Designing Focused Libraries Using MoSELECT

Valerie J. Gillet[*a], Peter Willett[a], Peter J. Fleming[b] and Darren V.S. Green[c]

[a]*Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom*

[b]*Department of Automatic Control and Systems Engineering, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom*

[c]*GlaxoSmithKline, Gunnels Wood Road, Stevenage, SG1 2NY, United Kingdom*

When designing a combinatorial library it is usually desirable to optimise multiple properties of the library simultaneously, and often the properties are in competition with one another. For example, a library that is designed to be focused around a given target molecule should ideally have minimum cost and also contain molecules that are bioavailable. In this paper, we describe the program MoSELECT for multiobjective library design that is based on a multiobjective genetic algorithm (MOGA). MoSELECT searches the product-space of a virtual combinatorial library to generate a family of equivalent solutions where each solution represents a combinatorial subset of the virtual library optimised over multiple objectives. The family of solutions allows the relationships between the objectives to be explored and thus enables the library designer to make an informed choice on an appropriate compromise solution. Experiments are reported where MoSELECT has been applied to the design of various focused libraries.

Keywords: combinatorial library design; focused libraries; multiobjective optimisation; MOGA; MoSELECT

---

[*] Author to whom correspondence should be sent. E-mail: v.gillet@sheffield.ac.uk

INTRODUCTION

Initial efforts in combinatorial library design were directed towards the design of diverse libraries; [1,2] however, there is a growing interest in the design of focused libraries and in the design of libraries optimised on multiple properties simultaneously. [3-5] Diverse libraries are designed to give broad coverage of chemistry space and are useful for screening against a range of structural targets. Focused libraries, on the other hand, are constrained to occupy restricted regions of chemistry space with the boundaries being defined by what is known about the biological target of interest. For example, when the 3D structure of the target is known, a virtual library can be screened against the target to eliminate molecules that cannot fit into the active site. Alternatively, if an active compound is known, the library could be constrained to contain molecules that are similar to the known active or the library could be designed to contain molecules that are predicted to be active according to a QSAR model.

It is now recognised that, even in diverse libraries, it is important that the physicochemical properties of the libraries are also optimised in order that compounds contained within the library constitute good start points for further optimisation. It is even more desirable to optimise multiple properties in focused library design since in addition to matching constraints related to the target molecule, other criteria are often required during lead optimisation, for example, bioavailability and cost of goods.

We recently reported the development of MoSELECT[6,7] for multiobjective library design in product-space and described its application to the design of libraries on multiple competing objectives. Examples were given of libraries that are simultaneously diverse while having drug-like physicochemical property profiles.

Here we describe the application of MoSELECT to focused libraries that are optimised on multiple objectives simultaneously.

METHODS

MoSELECT is a recent development of the earlier SELECT program for product-based combinatorial library design.[8] SELECT is based on a genetic algorithm (GA) and incorporates multiple objectives via a weight-sum fitness function. SELECT was developed following a study of the effectiveness of reactant versus product-based methods for library design which showed that greater structural diversity can be achieved by analysing product space.[9] The results have been subsequently confirmed in other studies.[10,11] The objectives to be optimised by SELECT are normalised and relative weights are specified by the user at run time. Experiments have been reported where SELECT has been applied to the design of libraries that are simultaneously diverse and have drug-like physicochemical properties.[8] Several other GA based programs developed in computational chemistry also use a weighted-sum fitness function, e.g., for combinatorial library design[12-14] and ligand docking programs.[15] Despite the wide spread use of the weighted-sum approach, there are several limitations associated with it as an approach to multiobjective optimisation. For example, setting appropriate weights can be a difficult task often requiring several trial and error experiments[16] and the weights chosen then determine the regions of the search space that will be explored. These limitations are described in more detail in Gillet et al.[6]

Multiobjective problems are often characterised by a family of solutions that each represent a compromise in terms of the individual objectives. The family of solutions

maps out a hypersurface in the search space. The weighted-sum fitness function used in a GA finds one solution within the family, with the position of the solution on the surface being determined by the relative weights assigned to the objectives. In MoSELECT,[6,7] the GA of SELECT is replaced by a MOGA (MultiObjective Genetic Algorithm).[17,18] The MOGA technique allows multiple objectives to be explored simultaneously without the need for summation and maps out the entire surface of solutions in a single run thus overcoming many of the limitations of using a weighted-sum fitness function.

