ARTICLE

OBC
www.rsc.org/obc

Optimising the EVA descriptor for prediction of biological activity†

Martyn Ford,*[a] Laurie Phillips*[a] and Adrian Stevens[a,b]

[a] *Centre for Molecular Design, IBBS and University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, Hampshire, UK PO1 2DY. E-mail: martyn.ford@port.ac.uk*

[b] *BioFocus Discovery Ltd., Chesterford Research Park, Saffron Walden, Essex, CB10 1XL*

Received 5th July 2004, Accepted 11th October 2004
First published as an Advance Article on the web 27th October 2004

EVA is a multivariate molecular descriptor for use in QSAR studies. It is constructed from vibrational eigenvalues derived from either a quantum theoretical or molecular mechanical treatment of molecular structure. This paper applies the method to biological-activity data using measures of the inotropic potential of a range of Calcium channel agonists. The performance of the descriptor, as both an explanatory and a predictive tool, is analysed in relation to the way in which it is constructed using a rigorous statistical treatment. Its capabilities are examined in relation to those of previously published methodology which used a composite descriptor. It is shown to have improved performance and several procedural advantages, such as ease of calculation and operation. It is a 3-D structural descriptor which does not require prior co-alignment of structures for a QSAR study.

1 Introduction

A Quantitative Structure-Activity Relationship (QSAR) is a statistical model relating a common property of a series of chemicals (such as their responses in an assay of a particular type of biological activity) to their molecular structure. The major goal is to use this relationship to discover new chemicals with properties optimally suited to a defined application in an area such as pharmaceuticals, agrochemicals or catalysts. The process entails relating structurally significant features of a set of chemicals to activity measurements (responses) that are the complex result of integration of many physical, chemical and metabolic processes.

The structural features are summarised numerically in a molecular descriptor, but it is often difficult to do so efficiently and effectively since these features may also be very complex. Furthermore, many studies are performed on data sets comprising limited numbers of compounds and associated responses and this restricts the number of variables that can be used to describe structure in the QSAR model. In order to maximise the value of this small number of variables, molecular descriptors that summarise efficiently a large amount of diverse chemical information are required. One recent example is the theoretically-based multivariate descriptor, EVA.[1–3]

1.1 The EVA (*E*igen*VA*lue) descriptor

The premise on which the EVA approach is based is that a molecule's normal modes of vibration encode a suitable description of chemical structure, summarising the features necessary for the production of robust, predictive, QSARs. The normal modes are readily derived, using computational techniques based on either Quantum Theoretical or Molecular Mechanical methods. When suitably encoded they may be used as robust structural descriptors in QSAR lead-optimisation studies. In the following section, an overview of the approach is provided; for a more complete discussion, the reader is referred to reference 1.

QSAR studies demand as complete a structural description of a molecule as possible, in numerical terms. Using Quantum Mechanics, the molecular wavefunction, Ψ, provides a thorough characterisation of the nuclear and electronic properties of a molecule (and in principle should also allow for the estimation of properties such as biological activity). However, extraction of useful information for a QSAR study is difficult. Following studies with experimentally-based Infra-red Spectra, *theoretically-derived normal modes of vibration* were identified as a suitable molecular descriptor encoding information on the identity of the constituent atoms, their bonding and spatial relationships (hence molecular shape and size), their vibrational modes and molecular electronic structure. Intuitively, such a descriptor should constitute a robust, predictive, structural representation.

Calculation of a molecule's normal modes is achieved using the molecular electronic potential energy function, V, which is derived from the molecular wavefunction. Normal Coordinate Analysis (NCA) is performed to characterise the molecular vibrational properties. This approach involves the pointwise estimation of the potential function to determine the equations of motion, yielding both normal coordinate *E*igen*VA*lues (the normal mode frequencies of vibration) and their associated eigenvectors (the atom vibration vectors). The EVA procedure makes use of the eigenvalues, the eigenvectors being retained for final back-transformation to the molecular structures.

The normal modes for a set of molecules cannot be used directly in a multivariate statistical analysis because current methods demand that the descriptor for each molecule in the training set is comprised of the same number of well-defined components, (*i.e.*, the molecular descriptor must be a vector of fixed dimension). Since the number of normal modes varies with the number of atoms N in a molecule (actually 3N-6 for a molecule without axial symmetry) each descriptor within a series will generally be of different dimension. Data transformation is therefore required to convert, without loss of information, a varying number of vibrational eigenvalues into the descriptor EVA, which is of fixed dimension.

This is done by projecting the eigenvalues (vibrational frequencies) onto a bounded frequency scale (BFS) with individual vibrations represented by points along this axis. The frequency range is chosen to be 0 cm^{-1} to 4000 cm^{-1} to encompass all fundamental molecular vibrations. Each of the 3N-6 vibrations is represented by an equivalent Gaussian curve, $G\{f(\mu), \sigma^2\}$, in which μ is the vibrational frequency; the area under each curve is assigned as unity. Proximate or coincident Gaussian functions are permitted to overlap and the "intensity" is summed. The choice of σ for each function defines the degree to which the vibrations overlap and is typically 10 cm^{-1} to 20 cm^{-1}. This smoothing operation introduces, deliberately, a "fuzziness" to the spectrum of vibrations and is a key step in the definition of EVA; a value that is too small will fail to

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b410053k

detect similarities where they exist, and one that is too large will result in overlaps that obscure significant proportions of the variation in the descriptor. The process results in a degree of serial correlation in the descriptor that causes inevitably some redundancy in the descriptor variables. However, it also enables the significance of the presence or absence of peaks to be assessed in the subsequent analysis, together with the monitoring of changes in peak position.

Once a Gaussian smoothing function has been applied, the resultant spectrum is sampled across its whole width in fixed increments (L cm^{-1}). The choice of increment determines the number of variables in the EVA descriptor; thus, for example, a 2 cm^{-1} increment results in a descriptor string of 2000 variables. Molecular vibrations are therefore depicted as "peaks" of intensity on a scale of frequency *versus* height. This produces the molecular descriptor, EVA, which retains the integrity of its constituent frequency data. The choice of the variables σ and L can have a marked effect on the quality of the subsequently derived QSAR/QSPR model, and this is discussed below (see section 3.2).

The EVA descriptor was originally formulated at Shell Research Ltd, Sittingbourne Research Centre, and was applied successfully in a series of agrochemical lead-optimisation studies. While much of the experimental data used in the development remain proprietary and hence are not available for publication, its utility has been demonstrated by modelling experimentally observed log$P$ values.[1] The intention of the present study is to illustrate the application of the method to biological data that are within the public domain, incidentally providing a comparison of the performance of the EVA descriptor with that of other, more familiar, descriptors. A convenient example lies in a set of calcium ion channel agonists and their associated responses in an assay of biological activity;[4] this represents a particularly attractive set as it has been analysed previously by Davis *et al.*[5,6] with another theoretically-based 3-D descriptor, GRID.[7] This paper compares the performance of EVA with the published results of this study.

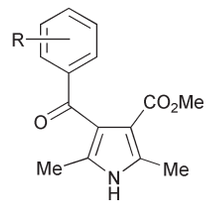
In the previous work,[5,6] the 3D structural descriptor, GRID,[7] was used to analyse the molecular features which account for the observed activities. The additional inclusion of selected physicochemical descriptors, and the consequences of variable scaling, was then considered in turn to assess whether the QSAR model could be improved upon. All results were cited in terms of the squared correlation coefficient of the fitted PLS regression model, r^2. The intention of the present studies is to conduct a parallel investigation to examine the performance of the EVA descriptor. A particular focus will be to examine the robustness of the predictive ability of the EVA based QSAR model. Results will therefore be presented in terms of the squared correlation coefficients of the *cross-validated* equations, q^2, rather than the fitted regression model. The parameter q^2 is commonly used as a measure of the predictive ability of the regression model, but in order to facilitate comparison between the two studies the values of r^2 will also be included as appropriate. This paper reports the results of an investigation to show how to optimise the EVA descriptor in order to increase the predictive power of a QSAR equation.

2 Experimental

2.1 The Calcium Channel Agonist data set

Table 1 shows the structures of the compounds used by Davis *et al.*[5,6] The compounds were assayed *in vitro* for their inotropic potency, *i.e.*, their ability to increase cardiac contractibility, using guinea pig atria paced at 1 Hz. This was expressed as the concentration of the drug that increased the tension developed in the preparation to 50% of the isoprenaline maximum (EC_{50}). Results were expressed relative to the EC_{50} of the standard calcium channel agonist Bay K 8644 (Fig. 1) and these values are given in Table 1.

