Use Of The R-Group Descriptor For Alignment-Free QSAR

Linda Hirons, John D. Holliday, Stephen P. Jelfs and Peter Willett[1]
Krebs Institute of Biomolecular Research and Department of Information Studies,
University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Peter Gedeck
Novartis Horsham Research Centre, Novartis Pharmaceuticals UK Ltd.,
Wimblehurst Road, Horsham, West Sussex, RH12 5AB, UK

Keywords Alignment-free, Molecular descriptor, R-group descriptor, Substituent

Abstract An R-group descriptor characterises the distribution of some atom-based property, such as elemental type or partial atomic charge, at increasing numbers of bonds distant from the point of substitution on a parent ring system. Application of PLS to datasets for which bioactivity data and R-group descriptor information are available is shown to provide an effective way of generating QSAR models with a high level of predictive ability. The resulting models are competitive with the models produced by established QSAR approaches, are readily interpretable in structural terms, and are shown to be of value in the optimisation of a lead series.

Type of manuscript Full paper

[1] To whom all correspondence should be addressed. Email p.willett@sheffield.ac.uk

INTRODUCTION

Methods for 3D QSAR are widely used for the discovery of novel bioactive molecules [1]. Although effective in operation, such methods suffer from the need to carry out a conformational analysis of the molecules in a dataset and then to align the resulting conformations. There is hence much interest in new QSAR approaches that do not require such 3D processing but that can still provide robust QSAR models with a high level of predictive ability. Examples of such approaches include EVA [2], GRIND [3], HQSAR [4], MaP [5], MoRSE [6] and WHIM [7]. In this paper we report the use of the R-group descriptor (RGD) in a new method for alignment-free QSAR studies. The descriptor was introduced in a study demonstrating that RDGs provide an effective way of distinguishing between bioisosterically equivalent and non-bioisosterically equivalent substituents [8]. Here, we describe the use of this descriptor for QSAR studies. The next section summarises the main features of the RGD and describes its application to QSAR. We then report its use with six literature datasets, comparing the results obtained with those from the existing HQSAR and EVA methods for QSAR and demonstrating the explanatory power of the approach. Finally, we illustrate its use in a simulated lead-optimisation programme.

THE DESCRIPTOR

Our approach is designed specifically for the analysis of sets of analogues, consisting of an invariant central ring scaffold that is substituted at one or more positions by various different groups of atoms. We define an R-group, *i*, in terms of a specific atomic property, *p*, and represent it by a descriptor D_i^p that is a vector of length *n* containing a series of values $d^p_{i,d}$. Each of the values in an RGD is the sum (or some other combination) of the chosen atomic property values at a distance of *d* bonds from the point of attachment of the substituent to the central ring scaffold, i.e.,

$$D_i^p := \{d_{i,d}^p\}_{d=1,n} = (d_{i,1}^p, \ldots, d_{i,n}^p)$$

The index *d* encodes the through-bond distance (i.e., the number of bonds) from the point-of-attachment of the R-group, so that small values of *d* describe positions on the

substituent close to the start, while larger values of *d* denote positions further away from the point of attachment to the common template. The furthermost atoms are assumed to be at a distance of *n* bonds from the point-of-attachment. This representation of molecular structure is similar in concept to the descriptor introduced by Martin *et al*. [9] for the design of structurally diverse combinatorial libraries, and is also related to the autocorrelation functions that have been reported by several workers [10-12].

The generation of an RGD is illustrated in Figure 1. The precise nature of the descriptor will depend on the particular properties that are used to characterise the atoms comprising a substituent. For example, Figure 2 shows substituent vector values for four different descriptors: atomic weight; atomic contribution to hydrophobicity; atomic contribution to molar refractivity; and hydrogen-bond acceptor count. In this case, the maximum through-bond distance is just three bonds; more generally, a vector will be as large as the largest substituent in the dataset that is being analysed, with the elements for smaller substituents being right zero-filled.

If multiple properties are used to characterise a molecule, as shown in Figure 2, then a molecule, *i*, is defined by a vector of the form

$$\mathrm{RGD}_i^{r,p} = \{\mathrm{RGD}_{i,d}^{r,p}\}_{d=1,n_{\max}^r}$$

where each vector is associated with a particular attachment point *r* and atomic property *p*, and where each element of the vector is associated with a particular distance *d*. The individual descriptor vectors (corresponding to each of the chosen properties) for each molecule are then appended to each other to form a complete representation for each molecule of the form:

$$\mathrm{RGD}_i = \{\mathrm{RGD}_i^{r,p}\}_{r=1,n_r;\,p=1,n_p}$$

where n_p is the total number of atomic properties studied and n_r is the total number of attachment points in the dataset. A dataset is thus represented by a descriptor matrix, with each row representing a particular compound *i* and each column containing the descriptor values for a particular attachment point *r*, atomic property *p* and distance *d*.

These matrices can then be scaled in order to give descriptors equal weight in the QSAR calculations. Our experiments (as discussed in the next section) involved auto-

scaling, in which each matrix column is recalculated to give a standard deviation of unity, and block-scaling, in which each block is recalculated to give a standard deviation of unity (with a block here being the vectors associated with a particular property). A QSAR analysis is carried out by using PLS to correlate the structural information encoded in the RGDs for the molecules in a dataset with the corresponding biological activity data. In our experiments, after scaling the matrix, each column is mean-centred and the resulting matrix then analysed using the PLS-1 algorithm defined by Geladi and Kowalski [13].

