J Comput Aided Mol Des (2006) 20:333–341
DOI 10.1007/s10822-006-9063-1

ORIGINAL PAPER

A novel workflow for the inverse QSPR problem using multiobjective optimization

Nathan Brown · Ben McKay · Johann Gasteiger

Received: 3 May 2006 / Accepted: 3 August 2006 / Published online: 21 September 2006

Abstract A workflow for the inverse quantitative structure–property relationship (QSPR) problem is reported in this paper for the de novo design of novel chemical entities (NCE) in silico through the application of existing QSPR models to calculate multiple objectives, including prediction confidence measures, to be optimized during the de novo design process. Two physical property datasets are applied as case studies of the inverse QSPR workflow (IQW): mean molecular polarizability and aqueous solubility. The case studies demonstrate the optimization of molecular structures to within a property range of interest; the optimized structures are then validated against QSPR models that are generated from sets of alternative descriptors to those used in the IQW. The paper concludes with a discussion of the results from the case studies.

Keywords Inverse QSPR problem · QSAR · Genetic algorithms · De novo design · Partial least squares regression

N. Brown · B. McKay
Avantium Technologies B.V., P.O. Box 2915, 1000 CX Amsterdam, The Netherlands

J. Gasteiger
Computer-Chemie-Centrum and the Institute for Organic Chemistry, University of Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

N. Brown (✉)
Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland
e-mail: nathan.brown@novartis.com

Introduction

Quantitative structure–property relationships (QSPRs) have been demonstrated to be effective in process screening through application of multivariate regression methods such as partial least squares (PLS) regression. Process research and development in the pharmaceutical industry is an integral component of the drug development life cycle in terms of the design of synthesis routes of candidate drugs such that a sufficient yield is attainable [1]. QSPRs have been shown to be applicable in modeling properties, performance of reagents and enantioselectivity in asymmetric catalysis [2].

Typically, the application of QSPRs is to the forward problem where, given a new molecular structure, a set of molecular descriptors can be generated and a prediction calculated from the model, together with statistics that indicate the quality of those predictions. It follows that the inverse problem is the design of a structure or set of structures that exhibit a desired property. The inverse QSPR problem, therefore, is the application of an existing QSPR for the design of novel chemical entities (NCEs) that satisfy a given property range or objective, an approach that is generally referred to as de novo design. Several different approaches [see e.g. 3–5] to the inverse QSPR problem have been published [see 6, 7 for reviews of this area].

Recently, we have demonstrated the application of the median molecule workflow (MMW) [8, 9] as a method of designing NCEs within a property range of interest taking advantage of the similar-property principle [10]. The molecules that were created in that study were, by design, structurally similar to the specified objective molecules since the approach was

intended to explore structure space. Therefore, a method that enables the direct exploration of property space, rather than structure space, is the logical next step in the development of a de novo molecular design system.

In this paper we define a novel workflow for the inverse QSPR problem using multiobjective optimization in an approach intended to constrain the search space to within the space that is covered by the model. The inverse QSPR workflow (IQW) reported in this paper optimizes three separate objectives for one or more QSPR models: the predictions (relative to a target), residual standard deviation (RSD), and leverage. The methods together with the calculations from existing models are described in Sect. ''The inverse QSPR workflow'' of this paper. The general methodology of the experiments reported in this paper is provided in Sect. ''Methodology''.

The MMW was demonstrated by application of two physical property datasets: the mean molecular polarizability [11] and aqueous solubility [12] datasets. For reasons of consistency, we apply the same two datasets in this study, thereby allowing a direct comparison of the methods and results used in both papers. The results from these experiments are provided in Sects. ''Case study one: mean molecular polarizability'' and ''Case study two: aqueous solubility'', respectively. The paper concludes with an overview of the workflow and a discussion of the results from the case studies.

The inverse QSPR workflow

The IQW that is defined in this paper is a modular system allowing for alternative software programs to be incorporated into an adapted workflow as the need arises.