Table 1 Clog$P$[a] cMR[a] and relative force of contraction, EC_{50},[b] values reported for 36 compounds in the Calcium Channel Agonist QSAR set[5,6]



Compound	R	Clog$P$	cMR	Relative force EC_{50}
1	2-Cl	3.174	7.664	0.0943
2	2-CF_3	3.652	7.683	0.27
3	2-OCH_3	2.458	7.79	0.0053
4	2-H	2.703	7.173	0.059
5	2-OCO-(2′-OH–C_6H_5)	4.509	10.49	0.34
6	2-CH_3	3.202	7.637	0.14
7	2-F	2.884	7.188	0.0093
8	2,5-Cl_2	3.905	8.156	0.33
9	2-I	3.584	8.479	0.22
10	2-Br	3.324	7.95	0.15
11	2-OCH_2Ph	4.226	10.301	1.13
12	2-Cl, 5-NO_2	3.025	8.39	0.16
13	2-CH_2Ph	4.62	10.148	35.5
14	2-Ph	4.591	9.684	0.174
15	2-SCH_2Ph	5.072	10.954	2.89
16	2-$SOCH_2$Ph	3.021	10.987	0.312
17	2-SO_2CH_2Ph	2.771	11.02	0.021
18	2-CH_2CH_2Ph	5.149	10.612	8
19	2-CH_3, 5-CH_3	3.701	8.1	0.0568
20	2-SPh	5.083	10.49	2.57
21	2-SOPh	2.792	10.523	0.34
22	2-NH–Ph	5.225	10.053	18.91
23	2-CH_2-(4′-NO_2–Ph)	4.363	10.873	4.31
24	2-CH_2-(2′-NO_2–Ph)	4.083	10.873	2.9
25	2-S-(2′-NO_2–Ph)	4.968	11.216	1.24
26	2-O-(2′-NO_2–Ph)	4.616	10.563	0.96
27	2-CH_2-(4′-NH_2–Ph)	3.393	10.517	0.0457
28	2-OSO_2-(4′-Me–Ph)	3.986	11.173	0.0072
29	2-OPh	4.637	9.837	2.9
30	2-NH–pyrid-2-yl	4.38	9.842	7.7
31	2-CH_2–C_6H_{11}	5.852	10.242	27.6
32	2-NH–C_6H_{11}	5.277	10.147	19.8
33	2-Br, 5-F	3.485	7.965	0.22
34	2-CH_2-(4′-F–Ph)	4.763	10.163	14
35	2-CH_2Ph, 5-F	4.8	10.163	19
36	2-CH_2(4′-F–Ph), 5-F	4.944	10.179	19

[a] Physicochemical parameters calculated using MedChem software v 3.54 B (1989). [b] Values measured relative to Bay K 8644.

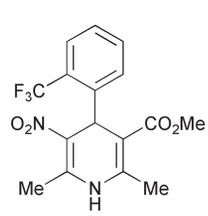


Fig. 1 The structure of Bay K 8644.

Also listed in Table 1 are values of the calculated log octanol/water partition coefficient (clog$P$) and the calculated molar refractivity (cMR) respectively, corresponding to the various structures. These were calculated using MedChem software v.3.54 B (1989). In the corresponding table in the paper by Davis *et al.* the quoted values for these parameters are different, having been modified by in-house fitting to a few experimentally determined values. In the present paper we use the standard values in order to maintain external consistency.

Table 2 PLS fitted r^2 regression models reported by Davis *et al.*[5,6] for the full 36 compound data set

PLS model (Block variances)	Regression model cumulative r^2				Overall r^2
	PLS 1	PLS 2	PLS 3	PLS 4	
clog*P* = 1.0 −log(EC_{50}) = 1.0	0.69 (q^2 = *0.66*[a])	—	—	—	0.69
GRID = 1458 −log(EC_{50}) = 1.0	0.42	*n/s*	*n/s*	*n/s*	0.42
GRID = 1458 cMR = 1.0 clog*P* = 1.0 −log(EC_{50}) = 1.0	0.42	*n/s*	*n/s*	*n/s*	0.42
GRID = 1.0 cMR = 1.0 clog*P* = 1.0 −log(EC_{50}) = 1.0	0.60	0.71	0.77 *n/s*	0.86	0.86

n/s – not significant *via* LGO cross-validation using 7 groups.

[a] Subsequently determined in present studies.

2.2 GRID descriptor

For the 3-D QSAR analysis of the data set, Davis *et al.* made use of the GRID descriptor.[7] The technique was developed originally as a method of probing proteins for potential binding sites and has been used successfully to predict sites of binding of small ligands in proteins,[8] in addition to its more recent role as a structure descriptor. A detailed discussion of the derivation of the GRID descriptor can be found in reference 7.

2.2.1 The inclusion of physicochemical descriptors. A limitation of GRID is that by generating the descriptors from the calculated interaction energy of a probe molecule, only the enthalpic component of the drug–receptor interaction can be modelled. This neglects other factors which can be, potentially, highly influential, such as the entropic changes resulting from the interaction. In recognition of this limitation, Davis *et al.*[5,6] included two physicochemical descriptors, clog*P* and cMR. Clog*P* represents a particularly useful descriptor since by accounting for lipophilicity and solvation, features that are influenced heavily by hydrogen bonding, it can give information on the entropic changes associated with a molecule's coming out of solution and binding at a receptor site.

2.3 EVA descriptor generation

EVA descriptors were generated for the 36 compounds in the calcium channel agonist set.[5,6] A typing error in the published table[5,6] cites incorrectly the substitution about the benzoylpyrrole backbone of compounds 8, 12, 19, 33, 35 and 36 as 2,4 rather than as 2,5.[9] The correct structures were used in the present study and were created using Sybyl 6.2 from Tripos Associates[10] on a Silicon Graphics Indy R4000 platform.

In the previous studies, Davis *et al.* constrained the energy minimisation of the structures in an attempt to minimise noise in the GRID descriptors and hence optimise the subsequent QSAR investigations.[5,6] In order to perform the calculation of the normal modes of vibration required for EVA, it is a prerequisite to ensure that the potential energy of a molecular conformation is at a stationary point. Slight deviations away from this energy minimum can lead to the introduction of rotational and translational modes during the determination of the equations of motion in the NCA. This in effect 'contaminates' the vibrational modes, resulting in the determination of vibrations that are not purely harmonic in nature and degrading significantly the quality of the calculated vibrational spectrum. A more detailed discussion of this is provided in reference 11. No attempt was made therefore to reproduce directly the conformational co-ordinates used by Davis *et al.*[5,6]

Molecular structures were optimised using the MOPAC 6.0 AM1 Hamiltonian[12] (keywords: SCFCRT = 1.D-12 GNORM = 0.05). Successfully minimised structures were then used to generate the NMs (additional keyword: FORCE). These were then converted into EVA descriptors in the manner described. It is acknowledged that the calculation of NMs using a semi-empirical route is less accurate than for example, by the use of Density Functional Theory methods.[13] However, the use of the MOPAC AM1 Hamiltonian has previously been shown to produce qualitatively good results[14]. Furthermore, EVA descriptors generated from the MOPAC AM1 Hamiltonian have been applied successfully in previously reported studies.[14] Consequently, it is believed that the EVA descriptors generated in this way are acceptable for the present investigation.

2.4 Statistical analysis

To generate useful QSARs between the activity, EC_{50}, and the EVA descriptor, the Partial Least Squares regression method (PLS) was adopted for the investigations; this is in line with the GRID-based studies of Davis *et al.*[5,6] PLS is a supervised approach that provides a reduced solution of new 'latent' variables that are linear combinations of the original variables and are well correlated with *Y*. In essence, it is a projection method that defines a hyperplane through the *x*-descriptor space. The latent variables represent the projected 'summaries' of the original variables onto this plane and each PLS component represents a co-ordinate dimension of the hyperplane. A more complete discussion of the PLS approach can be obtained from references 15 and 16.

The PLS regression approach is an accepted technique for handling situations where over-square matrices are encountered and has some advantages over unsupervised techniques such as, for example, Principal Components analysis. By constructing the components in a supervised fashion, it is possible to produce fewer numbers of components correlated with *Y*, thereby summarising the data more efficiently. With matrices such as EVA and GRID this is a clear advantage.

