroxide on carbon (5.71 g, 2.03 mmol) were combined in EtOAc (450 mL), and the resulting suspension was hydrogenated at 5 atm for 24 h. The reaction mixture was filtered, washed through with EtOAc, and the solvents

These are not the final page numbers! ↗↗

were removed under reduced pressure to afford 1-*tert*-butyl-3-ethyl-3-(chloromethyl)azetidine-1,3-dicarboxylate (22.59 g, 100%) as a clear, colourless oil. GC–MS: purity = 91%, base peak = 57, top ion = 222; ^{1}H NMR (400 MHz, CDCl$_3$): δ = 4.26 (q, J = 7.1 Hz, 2 H), 4.21 (d, J = 9.2 Hz, 2 H), 3.92 (s, 2 H), 3.84 (d, J = 9.0 Hz, 2 H), 1.45 (s, 9 H), 1.31 ppm (t, J = 7.0 Hz, 3 H); ^{13}C NMR (100 MHz, CDCl$_3$): δ = 170.99, 156.04, 80.15, 61.90, 47.06, 44.11, 28.27, 14.06 ppm.

tert-Butyl-3-(chloromethyl)-3-(hydroxymethyl)azetidine-1-carboxylate (3). Anhydrous powdered CaCl$_2$ (15.57 g, 140.27 mmol) was added to a stirred solution of 1-*tert*-butyl-3-ethyl 3-(chloromethyl)azetidine-1,3-dicarboxylate (19.48 g, 70.14 mmol) in specially dried EtOH (700 mL). After 15 min NaBH$_4$ (10.61 g, 280.55 mmol) was added, and the reaction was stirred for a further 2 h. TLC indicated reaction completion, and K$_2$CO$_3$(aq) (2 M, 100 mL) was added, and the mixture was stirred for a further 2 h. The EtOH was removed under reduced pressure, and the aqueous suspension was extracted with EtOAc. The organic phase was dried over MgSO$_4$, filtered, and evaporated to afford crude *tert*-butyl-3-(chloromethyl)-3-(hydroxymethyl)azetidine-1-carboxylate (15.65 g, 66.40 mmol, 95%) as a white solid. GC–MS: purity = 97.1%, base peak = 57, top ion = 235; ^{1}H NMR (400 MHz, CDCl$_3$): δ = 3.85 (s, 2 H), 3.79 (s, 2 H), 3.72 (app q, J = 7.9 Hz, 4 H), 1.85 (s, 1 H), 1.44 ppm (s, 9 H); ^{13}C NMR (100 MHz, CDCl$_3$): δ = 156.40, 79.81, 64.04, 47.18, 39.98, 28.32 ppm.

tert-Butyl-3-(chloromethyl)-3-formylazetidine-1-carboxylate (4). (COCl)$_2$ (4.64 mL, 53.03 mmol) was added to a solution of DMSO (4.52 mL, 63.64 mmol) under N$_2$ in CH$_2$Cl$_2$ (125 mL) in a three-necked flask at −70 °C over 5 min, and the mixture stirred for 45 min with venting. A solution of *tert*-butyl-3-(chloromethyl)-3-(hydroxymethyl)azetidine-1-carboxylate (5 g, 21.21 mmol) in CH$_2$Cl$_2$ (60 mL) was added over 15 min, and the resultant solution was stirred for a further 30 min at −70 °C prior to the addition of Et$_3$N (11.83 mL, 84.85 mmol) dropwise over 15 min, maintaining a temperature below −60 °C. After the addition of Et$_3$N, the reaction was stirred at −70 °C for 30 min and then allowed to warm to ambient temperature over 2.5 h to afford a pale, creamy yellow suspension. The reaction was quenched with NaOH (3.75 N, 30 mL) and diluted further with saturated NaHCO$_3$(aq) (100 mL). The organic layer was collected, and the aqueous layer was re-extracted further with CH$_2$Cl$_2$ (2×20 mL). The organic extracts were combined and washed with 5% KH$_2$PO$_4$(aq) (2×50 mL) and brine (1×30 mL) and dried over Na$_2$SO$_4$. The solvents were removed under reduced pressure to afford a yellow oil that was purified by flash silica chromatography (elution gradient: 10→50% EtOAc in isohexane). Pure fractions were evaporated to afford *tert*-butyl-3-(chloromethyl)-3-formylazetidine-1-carboxylate (4.37 g, 88%) as a colourless oil that was used immediately in the following reaction. ^{1}H NMR (400 MHz, CDCl$_3$): δ = 9.82 (s, 1 H), 4.13 (d, J = 9.1 Hz, 2 H), 3.95 (s, 2 H), 3.86 (d, J = 9.1 Hz, 2 H), 1.45 ppm (s, 9 H).