De novo design: CoG

The de novo design module of the IQW is satisfied in this instance by the compound generator (CoG) program. CoG is a genetic algorithm (GA) that operates directly on graph-based chromosomes that represent the molecules in a population. The chromosomes are graph-based where the nodes represent molecular atoms or substructures that are pre-defined in a fragment dictionary. The crossover and mutation operators implemented in CoG are constrained only by the valence bond model. CoG optimizes new molecules by iteratively scoring, sampling, and perturbing the current population of candidate molecules. Over a number of generations, which tends to be bound by the complexity of the search space, a set of molecules are designed that are suitable for the purpose for which they are required.

The reader is referred to Brown et al. [8] for details concerning the data structure, genetic operators and other algorithms implemented in the CoG de novo design program. Alternative GA-based de novo algorithms have been published in the past that operate on a number of molecular representations as the chromosome being optimized. Recently, Schneider and Fechner [7] published an extensive review of de novo design programs including many GA-based approaches.

Molecular descriptors: Fingal

The iterative nature of the IQW described herein renders it necessary to apply a molecular descriptor that is rapid to calculate, but which can also generate highly predictive models for the evaluation step. Fingal (*Fing*erprinting *Al*gorithm) is a molecular hash-key fingerprinting program developed for application to the MMW. However, since the Fingal descriptors can be generated rapidly and have also been demonstrated to be highly applicable to predictive modeling [13], the Fingal descriptors are the preferred choice for these initial studies of the IQW.

Partial least squares regression

PLS is a multivariate regression technique commonly used in Chemometric applications and in numerous QSPR/QSAR studies. It attempts to fit a linear model between, possibly correlated, input variables and an output variable (this single output version is commonly referred to as PLS1). PLS is of particular interest because, unlike Multiple Linear Regression, it can be used to analyze data sets with many strongly correlated input variables. PLS works by finding strong consistent variation in the input variables that also correlates with the output variable.

Consider an $(N \times K)$ input matrix $\mathbf{X}$ and an $(N \times 1)$ output vector $\mathbf{y}$, where N is the number of observations in the training set and K is the number of input variables. It is assumed that the $\mathbf{X}$ matrix is 'auto scaled' (each variable mean centered and scaled to unit variance) and that $\mathbf{y}$ is mean centered. For the purpose of defining the various entities referred to below, a PLS model may be expressed as follows in matrix notation:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{y} &= \mathbf{T}\mathbf{c}^{\mathrm{T}} + \mathbf{f} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{c}^{\mathrm{T}} + \mathbf{f} = \mathbf{X}\mathbf{b} + \mathbf{f} \end{aligned} \quad (1)$$

where $\mathbf{T}$ $(N \times K)$ is the input score matrix, $\mathbf{P}$ $(K \times A)$ and $\mathbf{c}$ $(1 \times A)$ are the input and output loadings respectively, $\mathbf{W}$ $(K \times A)$ is the input weight matrix, $\mathbf{b}$ $(N \times 1)$ is the vector of regression coefficients and $\mathbf{E}$ $(N \times K)$ and $\mathbf{f}$ $(N \times 1)$ are the input and output residual errors. PLS can be viewed as projecting both the $\mathbf{X}$ and $\mathbf{y}$ data onto a common low dimensional hyper plane. The values of the various model coefficients may be determined by means of the NIPALS algorithm. The reader is referred to [14] for a more thorough treatment.

Once a PLS regression model has been generated, it is typically applied to predict output values for new observations (as defined by a vector of input values). In addition to offering predictions of new observations, various model diagnostics can also be calculated, providing an indication of the confidence of the predictions. In this study we determine the RSDs and leverages of new observations. These are then considered together with model property predictions when determining the suitability of an evolved molecule during the de novo design process.

Predictions for new observations

The calculation of predicted property values of new 'observations' (i.e. evolved molecules) was required to indicate the extent to which evolved molecules satisfied the primary objective of the IQW. Once a PLS regression model had been generated (in this study, using SIMCA-P [15]), it was possible to extract the scaling parameters and regression coefficients, **b,** and calculate the predicted values of new observations by applying the necessary functions on the data.