As has been mentioned earlier in section 1.1, by having the ability to vary the form and dimension of the EVA descriptor (*i.e.*, σ and *L*), it is possible to change the manner in which the structural information is presented. In effect, this can be thought of as a form of data scaling, in line with log and variance scaling methods. This affects ultimately the regression analysis, particularly with respect to the co-variance with the activity data and the way in which the information is loaded onto the components. Therefore an essential stage in the application of EVA is the generation of descriptors with different σ and *L* values and their subsequent systematic validation to identify the optimum predictive model. Previous studies have shown that a σ value of between 10 and 20 cm^{-1} is a reasonable starting point for such investigations.[1]

2.4.1 Model validation and significance testing. A key step in regression analysis is the validation to determine the correct dimensionality of the model, *i.e.*, the number of significant PLS components to be included. Without this validation there is a risk of 'over-training', leading to apparently well-fitting models with effectively little predictive capacity. As the intention of the present study is to assess the predictive capability of the EVA descriptor, two forms of model validation have been incorporated. In line with the study by Davis *et al.*, Cross-Validation (CV) has been used to test each successive PLS component for significance. In addition, randomisation tests have been included to establish confidence limits on whether or not random correlations are being selected in place of real relationships.

CV represents one of the most favoured methods of testing the PLS components for significance. In essence, CV randomly divides the data set into groups of compounds. Each group is then omitted from the data set in turn and the remaining set used to construct a new regression model, which is then used to calculate the responses for the members of the omitted set. From the predicted *y*-values for each of the omitted 'test' sets, the sum of squares of the differences from actual values can be calculated. This yields the *PRESS* score (Predictive Residual Sum of Squares) and by weighting this against both the inclusion of successive components and the number of cases in the set, the standard error of cross-validation, S_{cv}, can be calculated. This provides a good test of the significance of the inclusion of each component upon the overall regression model and has some advantages over other procedures such as, for example, the *F*-ratio test. By simulating how well the model predicts new data, it provides an estimate of the robustness of the model and potentially can highlight models that are influenced by data points with high leverage (*e.g.*, outliers). As the underlying intention of a QSAR is to make predictions, then this is useful.

Initial analysis of the EVA descriptor models was performed using Leave-One-Out Cross-Validation (LOO CV). To determine the optimum regression model, the S_{cv} scores for the first 10 PLS components were calculated and the minimum S_{cv} score identified.[15] Previous experience has shown that the relevant information linking the response and descriptor variables is summarised in the first few components (typically ≤ 6).[1] The squared correlation coefficient of the cross-validated equation, q^2, was then cited for the inclusion of each successive component.

In a previous publication,[1] it was shown that the LOO CV q^2 obtained on a training set of 135 structures, using EVA as the descriptor and $\log P_{ow}$ as the measured response, was a very good indicator of subsequent predictive power for $\log P_{ow}$ in an unrelated test set of 76 compounds. However, work by Shao[17] has challenged the suitability of LOO CV for testing the predictive robustness of a QSAR model and suggests that Leave-*n*-Out Cross-Validation, where *n* is greater than 1, represents a more robust alternative. Consequently, in the present investigations we have compared LOO CV results with those obtained using Leave-Group-Out Cross-Validation (LGO CV) with $n = 7$ (see section 3.2); this is also in line with the validation approach applied by Davis *et al.*[5,6] However, while LOO CV represents an easily reproducible test, *i.e.*, each run will yield the same q^2 result, in LGO CV the cases assigned to each group to be left out are selected at random. This means that LGO CV must be repeated a large number of times (*e.g.*, 1000) in order to obtain a realistic estimate of the internal predictivity and is hence more computationally demanding than LOO CV. Our work shows that with this data set and analytical approach, LOO CV gives very similar results to LGO CV and so we have adopted the simpler procedure.

It must be acknowledged that neither LOO CV nor LGO CV can be regarded as a sufficient indicator of model validity; in addition, some measure of the probability that the result may be a chance occurrence is needed. This is particularly important in regression studies where there are a smaller number of cases than the dimensionality of the *x*-descriptor, *i.e.*, where an over-square matrix is present. In such instances, there is an increased possibility of the extraction of a random correlation. Consequently, in order to address this with regard to the EVA descriptor, a randomised permutation test was performed to examine the model for chance correlations; this is in line with previous studies that have applied the EVA descriptor.[2,18] The process involves:

1. Randomly assigning observations to structures,
2. Performing a regression analysis using EVA and then evaluating q^2 (or r^2),
3. Repeating steps 1 and 2 a large number of times (*e.g.*, 1000 times) to ensure thorough sampling,
4. Determining the frequency distribution of q^2 (or r^2),
5. Determining where in this frequency distribution the q^2 (or r^2) of the real assignment lies.

An actual result that is extreme in the upper confidence tail of the distribution is regarded as having a low probability of occurring by chance

3 Results and discussion

In the following section, the results are reported for the QSAR investigation of a set of 36 calcium channel agonists (Table 1), using EVA as the structural descriptor and PLS regression as the statistical approach. Except where stated otherwise, all results will be reported in terms of the squared correlation coefficient (q^2) of the Leave-One-Out Cross-Validated (LOO CV) equation.

To facilitate comparisons, Table 2 is taken directly from the work of Davis *et al.*, detailing the PLS regression models reported for the full 36 compound data set.[5,6]

3.1 Determination of optimum PLS regression model

3.1.1 Optimum EVA descriptor model. Initial identification of the optimum EVA model was performed by comparing the results obtained with a series of descriptors using σ values of 10, 20, 30 and 40 cm^{-1}, respectively. The results established a narrowed focus of attention around $\sigma = 20$ cm^{-1} and a further series of EVA descriptors was then generated ($16 \leq \sigma \leq 26$ cm^{-1}). Comparison of the predictive performance of the models identified the optimum EVA model as having $\sigma = 24$ cm^{-1}. Throughout, the BFS start point, *S*, was fixed at 1 cm^{-1} and *L* was defined as $\sigma/2$. This has been assumed

Table 3 Effect of altering the EVA descriptor model on the QSAR performance

EVA Model	Regression model cumulative q^2 (r^2)				Overall LOO CV q^2
	PLS 1	PLS 2	PLS 3	PLS 4	
{4,2}	0.230 (*0.648*)	0.266 (*0.876*)	*n/s*	*n/s*	0.266
{8,4}	0.233 (*0.505*)	0.281 (*0.762*)	*n/s*	*n/s*	0.281
{10,5}	0.230 (*0.469*)	0.296 (*0.693*)	*n/s*	*n/s*	0.296
{16,8}	0.215 (*0.402*)	0.307 (*0.565*)	0.383 (*0.751*)	0.413 (*0.881*)	0.413
{20,10}	0.205 (*0.374*)	0.298 (*0.513*)	0.384 (*0.730*)	0.477 (*0.836*)	0.477
{22,11}	0.201 (*0.363*)	0.292 (*0.491*)	0.374 (*0.719*)	0.493 (*0.817*)	0.493
{24,12}	**0.198 *(0.354)***	**0.286 *(0.472)***	**0.359 *(0.707)***	**0.498 *(0.802)***	**0.498**
{26,13}	0.196 (*0.347*)	0.282 (*0.456*)	0.342 (*0.693*)	0.492 (*0.789*)	0.492
{30,15}	0.193 (*0.335*)	0.276 (*0.431*)	0.306 (*0.663*)	0.459 (*0.767*)	0.459
{40,20}	0.189 (*0.317*)	0.274 (*0.399*)	*n/s*	*n/s*	0.274

n/s – not significant *via* cross-validation.