MOGA exploits the population nature of a GA in order to optimise a family of solutions simultaneously. The fitness of an individual is determined using the concept of *dominance* where an individual is *non-dominated* if an improvement in one of its objectives leads to a deterioration in one or more of the other objectives when compared to all other individuals in the population. The concept of dominance is illustrated in Figure 1 which shows a population of potential solutions to a two objective problem where the aim is to find solutions that represents minimum values of both objective functions f_1 and f_2. The non-dominated solutions are shown as solid circles and the dominated solutions are shown as unfilled circles. Solution *C* is dominated by solution *B* since *B* is better than *C* in both objectives f_1 and f_2. Solution *A*, however, is non-dominated since there is no solution in the population that is better than it in both objectives (solution *B* is better than *A* in terms of f_1, however, it is worse in terms of f_2). Similarly, solution *B* is non-dominated.

The GA in SELECT ranks the individuals in a population according to the weighted sum fitness function. In MoSELECT, however, ranking is based on dominance. A non-dominated individual is assigned rank of 0, an individual that is dominated by

one other individual is given rank 1, and so on as shown in Figure 1. Individuals are then selected for reproduction with a probability that is inversely proportional to their rank. This process is known as Pareto ranking and the non-dominated solutions map out what is known as the Pareto surface. Thus, all non-dominated solutions are treated as equivalent and have a higher probability of being selected for reproduction than do dominated solutions. The end result of a MOGA run is a family of non-dominated solutions spread out on what is known as the Pareto surface.

The technique of niching can be used to ensure that the entire Pareto surface is mapped and that an evenly spread family of solutions is found. Niching is implemented in an iterative procedure where the non-dominated solutions are examined one at a time. The first solution encountered is positioned at the centre of a hypervolume, or niche. Then, if the (absolute) difference in the objectives of the next solution and the objectives of any solution that already forms the centre of a niche is within a given threshold, for all objectives, the rank (or dominance) of the current solution is penalised, otherwise it forms the centre of a new niche. The threshold is also known as the *niche radius.*

MoSELECT allows competing objectives to be identified readily and by producing an entire family of solutions it is then the left to the library designer to choose the most appropriate solution based on additional criteria, such as chemical intuition. There are no significant overheads in terms of computing time for adopting Pareto ranking and a run of MoSELECT takes approximately the same time as a run of SELECT but has the advantage of finding a whole family of solutions. The following section describes the application of MoSELECT to the design of focused combinatorial libraries.

RESULTS

2-Aminothiazole library

A virtual library of 12850 product molecules was enumerated from 74 α-bromoketones and 174 thioureas, see Figure 2. Each set of reactants was extracted from the Available Chemicals Directory[19] and filtered using the ADEPT software[20]: reactants having molecular weight greater than 300 or more than 8 rotatable bonds were removed and a series of substructure searches were performed to remove reactants containing undesirable substructural fragments. The virtual library was enumerated and various properties were calculated for each of the product molecules including: 1024 Daylight fingerprints; molecular weight; number of rotatable bonds; number of hydrogen bond donors; number of hydrogen bond acceptors; and cost, which is based on summing the costs of each reactant from which the product is comprised.

A target compound was selected from the virtual library at random, shown in Figure 3a, and SELECT was run to find a 15×30 combinatorial subset focused around the target by maximising the normalised sum of similarities of the compounds in the subset with the target. Similarity was measured using Daylight fingerprints and the Tanimoto coefficient. Over 5 runs, solution libraries were found to have an average normalised sum of similarities of 0.832 (standard deviation 0.002). The cost of these libraries (as calculated from the price/g quoted in the ACD database) was found to range from $37436 to $64696 with an average cost of $48289.4 (standard deviation 10892.3). SELECT was then run to optimise libraries on cost alone and the minimum cost library averaged over 5 runs was found to be $1675.2 (standard deviation 184.7),

with these libraries having a normalised sum of similarities of 0.696 (standard deviation 0.01). Thus it can be seen that there is a high level of conflict between the two objectives with high similarity corresponding to relatively high cost and conversely low similarity corresponding to low cost. It is likely that a compromise library would be preferable to either of the two extremes. One way to reach a compromise would be to run SELECT using the weighted-sum fitness function with weights chosen to reflect the importance of each objective. However, in practise it is not easy to choose appropriate weights, especially for non-commensurate objectives like similarity and cost.

Next, MoSELECT was run to find a family of solution libraries focused around the same target while simultaneously minimising cost. Figure 3b shows cost plotted against normalised sum of similarities on initialisation of MoSELECT for a population size of 50, i.e., for 50 randomly selected combinatorial subsets. Non-dominated individuals, i.e., every individual for which there are no other individuals equal to or better than it in all objectives, are shown as solid circles, the dominated solutions are shown as crosses. The direction of the y axis has been reversed so that the direction of improvement in both objectives is towards the bottom left hand corner of the graph.

