Journal of Computer-Aided Molecular Design **18:** 761–771, 2004.
DOI 10.1007/s10822-004-6986-2

The *de novo* design of median molecules within a property range of interest

Nathan Brown[a], Ben McKay[a,*] & Johann Gasteiger[b]
[a]*Avantium Technologies B.V., P.O. Box 2915, 1000 CX Amsterdam, The Netherlands;* [b]*Computer-Chemie-Centrum and the Institute for Organic Chemistry, University of Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany*

Received 4 October 2004; accepted in revised form 30 November 2004

Key words: aqueous solubility, *de novo* design, genetic algorithms, graph theory, mean molecular polarisability, median molecules, multiobjective optimisation, pareto ranking, similar-property principle

Summary

In this paper an application is presented of the median molecule workflow to the *de novo* design of novel molecular entities with a property profile of interest. Median molecules are structures that are optimised to be similar to a set of existing molecules of interest as an approach for lead exploration and hopping. An overview of this workflow is provided together with an example of an instance using the similarity to camphor and menthol as objectives. The methodology of the experiments is defined and the workflow is applied to designing novel molecules for two physical property datasets: mean molecular polarisability and aqueous solubility. This paper concludes with a discussion of the characteristics of this method.

Introduction

In the fields of chemical process development and pharmaceutical discovery research, an automated computational method to suggest molecular entities that fit a given physical property profile is highly desirable. Typically this involves virtual screening campaigns using similarity searching, to mine an existing corporate collection of structural data [1–3]. However, this method is only capable of addressing the problem by suggesting structures that already exist in the collection itself and does not propose new structures directly.

Recently, we have reported the median molecule workflow (MMW) as an approach for the *de novo* design of molecules based on the optimisation of structural similarity of designed structures to extant molecules of interest [4]. Here, we suggest an application of the MMW, to the automated design of novel molecular entities that are similar to existing structures as an approach to suggesting molecules that will tend to share the same property profile. Each set of evolved structures is then validated against two predictive models to investigate the extent to which the new molecules fall within the property range of interest.

The median molecule workflow

The MMW is an approach that is intended to assist in the design of a set of novel molecular entities that lie between a set of existing molecules as a constrained method of exploring chemical space and automatically suggesting new structures.

The MMW relies essentially on three modular components: (1) molecular perturbation, (2) scoring and (3) ranking; the flowchart for the MMW is provided in Figure 1. Each of these three processes is encapsulated in a separate piece of software with all communication achieved between the program calls through files, thereby eschewing software

*To whom correspondence should be addressed. Tel.: +31-20-586-8041; E-mail: ben.mckay@avantium.com

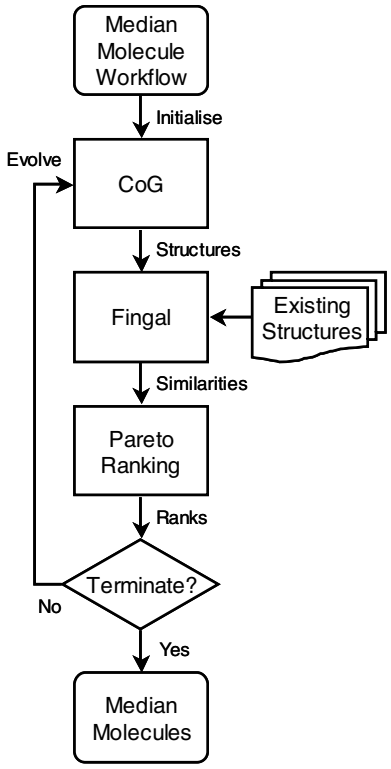

Figure 1. Flowchart of the median molecule workflow, including the Compound Generator (CoG) *de novo* design software and the *Fing*erprinting *Al*gorithm (Fingal) software.

dependent communication issues. The three software components developed for and applied in this study are defined below – although equivalent software components may also be used.

De novo *design module*

The *de novo* design module of the MMW takes the form of the Compound Generator (CoG) software. CoG is a genetic algorithm (GA) that operates directly on graph-based chromosomes that represent the molecules in a population. The nodes of the chromosomes can be adapted as required by the application to encode substructures of interest or simply elemental atom types. The only chemical constraints applied to the formation and perturbation of molecules in the GA is the valence bond model. The reader is referred to Brown et al. [4] for a more thorough discussion of the data structure, genetic operators and other algorithms implemented in the CoG program; a literature review of similar approaches that apply GAs and multiobjective optimisation is also provided in this paper.

Molecular scoring module

The degree of similarity of new, evolved molecules to the stated molecules of interest is determined by the application of molecular fingerprints and similarity coefficient calculations. To achieve this the *Fing*erprinting *Al*gorithm (Fingal) software was implemented, which generates hash-key fingerprints and calculates molecular similarities. Although hash-key fingerprints are not ideal molecular descriptors in all cases, they are rapid to calculate and have proved to be effective for the evolution of median molecules applying the workflow described here. The Fingal software also contains a variety of novel enhancements to encode geometry information, as described in Brown et al. [5].