Table 4 Effect of model transformations and variance scaling on the EVA {24,12} *vs.* EC_{50} PLS model

PLS Model	Var. scaling	Regression Model cumulative q^2 (r^2) PLS 1	PLS 2	PLS 3	PLS 4	Overall LOO CV q^2
No EVA Scaling						
EVA {24,12}	—	0.198	0.286	0.359	0.498	0.498
EC_{50}	None	(*0.354*)	(*0.472*)	(*0.707*)	(*0.802*)	
EVA {24,12}	—	0.198	0.286	0.359	0.498	0.498
EC_{50}	1.0	(*0.354*)	(*0.472*)	(*0.707*)	(*0.802*)	
EVA {24,12}	—	0.194	*n/s*	*n/s*	*n/s*	0.194
$-\log(EC_{50})$	None	(*0.332*)				
EVA {24,12}	—	0.194	*n/s*	*n/s*	*n/s*	0.194
$-\log(EC_{50})$	1.0	(*0.332*)				
Autoscaling						
EVA {24,12}	1.0	0.019	*n/s*	*n/s*	*n/s*	0.019
EC_{50}	1.0	(*0.365*)				
EVA {24,12}	1.0	0.165	*n/s*	*n/s*	*n/s*	0.165
$-\log(EC_{50})$	1.0	(*0.376*)				
Blockscaling						
EVA {24,12}	1.0	0.198	0.286	0.359	0.498	0.498
EC_{50}	1.0	(*0.354*)	(*0.472*)	(*0.707*)	(*0.802*)	
EVA {24,12}	1.0	0.194	*n/s*	*n/s*	*n/s*	0.194
$-\log(EC_{50})$	1.0	(*0.332*)				

n/s – not significant *via* cross-validation.

previously[1] to represent a reasonable sampling frequency for a given value of σ and is justified below (section 3.2). The results for the full range of EVA models considered are included in Table 3.

3.1.2 Model transformations and variance scaling effects. As was demonstrated earlier,[5,6] PLS analysis can be highly sensitive to the relative scales of variance between the *x*-block and *y*-variable (this is especially significant when different *x*-descriptors, each with its own unit and scale, are included in the regression equation, although this situation does not apply to EVA). A series of transformations and scaling procedures was applied to both the activity and EVA variables, to investigate the impact on performance.

Scaling and transformation tests were conducted in two parts. In the case of the activity variable, EC_{50}, both log scaling and variable standardisation were considered, as in the study of Davis *et al.*[5,6] For the EVA descriptor, two forms of variance scaling were applied, Autoscaling and Blockscaling. Autoscaling standardises each individual variable column to zero mean and unit variance, while Blockscaling scales the entire descriptor block to zero mean and unit variance. The impact of these two scaling methods is of particular interest, because differences in column variance across the descriptor are considered to contain structurally significant information.

The full results are reported in Table 4. Variance scaling with the EC_{50} variable has no impact on the PLS regression model (4*c* model, $q^2 = 0.498$). Similarly, blockscaling the EVA descriptor does not affect the performance. However, $\log_{10}$ transformation of the activity values does result in a significant degradation of the PLS model (1*c* model, $q^2 = 0.194$). Notably, use of the autoscaled EVA descriptor yields the worst model of all.

3.2 EVA QSAR model testing and stability

As discussed above (section 2.4.1), with over-square *x*-matrices the PLS regression approach, potentially, can extract chance correlations with the *y*-variable. In the case of the EVA descriptor, an extreme example of an over-square matrix, then the likelihood of such an occurrence may be increased. Therefore, a series of rigorous validation exercises was applied to establish confidence in the QSAR model obtained. These include sensitivity tests on the effects of changes in the EVA *L* variable, LGO CV (in line with references 5 and 6) and randomised activity permutation tests.

Fundamentally, two variables, namely σ (Gaussian standard deviation) and *L* (the BFS sampling interval), determine the form and dimension of the EVA descriptor when derived from the normal coordinate spectrum. Potentially, the value chosen for *L* can impact on the information content retained in the final descriptor model and an extreme illustration of this would occur if the sampling interval were so large that not all of the intensities under the Gaussian curves were to be sampled. As outlined earlier in the initial determination of the optimum EVA model, *L* was defined as $\sigma/2$. This ensures that all structural information is retained, while keeping the associated computational time within reason. Validation of this procedure was carried out by comparing the effects of using a range of sampling intervals ($\sigma/4 \leq L \leq 4\sigma$).

The full set of results is shown in Table 5 and illustrated in Fig. 2. For sampling intervals where $L \leq \sigma$, the information content remains intrinsically stable, with no significant impact on the statistical performance of the EVA models. However when $L > \sigma$, decreasing stability in the performance in the EVA model is observed. The results confirm that $L = \sigma/2$ is a reasonable rule.

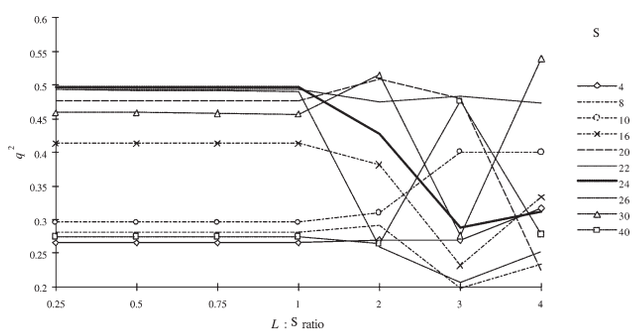


Fig. 2 Plot of the effect on q^2, as a function of the $L:\sigma$ ratio.

To test rigorously the predictive performance of the optimum EVA model (EVA{24,12}), LGO CV was used and compared with the results of LOO CV. In line with the LGO CV in references 5 and 6, the 36 compounds were divided randomly into 7 groups. Repeated runs of the LGO CV (1000 iterations) were carried out with different randomisations to ensure that the range of possibilities was thoroughly sampled.

The results for the LGO CV study are presented in Table 5. These show that the 4-component model occurs most frequently (560 iterations; Fig. 3a), in agreement with LOO CV. The range

Table 5 Effect of altering the ratio of L to σ on the PLS regression model LOO CV q^2 (r^2)

	$L:\sigma$ ratio						
σ	0.25	0.50	0.75	1	2	3	4
4	0.265 (*0.876*)	0.265 (*0.876*)	0.265 (*0.876*)	0.265 (*0.876*)	0.270 (*0.878*)	0.269 (*0.853*)	0.317 (*0.825*)
8	0.281 (*0.762*)	0.281 (*0.762*)	0.281 (*0.762*)	0.281 (*0.762*)	0.291 (*0.745*)	0.198 (*0.475*)	0.234 (*0.489*)
10	0.296 (*0.693*)	0.296 (*0.693*)	0.296 (*0.693*)	0.296 (*0.693*)	0.310 (*0.696*)	0.401 (*0.875*)	0.400 (*0.804*)
16	0.413 (*0.881*)	0.413 (*0.881*)	0.413 (*0.881*)	0.413 (*0.881*)	0.381 (*0.749*)	0.232 (*0.535*)	0.333 (*0.608*)
20	0.477 (*0.836*)	0.477 (*0.836*)	0.477 (*0.836*)	0.477 (*0.836*)	0.509 (*0.845*)	0.480 (*0.773*)	0.223 (*0.451*)
22	0.494 (*0.817*)	0.493 (*0.817*)	0.493 (*0.816*)	0.493 (*0.816*)	0.475 (*0.790*)	0.483 (*0.788*)	0.474 (*0.763*)
24[a]	**0.498** ***(0.802)***	**0.498** ***(0.802)***	**0.498** ***(0.802)***	0.497 (*0.801*)	0.288 (*0.796*)	0.288 (*0.397*)	0.312 (*0.477*)
26	0.493 (*0.789*)	0.492 (*0.489*)	0.492 (*0.789*)	0.491 (*0.789*)	0.259 (*0.424*)	0.207 (*0.404*)	0.253 (*0.600*)
30	0.459 (*0.767*)	0.459 (*0.767*)	0.458 (*0.767*)	0.457 (*0.766*)	0.515 (*0.780*)	0.276 (*0.429*)	0.540 (*0.751*)
40	0.274 (*0.399*)	0.274 (*0.399*)	0.274 (*0.399*)	0.274 (*0.399*)	0.264 (*0.394*)	0.476 (*0.727*)	0.278 (*0.389*)

[a] Optimum regression mode.

of q^2 scores for the LGO CV 4 component model is approximately normal in distribution (Fig. 3b and 3c). Determination of the mean q^2 reveals that q^2_{LGO} compares closely with q^2_{LOO} (Fig. 2c: $q^2_{LGO} = 0.474$, $q^2_{LOO} = 0.498$), which gives confidence in the use of LOO CV to test the range of EVA models.

Figs. 4 and 5 show the results of the randomised activity permutation tests on the same data set, obtained by using 1000 permutations of the response data with respect to structure. Fig. 3 shows the frequency distribution and normalised plot of q^2, obtained from the LOO CVs; the actual experimental result of 0.498 lies well outside the upper 95% confidence limit (>99%). A similar result is shown in Fig. 4 for the fitted r^2 distribution. Table 6 summarises these results.