Figure 3c shows the non-dominated solutions after 5000 iterations of MoSELECT on the same scale as Figure 3b. Here it can be seen that the entire surface of non-dominated solutions has moved towards the bottom left hand corner of the plot. The non-dominated solutions are shown on an expanded scale in Figure 3d. where it can be seen that the solutions are spread out over a surface representing a range of

different values for each of the objectives. The percentage of the population that is non-dominated has increased from 12% at initialisation to 88%.

The dashed horizontal and vertical lines in Figure 3d show the optimum values achieved when each of the objectives are optimised independently. In this example solutions at the extremes are not found by MoSELECT. However, Figure 3e shows the non-dominated solutions found when MoSELECT is run with niching. The niche radius was set dynamically throughout the run at 10% of the range of values that exist for each objective on the current Pareto frontier. A much wider spread of solutions is found, as shown by the unfilled diamonds. The solutions of Figure 3d are superimposed and can be seen to occupy the central portion of the Pareto surface. Thus niching can be used to control the range of solutions found.

The conflict between cost and normalised sum of similarities is clearly evident, with libraries that are more tightly focused on the target corresponding to those of highest cost and conversely lower cost libraries tending to have lower normalised sums of similarities. The entire family of solutions is found in a single run (whereas a run of SELECT produces a single solution only) without the need to assign relative weights to the two objectives. In the absence of any further information, the family of non-dominated solutions are all equivalent and the library designer is then able to make an informed decision on what would be an appropriate compromise between the two objectives.

MoSELECT can be used to optimise any number of objectives and the next experiment was designed to optimise six objectives simultaneously: similarity to the known target; cost; and profiles of the following physicochemical properties:

hydrogen bond donors; hydrogen bond acceptors; rotatable bonds; and molecular weight. The physicochemical property profiles are optimised by minimising the difference between the distribution in the library and the distribution in the World Drug Index.[21]

When there are more than two objectives the results can be displayed by a parallel graph as shown in Figure 4 where the non-dominated solutions are shown after 5000 iterations of MoSELECT. The objectives have been scaled to allow them to be plotted on the same graph. Scaling was achieved by finding maximum and minimum values for each objective independently using SELECT and adjusting the values for each potential solution accordingly. Thus zero on the y axis, labelled Penalty, represents the best value that can be achieved when an objective is optimised independently with larger values indicating the degree to which an objective is compromised. Note that similarity is plotted as 1-SIM so that zero indicates maximum sum of similarities. Each line in the parallel graph represents one solution found by MoSELECT and crossing lines indicate objectives that are in competition with one another. It can be seen that there is significant competition between all of the pairs of adjacent objectives in the graph. Thus, if the overall aim is to design and synthesise libraries that are focused on a target compound, that have minimum cost and that have drug-like physicochemical properties profiles then a compromise solution should be selected. MoSELECT allows the full range of potential solutions to be visualised, thus allowing an informed choice to be made on where that compromise should lie.

Amide library

The second library to be studied was an amide library. The library has been used in a recent comparison of the PLUMS program with a GA based program called VOLGA and a dynamic monomer frequency analysis program (DMFA).[16] The library consists of 100 diverse carboxylic acids and 100 diverse amines extracted from the MedChem database, representing a virtual library of 10K products.

The virtual library was enumerated and the following properties were calculated for the products: 1024 Daylight fingerprints; number of rotatable bonds; and number of hydrogen bond donors. In addition, the bioavailability of each compound was predicted using the QSAR model recently published by Yoshida and Topliss.[22] Bioavailability is represented in the model as the percentage of an administered dose of the compound that reaches systemic circulation after oral administration. Compounds were given a bioavailability rating as follows: those with a bioavailability prediction ≤ 20% were rated 1; compounds in the range 20-49% were rated 2; compounds in the range 50-79% were rated 3; and compounds with predicted bioavailability ≥ 80% were rated 4. There were 649 compounds in the virtual library for which predictions could not be calculated due to missing parameters and these were assigned the rating 0. This demonstrates a further advantage of MoSELECT over weighted-sum methods, in that there is the flexibility to handle different types of objective simultaneously such as classifiers, ranges, profiles and so on.

The first experiment was based on designing libraries focused around a target compound while simultaneously optimising bioavailability over the library as a whole. A compound was selected at random from the virtual library as the target

compound and libraries were focused on the target by maximising the normalised sum of similarities to the target. Bioavailability was optimised by maximising the sum of bioavailability ratings over all compounds in the library.

As in the 2-aminothiazole example, SELECT was run to find optimum values for 10×10 combinatorial subsets when each of similarity and bioavailability are optimised independently. The average maximum normalised sum of similarities over 5 runs was 0.502 (standard deviation 0.001), with these values being found for libraries having an average bioavailability score of 215 (standard deviation 7.8). Conversely, the maximum bioavailability score was an average of 400 (standard deviation 0.0) over 5 runs with corresponding normalised sum of similarities of 0.285 (standard deviation 0.013). An optimum library would be one that had maximum similarity to the target while also having a maximum bioactivity score. Thus, the two objectives are in conflict and once again a compromise between the two objectives may provide the most appropriate solution.