Multiobjective ranking module

The final stage of the MMW ranks each of the candidate molecules using the multiobjective optimisation technique called Pareto ranking [6] and was first applied to problems in chemistry by Handschuh et al. [7]. The approach calls for the determination of a rank for each individual in a

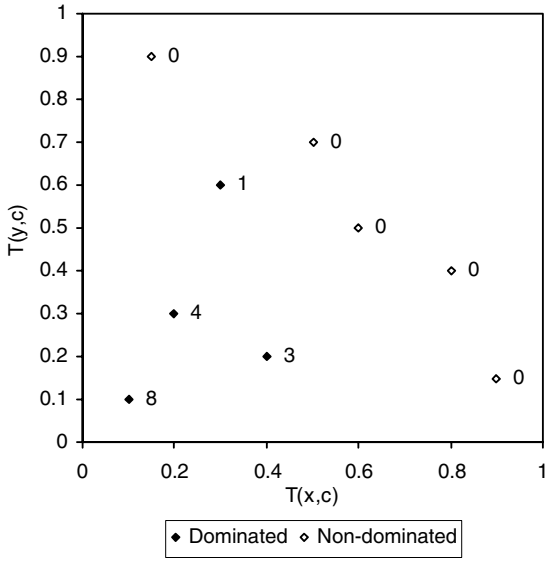

Figure 2. Example of a set of solutions that have been ranked using the Pareto ranking method, labelled by their respective rank positions. The Tanimoto similarity (T) of the set, c, to the two starting structures (x and y) is illustrated.

particular population based on the number of other individuals that dominate this individual in each of the objectives. In Figure 2, nine individuals (from the set, *c*) are plotted according to their score in the two objectives – *x* and *y*, respectively – where 1.0 is the optimal score in both cases. The rank of each object can then be determined simply by counting the number of other objects that have higher scores in each objective. If this rank is greater than zero then the individual is one that is dominated, whereas if the rank equals zero then the solution is said to be non-dominated; the set of non-dominated solutions is said to lie on the Pareto front.

The application of Pareto ranking has been shown to be highly effective at balancing competing objectives compared to the more typical weighted-sum approach; the latter tends towards solutions that are skewed towards a single or a small subset of objectives. Conversely, optimisation algorithms that apply the Pareto ranking method tend to evolve an evenly spread set of solutions on the Pareto frontier allowing the user subsequently to select from a set of equally valid solutions, depending on the importance of the relevant objectives.

Median molecules: an example

To demonstrate the concept of median molecules, here we report an exemplar run of the MMW with CoG and Fingal as the perturbation and scoring modules, respectively, and the similarity of the evolved structures to two molecules as objectives: menthol and camphor (Figure 3). Nachbar [8] first suggested these structures in the experiments related to the generation of the average chemical structure.

A single execution of the MMW was performed with a population of 2000 individuals and a total of 5000 generations. The initial population was generated randomly by a simple graph generation method: adding nodes and edges iteratively, ensuring that the graph is connected and satisfies valence bond laws, with a final step of adding further bonds randomly throughout the population to introduce some cyclic substructures. The nodes used in this experiment represented the elements, C and O, since these are the only atoms present in the starting structures.

The final generation of the workflow run is presented in Figure 4, plotted as the similarities to the two starting structures, respectively. One can observe readily that, even in the final population, there is a substantial degree of diversity present in the population in terms of the similarity to the starting structures. The Pareto frontier in the upper-right of Figure 4 contains 130 median molecules in the final generation; this frontier includes menthol and camphor, which were also evolved. A six-member representative, diverse subset of the evolved 130 median molecules is provided in Figure 5 to illustrate the type of structures that have been evolved, together with their Tanimoto similarities to each of the two starting structures. The evolution rate of this MMW run is provided in Figure 6 in terms of the size of the current frontier and the number of these solutions that are also present in the final Pareto frontier, at 100-generation intervals.

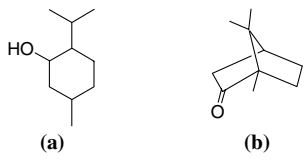

Figure 3. The two molecular targets used in the exemplar experiment: menthol (a) and camphor (b).

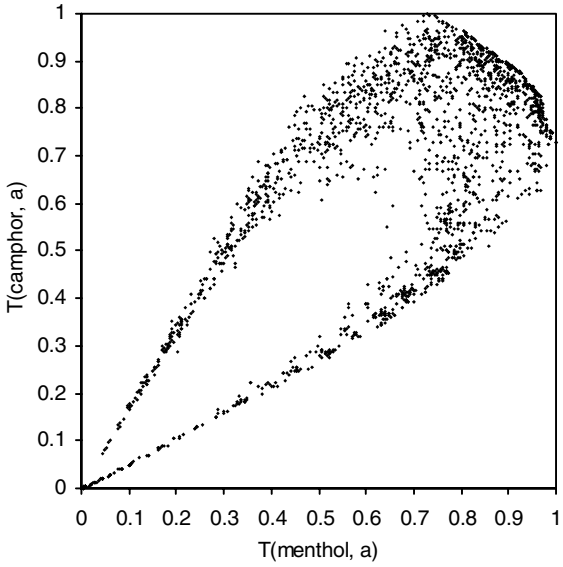

Figure 4. Plot of the continuous Tanimoto similarity of each evolved median molecule (a) in the final generation with respect to the structures menthol and camphor.