3.3 Augmentation of EVA {24,12} with the physicochemical descriptors clogP and cMR

Davis *et al.*[5,6] showed that the 3D molecular descriptor GRID did not by itself encode sufficient information to give a high-quality QSAR regression model of $-log\ EC_{50}$ ($r^2 = 0.42$). However, inclusion of clog*P* and cMR, in linear combination with GRID, gives a significant improvement ($r^2 = 0.86$). It is therefore of interest to examine the influence of including clog*P* and cMR with EVA. The results are given in Table 6 and show that inclusion of these parameters (either singly or in combination) degrades the quality of the fit (r^2) and also the predictive power (q^2).

3.4 EVA and analysis of the data of Davis *et al.*[5,6]

QSAR methods that relate biological response to molecular structure are traditionally based on linear free energy relationships, as illustrated in the pioneering work of Hansch *et al.*[19] Typically, these methods attempt to draw correlations between empirical descriptors which supposedly represent generalised forms of enthalpic and entropic effects associated with a drug molecule's transport to, and interaction with, a receptor or enzyme active site. The current research is aimed at further developing a new molecular descriptor, EVA, which encodes a large amount of diverse chemical information and can be used more effectively than the previous empirical forms.

Analysis of the predictive performance of a range of EVA PLS models (Table 3) identifies the optimum descriptor as having a Gaussian sigma; = 24 cm^{-1} and a sampling interval $L = 12$ cm^{-1}, EVA {24,12}. Cross-Validation (using both LOO CV and LGO CV methods) suggests that the first 4 components are significant, yielding a QSAR that is able to account for *circa.* 80% of the variation in the EC_{50} activity (fitted $r^2 = 0.801$). Rigorous testing of the internal predictivity of the EVA QSAR model (Tables 3 and 5) shows that for structures outside the training set *circa.* 50% of the activity variation may be predicted with confidence, an experimentally useful result (4-component models, $q^2_{LGO} = 0.474$; $q^2_{LOO} = 0.498$). The close agreement between the two validation methods, illustrated in Fig. 2, gives considerable confidence in the use of LOO CV to validate the chosen EVA model and indicates that EVA is robust as a predictive QSAR tool, *i.e.*, the model is stable.

Since Davis *et al.*[5,6] analysed their data with log-transformation of the EC_{50} values (*i.e.*, the *y*-variate was $-\log(EC_{50})$, Table 1) and their original data are not available to us, an exact comparison of the two sets of results is difficult. Moreover, cross-validation tests were not reported so it is only possible to compare values of r^2, obtained by fitting the full-regression equation, rather than q^2.

The best model of Davis *et al.*[5,6] uses a linear combination of GRID, clog*P* and cMR. (GRID alone gives a 1-component model, $r^2 = 0.42$, clog*P* alone a 1-component model, $r^2 = 0.69$, and GRID in combination with clog*P* and cMR gives a 4-component model, $r^2 = 0.86$). The values of $-\log(EC_{50})$ predicted by this model are not quoted, but a comparison with experimentally observed values is given graphically in Fig. 5 of reference 5; by careful measurement, it is possible to obtain reliable estimates for the predicted $-\log(EC_{50})$ values. Taking antilogarithms yields values for predicted EC_{50} and the regression statistics may be calculated and compared with those obtained for the EVA model. Overall, the full EVA model ($r^2 = 0.80$, standard error = 4.1) gives a considerably better fit to the untransformed experimental data than does the GRID + clog*P* + cMR model ($r^2 = 0.42$, standard error = 7.0). The results are shown in Table 7 and are further illustrated in Fig. 6 which shows plots of residuals for both treatments. For small values of EC_{50} (typically less than 4.1, the standard error of the EVA treatment) the two models (not surprisingly) cannot be distinguished; however, for higher values of EC_{50} EVA consistently outperforms the other model by an impressive margin.

A model-validation test using randomised activity permutations establishes confidence intervals for both the predictive (Fig. 4) and fitted (Fig. 5) regression models and demonstrates that, although the EVA descriptor presents an over square *x*-matrix in the PLS regression model, random correlations

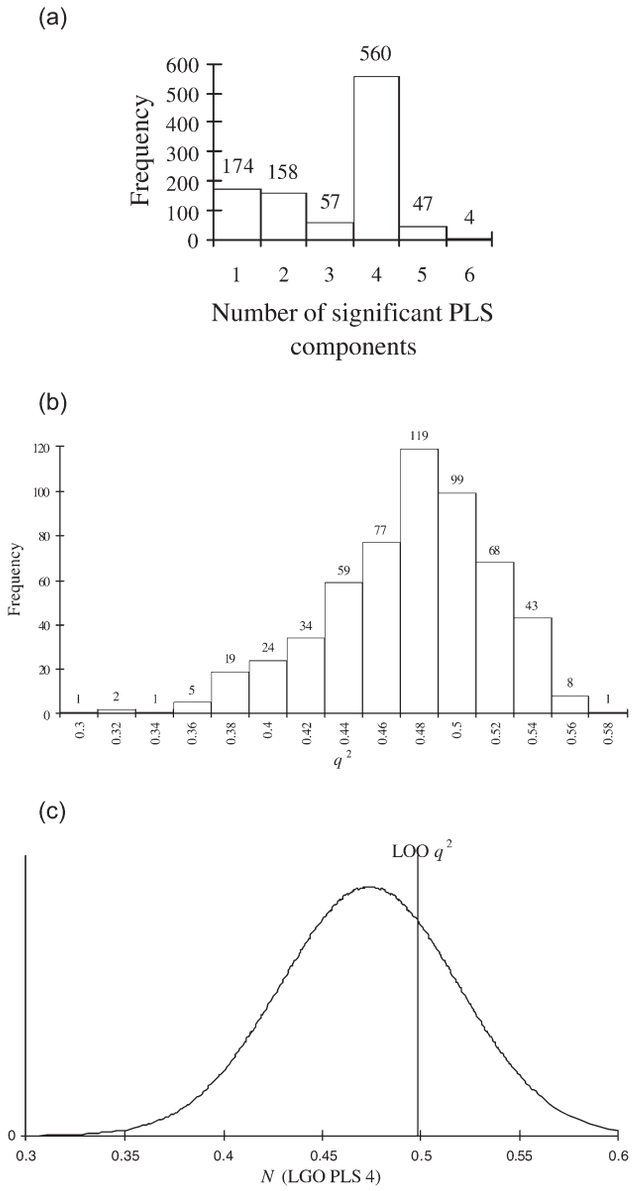


Fig. 3 Leave-Group-Out Cross Validation analysis (LGO CV) of EC_{50} with EVA{24,12}, showing a) the frequency distribution of the number of significant component extracted from 1000 LGO CV runs, b) the frequency distribution of q^2 scores for the 4 component models and c) the normalised distribution of these scores, illustrating that the Leave-One-Out (LOO) result compares closely with the mean.

with the activity data are highly unlikely (*>99%* probability that the correlations have not arisen by chance). EVA contains structurally significant features that account well for the observed biological activity. It should also be emphasised that EVA is a robust descriptor of 3D molecular structure, which does not require prior alignment of structures in order for a QSAR study to be performed. This is especially useful when dealing with non congeneric series, where alignment about a common skeletal feature is not feasible.[1,3]

Although the relevance, multicolinearity and redundancy of the EVA descriptors have not been addressed in this study, they probably provide the most likely explanation for the relatively small values obtained for the estimates of LOO q^2. These aspects will be reported in a future publication. Clearly, a model that yielded a q^2 of >50% would be even more useful. Plots of the predicted *vs.* actual activity for both the fitted and cross-validated regression models (Figs. 7 and 8) highlight problems with outliers, particularly those compounds with EC_{50} values which are essentially zero; these structures can only contribute noise to the analysis, and the consequences are addressed in detail in section 3.7.