MoSELECT was then run to optimise 10×10 combinatorial subsets on similarity to the target simultaneously with predicted bioavailability and hence to find a family of equivalent solutions. The results are shown in Figure 5. The solid circles show the MoSELECT solutions and the dashed lines represent optimum values when each objective is optimised independently. The competition between the two objectives is clear with relatively good values of similarity corresponding to relatively poor values of bioavailability.

Next, MoSELECT was run to include additional objectives, namely, profiles of hydrogen bond donors and rotatable bonds. The non-dominated solutions are shown

in Figure 6 after 5000 iterations. Again in the parallel graph the objectives have been scaled to allow them to be compared on the same graph. The competing nature of similarity to the target and bioavailability is evident from the crossing lines. There are some solutions with parallel lines passing through BIO and HBD indicating that these two objectives are correlated (as would be expected); however there are also some solutions where lines between these two objectives cross, indicating that the relationship becomes more complex as additional objectives are optimised simultaneously. One possible compromise solution is indicated by the bold line. This solution has near optimum sum of similarities and bioavailability score at the expense of hydrogen bond donor and rotatable bond profiles. Other compromise solutions may be equally appropriate.

The next series of experiments with the amide library were based on the PLUMS study. Filtering techniques based on various physicochemical properties (molecular weight, CMR, number of rotatable bonds, maximal binding energy, and complexity) were used to identify 409 molecules within the virtual library as having favourable properties. Analysis of the 409 molecules revealed that they were constructed from 67 acids and 71 amines and hence the combinatorial library that contains all 409 molecules consists of 4757 molecules. The aim was to design 10×10 combinatorial subsets that contain as many of the 409 favourable molecules as possible.

The maximum number of favourable compounds achievable in a 10×10 subset was found to be 69 using PLUMS, VOLGA and DMFA.[16]

Initially, SELECT was run to optimise the number of favourable compounds in a 10×10 subset and was also able to find libraries containing 69 of the preferred

compounds. When MoSELECT was run to optimise this single objective, four non-dominated solutions were found with each library containing a different set of 69 of the preferred compounds (between 61 and 63 of the preferred compounds were identical when the libraries were compared with one another).

The filters used to select the preferred compounds define boundaries in physicochemical property space within which it is desirable that the bulk of the library resides. Within these boundaries, however, it is usually desirable that the compounds are widely spread, and the less sophisticated DMFA and PLUMS algorithms are not able to meet this requirement. Thus an attempt was made to optimise the number of preferred compounds simultaneously with diversity (where diversity was measured as the sum of pairwise dissimilarities using Daylight fingerprints and the cosine coefficient). The non-dominated solutions found after 5000 iterations of MoSELECT are shown in Figure 7. It can be seen that the maximum number of preferred compounds corresponds to minimum diversity within the subset and, conversely, maximum diversity in the subset corresponds to minimum number of preferred compounds. The maximum number of preferred compounds found in a non-dominated solution is 64 for a library with diversity 0.510.

Extending the optimisation problem to include a third objectives, namely, drug-like molecular weight profile, highlights the need for further compromise as shown in the parallel graph of Figure 8. The number of preferred compounds and diversity are plotted as 1-PREF and 1-DIV, respectively, so that optimum solutions are those nearest zero for all objectives. The solution with a near drug-like molecular weight profile, has diversity of 0.544 but it contains only 6 of the preferred compounds,

whereas the solution with the highest number of preferred compounds (48), shown in bold, has similar diversity but an unfavourable molecular weight profile.

CONCLUSION

MoSELECT is a program for the design of combinatorial libraries optimised on multiple objectives simultaneously. Here, it has been applied to the design of focused libraries. In focused library design it is often desirable to optimise several objectives, for example, similarity to a target compound simultaneously with bioavailability. MoSELECT generates a family of equivalent solutions in a single run so that the library designer can make an informed decision about which library to progress to synthesis. A further advantage of MoSELECT is that it allows the relationships between objectives to be readily identified.

Acknowledgement. We thank GlaxoSmithKline for funding and Daylight Chemical Information Systems Inc. for software support. We thank Anne Hersey and Sandeep Modi of GSK for the provision of the Topliss bioavailability predictions. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES

1. Dean, P.M., and Lewis, R.A. Eds. *Molecular Diversity in Drug Design*. Kluwer, Dordrecht, 1999.

2. Willett, P. Ed. Computational Methods for the Analysis of Molecular Diversity. *Perspect. Drug Discov. Design*. 1997, *7/8*.