First, the descriptor variables of a new observation, $\mathbf{x}_i$ $(1 \times K)$, were scaled using the mean and standard deviation of the variables in $\mathbf{X}$:

$$x_{s,i,j} = (x_{i,j} - \bar{x}_j)/\sigma_j \tag{2}$$

The predicted property values, $\hat{y}'$, of new observations were then calculated as:

$$\begin{aligned} \hat{y}_i &= \mathbf{x}_i\mathbf{b} \\ \hat{y}_i &= \sum_{j=1}^{K} (b_j \cdot x_{s,i,j}) \\ \hat{y}'_i &= \bar{y} + \hat{y}_i \end{aligned} \tag{3}$$

Residual standard deviation

The RSD is a model diagnostic indicating the distance of an observation from the hyper plane that has been defined by the regression. As such, it is a measure of the degree of extrapolation. The RSD of an observation is closely correlated with the *DModX* (Distance to the Model in the input space) as implemented by Umetrics [16]. A high RSD indicates a new observation that is far from the plane of the model and warns that model predictions may be less reliable. The RSD is therefore applied as an additional objective in the IQW.

The RSD of an observation, in the input space, is calculated as:

$$\mathrm{RSD}_i = \sqrt{\frac{\sum_{k=1}^{K} (e_{ik}^2)}{K - A}} \tag{4}$$

where e_{ik} is an element of the error vector for a new observation, $\mathbf{e_i}$ $(1 \times K)$, calculated as:

$$\mathbf{e_i} = \mathbf{x_i} - \mathbf{x_i} \cdot \mathbf{W}(\mathbf{P}^{\mathrm{T}} \cdot \mathbf{W})^{-1} \cdot \mathbf{P} \tag{5}$$

Leverage

The leverage is an additional model diagnostic that is complementary to the RSD, since it is the distance of a new observation from the model center when projected onto the hyper plane derived by the regression. The leverage is directly proportional to the Hotelling's T^2 statistic [17]. A high leverage indicates a new observation that is far from the center of the model and warns that model predictions may be less reliable. Therefore, the leverage of a new observation is applied as an additional objective to impose selection pressure in favor of those observations that have a higher confidence of an accurate prediction. The leverage of an observation (h_i) is calculated as follows:

$$h_i = \mathbf{t}_i \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{t}_i^{\mathrm{T}} \tag{6}$$

It is recognized that for complex (e.g. highly clustered) and distinctly non-Gaussian distributions of $\mathbf{X}$ data, the leverage statistic may not be a sufficient indication of extrapolation.

Multiobjective optimization: Pareto ranking

The simultaneous optimization of the prediction, RSD and leverage of new candidate solutions required the application of a multiobjective optimization strategy to this problem. To achieve a balanced optimization of all three objectives the Pareto ranking method [18] was

adopted, the same approach the authors have successfully applied to de novo molecular design [8, 9] and that is commonly used in other multiobjective optimization problems in Chemoinformatics [19–22]. Pareto ranking determines a rank position for each candidate solution in a population according to the number of other solutions that dominate it in all objectives. A solution is said to be non-dominated when no other solution exists in the current population that is more suitable in all of the objectives.

To prevent extrapolation, the RSD and leverage model diagnostics should be optimized to be as low as possible given the multiobjective constraints on the trade-off surface. The targets for these objectives are therefore set to zero. The distance of a predicted value (e.g. property value) for a candidate solution (i.e. evolved molecule) from a target value could simply be calculated as the absolute difference (or squared difference) between the predicted response of the candidate solution and the target value. However, this leads to issues regarding the relative dominance of predicted values that are above or below the target value. Therefore, in this work, predicted values that are less than the target and those that are greater than the target are considered as separate Pareto fronts.

Methodology

In this section the general methodology that is applied in each of the subsequent experiments in this paper is defined.

The IQW in each case is run for 10,000 generations with a population of 4000 molecule chromosomes, which are generated randomly within CoG from the atoms C, N, O, and Cl. The Pareto front in each generation is restricted such that it can constitute no more than half of any generation (i.e. 2000 individuals) Fig. 1.

The entire physical property dataset in each experiment is referred to as the *dataset*, while each dataset is also partitioned into a *training set* and at least one *test set*. To simulate a missing property range, a subset of the training set within a specified physical property range is excluded, which is referred to as the *removed subset*. Lastly, the set of evolved structures on the Pareto frontier in each run of the IQW is called the *Evolved set*. A schematic of this strategy is provided in Fig. 2.