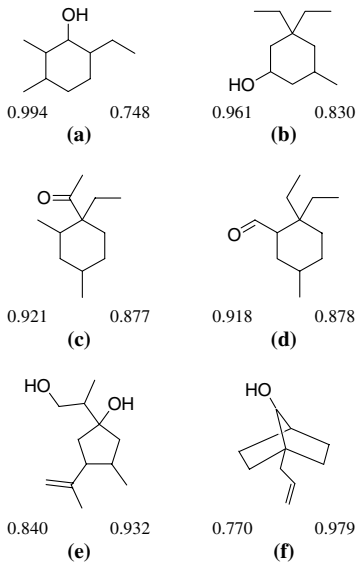

Figure 5. A six-member representative, diverse subset of the evolved median molecules using similarity to menthol (first similarity value) and camphor (second similarity value) as the objectives along with their respective similarities to those structures.

Essentially, when considering two objectives, the median molecule concept may be described as the drawing of a line between two points in a high-dimensional representation of chemistry space and sampling a number of novel molecular entities that are proximate to that line.

Methodology

The similar-property principle conjectures that molecules that are structurally similar to one another will also tend to exhibit similar physicochemical and biological properties [9]. Therefore, since the median molecule concept is intended to explore "structure space" between known molecules of interest, it follows that the evolved structures will exhibit properties similar to the targets. This is of course dependent on the degree and measure of similarity captured by the descriptors, together with the extent to which the similar-property principle holds true.

To investigate the suitability of the median molecule concept to *de novo* design, the experiments presented in this paper are intended to evolve median molecules that optimise their similarity to two starting structures from a larger dataset for which physical property data are available. In these experiments, multiple runs of the MMW are executed with the molecular similarity objectives drawn from a selected physical property range. The evolved median molecules from each experiment are then combined and rationalised to remove duplicates.

We refer to the entire physical property dataset as the *dataset*, while each dataset is also partitioned into a *training set* and at least one *test set*. To simulate the missing property range, we artificially exclude a subset of the training set within a specified physical property range, referred to as the *removed subset*. Lastly, the set of evolved structures from each run of the MMW is called the *median molecule set*; a schematic of this strategy is provided in Figure 7.

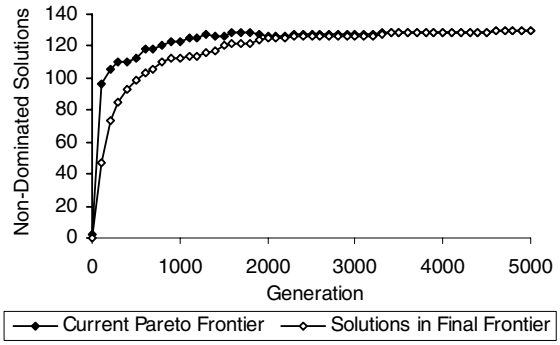

Figure 6. The size of the current Pareto frontier at 50-generation intervals and the proportion of each population that is also on the final Pareto frontier after 5000 generations from the menthol and camphor experiment.

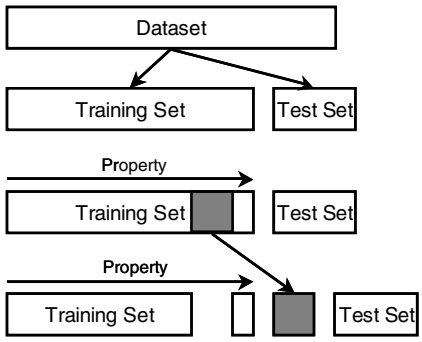

Figure 7. Schematic of the partitioning methodology applied in this study. The dataset is first partitioned into a training set and at least one test set. The remaining training set is then ordered by property value and a contiguous subset removed, the removed subset, which is highlighted.

In the absence of physical property values for the median molecule set, the method applied to validate these structures is to use them as queries to quantitative structure–property relationships (QSPRs). These models are referred to throughout this paper as the *validation models*. The QSPR validation models, where necessary, were developed using partial least squares (PLS) regression in the SIMCA-P 10.5 software [10]. The following performance statistics are reported for each of the validation models generated:

- R^2, Q^2_{cum} (or cross-validated R^2 with 7-fold leave-one-out as calculated by SIMCA-P) and RMSEE (root-mean-square error of estimation) statistics of the training set;
- R^2_{Pred} and RMSEP (root-mean-square error of prediction) of each of the test sets; and
- R^2_{Pred} and RMSEP of the removed subset.

Note that the removed subset will tend to have a significantly lower R^2_{Pred} since it has a very narrow physical property value range in each case. The number of latent variables (LVs) of each validation model is also provided with the model statistics above.