tert-Butyl-6-benzyl-2,6-diazaspiro[3.3]heptane-2-carboxylate (5). Benzylamine (18.32 mL, 167.74 mmol) and *tert*-butyl-3-(chloromethyl)-3-formylazetidine-1-carboxylate (39.2 g, 167.74 mmol) were combined in toluene/MeOH (3:1, 150 mL) and stirred for 20 min at 40 °C. The solvents were removed under reduced pressure, and the resulting oil was evaporated with the same solvent mixture (3×100 mL). The imine was dissolved in MeOH (300 mL), cooled to 0 °C, and NaBH$_4$ (7.77 g, 201.29 mmol) was added. The reaction was stirred overnight and then quenched with aq NH$_3$ (d = 0.880 kg L^{-1}) and stirred for a further 2 h. Solvents were removed in vacuo, and the resultant white solid was partitioned between EtOAc (600 mL) and H$_2$O (100 mL). The organic layer was isolated and washed with 2 M K$_2$CO$_3$(aq) (2×50 mL), brine (1×50 mL), dried over Na$_2$SO$_4$, filtered, and the solvents were removed under re-

duced pressure. The residue was dissolved in *N,N*-dimethylformamide (DMF; 250 mL) and H$_2$O (62.5 mL), and the mixture was heated at 95 °C for 4 h. After this time, further H$_2$O (60 mL) was added, and the heating continued for a further 5 h. Further H$_2$O (25 mL) was added, and heating continued for a further 4 h, after which time the reaction was cooled slightly and then solvents removed under reduced pressure to afford a creamy white solid. The product was partitioned between Et$_2$O (500 mL) and 2 M K$_2$CO$_3$(aq) (100 mL). The organic phase was separated and washed with further 2 M K$_2$CO$_3$(aq) (2×50 mL), H$_2$O (4×20 mL), brine (2×50 mL), and dried over Na$_2$SO$_4$. The suspension was filtered, and solvents were removed under reduced pressure to afford *tert*-butyl-6-benzyl-2,6-diazaspiro[3.3]heptane-2-carboxylate (47.7 g, 97%) as a pale yellow oil. HPLC–MS: purity = 99.68% at 220 nm, [M+H]$^+$ m/z: 289 (calcd = 289.1916) (MultiMode$^+$); ^{1}H NMR (300 MHz, CDCl$_3$): δ = 7.38–7.19 (m, 5 H), 3.99 (s, 4 H), 3.56 (s, 2 H), 3.32 (s, 4 H), 1.42 ppm (s, 9 H); ^{13}C NMR (75 MHz, CDCl$_3$): δ = 156.09, 137.74, 128.43, 128.39, 127.15, 79.46, 64.26, 63.57, 33.44, 28.36 ppm.