The IQW requires a QSPR model (or models) to evaluate the candidate molecules that are generated by CoG, referred to throughout as the *query model*. In the absence of physical property values for the Evolved

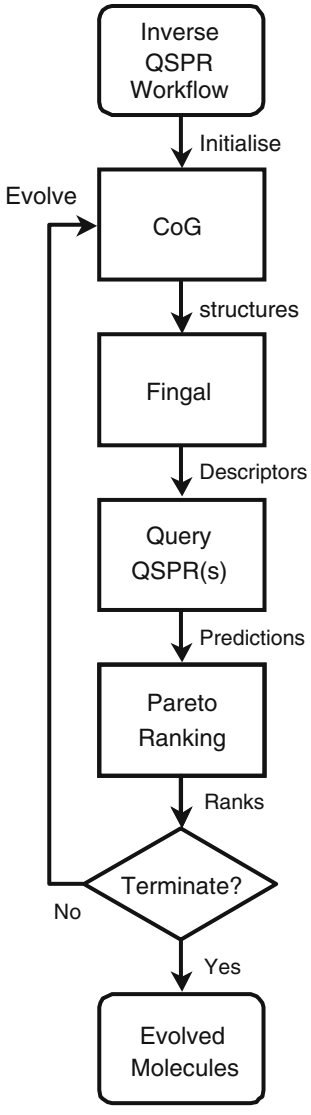

Fig. 1 Flowchart of the inverse quantitative structure–property relationship (QSPR) workflow

set, the method applied to validate these molecules is to apply them as queries to a QSPR model that has been derived using descriptors generated by Dragon [23]. These models are referred to throughout this paper as *validation models*, and are the same models as applied in [9]. The Dragon descriptors are used in the validation since they represent diverse sets of molecular descriptors that are substantially different from the Fingal descriptors. This suggests that similar predictions between the models will more likely correspond to the actual physical property value of the molecule under consideration.

The query and validations models, in each case, were derived using the PLS regression method in SIMCA-P [15]. The following performance statistics are reported for each of these models:

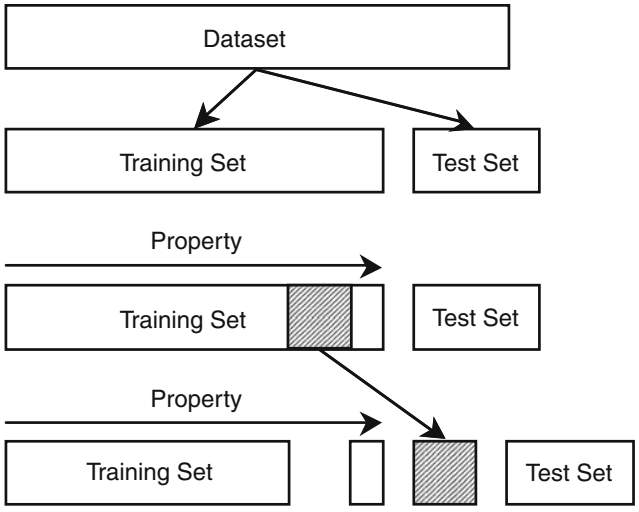

Fig. 2 Schematic of the partitioning methodology applied in this study. The dataset is first partitioned into a training set and at least one test set. The remaining training set is then ordered by property value and a contiguous subset removed, the removed subset, which is highlighted

- R^2, Q^2_{cum} (or cumulative cross-validated R^2 obtained by 7-fold CV) 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 response value range in each case. The number of LVs of each model is also provided with the model statistics detailed above.

Once the query models have been generated the required model coefficients and parameters were extracted from SIMCA-P.

Two programs were implemented to calculate the predictions, RSDs and leverages of new observations: *setupModel* and *queryModel*. The *setupModel* program calculates the mean and standard deviations of each descriptor variable together with the $(\mathbf{T}^T \cdot \mathbf{T})^{-1}$ matrix since the results of these calculations are required frequently in the calculation of the required values for new observations. The *queryModel* program is used to calculate the predicted value, RSD and leverage of each new observation in every generation.