Case study one: mean molecular polarisability

The mean molecular polarisability ($\bar{\alpha}$) dataset applied in this study was published by Miller [11], and contains 290 molecules. The hydrogen molecule was removed from the dataset since the Fingal descriptors operate on hydrogen-depleted molecular graphs. An additional four structures were removed as extreme model outliers, provided here with their Chemical Abstracts Service (CAS) registration numbers: coronene (191-07-1), difluoroenyl (1530-12-7), 1,2:5,6-dibenzanthracene (224-41-9), and 2,3:4,5-dibenzophenazine (226-47-1). From the remaining dataset of 285 structures, a training and test set partition was created randomly with 203 and 58 in each set, respectively. To simulate a missing property range 24 structures were removed from the training set between the $\bar{\alpha}$ values of 9.84 and 10.93. From the removed subset, two further subsets were selected randomly from subranges at each end of the overall range to act as the similarity objectives for the MMW; the structures are provided in Figure 8 of the lower (a–c) and higher (d–f) $\bar{\alpha}$ range together with the CAS registry numbers and $\bar{\alpha}$ values. The two molecular similarity objectives of each run were then selected as all of the unique pairwise combinations of the two 3-member subsets (a, d; a, e; a, f; b, d; . . . ; c, f), resulting in a total of nine individual experiments.

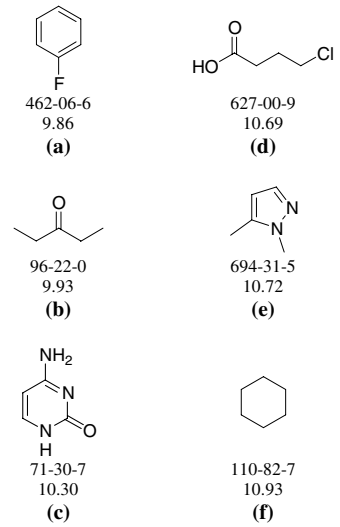

Figure 8. The structures, CAS numbers and $\bar{\alpha}$ values of the three-member subsets at the lower (a–c) and upper (d–f) extremities of the $\bar{\alpha}$ subset property range, respectively.

Once the median molecule sets from all nine experiments were combined and duplicates removed, 519 unique structures remained in the median molecule set; a randomly chosen

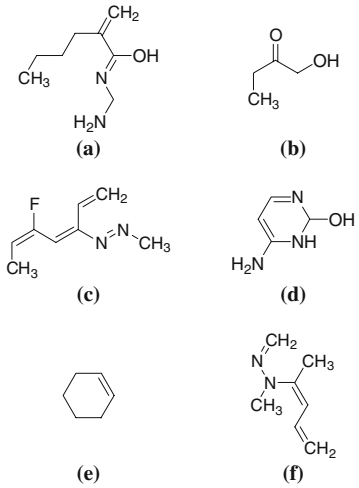

Figure 9. A six-member representative, diverse subset of the evolved median molecule set that are predicted by the validation models to lie within the $\bar{\alpha}$ property range under consideration.

six-member subset is provided in Figure 9. Two validation models were applied in this case study: the atom additivity scheme proposed by Miller [11] and implemented in PETRA [12], and a PLS regression model generated using Dragon descriptors [13]. The model statistics for the two models are provided in Tables 1 and 2. The high Q^2_{cum} value together with the low RMSEP on the unseen test set suggests that this model has substantial predictive power. The molecular descriptors of the median molecule set were also calculated with Dragon; the descriptors were then loaded into SIMCA-P to calculate the predicted values. A plot of the actual vs. predicted $\bar{\alpha}$ values with this validation models is provided in Figures 10 and 11, respectively.

The histograms in Figures 12 and 13 illustrate the degree to which the predicted values of the median molecule set fall within the target physical property range when predicted through application of the two validation models. Both histograms show distributions of the actual values of the training set, the predicted values of the removed subset and the predicted values of the median molecule set. One may observe that the distributions in both Figures 12 and 13 cover the $\bar{\alpha}$ range of the starting structures, but that they are also centred to the right of this range.

Table 1. Correlation and residual error statistics for the calculated $\bar{\alpha}$ values.

Training set (203)	R^2	0.997
	RMSEE	0.428
Test set (58)	R^2_{Pred}	0.998
	RMSEP	0.336
Removed subset (24)	R^2_{Pred}	0.584
	RMSEP	0.299

The number of structures in each set is given in brackets.

Table 2. PLS model statistics for the $\bar{\alpha}$ validation model developed in SIMCA-P with molecular descriptors generated using Dragon together with the number of latent variables (LVs) in the model.

Training set (203)	R^2	0.994
	Q^2_{cum}	0.990
	RMSEE	0.558
	LVs	3
Test set (58)	R^2_{Pred}	0.993
	RMSEP	0.562
Removed subset (24)	R^2_{Pred}	0.273
	RMSEP	0.510

The number of structures in each set is given in brackets.