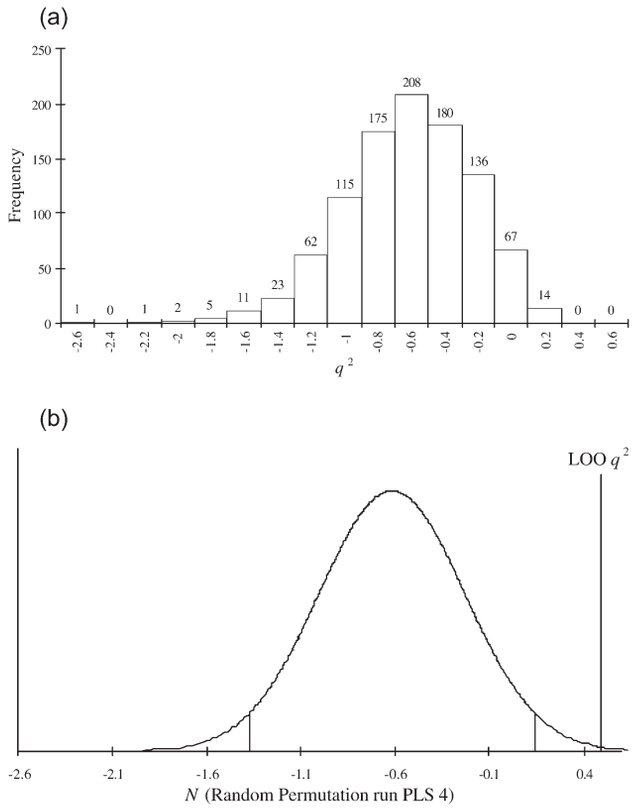


Fig. 4 The randomised-activity permutation analysis of the EVA {24,12} Leave-One-Out Cross Validation model (LOO CV) with EC_{50}, showing a) the frequency distribution of the q^2 scores for the randomised LOO CV regression models and b) the normalised distribution of these scores, to illustrate that real LOO CV result lies to the extreme right of the distribution.

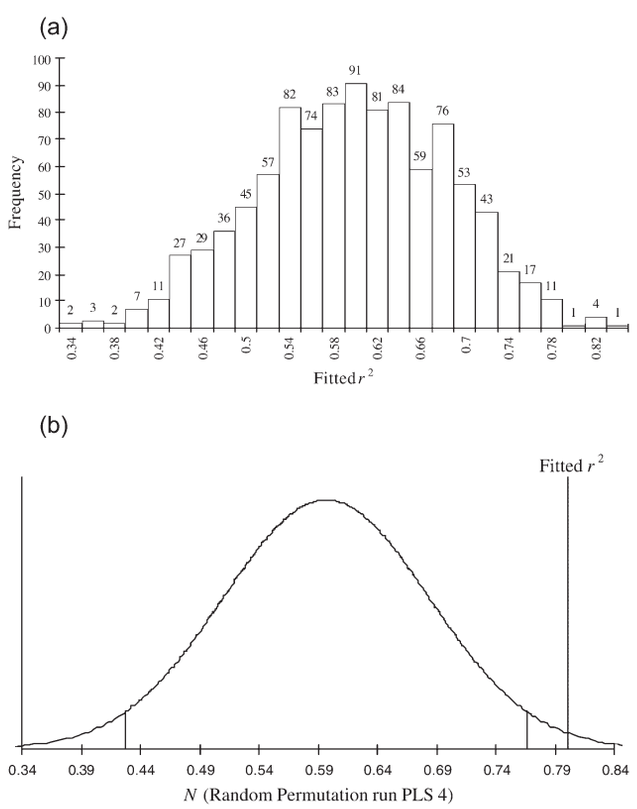


Fig. 5 The randomised activity permutation analysis of the EVA {24,12} fitted PLS model with EC_{50}, showing a) the frequency distribution of the r^2 scores for the randomised fitted regression model and b) the normalised distribution of these scores, to illustrate that real fitted r^2 PLS result lies in the upper confidence limit.

Table 6 LGO (7 groups) cross-validation and randomised EC_{50} permutation validation of EVA {24,12} PLS regression model

Validation method	No. of runs	c	Mean q^2 or r^2	Standard deviation	95% confidence limits
LGO CV					
EVA {24,12} EC_{50}	1000	4	$q^2 = 0.474$	s.d. = 0.045	
Randomised Activity Permutations					
EVA {24,12} Random (EC_{50})	1000	4	$q^2 = -0.616$	s.d. = 0.387	$-1.374 < q^2 < 0.141$
EVA {24,12} Random (EC_{50})	1000	4	$r^2 = 0.597$	s.d. = 0.087	$0.427 < r^2 < 0.767$

Table 7 Predicted *vs.* observed EC_{50} for the calcium channel agonist set. Comparison between the GRID + Hansch and the EVA regression models as EC_{50} predictors for the full 36 compound data set

		GRID + Hansch *vs.* EC_{50}		EVA 36 set model *vs.* EC_{50}	
Structure	Observed EC_{50}	Predicted EC_{50}	Residuals	Predicted EC_{50}	Residuals
13	35.5	6.2	29.3	18.4	17.1
31	27.6	69.2	−41.6	25.6	2.0
32	19.8	13.8	6.0	22.9	−3.1
35	19.0	11.2	7.8	19.5	−0.5
36	19.0	11.2	7.8	22.1	−3.1
22	18.9	56.2	−37.3	15.3	3.6
34	14.0	7.6	6.4	15.6	−1.6
18	8.0	4.1	3.9	9.0	−1.0
30	7.7	6.6	1.1	5.3	2.4
23	4.3	0.6	3.7	4.7	−0.4
24	2.9	3.2	−0.3	2.1	0.8
29	2.9	6.0	−3.1	6.0	−3.1
15	2.9	3.7	−0.8	5.2	−2.3
20	2.6	15.1	−12.6	3.8	−1.2
25	1.2	1.7	−0.5	4.6	−3.4
11	1.1	0.5	0.7	4.7	−3.6
26	1.0	0.8	0.2	0.8	0.1
5	0.3	0.3	0.1	−3.2	3.6
21	0.3	0.3	0.0	−2.8	3.2
8	0.3	0.3	0.0	−1.0	1.3
16	0.3	0.1	0.2	1.5	−1.2
2	0.3	0.3	0.0	−2.9	3.1
9	0.2	0.2	0.0	−1.0	1.3
33	0.2	0.2	0.0	1.9	−1.7
14	0.2	0.3	−0.2	−1.3	1.5
12	0.2	0.1	0.1	−0.8	1.0
10	0.2	0.1	0.1	0.5	−0.3
6	0.1	0.1	0.1	0.5	−0.3
1	0.1	0.1	0.0	3.5	−3.4
4	0.1	0.0	0.0	2.7	−2.6
19	0.1	0.2	−0.2	−2.7	2.8
27	0.0	0.2	−0.2	9.7	−9.7
17	0.0	0.1	0.0	3.8	−3.8
7	0.0	0.0	0.0	1.1	−1.1
28	0.0	0.0	0.0	−2.6	2.7
3	0.0	0.0	0.0	−0.9	0.9

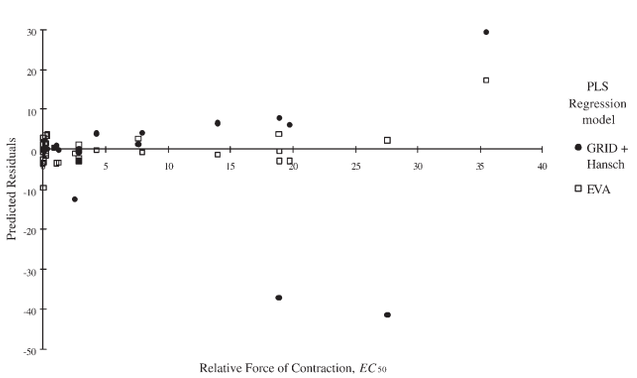


Fig. 6 Comparison between of the predicted EC_{50} residuals for the EVA {24,12} *vs.* GRID + Hansch fitted PLS regression models.

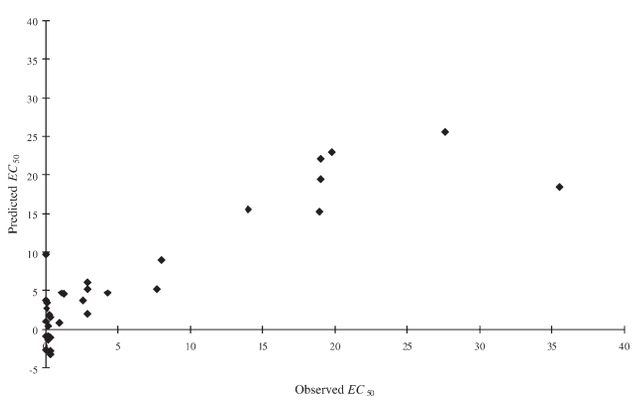


Fig. 7 Plot of predicted *vs.* observed EC_{50} for the EVA {24,12} fitted regression model with the full 36 structure set.