Case study one: mean molecular polarizability

The mean molecular polarizability (MMP or $\bar{\alpha}$) 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 4 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 all structures (24) were removed from the training set between the $\bar{\alpha}$ values of 9.84 and 10.93.

The validation model applied in this case study is a PLS regression model generated using Dragon descriptors. The model statistics for the models are provided in Table 1. 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 *evolved set* were also calculated with Dragon; the descriptors were then loaded into SIMCA-P to calculate the predicted values.

A single model

Each of the two Fingal models was first applied in isolated runs of the IQW to optimize molecules that fit the specified property range. The sets of evolved molecules were then used to generate Dragon descriptors and predictions from the Dragon validation model based on these descriptors. The results from the two IQW runs with the two Fingal query models are provided in Figs. 3 and 4, respectively. In these figures, the actual distributions of the training set property values used to train both the query and validation models are provided together with the predicted property values using the validation model. The

Table 1 Partial least squares (PLS) model statistics for the mean molecular polarizability validation model derived in SIMCA-P with molecular descriptors generated using Fingal and Dragon, respectively

Dataset (size)		Fingal 1	Fingal 2	Dragon
Training set (203)	R^2	0.942	0.962	0.994
	Q^2	0.917	0.942	0.990
	RMSEE	1.755	1.426	0.558
	LVs	4	6	3
Test set (58)	R^2_{Pred}	0.878	0.913	0.993
	RMSEP	2.387	2.01	0.562
Removed subset (24)	R^2_{Pred}	0.015	0.055	0.273
	RMSEP	1.729	1.795	0.510

The number of structures in each set is given in brackets

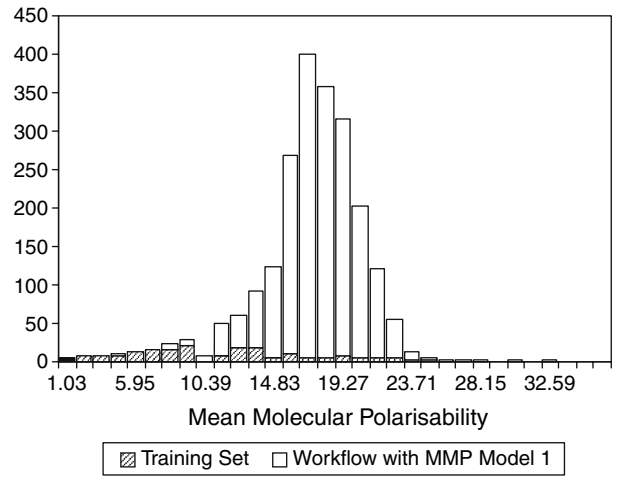

Fig. 3 A histogram of the actual mean molecular polarizability ($\bar{\alpha}$) for the training set together with the predicted values, using the Dragon validation model, for the set of evolved molecules from the workflow using the first $\bar{\alpha}$ Fingal model

distributions in both of these Figs. 3 and 4 indicate that the IQW is taking advantage of limitations in each of the Fingal query models that are not replicated in the Dragon validation model. This leads to structures that are predicted by the validation model to fall largely outside our property range of interest.

Two models

The limitations of using only one query model to evaluate the suitability of candidate molecules evolved in the IQW workflow are evident from Figs. 3 and 4. Therefore, in an attempt to overcome these limitations, the individual workflows from Sect. ''A single model'' were combined to utilize both models. The rationale for applying both models in the optimization stage is that each the models with be constrained by each other leading to more reliable predictions since the GA will not be able to take advantage of limitations in one model as the same limitations are unlikely to be present in the other model. This is similar to a jury or ensemble modeling approach [24]. The results from this run of the combined IQW are provided in Fig. 5 and some examples of the optimized structures are given in Fig. 6. It can be observed readily that the optimized structures from this run of the IQW predict well against the Dragon validation model when considering the target property range of interest with substantially more optimized structures being independently predicted to fall within the property range of interest.