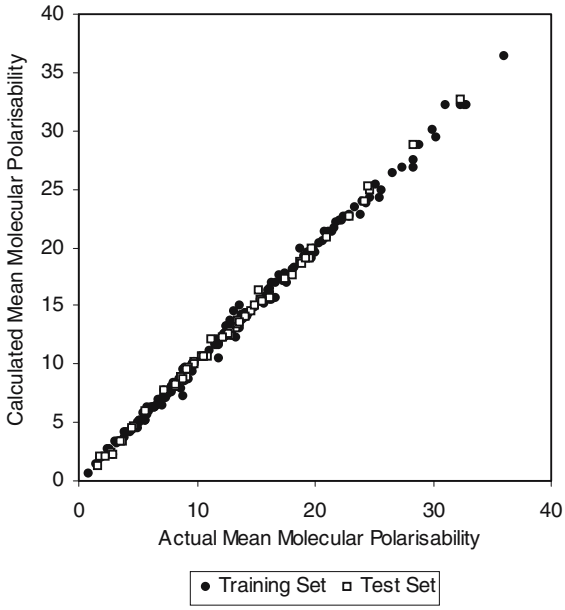

Figure 10. The actual vs. calculated values the $\bar{\alpha}$ atom additivity scheme implemented in PETRA.

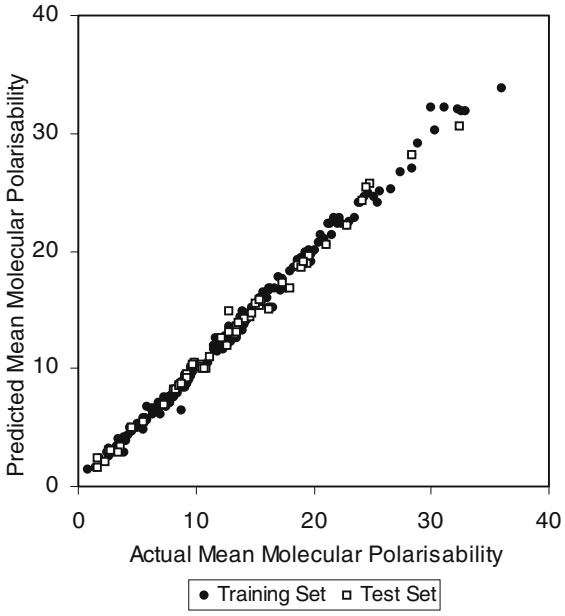

Figure 11. The actual vs. predicted values from the $\bar{\alpha}$ validation model using Dragon descriptors.

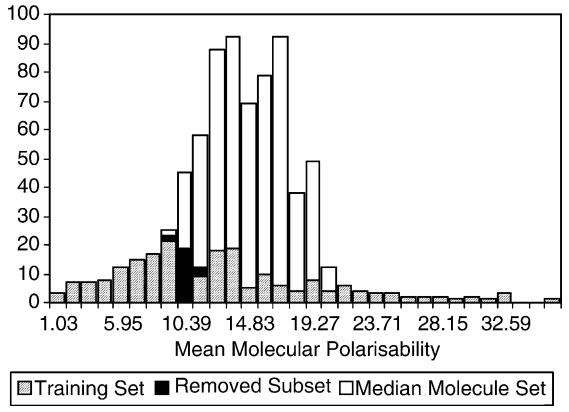

Figure 12. A histogram of the actual $\bar{\alpha}$ for the training set together with the calculated values for the removed subset and the median molecule set using the $\bar{\alpha}$ atom additivity scheme implemented in PETRA.

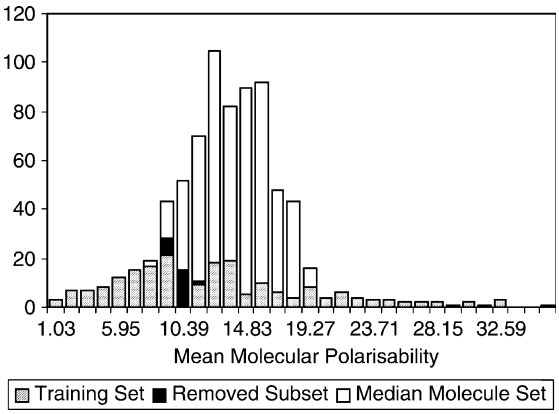

Figure 13. A histogram of the actual $\bar{\alpha}$ for the training set together with the predicted values for the removed subset and the median molecule set using the Dragon validation model.

Of the 519 unique structures in the median molecule set evolved in this experiment, 26 and 37 are predicted to fall within the target physical property range by the atom additivity scheme and Dragon model, respectively, whereas the original dataset contained 24 structures within this range. Furthermore, none of the median molecule set is already present in the $\bar{\alpha}$ dataset, suggesting that this method is truly designing novel structures. Therefore, although the distribution of the median molecule predictions is not centred on the target $\bar{\alpha}$, this case study does demonstrate that this method is applicable to suggesting new molecules for library augmentation, at least in terms of $\bar{\alpha}$.

Case study two: aqueous solubility

The second dataset considered in this study is the Huuskonen dataset [14] of aqueous solubility (log *S*) for 1342 structures that has been published as three partitions: 1201 in the training set and 21 and 120, respectively, in each of the test sets. An additional four structures were removed from the training set as extreme model outliers, given here with their CAS registration numbers: ajmaline (4360-12-7), benzo[g,h,i]perylene (191-24-2), brucine (357-57-3), and strychnine (57-24-9). In this experiment we partition the remaining training set of 1197 structures further, to create the removed subset, by removing an additional 125 structures in the log *S* range from −2.22 through −1.76 to simulate a missing physical property value range. From the removed subset, six structures were selected randomly from each extremity of the range to act as the similarity objectives for the MMW; these structures are given in Figures 14 and 15, respectively, together with the CAS registry numbers and log*S* values. The two similarity objectives of each run were then selected as all unique pairs of the two 6-member subsets, resulting in a total of 36 separate experiments.