3. Bohm, H.-J., Schneider, G., Eds. *Virtual Screening for Bioactive Molecules*. Wiley-VCH, Weinheinm, 2000.

4. Valler, M.J., and Green, D. Diversity screening versus focussed screening in drug discovery, *Drug Discovery Today*. 2000, **5**, 286-293.

5. Martin, E.J., and Crichlow, R. W. Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem*. 1999, **1,** 32-45

6. Gillet, V.J., Khatib, W., Willett, P., Fleming, P.J., and Green D.V.S. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. . *J. Chem. Inf. Comput. Sci.* In the Press.

7. UK Patent Application No. 0029361.

8. Gillet, V.J., Willett, P., Bradshaw, J. and Green D.V.S. Selecting combinatorial libraries to optimise diversity and physical properties. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 167-177.

9. Gillet, V.J.; Willett, P.; Bradshaw, J. The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 731-740.

10. Gillet, V.J.; Nicolotti, O. Evaluation of reactant-based and product-based approaches to the design of combinatorial libraries. *Perspect. Drug Discov. Design*, 2000, **20**, 265-287.

11. Jamois, E.A.; Hassan, M.; Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.,* 2000, **40**, 63-70.

12. Zheng, W., Hung, S.T., Saunders, J.T., and Seibel, G.L. PICCOLO: A tool for combinatorial library design via multicriterion optimization. *In Pacific Symposium on Biocomputing 2000*; Atlman, R. B., Dunkar, A. K., Hunter, L., Lauderdale K., Klein, T.E., Eds. World Scientific: Singapore, 2000, pp588-599.

13. Brown, J.D., Hassan, M., and Waldman, M. Combinatorial library design for diversity, cost efficiency, and drug-like character. *J. Mol Graphics Modell*. 2000, **18**, 427-437.

14. Rassokhin, D.N., and Agrafiotis, D.K. Kolmogorov-Smirnov statistic and its application in library design. *J. Mol Graphics Modell.* 2000, **18**, 368-382.

15. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol*. 1997, **267**, 727-748.

16. Bravi, G., Green, D.V.S,, Hann, M.A., and Leach, A.R. PLUMS: A program for the rapid optimization of focused libraries. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 1441-1448.

17. Fonseca, C.M., and Fleming, P.J. An overview of evolutionary algorithms in multiobjective optimization, In *Evolutionary Computation*; De Jong, K., Ed.; The Massachusetts Institute of Technology, 1995, **3**, pp. 1-16.

18. Fonseca, C.M., and Fleming, P.J. Genetic algorithms for multiobjective optimization: formulation, discussion and generalisation, In *Genetic Algorithms: Proceedings of the Fifth International Conference*; Forrest, S. Ed.; Morgan Kaufmann, San Mateo, CA, 1993, pp. 416-423.

19. The Available Chemicals Directory is available from MDL Information Systems, Inc., 146000 Catalina Street, San Leandro, CA 94577

20. Leach, A.R., Bradshaw, J., Green, D.V.S., Hann, M., and Delany III, J.J. Implementation of a system for reagent selection, library enumeration, profiling and design. *J. Chem. Inf. Comput. Sci*. 1999, **39**, 1161-1172.

21. The World Drug Index is available from Derwent Information, 14 Great Queen St., London WC2B 5DF, UK.

22. Yoshida F., and Topliss J.G. QSAR model for drug human oral bioavailability. *J. Med. Chem*., 2000, **43**, 2575-2585.

FIGURE CAPTIONS

Figure 1. A population of individuals in a two-objective (f_1 and f_2) minimisation problem is shown. The solid circles (including those labelled *A* and *B*) represent non-dominated solutions, where a solution is non-dominated if there is no other individual in the population better than it in both objectives. The unfilled circles represent dominated solutions, for example, solution *C* is dominated by solution *B* which is better than it in both f_1 and f_2. Each individual is assigned a rank that reflects the number of individuals by which it is dominated.

Figure 2. 2-Aminothiazole library.

Figure 3a. The target compound used for library focusing.

Figure 3b. MoSELECT was configured to find 15×30 2-aminothiazole subsets optimised on similarity to a target compound simultaneously with cost. The population is shown on initialisation. The dominated solutions are shown by crosses and the non-dominated solutions as filled circles.

Figure 3c. The population is shown after 5000 iterations of MoSELECT. The whole population has moved towards the left-hand corner of the graph indicating improvement in both objectives and the number of non-dominated solutions in the population has increased.

Figure 3d. The non-dominated solutions in Figure 2b are shown on an expanded scale where it can be seen that the two objectives (similarity and cost) are in conflict.