Case study two: aqueous solubility

The second dataset considered in this study is the Huuskonen dataset [12] of aqueous solubility (logS) for 1342 structures that has been published as 3 partitions: 1201 in the training set and 21 and 120, respectively, in each of the test sets. An additional 4 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,

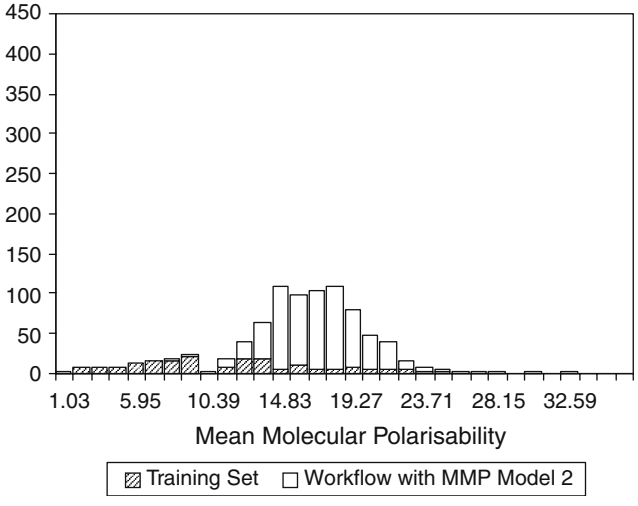

Fig. 4 A histogram of the actual mean molecular polarizability ($\bar{\alpha}$) for the training set together with the predicted values, using the Dragon validation model, for the set of evolved molecules from the workflow using the second $\bar{\alpha}$ Fingal model

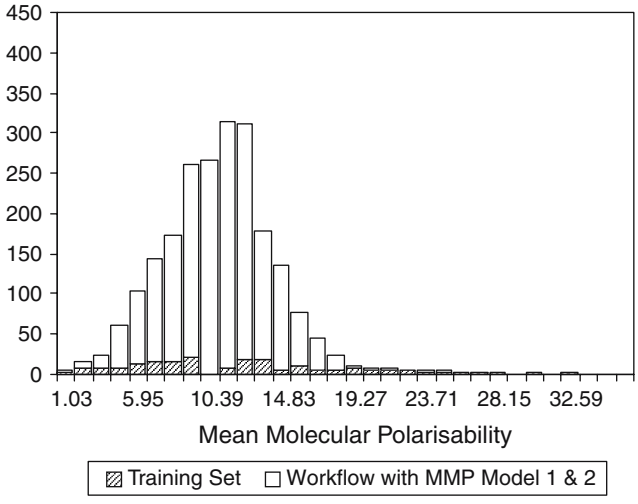

Fig. 5 A histogram of the actual mean molecular polarizability ($\bar{\alpha}$) for the training set together with the predicted values, using the Dragon validation model, for the set of evolved molecules from the workflow using both $\bar{\alpha}$ Fingal models

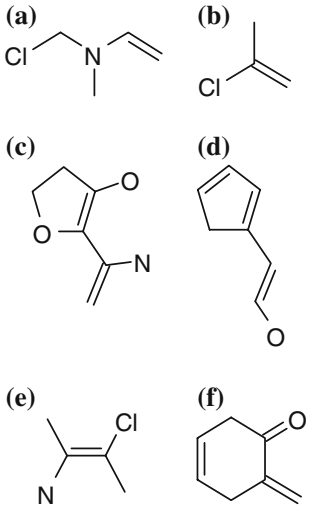

Fig. 6 Some examples of the types of molecules evolved with the inverse quantitative structure–property relationship (QSPR) workflow for median molecule workflow (MMW) with two models

by removing all 125 structures in the logS range from –2.22 to –1.76, to simulate a missing physical property value range.

The validation model was developed with PLS regression using the Dragon descriptors. The model exhibits 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 Table 2. These statistics provide a sufficient level of confidence that indicative predictions may be drawn from the model.

One model

The results of the runs of the IQW for logS with the two individual Fingal query models are given in Figs. 7 and 8, respectively. With these experiments it can once again be observed that, in Fig. 7 particularly, significantly fewer optimized molecules fall within the property range that was set as the target. However, it may also be observed that the second run of the IQW (Fig. 8) provides a substantial number of optimized structures that are also validated to fit our property range of interest.