The 36 runs executed for the log *S* dataset resulted in 1364 unique structures in the median molecule set, a six-member subset is provided in Figure 16. Two validation models were developed with PLS regression using the PETRA descriptors reported by Yan and Gasteiger [15] and the Dragon descriptors, respectively. Both models exhibit good predictive characteristics indicated by the high Q^2_{cum} and also low RMSEP values when applied to the two external test sets; the model statistics are provided in Tables 3 and 4. These statistics provide a sufficient level of confidence that indicative predictions may be drawn from the model. A plot of the actual vs. predicted $\bar{\alpha}$ values with the PETRA and Dragon validation models are provided in Figures 17 and 18, respectively.

The application of the molecular descriptors from the median molecule set to the PETRA and Dragon validation models resulted in predictions that were then plotted as histograms provided in Figures 19 and 20, respectively. The histogram shows the actual values of the training set used in the validation model along with the predicted values of the removed subset and the median molecule set. Of the 1364 median molecules that

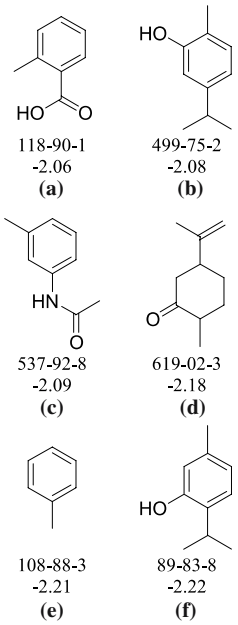

Figure 14. The structures, CAS numbers and log *S* values of the six-member subset at the lower end of the log *S* subset property range.

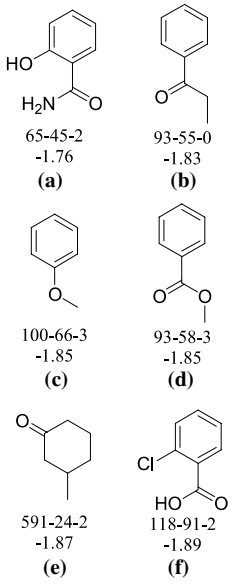

Figure 15. The structures, CAS numbers and log *S* values of the six-member subset at the upper end of the log *S* subset property range.

were evolved, the PETRA and Dragon validation models predict 390 and 337, respectively, to fall within the target physical property range compared with 125 in the removed subset. Again, there

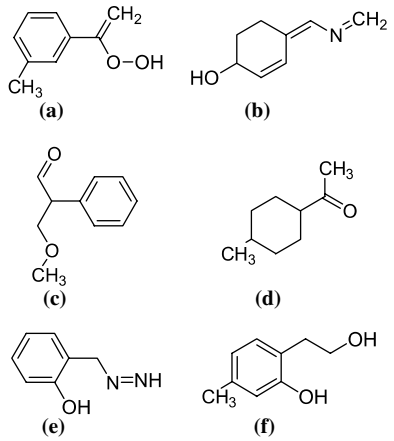

Figure 16. A six-member representative, diverse subset of the evolved median molecule set that are predicted by the validation models to lie within the log *S* property range under consideration.

Table 3. PLS model statistics for the aqueous solubility validation model developed in SIMCA-P with molecular descriptors generated using PETRA together with the number of latent variables (LVs) in the model.

Training set (1072)	R^2	0.835
	Q^2_{cum}	0.827
	RMSEE	0.825
	LVs	4
Test set 1 (21)	R^2_{Pred}	0.732
	RMSEP	1.059
Test set 2 (120)	R^2_{Pred}	0.905
	RMSEP	0.925
Removed subset (125)	R^2_{Pred}	0.021
	RMSEP	0.725

The number of structures in each set is given in brackets.

Table 4. PLS model statistics for the aqueous solubility validation model developed in SIMCA-P with molecular descriptors generated using Dragon together with the number of latent variables (LVs) in the model.

Training set (1072)	R^2	0.892
	Q^2_{cum}	0.866
	RMSEE	0.671
	LVs	6
Test set 1 (21)	R^2_{Pred}	0.825
	RMSEP	0.882
Test set 2 (120)	R^2_{Pred}	0.915
	RMSEP	0.852
Removed subset (125)	R^2_{Pred}	0.008
	RMSEP	0.685

The number of structures in each set is given in brackets.