PREDICTIVE ABILITY OF THE DESCRIPTOR

The effectiveness of the approach was evaluated using four datasets, the scaffolds of which are shown in Figure 3. The benzodiazepine dataset [14] contained 57 benzodiazepin-2-ones represented by a common core with five positions of structural variation, and with binding affinities (log IC50) for the benzodiazepine $GABA_A$ receptor. The triazine dataset [15] contained 54 triazines represented by a common core with three positions of structural variation, and with anticoccidial potencies (log(1/MEC)). The tropane dataset [16] contained 62 phenyltropanes represented by a common core with three positions of structural variation, and with data for three transporters: seronin (5-HT), dopamine (DA) and norepinephrine (NA). The serotonin dataset contained 58 serotonin 5HT-3 ligands selected from a set of 75 described by Bureau *et al*. [17], so that all of the chosen molecules had a common core with two positions of structural variation.

The datasets encompass a range of possible core and substituent sizes, varying from the relatively large benzodiazepine core structure in the first dataset to the relatively small piperazine core structure of the serotonin dataset. In terms of the average number of heavy atoms, the benzodiazepine core structure is 1.3 times larger than the dataset's substituents, whereas the serotonin substituents are 3.4 times larger than the piperazine core structure. These relative size values could affect the models due to the common core's ability to fix the relative spatial orientation of the corresponding R-groups, i.e., relatively large substituents could be detrimental to the resultant QSAR model.

QSAR models were generated for a total of six biological targets across the four datasets. The substituents were characterised using a total of eight atomic properties: atomic weight; atomic positive charge; atomic negative charge; hydrogen-bond donor count (HBD); hydrogen-bond acceptor count (HBA); atomic contribution to molar refractivity (MR); atomic contribution to hydrophobicity (logP); and atomic contribution to polar surface area (PSA). These were calculated as described by Holliday *et al*. [8]

Each dataset was divided into both a training-set and a test-set: the training-sets were used to derive the QSAR models, which were subsequently validated using the associated test-sets. For the tropane dataset, the test-sets described in the original publication were also used here. For the remaining datasets, the test-set compounds were selected using an activity-based method, with compounds selected using the following steps:

1. Rank the dataset compounds in order of increasing activity;
2. Select the initial compound at the rank position closest to $0.5\times {}^{n}\!/_{n_{test}}$, where n is the number of compounds in the dataset and n_{test} is the final number of compounds in the test-set.
3. Select additional compounds at ${}^{n}\!/_{n_{test}}$ intervals along activity-rank distribution.

The resultant test-sets contained between 10 and 15% of the original compounds.

The resulting models were assessed using the statistical measures r^2, s and F, with leave-one-out cross-validation being used to estimate the predictive ability of the models in terms of q^2 and s_{CV}. Finally, the predictive ability was also estimated by calculating the predictive r^2, pr^2, for the test set compounds. In the experiments, autoscaling proved to be superior to blockscaling, and we have hence included only the results from the former approach; however, both autoscaling and blockscaling were noticeably superior to the use of raw data.

The results obtained are detailed in Table 1, in the rows marked RGD. These values have been obtained using both the internal and the external estimates of predictive-

ability for the selection of an optimum number of latent variables for each QSAR model. Specifically, an increasing number of latent variables was used until there was no significant improvements (>5%) in the q^2 and pr^2 values. The maximum number of latent variables was limited to six in order to minimise the chance of over-fitting. Inspection of the table demonstrates that, excluding the serotonin dataset, all of the models show a very good fit to the training-set data, with r^2 values greater than 0.81 and a negligible probability of r^2 equalling zero ($F >> F_{0.01}$). The predictive abilities of these models are also very good, with q^2 values greater than 0.55 and pr^2 values greater than 0.63. For the serotonin dataset, the poorer results support the previously expressed view that the descriptor may not be ideally suited to such datasets; even so, the small number of latent variables and the reasonable pr^2 value suggest that even this model is far from useless.

The results obtained using the RGDs have been compared with those obtained using two existing methods for alignment-free QSAR: EVA (for EigenValue Analysis) [2] and HQSAR (for Hologram QSAR) [4]. EVA characterises the 3D structure of a molecule in terms of the vibrational frequencies encoded in its infra-red spectrum. Specifically, a standardised spectral profile is generated, based on summed Gaussian kernels, that is then sampled to give the final descriptor. The descriptors for the molecules in a dataset can then be correlated with biological activity using partial least squares (PLS). HQSAR is a 2D technique, which uses structural fingerprints that take account not just of the presence of fragments in a molecule (as in a conventional fragment bit-string) but also of the frequency of occurrence of those fragments. A hashing procedure is used to map fragments to positions in the hologram, and PLS is again used to correlate this structural information with the biological activity data. The Tripos implementations of both EVA and HQSAR were used in our experiments [18].

For the HQSAR models, molecular holograms were derived from fragments containing between four and seven adjacent atoms. These were defined in terms of: their constituent atoms; their constituent atoms and bonds; their constituent atoms, bonds and the connectivity of the atoms; and their constituent atoms, bonds, the connectivity, and also hydrogen-bonding features. The resultant descriptors were encoded using the default hologram lengths of 97, 151, 199, 257, 307 and 353. In

each case, the optimum settings (fragment type and hologram length) were selected automatically via the QSAR model that produced the minimum value of s_{CV}, overall. The vibrational frequencies in the EVA tests were calculated by normal coordinate analysis using the semi-empirical AM1 method. The best results were obtained by sampling the vibrational spectrum from 200 to 4000 cm^{-1} and at $5cm^{-1}$ intervals, with the individual frequencies represented by Gaussians of width $10cm^{-1}$.