Figure 3e. The same library design problem was run with niching to force MoSELECT to find solutions corresponding to extreme values in each objective.

Figure 4. 2-Aminothiazole library simultaneously optimised on six objectives: similarity to the target (SIM); cost in $/g (COST); and profiles of molecular weight (AMW); hydrogen bond donors (HBD); hydrogen bond acceptors (HBA) and rotatable bond (RB) profiles. The y axis, labelled Penalty, represents the relative value achieved for an objective with a Penalty of zero representing the best that can be achieved when an objective is optimised independently. Similarity to the target is plotted as 1-SIM so that the direction of improvement in all the objectives is towards zero on the y-axis.

Figure 5. 10×10 amide subsets optimised on similarity to a target compond and bioavailability (*BIO*).

Figure 6. Selecting amide subsets optimised on similarity to the target (SIM), bioavailability (BIO) and profiles of hydrogen bond donors (HBD) and rotatable bonds (RB). Simlarity to the target and bioavailability are plotted as 1-SIM and 1-BIO, respectively, so that the direction of improvement in all the objectives is towards zero on the y-axis. Penalty is as described in Figure 4.

Figure 7. The non-dominated solutions are shown for selecting 10×10 amide subsets optimised on the number of preferred compounds and diversity, simultaneously. The maximum number of preferred compounds found is 64 for a library with relatively low diversity.

Figure 8. The non-dominated solutions are shown for selecting 10×10 amide subsets optimised on the number of preferred compounds (PREF), diversity (DIV) and

molecular weight profile (AMW), simultaneously. PREF and DIV are plotted as 1-PREF and 1-DIV, respectively, so that the direction of improvement in all the objectives is towards zero on the y-axis. Penalty is as described in Figure 4. The maximum number of preferred compounds found for this run of MoSELECT is 46.

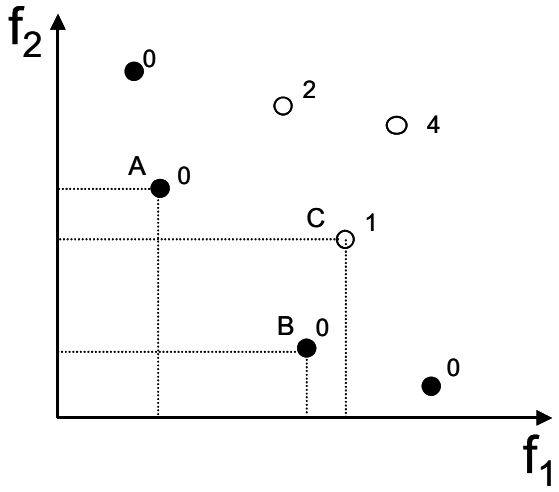

Figure 1.

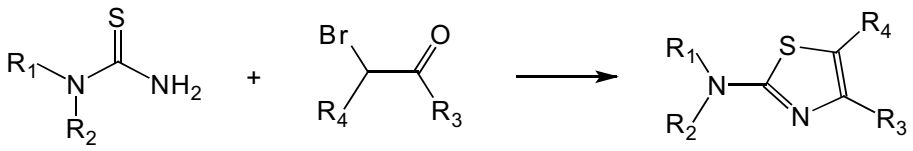

Figure 2.

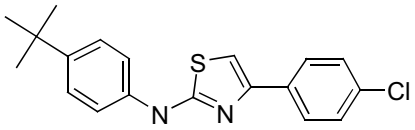

Figure 3a.

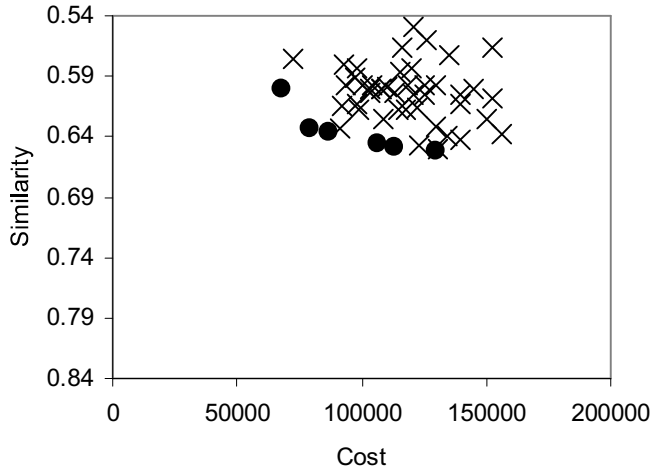

Figure 3b.