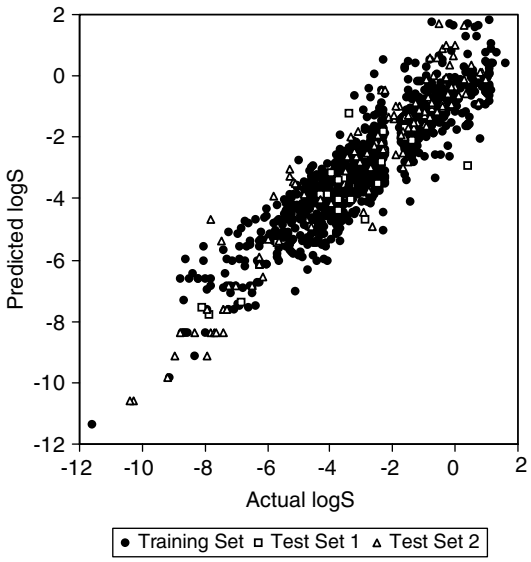

Figure 17. The actual vs. predicted values from the log *S* validation model using PETRA descriptors.

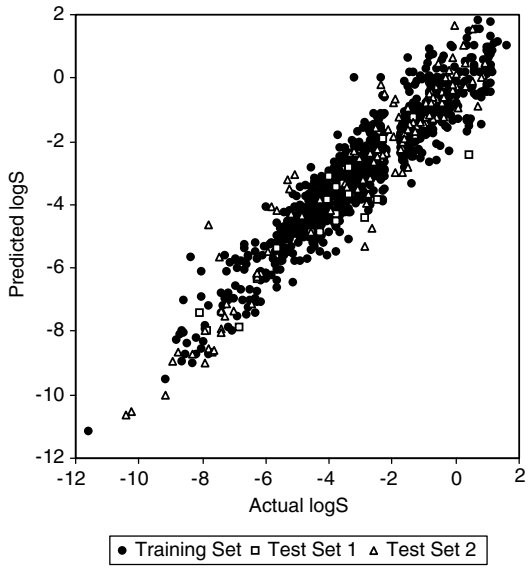

Figure 18. The actual vs. predicted values from the log *S* validation model using Dragon descriptors.

is no overlap between the log *S* dataset and the median molecule set, adding further credence to the position that this method is capable of designing novel structures. Furthermore, the distribution of the predicted log *S* values of the median molecule set is also centred within the target property range. Therefore, this suggests that this method is highly suitable for the design of

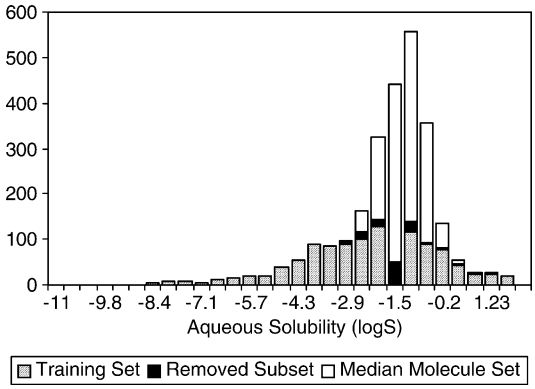

Figure 19. A histogram of the actual log *S* for the training set together with the predicted values for the removed subset and the median molecule set using the PETRA validation model.

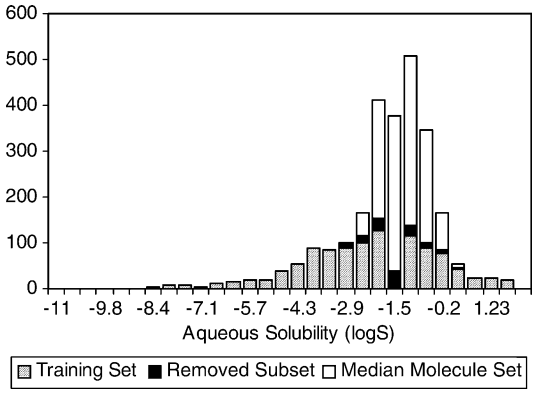

Figure 20. A histogram of the actual log *S* for the training set together with the predicted values for the removed subset and the median molecule set using the Dragon validation model.

novel molecules within an aqueous solubility range of interest.

Discussion and conclusions

The intention of the MMW case studies above was to attempt to evolve molecules that are similar to existing molecules of interest with the anticipation that, if the similar-property principle holds true, these molecules will also share the characteristics of the molecular similarity objectives.

It is evident by a comparison of the histograms of $\bar{\alpha}$ (Figures 12 and 13) and log *S* (Figures 19 and 20) that the distribution of log *S* values for the

evolved median molecules is centred within the target log S range, while the $\bar{\alpha}$ distribution is skewed to the right of the target range. Furthermore, while the predicted $\bar{\alpha}$ values for the removed subset have a low residual error, the predictions for the log S removed subset have a higher residual error that results in a similar distribution to that of the predictions of the evolved median molecules. This suggests that either, in the case of the $\bar{\alpha}$ experiments, the Fingal descriptors do not capture the structural characteristics that are important for the property, or that the similar-property principle does not hold for this property to the extent exhibited by the log S dataset.

To test the first hypothesis, we developed a model from the Fingal descriptors of the $\bar{\alpha}$ dataset to investigate whether or not the descriptors are capturing sufficient information to correlate with the experimental $\bar{\alpha}$ values. This resulted in a model with a Q^2_{cum} of 0.917, an RMSEP of 1.729 on the removed subset, and the plot of actual vs. predicted $\bar{\alpha}$ values in Figure 21. This model suggests that, while not as effective as the calculated $\bar{\alpha}$ values or the Dragon validation model for this property, the Fingal descriptors do exhibit a high correlation with the $\bar{\alpha}$.