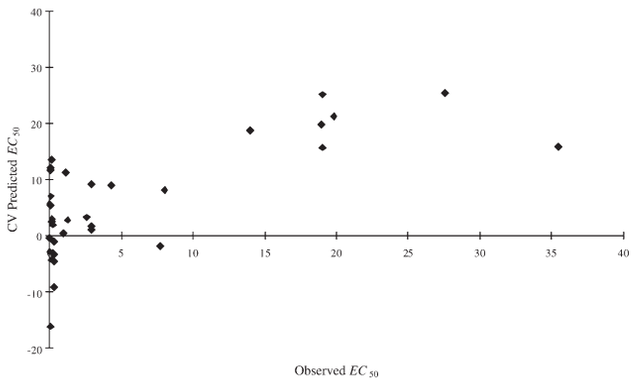


Fig. 8 Plot of predicted *vs.* observed EC_{50} for the EVA {24,12} cross-validated regression model with the full 36 structure set.

3.5 Data scaling

In studies where the *x*-variate function is a vector consisting of a number of variables such as molecular properties with different dimensionalities and variances, careful tailoring of the variable variances is critical in extracting an optimum QSAR model. This is true, for example, of the work by Davis *et al.*[5,6] In the present study however, Table 4 shows that blockscaling and autoscaling of the *x*-variate function does not improve the EVA QSAR model.

In particular, autoscaling (which scales columns independently to zero mean and unit variance), performs significantly worse. In the EVA descriptor, all variable columns are presented on the same scale with the differences in variance between them intrinsically encoding structurally relevant information. Consequently, individual columns should not be standardised since this would result in giving columns with little variation and therefore little information, undue weighting. Indeed, in-house development with proprietary data sets (Shell Research Ltd., the former Sittingbourne Research Centre) has in the past demonstrated that scaling typically does not improve the performance of EVA-based QSAR models.

Log transformation of the *y*-variate, EC_{50}, leads to a substantially worse EVA model. This contrasts markedly with the results of a similar transformation in the work of Davis *et al.*, which gave the best result. This is investigated further in section 3.7, below.

3.6 Inclusion of the physicochemical descriptors

It is reasonable to ask the question as to whether or not the inclusion of the physicochemical parameters clog*P* and cMR would improve the EVA model. The results of doing so are shown in Table 8 and provide an insight into the information content of EVA. It is clear that the inclusion of either clog*P* or cMR, or both, does not lead to an improved QSAR model, even with scaling of the variables in this more complex *x*-variate vector. The implication may be that the information contained within the physicochemical parameters is already encoded intrinsically within the EVA descriptor.

In order to test this hypothesis, namely that EVA encodes physicochemical information, regression models were constructed for clog*P* and cMR, respectively. The results (Table 9) show that while EVA accounts for the variation in cMR extremely well ($r^2 = 0.983$, $q^2 = 0.963$), it does not do so at all well for clog*P* ($r^2 = 0.415$, $q^2 = 0.281$).

The result with clog*P* is surprising, since in previous work it has been shown that EVA QSAR can explain the variation in experimentally measured log*P* values across a wide variety of structurally-diverse, non-related compounds and be usefully predictive (135 compounds, $r^2 = 0.96$, $q^2 = 0.68$).[1] Why should the approach be less successful with the congeneric set of structures described in Table 1? The most likely reason for this anomaly is in the variance in clog*P*; the range of values is relatively small, *i.e.*, only 3 log units in comparison to 6 log units in the log*P* study.

3.7 Investigation of outliers

It was noted that eight of the structures in the data set have corresponding values of EC_{50} that were approaching zero, ranging from 0.01 to 0.09 (see Table 1). Given the nature of the bioassay from which they have been derived, within experimental error, they probably may not be significantly different from zero. Inclusion of these data has the consequence of biasing the activity data set (*y*-variate) heavily, and introducing noise in the form of irrelevant descriptor information (*x*-variate). The consequence of removing these data to leave a set of 28 structures and activity scores was investigated.

Table 10 compares the regression results with those obtained for the original set of 36 structures. For the untransformed EC_{50} data, there was a slight improvement in both the fit of the full regression equation and in the predictive power, albeit with the addition of a further 2 components in the regression model (6 component model, $r^2 = 0.939$, $q^2 = 0.527$, *cf.* a 4-component model, $r^2 = 0.802$, $q^2 = 0.498$). In contrast, there was a highly significant change in the performance of the regression results for the $-\log(EC_{50})$ scores, Figs. 9 and 10 (5-component model, $r^2 = 0.963$, $q^2 = 0.721$, *cf.* a 1-component model, $r^2 = 0.332$, $q^2 = 0.194$). This compares favourably with the results of a similar transformation in the work of Davis *et al.* (4-component model, $r^2 = 0.86$). More importantly, the q^2 result would provide significant confidence in support of the application of this model in a predictive role. However, it should be recognised that chemical interpretability will be non-trivial, given the Gaussian transformation and subsequent sampling of the normal modes of vibration.

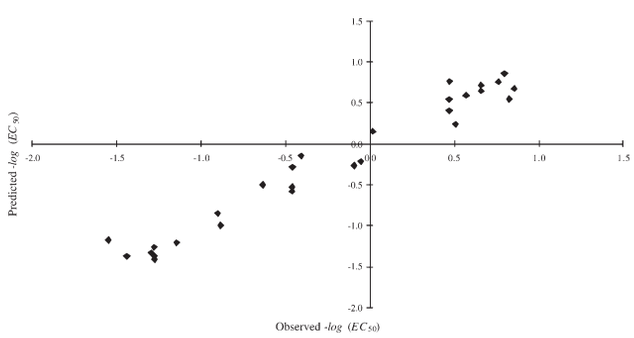


Fig. 9 Plot of predicted *vs.* observed $-\log(EC_{50})$ for the EVA {24,12} fitted regression model with the 28 structure set.

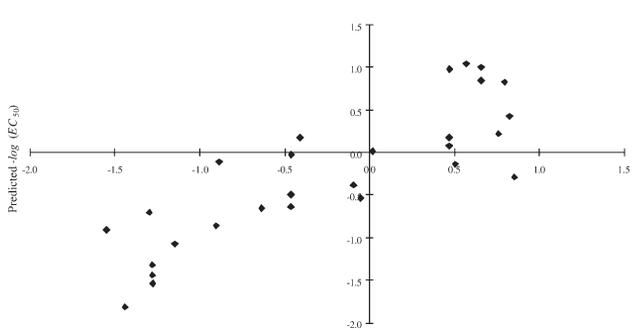


Fig. 10 Plot of predicted *vs.* observed $-\log(EC_{50})$ for the EVA {24,12} cross-validated regression model with the 28 structure set.

The degree of improvement in the log transformed EC_{50} was marked, however, it was not unexpected. As was highlighted earlier, PLS analysis can be highly sensitive to the relative scales of variance between the *x*-block and *y*-variable. Certainly, for the log-transformed *y*-variable set, this variance would have been significant. Furthermore, given the likely experimental error in the bioassay, it is reasonable to postulate that relative

Table 8 Effects upon the QSAR model of augmenting the EVA {24,12} with Hansch-type physicochemical descriptors

PLS model	Var. scaling	Regression Model cumulative q^2 (r^2) PLS 1	PLS 2	PLS 3	PLS 4	PLS 5	Overall LOO CV q^2
EVA {24,12}	—	0.198	0.286	0.359	0.498	*n/s*	0.498
EC_{50}	None	(*0.354*)	(*0.472*)	(*0.707*)	(*0.802*)		
EVA {24,12} clog*P*	None	0.134	0.394	*n/s*	*n/s*	*n/s*	0.394
EC_{50}		(*0.389*)	(*0.517*)				
EVA {24,12} cMR	None	0.006	0.269	0.299	0.433	0.473	0.473
EC_{50}		(*0.082*)	(*0.389*)	(*0.615*)	(*0.742*)	(*0.819*)	
EVA {24,12} cMR clog*P*	None	0.174	0.309	0.412	*n/s*	*n/s*	0.412
EC_{50}		(*0.264*)	(*0.400*)	(*0.537*)			

n/s – not significant *via* cross-validation.