[1] a) R. Macarrón, R. P. Hertzberg, *Methods in Molecular Biology, Vol. 190*, Humana Press, Totowa, **2002**, pp. 1–2; b) J. A. Landro, I. C. Taylor, W. G. Stirtan, D. G. Osterman, J. Kristie, E. J. Hunnicutt, P. M. Rae, P. M. Sweetnam, *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 273–289; c) P. Gribbon, A. Sewing, *Drug Discovery Today* **2005**, *10*, 17–22; d) D. A. Pereira, J. A. Williams, *Br. J. Pharmacol.* **2007**, *152*, 53–61; e) K. R. Oldenburg, *Annu. Rep. Med. Chem.* **1998**, *33*, 301–311.

[2] a) S. L. McGovern, E. Caselli, N. Grigorieff, B. K. A. Shoichet, *J. Med. Chem.* **2002**, *45*, 1712–1722; b) B. Y. Feng, A. Shelat, T. N. Doman, R. K. Guy, B. K. Shoichet, *Nat. Chem. Biol.* **2005**, *1*, 146–148; c) B. C. Pearce, M. J. Sofia, A. C. Good, D. M. Drexler, D. A. Stock, *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.

[3] G. M. Rishton, *Drug Discovery Today* **1997**, *2*, 382–384.

[4] C. A. Lipinski, *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.

[5] B. Y. Feng, A. Simeonov, A. Jadhav, K. Babaoglu, J. Inglese, B. K. Shoichet, C. P. Austin, *J. Med. Chem.* **2007**, *50*, 2385–2390.

[6] A. Tinker, N. Tomkinson, Third Joint Sheffield Conference on Chemoinformatics, April 21–23, **2004**.

[7] a) J. Provis, *J. Assoc. Lab. Autom.* **2005**, *10*, 124–129; b) A. M. Davis, D. J. Keeling, J. Steele, N. P. Tomkinson, A. C. Tinker, *Curr. Top. Med. Chem.* **2005**, *5*, 421–439.

[8] a) P. D. Leeson, B. Springthorpe, *Nat. Rev. Drug Discovery* **2007**, *6*, 881–890; b) T. I. Oprea, A. M. Davis, S. J. Teague, P. D. Leeson, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315; c) M. M. Hann, T. I. Oprea, *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–263; d) M. M. Hann, A. R. Leach, J. N. Burrows, E. Griffen, *Compr. Med. Chem. II* **2007**, *4*, 435–458.

[9] S. J. Teague, A. M. Davis, P. D. Leeson, T. Oprea, *Angew. Chem.* **1999**, *111*, 3962–3967; *Angew. Chem. Int. Ed.* **1999**, *38*, 3743–3748.

[10] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

[11] a) P. D. Leeson, A. M. Davis, J. Steele, *Drug Discovery Today Technol.* **2004**, *1*, 189–195; b) M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis, P. D. Leeson, *J. Med. Chem.* **2003**, *46*, 1250–1256; c) M. Vieth, M. G. Siegel, R. E. Higgs, I. A. Watson, D. H. Robertson, K. A. Savin, G. L. Durst, P. A. Hipkind, *J. Med. Chem.* **2004**, *47*, 224–232.

[12] Daylight Chemical Information Systems Inc., ClogP Version 4.71.

[13] a) I. Akritopoulou-Zanze, J. T. Metz, S. W. Djuric, *Drug Discovery Today* **2007**, *12*, 948–952; b) W. H. B. Sauer, M. K. Schwarz, *Chimia* **2003**, *57*, 276–283; c) W. H. B. Sauer, M. K. Schwarz, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987–1003; d) R. Nilakantan, F. Immermann, K. Haraki, *Comb. Chem. High Throughput Screening* **2002**, *5*, 105–110; e) G. Harper, S. D. Pickett, D. V. S. Green, *Comb. Chem. High Throughput Screening* **2004**, *7*, 63–71.

[14] C. G. Wermuth, *J. Med. Chem.* **2004**, *47*, 1303–1314.