The optimum models obtained with these two QSAR methods are also included in Table 1. Inspection of this table demonstrates that the RGD method is fully competitive with the established EVA and HQAR methods across the full range of statistical parameters. Thus, the average value of q^2 for the RGD models of Table 1 is 0.58, the average r^2 is 0.81 and the average pr^2 is 0.78. The corresponding three figures for the EVA models are 0.63, 0.95 and 0.70, respectively, while those for the HQSAR models are 0.60, 0.82 and 0.65, respectively.

EXPLANATORY ABILITY OF THE DESCRIPTOR

The results presented in Table 1 are very satisfying, given the simplicity of the RGDs that we have used. However, if a QSAR method is to be of general applicability, then the models resulting from its use must not only offer a high level of predictive ability but they must also be interpretable, so as to describe qualitatively the important structural relationships and so as to facilitate the design of new analogues that can further increase potency. To be able to do this, we need to look at the nature of the RGD in more detail.

All of the results thus far have been based on the use of all eight atomic properties described above. However, our previous similarity study [8] has shown the very different contributions that the various properties can make to the similarities between substituents and we would expect such effects to be much greater here, where specific types of activity are dependent on specific types of molecular property. We have hence developed a procedure based on the generating optimal linear PLS estimates (GOLPE) methodology of Baroni *et al*. [19]. GOLPE essentially removes variables that have a detrimental effect on the predictive ability of a model. The effect of each

variable being either included or excluded from a model is assessed in terms of the average change inflicted upon the standard error of prediction s_{DEP}:

$$E = SDEP_{+} - SDEP_{-}$$

where, for a given variable, $SDEP_{+}$ is the average value of s_{DEP} for the models that include the variable and $SDEP_{-}$ is the average value of s_{DEP} for the models that exclude the variable. Our approach simply extends this idea by removing blocks from the descriptor matrix that are associated with an atomic property, rather than individual values. The assessment of properties is made by deriving multiple models based on different, reduced sets of the properties. The result of deriving these multiple models, 255 in all, is an average effect *E*, which can be used to rank the properties in terms of predictive ability. The results are shown in Table 2 and Figure 4, where properties that are beneficial have negative *E* values and those that are detrimental have positive *E* values. With the exception of the serotonin dataset (discussed previously), logP appears to have the most beneficial effect overall, closely followed by atomic weight and molar refractivity; polar surface area seems to be slightly detrimental, although it is not obvious why this is so..

Optimised QSAR models, using only the beneficial atomic properties, were generated for each dataset. The results were not significantly different in terms of predictive ability to those in Table 1 and have thus not been included here; however, their greater simplicity should imply that they are both more robust and easier to interpret than the conventional RGD models discussed thus far. Specifically, the regression coefficients associated with these simplified QSAR models have been standardised by multiplying the coefficient by the standard deviation of the associated descriptor matrix column; this is similar to the methodology employed for generating CoMFA plots and results in a diagram such as that shown in Figure 5 [20]. The standardised plots here are organised so that coefficients implying that an increase in a particular variable leads to an increase in activity, are shown on the upper side of the plot (i.e. irrespective of the unit-of-measurement employed for the activity). The bars representing the individual coefficients are also coloured by atomic property and sorted in terms of their associated distance and attachment point. A typical example of such a plot, for the benzodiazepines, is shown in Figure 5.

An inspection of Figure 5 suggests that substituents in the R_1 position have little influence on activity. However, the logP, MR, and weight coefficients suggest that unsubstituted compounds in this position are slightly preferred: with the coefficient values next to the attachment point (d=1) favouring an increase in these steric properties (i.e., due to a hydrogen being attached to the dummy atom at the point of attachment), while the coefficient values further along these groups (d>1) favour a decrease in the steric properties. For the R_3 and R_5 positions, the substituent properties also appear to have little influence on activity. However, substitution of these positions is not explored particularly well by the dataset, with the majority of compounds being unsubstituted.

At the R_2 position, the compounds in the dataset consist of meta-substituted phenyl groups. The high –Charge coefficient (marked on the figure at 1.) indicates that a positive or neutral atom is favoured. However, any difference in charge at this position (d=1) is due entirely to the inductive effects caused by the meta substituents of the phenyl groups. Implicitly, this suggests that the mono- or di-substituted phenyl groups with either the chlorine or fluorine halogens will be favoured.

Overall, the R_4 position appears to affect the activity to the greatest extent, and it is also the position that is most thoroughly explored by the dataset. For the first atom (d=2) of the substituents (marked at 2.) a high logP and atomic weight and a low HBD are favoured; in other words, any heteroatom in this position is preferred with the exception of the amine group (a hydrogen-bond donor). Heavy atoms adjacent to this position (marked at 3.) are also favoured, with the nitro group being the optimum choice overall.

RETROSPECTIVE SERIES DESIGN

We have shown that the RGD is able to produce models that have both predictive and descriptive abilities: we now demonstrate the extent of these abilities by means of a simulated lead optimisation study. This study used three datasets that had been assayed in GCPR (G-coupled protein receptor), kinase and PDE (phosphodiesterase) programmes at Novartis: the precise natures of the compounds are proprietary, but

some summary details of these three datasets are included in Table 3. Each dataset contained between 500 and 600 compounds, and was based upon a common core structure. The activity of each compound was measured in pIC_{50} units, the values ranging from around 4.0 (for inactive compounds) to around 9.0 (for active compounds). Since the values for the extremely inactive compounds (over 13% and over 40% of the kinase and PDE datasets, respectively) were subject to a high degree of experimental error, they were assigned an arbitrary cut-off value of 4.0. The structures were characterised by RGDs based on the eight atomic properties that have been discussed previously, with autoscaling being applied prior to the derivation of the QSAR models. For comparison, molecular holograms of length 53 and 151 were generated with fragment definitions based upon the constituent elements, bond types and the connectivity of the atoms within each of the structures.