Table 9 The ability of EVA to explain the variance in the Hansch-type physicochemical descriptors

PLS model	Var. scaling	Regression Model q^2 (r^2) PLS 1	PLS 2	PLS 3	PLS 4	Overall LOO CV q^2
EVA {24,12} clog*P*	None	0.281 (*0.415*)	*n/s*	*n/s*	*n/s*	0.281
EVA {24,12} clog*P*	None 1.0	0.281 (*0.415*)	*n/s*	*n/s*	*n/s*	0.281
EVA {24,12} cMR	None	0.873 (*0.893*)	0.906 (*0.936*)	0.959 (*0.976*)	0.963 (*0.983*)	0.963
EVA {24,12} cMR	None 1.0	0.873 (*0.893*)	0.906 (*0.936*)	0.959 (*0.976*)	0.963 (*0.983*)	0.963

n/s – not significant *via* cross-validation.

Table 10 The effects of the removal of compounds with very low relative EC_{50} scores (where $EC_{50} < 0.1$) upon the PLS regression performance of the EVA {24,12} descriptor

PLS model	Var. scaling	Regression Model cumulative q^2 (r^2) PLS 1	PLS 2	PLS 3	PLS 4	PLS 5	PLS 6	Overall LOO CV q^2
Full Structure Set								
EVA {24,12}	None	0.198	0.286	0.359	0.498	*n/s*	*n/s*	0.498
EC_{50}	—	(*0.354*)	(*0.472*)	(*0.707*)	(*0.802*)			
EVA {24,12}	None	0.194	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	0.194
$-\log(EC_{50})$	—	(*0.332*)						
28 Set (>0.1 EC_{50})								
EVA {24,12}	None	0.287	0.356	0.443	0.458	0.483	0.527	0.527
EC_{50}	—	(*0.421*)	(*0.589*)	(*0.734*)	(*0.818*)	(*0.902*)	(*0.939*)	
EVA {24,12}	None	0.483	0.567	0.611	0.669	0.721	*n/s*	0.721
$-\log(EC_{50})$	—	(*0.571*)	(*0.808*)	(*0.843*)	(*0.919*)	(*0.963*)		

n/s – not significant *via* cross-validation.

differences in the log values for these compounds would have been subject to significant error. Consequently, in attempting to generate a model, these data would have had a greater bias in the log-transformed set.

It remains interesting to speculate on whether the sensitivity in model formation was also in part a product of the choice of an essentially linear regression model. While outside the scope of this study, it remains an appealing point for future studies to test the performance of the EVA descriptor across a series of non-linear statistical methods.

3.8 Chemical similarity searching

The formulation of the EVA descriptor as a bit string opens up the possibility of using the encoded vibrational mode information for chemical similarity searching. There is a substantial body of literature that describes the procedures and software available for this purpose and which points out the strengths and pitfalls of the approach.[20–22] Current procedures are based principally on the use of chemical fragments represented as topological maps, atom counts and bond types. Use of EVA for similarity searches would provide useful additional information to complement and extend this structural data. Applications might include selective compound acquisition[20] and identification of compounds with similar biological activities.[21]

4 Conclusions

Within the context of the present study, EVA is well suited to the production of usefully predictive QSARs with both biological and physicochemical data. The absence of a requirement for additional descriptors demonstrates that EVA has a very high information content encompassing, intrinsically, physicochemical descriptors such as clog*P* and

cMR. In direct comparison with a composite descriptor (GRID + clog*P* + cMR), EVA has been shown to have superior predictive performance.

As a general tool for QSAR studies, EVA advantageously may be applied without the prior alignment of structures required for other 3-D descriptors such as COMFA and GRID. However, optimisation of the EVA descriptor by adjusting how it is constructed can influence the predictive power of a QSAR equation.

Further studies with other sets of biological data will be necessary in order to demonstrate the robustness of these conclusions.

Abbreviations

QSAR, Quantitative Structure-Activity Relationship; EVA {σ, *L*, *S*}, *E*igen *VA*lue molecular spectral descriptor model where: σ (sigma) = standard deviation of the EVA Gaussian curve, *L* = sampling interval of summed intensities and *S* = sampling interval start point on the bounded frequency scale; BFS, Bounded Frequency Scale (0 to 4000 cm^{-1}); NCA, Normal Coordinate Analysis; NMs, Normal Modes (of Vibration); EC_{50}, relative force of contraction; clog*P*, calculated log octanol/water partition coefficient; cMR, calculated molar refractivity; PLS, Partial Least Squares regression analysis; r^2, squared correlation coefficient; CV, cross-validation; LOO, 'Leave One Out' cross-validation; LGO, 'Leave One Group Out' cross-validation; q^2, squared correlation coefficient of the cross-validation equation; *PRESS*, Predictive REsidual Sum of Squares; S_{cv}, cross-validated standard error, where $S_{cv} = \sqrt{\frac{PRESS}{(n-c-1)}}$; *n*, number of compounds in data set; *c*, number of PLS components included in QSAR model; *F*-ratio, weighted ratio of r^2 to $1.0 - r^2$, where $F = \frac{r^2/A}{(1-r^2)/(n-c-1)}$

Acknowledgements

The EVA method was originally developed by Shell Research Limited, at the Sittingbourne Research Centre. Subsequent development of the EVA descriptor, in collaboration with Glaxo Wellcome, Tripos Associates and the University of Sheffield, was funded at Portsmouth by the BBSRC ROPA award scheme (Grant number 322/MOL04580; the EVA methodology has been incorporated into Tripos Sybyl since version 6.4. Additional acknowledgement is made to Dr Andy Davis (AstraZeneca Research Loughborough) with respect to information regarding the original GRID study.

References

1 A. M. Ferguson, T. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan and P. J. Snaith, *J. Comput.-Aided Mol. Des.*, 1996, **11**, 143–152.
2 P. Jonathon, W. V. Mcarthy and A. M. I. Roberts, *J Chemom.*, 1996, **10**, 189–213.
3 C. Ginn, D. B. Turner and P. Willett, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 18.
4 A. J. G. Baxter, J. Dixon, F. Ince, C. N. Manners and S. J. Teague, *J. Med. Chem.*, 1993, **36**, 2739–2744.
5 A. M. Davis, N. P. Gensmantel, E. Johansson and D. P. Marriott, *J. Med. Chem.*, 1994, **37**, 963–972.
6 A. M. Davis, Advanced Computer-Assisted Techniques in Drug Discovery, in *Methods and Principles in Medicinal Chemistry*, ed. H. van der Waterbeemd, VCH, Weinheim, 1995, vol. 3, pp. 39–60.
7 P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849–857.
8 M. Itzstein, W. W. Yang, G. B. Kok, M. S. Pegg, J. C. Dyason, B. Jin, T. V. Phan, M. L. Smythe, H. F. White, S. W. Oliver, P. M. Colman, J. N. Varghese, D. M. Ryan, J. M. Woods, R. C. Bethell, V. J. Hotham, J. M. Cameron and C. R. Penn, *Nature*, 1993, **363**, 418–423.
9 A. M. Davis, private communication, 1996.
10 Tripos Associates Inc., 1699 South Hanley Road, Suite 303 St. Louis, Missouri, 63144, USA.
11 *MOPAC manual, 6th Edition*, J. J. P. Stewart, Quantum Chemistry Program Exchange no. 455 (1983), version 6, 1990.
12 *MOPAC: A General Molecular Orbital Package*, J. J. P. Stewart, Quantum Chemistry Program Exchange no. 455 (1983), version 6.0, 1990.
13 A. P. Scott and L. Radom, *J. Phys. Chem.*, 1996, **100**, 16502–16513.
14 M. B. Coolidge, J. E. Marlin and J. J. P. Stewart, *J. Comput. Chem.*, 1991, **12**, 948–952.
15 S. Wold, Chemometric Methods in Molecular Design, in *Methods and Principles in Medicinal Chemistry*, ed. R. Mannhold, P. Krogsgaard-Larson and H. Timmerman, 1995, vol. 2, pp. 195–218.
16 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1–17.
17 J. Shao, *J Am. Stat. Assoc.*, 1993, **88**, 486–494.
18 D. B. Turner, *An Evaluation of a Novel Molecular Descriptor (EVA) for QSAR Studies and the Similarity Searching of Chemical Structure Databases*, PhD thesis, University of Sheffield, 1996.
19 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616–1626.
20 N. Rhodes, P. Willett, J. B. Dunbar Jr. and C. Humblet, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 210–214.
21 L. Xue, J. W. Godden and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 881–886.
22 D. R. Flower, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 379–386.