[15] For some recent examples of incorporation of spirocyclic diamines, see: a) I. Shamovsky, S. Connolly, L. David, S. Ivanova, B. Norden, B. Spring-thorpe, K. Urbahns, *J. Med. Chem.* **2008**, *51*, 1162–1178; b) P. E. Finke, L. C. Meurer, B. Oates, S. K. Shah, J. L. Loebach, S. G. Mills, M. MacCoss, L. Castonguay, L. Malkowitz, M. S. Springer, S. L. Gould, J. A. DeMartino, *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2469–2473; c) M. S. Smyth, J. Rose, M. M. Mehrotra, J. Heath, G. Ruhter, T. Schotten, J. Seroogy, D. Volkots, A. Pandey, R. M. Scarborough, *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1289–1292; d) J. Jiang, M. Hoang, J. R. Young, D. Chaung, R. Eid, C. Turner, P. Lin, X. Tong, J. Wang, C. Tan, S. Feighner, O. Palyha, D. L. Hreniuk, J. Pan, A. W. Sailer, D. J. MacNeil, A. Howard, L. Shearman, S. Stribling, R. Cama-cho, A. Strack, L. H. T. Van der Ploeg, M. T. Goulet, R. J. DeVita, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5270–5274; e) T. P. Culbertson, J. P. Sanchez, L. Gambino, J. A. Sesnie, *J. Med. Chem.* **1990**, *33*, 2270–2275.

[16] For a recent discussion of the potential substitution for piperidine, see: J. Burkhard, E. M. Carreira, *Org. Lett.* **2008**, *10*, 3525–3526.

[17] D. Hamza, M. J. Stocks, A. Décor, G. Pairaudeau, J. P. Stonehouse, *Synlett* **2007**, 2584–2586.

[18] D. Bartholomew, M. J. Stocks, *Tetrahedron Lett.* **1991**, *32*, 4795–4798.

[19] Y. Tomio, M. Mitsuo, N. Hiroyuki, *Bull. Chem. Soc. Jpn.* **1991**, *64*, 2730–2734.

[20] A. J. Mancuso, D. Swern, *Synthesis* **1981**, 165–185.

[21] All compound libraries are enumerated with a set of protocols created by the "Pipeline Pilot" software tool and made available through various chemically aware applications (e.g. Excel, Spotfire), so that the individual chemist can access these from the software with which they are most familiar. The process described below is completed in discrete stages ("Protocols" within the software), and after each stage the results are saved to disk and available for inspection by the chemist. In this manner the entire library is available at any point during the process, and should changes to the design become necessary, all of the information is immediately available for reuse with differing criteria. Accelrys Software Inc., Pipeline Pilot, Version 6.1.

[22] a) Y. C. Martin, J. L. Kofron, L. M. Traphagen, *J. Med. Chem.* **2002**, *45*, 4350–4358; b) Y. C. Martin, R. D. Brown, M. G. Bures, *Comb. Chem. Mol. Diversity Drug Discovery* **1998**, 369–385; c) Y. C. Martin, M. G. Bures, R. D. Brown, *Pharm. Pharmacol. Commun.* **1998**, *4*, 147–152.

[23] G. R. H. Wilden, Fourth Joint Sheffield Conference on Chemoinformatics, June 18–20, **2007**.

FULL PAPERS

M. J. Stocks,* G. R. H. Wilden,
G. Pairaudeau, M. W. D. Perry, J. Steele,
J. P. Stonehouse

A Practical Method for Targeted Library Design Balancing Lead-like Properties with Diversity
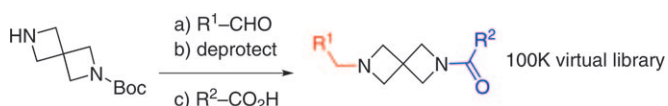
Choosing the right compounds to synthesise from large virtual combinatorial libraries is a current challenge for the pharmaceutical industry. Herein we describe a highly optimised method that aligns lead-like properties with compound diversity. The methods are illustrated by considering a two-dimensional library based on the interesting spirocyclic *bis*-azetidine template.