Table 2 Partial least squares (PLS) model statistics for the aqueous solubility validation model derived in SIMCA-P with molecular descriptors generated using Fingal and Dragon, respectively

Dataset (size)		Fingal 1	Fingal 2	Dragon
Training set (1,072)	R^2	0.866	0.864	0.892
	Q^2	0.813	0.812	0.866
	RMSEE	0.747	0.752	0.671
	LVs	9	10	6
Test set 1 (21)	R^2_{Pred}	0.715	0.662	0.825
	RMSEP	1.137	1.286	0.882
Test set 2 (120)	R^2_{Pred}	0.888	0.902	0.915
	RMSEP	0.965	0.901	0.852
Removed subset (125)	R^2_{Pred}	0.051	0.043	0.008
	RMSEP	0.829	0.801	0.685

The number of structures in each set is given in brackets

Two models

Although the second run of the single query model IQW presented a large number of molecules that validated correctly against the Dragon validation model, for consistency and general investigation of applying two models in a single IQW we used both Fingal logS query models in a single IQW (Fig. 9), while six examples of the molecules optimized with this workflow are given in Fig. 10. The results here were somewhat less convincing than in the previous case, however a substantial number of NCEs were still independently predicted to fit the property range of interest.

Discussion and conclusions

The IQW has been presented in this paper as a new method of computer-aided molecular design (CAMD) that evolves NCEs that exhibit properties in a desired range. This is an advance on previous work by the authors [8, 9] where NCEs were evolved only to be structurally similar to a set of target molecules. The

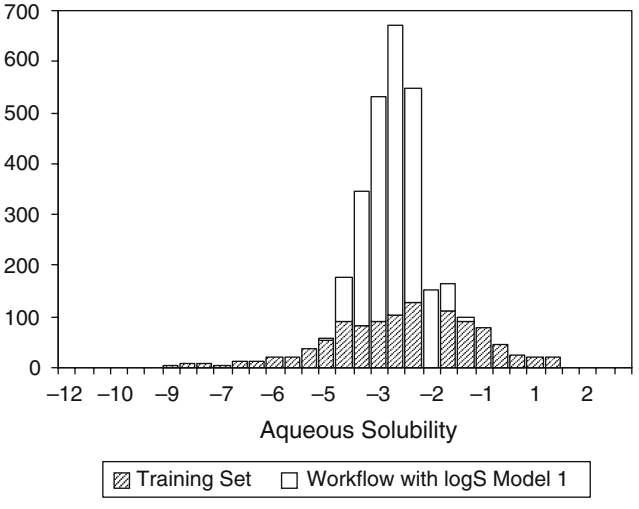

Fig. 7 A histogram of the actual aqueous solubility (logS) for the training set together with the predicted values, using the Dragon validation model, for the set of evolved molecules from the workflow using the first logS Fingal model

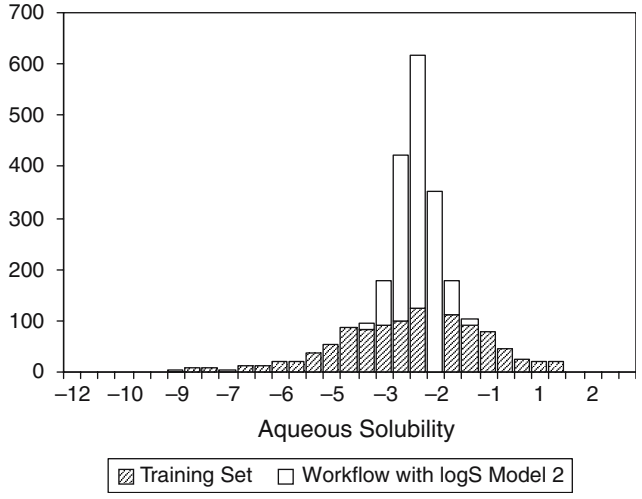

Fig. 8 A histogram of the actual aqueous solubility (logS) for the training set together with the predicted values, using the Dragon validation model, for the set of evolved molecules from the workflow using the second logS Fingal model

results suggest that multiple QSPR models can be effectively integrated into a CAMD strategy. Further, the simultaneous optimization of multiple objectives, including model extrapolation measures, appears to provide an effective means for evolving NCEs.