The simulated lead optimisation procedure began with the selection of compounds for the initialisation step followed by a series of optimisation steps, in each of which further compounds were retrieved from the particular dataset under investigation. In all cases, 50 compounds were retrieved in both the initialisation and optimisation steps, with the exception of the final optimisation step (in which only the remaining compounds could be retrieved). This resulted in around ten optimisation steps being performed for each experiment.

The initialisation steps were performed using two different compound initialisation methods, referred to as *chronological* and *diverse* selection. Chronological initialisation was based upon the historical order of the compounds within the corporate archive and, in each case, involved retrieving the earliest recorded compounds. This initialisation is likely to be relatively focussed, containing only a few lead compounds and limited amounts of information, so that the subsequent QSAR models are expected to be relatively poor. Diverse initialisation, conversely, was based upon the identification of 50 structurally diverse compounds from the entire dataset, using a Pipeline Pilot version 3.0 utility based on structural fingerprints and the Tanimoto coefficient [21]. This approach hence takes account of the full range of structures that were considered in the programme and should thus increase the predictive ability of the QSAR models resulting from their use [22].

The optimisation steps were performed using three different approaches to compound selection. The principal approach, and the focus of the study here, is to base the selection of new compounds on QSAR models derived from the compounds that have been analysed thus far in the programme, with the models being used to prioritise the hereto untested compounds. The optimum number of latent variables used in the QSAR models used to select further compounds was determined automatically. This was achieved by generating all possible models with between one and five latent variables and then selecting the model with the lowest value of s_{CV}. It should be noted that this procedure resulted in models containing relatively few latent variables, thus minimising the chances of over-fitting the previously retrieved compounds. For comparison, *chronological* and *random* selection methods were also employed. As with the initialisation step, chronological selection involved retrieving the earliest recorded compounds present in the remainder of the dataset. Random selection simply involved retrieving further compounds at random, and was repeated three times for each of the initialisation methods and datasets under investigation.

The effectiveness of the selections was assessed in terms of the median activities of all of the previously selected compounds up to and including each step of an experiment. In each of the plots here, the Y axis denotes the median activity of the selected compounds, and the X axis denotes the current step of the simulation. The various selection methods that were tested are represented as follows: chronological is a grey solid line; random is a grey dotted line (in fact, three grey dotted lines as the random simulations were performed three times); RGD is a red solid line; hologram-51 is a blue solid line; and hologram-151 is a green solid line.

Given an initialisation method, all of the selection methods will have the same initial median value, i.e., the median of the activities for the set of compounds selected using the chosen initialisation method. The median of the final optimisation step will be identical for all simulations, this being the median activity for all of the compounds comprising a dataset. The median values for the intervening steps should be greater than both the initial median (first step) and the overall median (last step): this indicates that a particular selection method is performing well, i.e., it is able to enrich the currently selected compounds. Furthermore, it is the hope that the median values for the QSAR-based selection methods (i.e., those that use the RGDs and the

holograms) should be greater than the values for random and chronological selection. The median plots that were obtained for the three datasets are shown in Figure 6.

The GCPR simulations are shown in Figure 6(a). When chronological initialisation is used, all of the three QSAR-based selection methods out-perform the random and chronological selections after just a few steps; this is the case after the third step for the RGDs, the fourth step for Hologram-151 and the sixth step for Hologram-53. It is clear, however, that the increase in the median values is delayed by the poor selections made by the QSAR models during the first two or three steps. Thereafter, the selected compounds appear to be representative enough to allow reasonable QSAR models to be derived. The results are still more encouraging when diverse initialisation is used, as all the QSAR-based methods markedly out-perform random selection throughout the entire optimisation.

The kinase simulations are shown in Figure 6(b). With chronological initialisation, a clear degree of optimisation is apparent for all the methods from the second step (the first optimisation step); however, many more steps are required until the majority of the active compounds have been retrieved. Indeed, two peaks are observed for all of the QSAR-based selection methods, rather than the ideal of a single near to the start of the simulation: this suggest that the QSAR models are not adequately describing all of the structural classes in the dataset until much later in the optimisation. Overall, however, the three QSAR-based methods out-perform both chronological and random selection, and this distinction is still more evident when diverse initialisation is used, with the QSAR-based plots rapidly pulling away from the three random selection plots.

The PDE simulations are shown in Figure 6(c). The results here are much worse than for the other two datasets when chronological initialisation is employed; we believe that this is due to this dataset having far more inactive compounds, both within the initial selection and within the dataset as a whole. The QSAR-based selections here are not obviously superior to the random selections, although far better than chronological selection. This is because the dataset contained a number of very potent compounds that were structurally very different to the compounds studied at the start and which were only found at a late stage of the program. With diverse

selection, however, the QSAR-based selections markedly out-perform random selection even by the second step (i.e., the first optimisation step), demonstrating the greater power of diversity-based screening.