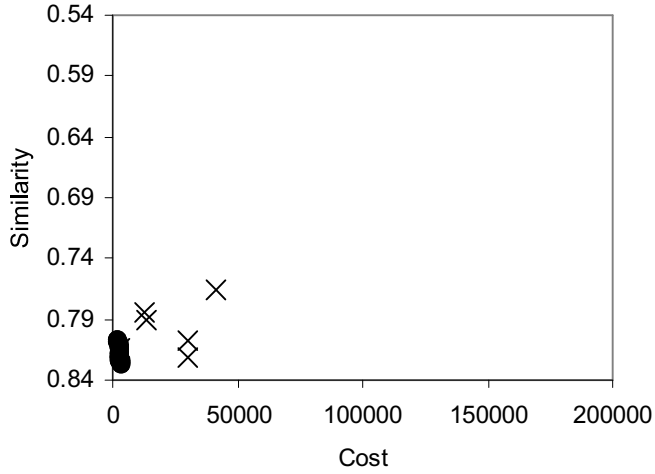

Figure 3c.

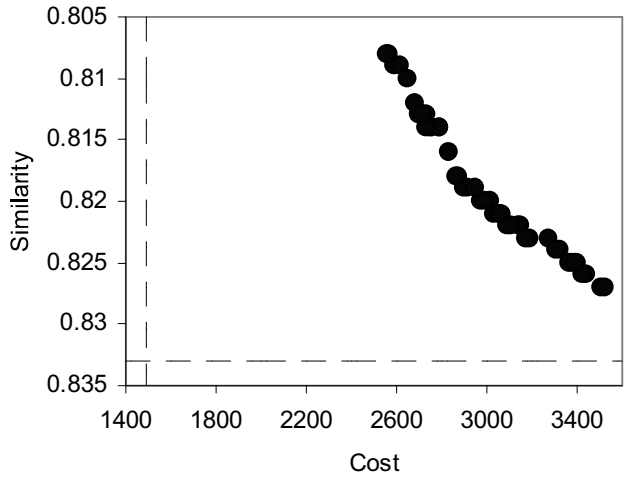

Figure 3d.

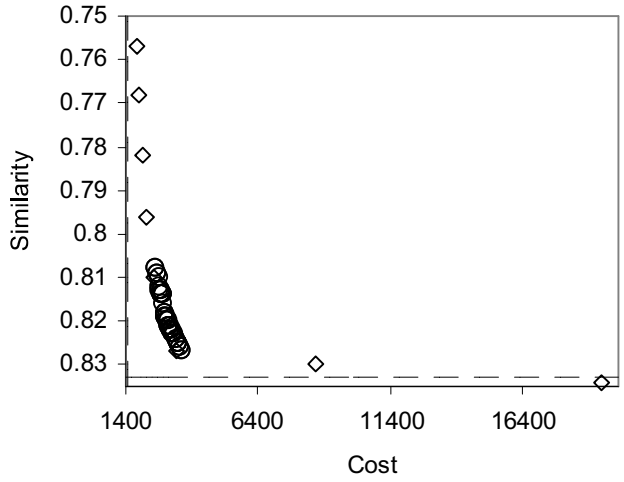

Figure 3e.

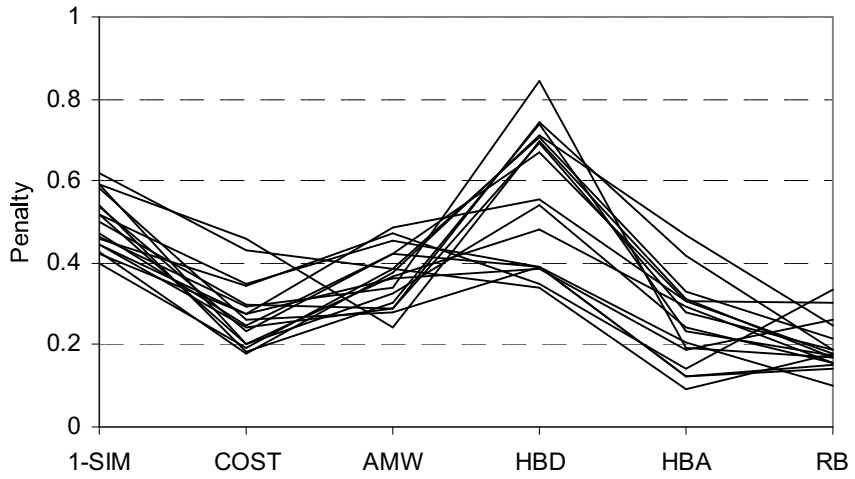

Figure 4.