Using two query models in a single IQW has been found to provide greater consistency in returning molecules that still meet the targets when tested against an independent validation model. Although the results for logS are somewhat inconclusive, it is nevertheless anticipated that, given the approach reported here and current thinking in the literature regarding

Fig. 9 A histogram of the actual aqueous solubility (logS) for the training set together with the predicted values, using the Dragon validation model, for the set of evolved molecules from the workflow using both the logS Fingal models

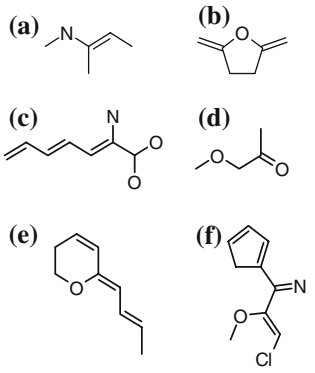

Fig. 10 Some examples of the types of molecules evolved with the inverse quantitative structure–property relationship (QSPR) workflow for logS with two models

the use of ensemble models and jury voting, the results of using this combined optimization system will still lead to more consistent results.

Although in this work we only consider models of a single property in any single workflow there is no reason why, given the Pareto ranking multiobjective scheme applied in these studies, that models for multiple properties cannot also be utilized in a single workflow. We anticipate that this will lead to inverse QSPR workflows that permit for the simultaneous optimization in multiple properties resulting in NCEs that will satisfy multiple property or performance criteria.

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. Federsel H-J (2003) Curr Opin Drug Discov Dev 6:838–847
2. McKay B, Hoogenraad M, Damen EWP, Smith AA (2003) Curr Opin Drug Discov Dev 6:966–977
3. Venkatasubramanian V, Chan K, Caruthers JM (1995) J Chem Inf Comput Sci 35:188–195
4. Skvortsova MI, Baskin II, Slovokhotova OL, Palyulin VA, Zefirov NS (1993) J Chem Inf Comput Sci 33:630–634
5. Kamphausen S, Höltge N, Wirsching F, Morys-Wortmann C, Riester D, Goetz R, Thürk M, Schwienhorst A (2002) Comput-Aid Mol Design 16:551–567
6. de Julián-Ortiz JV (2001) Combinator Chem High Throughput Screen 4:295–310
7. Schneider G, Fechner U (2005) Nat Rev Drug Discov 4:649–663
8. Brown N, McKay B, Gilardoni F, Gasteiger J (2004) J Chem Inf Comput Sci 44:1079–1087
9. Brown N, McKay B, Gasteiger J (2004) J Comput-Aid Mol Des 18:761–771
10. Johnson MA, Maggiora GM (1990). Concepts and applications of molecular similarity. Wiley, New York, NY

11. Miller KJ (1990) J Am Chem Soc 112:8533–8542
12. Huuskonen J, Salo M, Taskinen J (1998) J Chem Inf Comput Sci 38:450–456
13. Brown N, McKay B, Gasteiger J (2005) QSAR Comb Sci 24:480–484
14. Geladi P, Kowalski BR (1986) Anal Chim Acta 185:1–17
15. The SIMCA-P software is available from Umetrics, A.B. at http://www.umetrics.com/
16. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (xxxx) Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS)
17. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P (2003) Environ Health Perspect 111:1361–1375
18. Fonseca CM, Fleming PJ (1993) In: Forrest S (ed) Genetic algorithms: proceedings of the fifth international conference. San Mateo, CA, Morgan Kaufmann, pp 416–423
19. Handschuh S, Wagener M, Gasteiger J (1998) J Chem Inf Comput Sci 38:220–232
20. Agrafiotis DK (2001) IBM J Res Dev 45:545–566
21. Wright T, Gillet VJ, Green DVS, Pickett SD (2003) J Chem Inf Comput Sci 43:381–390
22. Cottrell SJ, Gillett VJ, Taylor R, Wilton DJ (2004) J Comput-Aided Mol De 18:665–682
23. The Dragon software is available from Talete, Srl. at http://www.talete.mi.it/
24. Sutherland JJ, O'Brien LA, Weaver DF (2004) J Med Chem 47:5541–5554