Inspection of the median plots suggests that there is little obvious difference in performance between the three QSAR-based selection methods for the kinase and PDE datasets, and that the RGD selections out-perform the two hologram selections for the GCPR dataset. These observations are broadly supported by the figures in Table 4, which summarises the QSAR models generated in the final step of the simulations. The table shows that all three methods produce predictive QSAR models even with these relatively large datasets containing many inactive compounds.

CONCLUSIONS

In this paper, we have described the use of the R-group descriptor for QSAR studies. Application of PLS to datasets for which bioactivity data and R-group descriptor information are available is shown to provide an effective way of generating QSAR models with a high level of predictive ability. The resulting models are competitive with the models produced by established QSAR approaches, are readily interpretable in structural terms and are shown to be of value in the optimisation of a lead series. It should be noted that the R-group descriptor approach is an entirely general one, in the sense that it can be extended to more complex molecular representations. The descriptors used here are based on the underlying molecular topology, with distances calculated on the basis of numbers-of-bonds separations between pairs of atoms. An entirely comparable descriptor can be generated using Euclidean (i.e., through-space rather than through-bond) inter-atomic separations, or the Euclidean separations between points on a molecular surface [23]. However, even the simple topological representations used here demonstrate that the RGD concept provides a conceptually simple, and computationally effective, way of investigating structure-activity relationships.

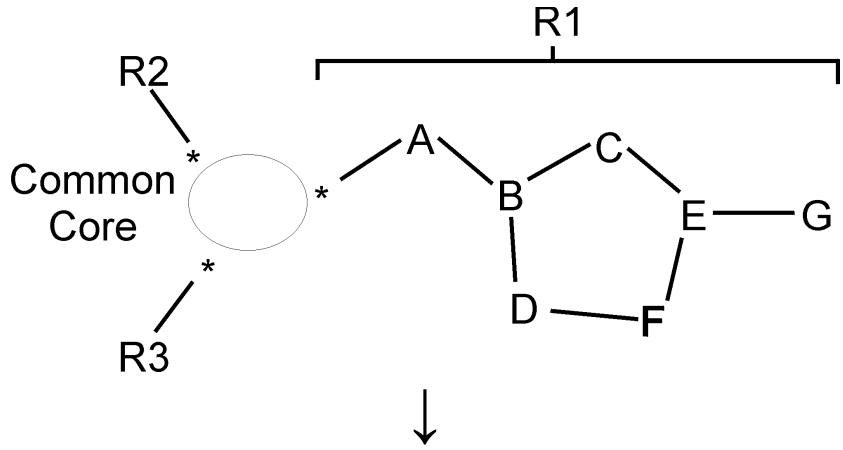
Acknowledgements. We thank Novartis Pharmaceuticals UK Ltd and the Biotechnology and Biological Sciences Research Council for funding SJ, the

Engineering and Physical Sciences Research Council for funding LH, and the Royal Society, Tripos Inc. and the Wolfson Foundation for software and hardware support. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES

1. H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), *3D QSAR in Drug Design.* Kluwer/ESCOM, Leiden, **1998**.
2. A. M. Ferguson, T. W. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan, P. J. Snaith, *J. Comput.-Aid. Mol. Design* **1997**, *11*, 143-152.
3. M. Pastor, G. Cruciani, I. Mclay, S. Pickett, S. Clementi, *J. Med. Chem.* **2000**, *43*, 3233-3243.
4. W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage, D. M. Sheehan, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669-677.
5. N. Stiefl, K. J. Baumann, *J. Med. Chem*. **2003**, *46*, 1390-1407.
6. M. Wagener, J. Sadowski, J. Gasteiger, J. Am. Chem. Soc. **1995**, *117*, 7769-7775.
7. R. Todeschini, M. Lasagni, E. Marengo, *J.Chemomet.* **1994**, *8*, 263-273.
8. J. D. Holliday, S. P. Jelfs, P. Willett, P. Gedeck, *J. Chem. Inf. Comput. Sci*.**2003**, *43*, 406-411.
9. E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, W. H. Moos, *J. Med. Chem*. **1995**, *38*, 1431–1436.
10. P. Broto, G. Moreau, C. Vandycke, *Eur. J. Med. Chem*. **1984**, *19*, 66-70.
11. G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chemie Int. Ed*., **1999**, *38*, 2894-2896.
12. C. T. Klein, D. Kaiser, G. Ecker, *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 200-209.
13. P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1-17.
14. D. J. Maddalena, G. A. R. Johnston, *J. Med. Chem.* **1995**, *38*, 715-724
15. J. W. McFarland, *J. Med. Chem.* **1992**, *35*, 2543-2550.
16. I. C. Muszynski, L. Scapozza, K. A. Kovar, G. Folkers, *Quant. Struct.-Act. Relat.* **1999**, *18*, 342-353.

17. R. Bureau, C. Daveu, I. Baglin, J. S.-D. O. Santos, J. C. Lancelot, S. Rault, *J. Chem. Inf. Comput. Sci. 2001*, 41, 815-823.
18. The EVA and HQSAR software packages are available from Tripos Inc. at http://www.tripos.com
19. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, *Quant. Struct.-Act. Relat.* **1993**, *12*, 9-20.
20. R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
21. The Pipeline Pilot software package is available from SciTegic at http://www.scitegic.com
22. A. Golbraikh, A. Tropsha, *J. Comput.-Aid. Mol. Design* **2002**, *16*, 357--369.
23. S. P. Jelfs, *Development of a Novel Descriptor Targeted to High-Throughput Analysis in Lead Exploration and Combinatorial Library Design.* PhD thesis, University of Sheffield, **2004**.