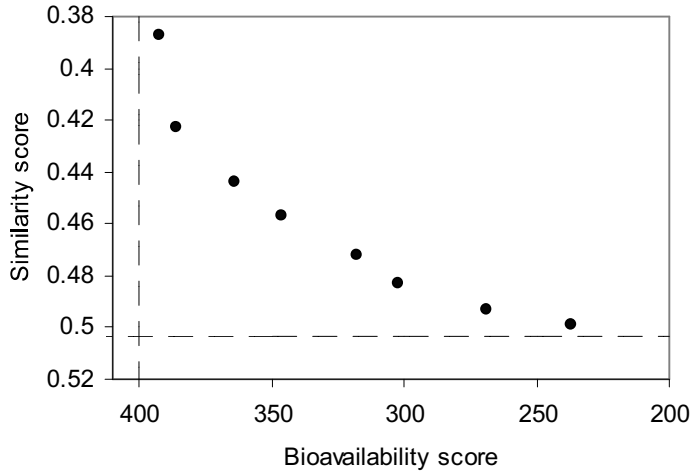

Figure 5.

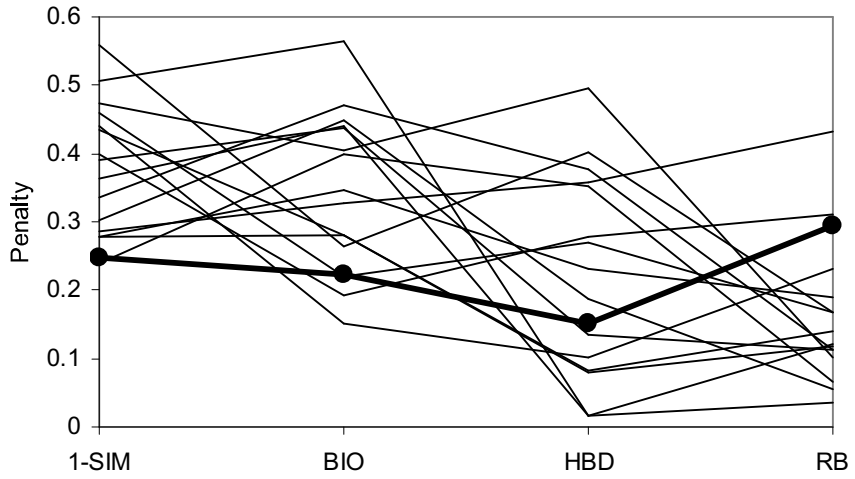

Figure 6.

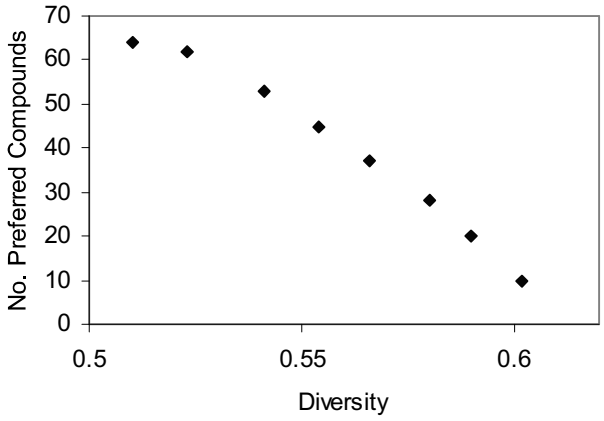
70
60
50
40
30
20
10
0
No. Preferred Compounds
0.5
0.55
0.6
Diversity

Figure 7.

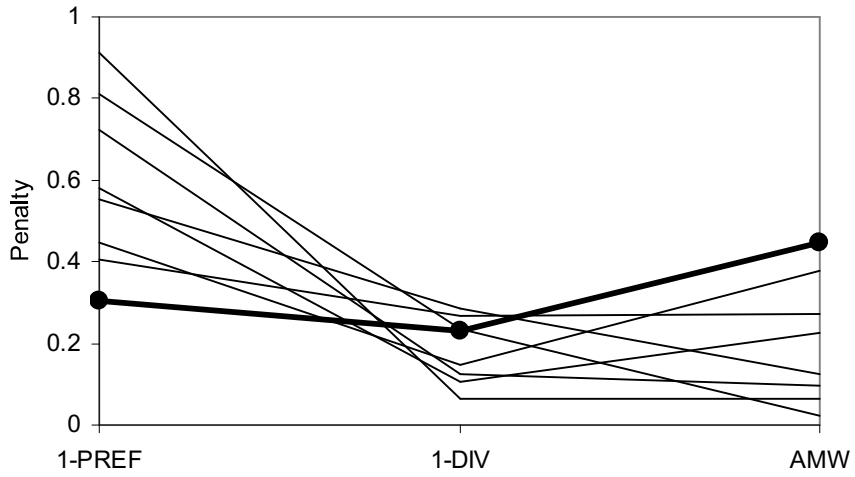

Figure 8.