There is no reason to believe that the similar-property principle, given its heuristic nature, will hold to the same extent in all cases. Indeed, the degree to which the similar-property principle holds true has already been investigated in the literature [16, 17] finding that a number of exceptions to the heuristic do exist.

Even given the limitations discussed here, both case studies still suggest more novel molecules, predicted to fall within the property range of interest by two validation models, than existed in the dataset itself: 26 and 37 molecules, respectively, against 24 in the case of $\bar{\alpha}$; and 390 and 337, respectively, against 125 in the case of log S. However, it is evident from these studies that it is good practice to apply at least one additional validation step to allow the focus to remain on the subset of median molecules that are likely to be of most interest.

The method by which median molecules are designed emphasises similarity to a small set of starting structures to explore structure space. The application published in this paper uses similarity as a means of investigating property space indirectly using the Tanimoto similarity of Fingal molecular hash-key fingerprints; although it is expected that other similarity coefficients and molecular hash-key fingerprints would provide comparable results. However, given the apparent degeneracies of the similar-property principle in some cases, it is likely that the direct exploration of property space will result in more structures that are predicted to lie within the physical property range of interest.

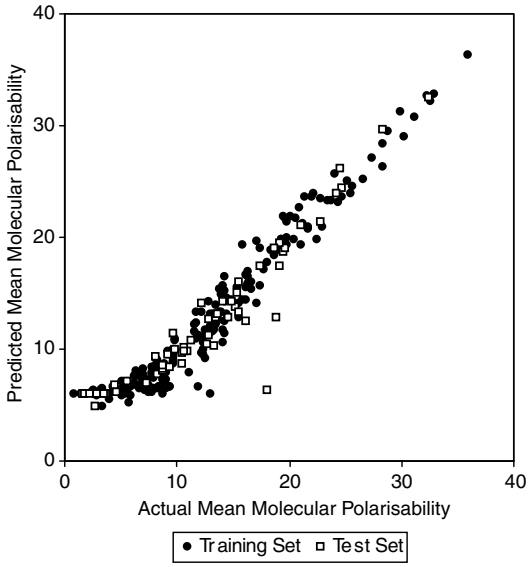

Figure 21. The actual vs. predicted values from the $\bar{\alpha}$ validation model using Fingal descriptors.

One method of exploring property space is the inverse problem [18], where an existing predictive model, in our case a QSPR, is provided with a target value. The system will then proceed to optimise structures that exhibit this property. Currently, we are developing a novel inverse QSPR workflow integrating CoG and Fingal to design novel structures within a physical property range of interest.

Acknowledgements

This research has been supported by a Marie Curie Fellowship of the European Community programme "Exploring leads in combinatorial catalysis for novel clean pharmaceutical/fine chemical processes" under contract number HPMI-CT-2001-00108.

References

1. Carhart, R.E., Smith, D.E. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.
2. Willett, P., Winterman, V. and Bawden, D., J. Chem. Inf. Comput. Sci., 26 (1986) 36.
3. Willett, P., Barnard, J.M. and Downs, G.M., J. Chem. Inf. Comput. Sci., 38 (1998) 983.
4. Brown, N., McKay, B., Gilardoni, F. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 44 (2004) 1079.
5. Brown, N., McKay, B. and Gasteiger, J., The 15th European Symposium on Quantitative Structure–Activity Relationships, 5–10 September 2004, Istanbul, Turkey.
6. Fonseca, C.M. and Fleming, P.J., In Forrest, S. (Ed.), Genetic Algorithms: Proceedings of the Fifth International Conference. Morgan Kaufmann, San Mateo, CA, 1993, pp. 416–423.
7. Handschuh, S., Wagener, M. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 38 (1998) 220.
8. Nachbar, R.B., Genet. Program. Evolvable Mach., 1 (2000) 57.
9. Johnson, M.A. and Maggiora, G.M. Concepts and Applications of Molecular Similarity. Wiley, New York, NY, 1990.
10. The SIMCA-P 10.5 software is available from Umetrics at http://www.umetrics.com/.
11. Miller, K.J., J. Am. Chem. Soc., 112 (1990) 8533.
12. The PETRA software is available from Molecular Networks, GmbH at http://www.mol-net.com.
13. The Dragon 4 software is available from Talete, Srl at http://www.talete.mi.it/.
14. Huuskonen, J., Salo, M. and Taskinen, J., J. Chem. Inf. Comput. Sci., 38 (1998) 450.
15. Yan, A. and Gasteiger, J., QSAR Comb. Sci., 22 (2003) 821.
16. Kubinyi, H. Perspect. Drug. Discov. Des., 11 (1998) 225.
17. Martin, Y.C., Kofron, J.L. and Traphagen, L.M., J. Med. Chem., 45 (2002) 4350.
18. de Julián-Ortiz, J.V., Comb. Chem. High Throughput Screen., 4 (2001) 295.