	Distance				
	1	2	3	4	5
Sum of atomic values	*a*	*b*	*c+d*	*e+f*	*g*

R1 RGD = {*a*, *b*, *c+d*, *e+f*, *g*}

Figure 1. Generation of the RGD for the substituent R1, where *a-g* are the atomic property values for the atoms A-G.

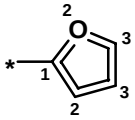
		Distance		
	Property	1	2	3
	Atomic Weight	12.01	29.02	26.04
	Hydrophobicity	0.08	0.44	0.56
	Molar Refractivity	3.24	5.49	8.81
	Hydrogen Bond Acceptor	0.00	1.00	0.00

Figure 2. R-group descriptors based on four different atomic properties.

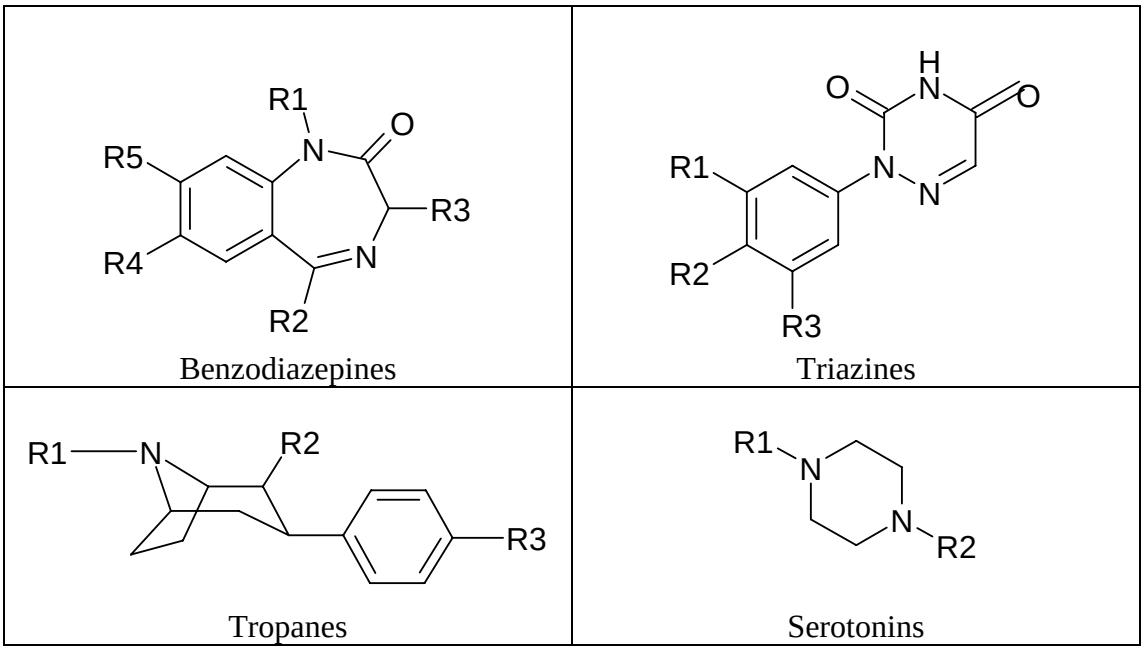


Figure 3. Ring scaffolds for the four datasets.

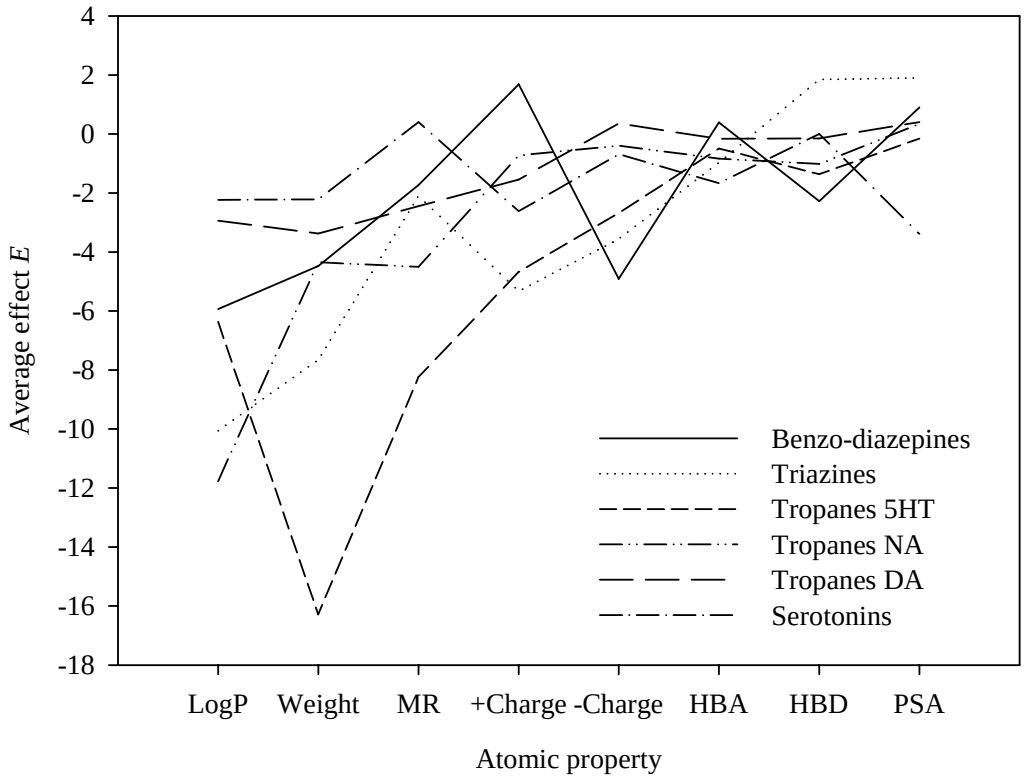


Figure 4. Average effect *E* of individual atomic properties (negative values are beneficial, positive values are detrimental). Properties are ordered by average of *E*.

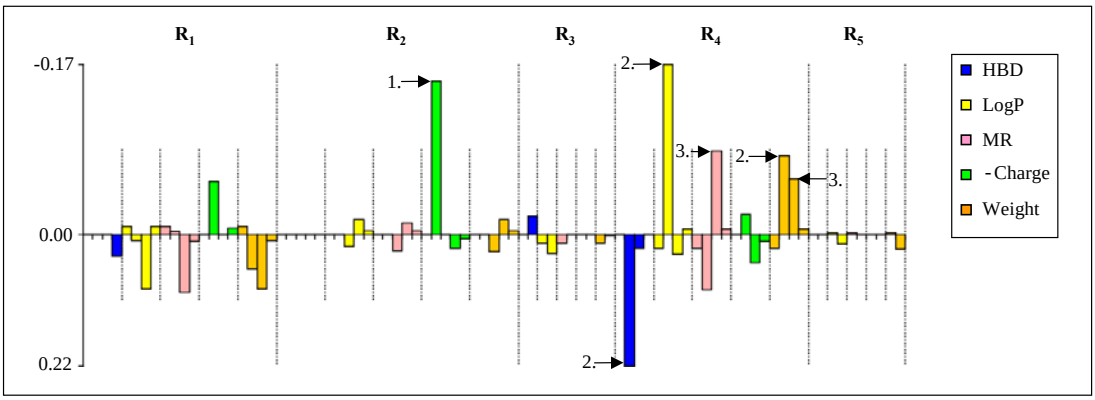
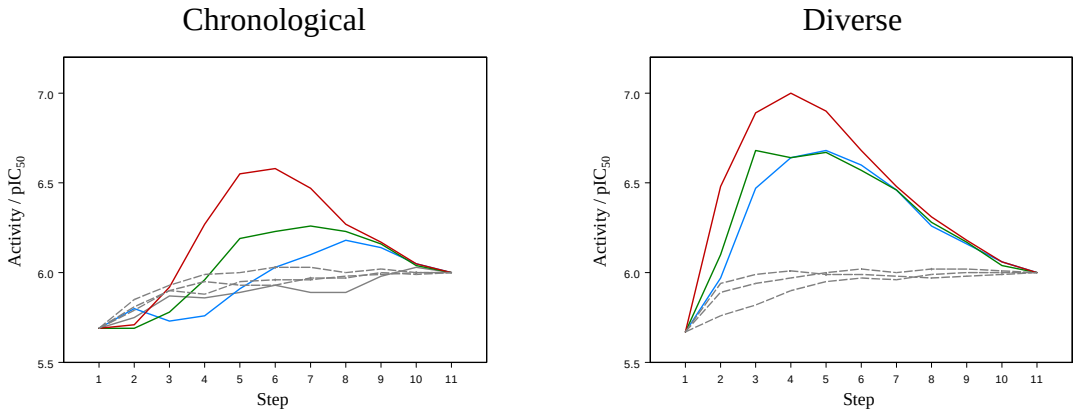
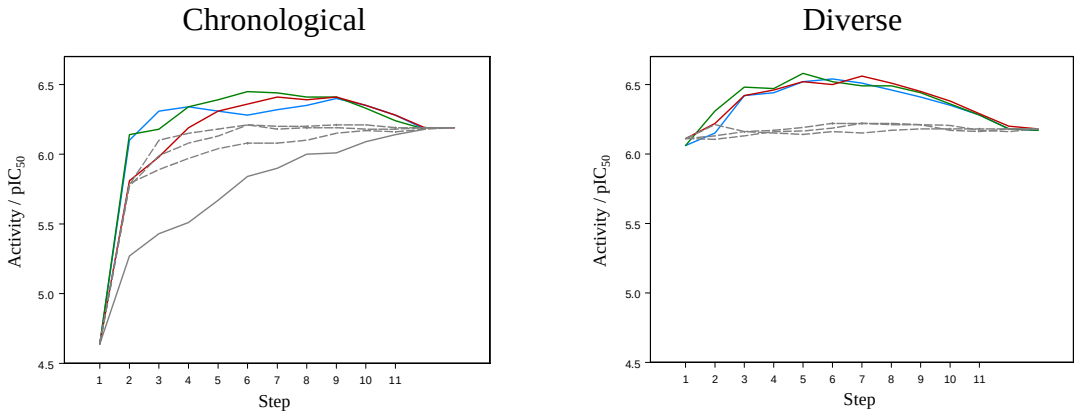


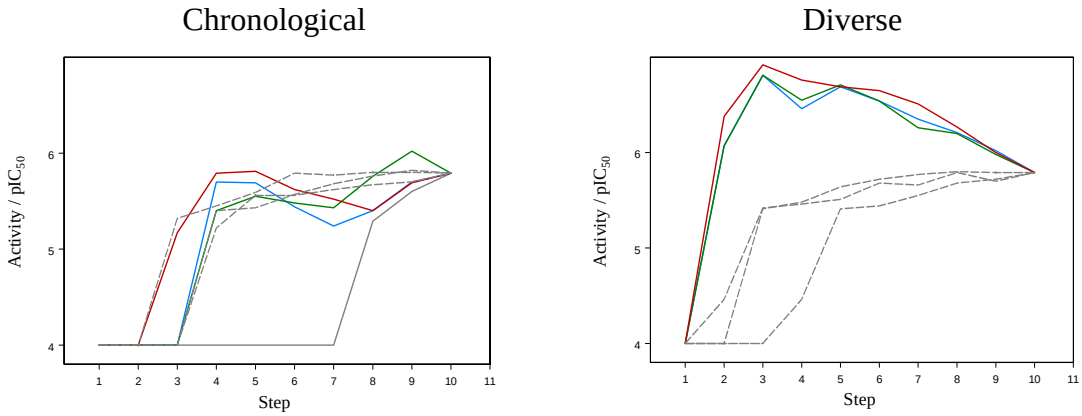
Figure 5. Visualisation and interpretation of the QSAR model for the benzodiazepine dataset. See text for explanation.



6(a)



6(b)



6(c)

Figure 6. Median activity plots for the (a) GCPR, (b) kinase and (c) PDE datasets. In these plots, chronological is a grey solid line; the three random experiments in each case are the three grey dotted lines, RGD is a red solid line, hologram-51 is a blue solid line, and hologram-151 is a green solid line.

Dataset	Method	N	r^2	s	F	q^2	s_{CV}	pr^2
Benzodiazepines	RGD	5	0.86	0.28	51	0.56	0.49	0.64
	HQSAR	6	0.86	0.29	42	0.61	0.47	-0.16
	EVA	4	0.92	0.22	118	0.53	0.50	0.60
Serotonin	RGD	2	0.55	1.44	29	0.26	1.84	0.61
	HQSAR	5	0.80	1.00	35	0.52	1.53	0.72
	EVA	6	0.91	0.67	75	0.50	1.58	0.55
Triazines	RGD	4	0.86	0.40	63	0.61	0.66	0.91
	HQSAR	2	0.72	0.55	55	0.51	0.72	0.85
	EVA	7	0.98	0.16	240	0.62	0.67	0.69
Tropanes 5-HT	RGD	4	0.93	0.41	163	0.85	0.60	0.84
	HQSAR	6	0.95	0.36	143	0.88	0.54	0.96
	EVA	5	0.98	0.24	388	0.89	0.51	0.90
Tropanes NA	RGD	4	0.83	0.36	60	0.63	0.53	0.80
	HQSAR	6	0.81	0.40	32	0.51	0.63	0.88
	EVA	5	0.89	0.30	75	0.56	0.59	0.79
Tropanes DA	RGD	6	0.82	0.35	37	0.56	0.56	0.85
	HQSAR	6	0.79	0.39	30	0.55	0.57	0.60
	EVA	6	0.92	0.25	85	0.57	0.56	0.57

Table 1. Summary of QSAR models obtained using RGD, HQSAR and EVA methods.

	Dataset					
Atomic Property	Benzo-diazepines	Triazines	Tropanes 5HT	Tropanes NA	Tropanes DA	Serotonins
LogP	–5.93	–10.06	–6.36	–11.77	–2.94	–2.24
Weight	–4.47	–7.66	–16.28	–4.35	–3.38	–2.22
MR	–1.73	–2.11	–8.24	–4.51	–2.45	+0.40
+Charge	+1.68	–5.33	–4.68	–0.73	–1.55	–2.62
-Charge	–4.91	–3.56	–2.68	–0.40	+0.35	–0.69
HBA	+0.39	–0.97	–0.49	–0.84	–0.17	–1.67
HBD	–2.28	+1.84	–1.37	–1.02	–0.15	0.00
PSA	+0.90	+1.90	–0.16	+0.34	+0.40	-3.39

Table 2. Average effect *E* values for individual atomic properties (negative values are beneficial, positive values are detrimental). Properties are ordered by average of *E*.

Dataset	Core Size	Substituent Position	Number	Mean Size
GCPR	5	R1	278	10.9
		R2	86	14.8
		R3	48	1.3
Kinase	5	R1	224	11.4
		R2	228	7.4
		R3	3	1.0
PDE	11	R1	60	4.6
		R2	116	4.4
		R3	155	9.2
		R4	32	1.3
		R5	12	0.8

Table 3. Summary statistics for the three datasets used in the simulations. Each row details the number and the mean size (in terms of numbers of atoms) at a particular point of substitution on the central core, e.g., at position R1 in the GCPR dataset, there was a total of 278 substituents with a mean size of 10.9 atoms. In all three datasets, only some of the substituent positions show a significant level of structural variation, with some positions (most obviously R3 in the kinase dataset and R5 in the PDE dataset) having only a very limited range of substituent types present.

Dataset	Descriptors	N	r^2	s	F	q^2	s_{CV}
GCPR	Hologram-53	3	0.41	0.81	121	0.38	0.83
	Hologram-151	5	0.60	0.67	155	0.49	0.75
	RGD	5	0.70	0.58	239	0.57	0.70
Kinase	Hologram-53	5	0.46	0.83	101	0.39	0.88
	Hologram-151	5	0.61	0.71	182	0.53	0.77
	RGD	5	0.61	0.71	186	0.51	0.80
PDE	Hologram-53	5	0.56	1.01	124	0.51	1.07
	Hologram-151	5	0.59	0.98	139	0.52	1.06
	RGD	5	0.70	0.84	225	0.57	1.01

Table 4. Leave-*n*-out QSAR models generated in the final step of the simulations.