J-CAMD 382

EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis

A.M. Ferguson*, T. Heritage*, P. Jonathon**, S.E. Pack***, L. Phillips****, J. Rogan***** and P.J. Snaith******

Shell Research Ltd., Sittingbourne Research Centre, Sittingbourne, Kent ME9 8AG, U.K.

Received 4 July 1996
Accepted 9 October 1996

Keywords: Molecular descriptor; EVA; QSAR; QSPR; Drug discovery; log P_{ow}; Normal coordinate analysis

Summary

A new descriptor of molecular structure, EVA, for use in the derivation of robustly predictive QSAR relationships is described. It is based on theoretically derived normal coordinate frequencies, and has been used extensively and successfully in proprietary chemical discovery programmes within Shell Research. As a result of informal dissemination of the methodology, it is now being used successfully in related areas such as pharmaceutical drug discovery. Much of the experimental data used in development remain proprietary, and are not available for publication. This paper describes the method and illustrates its application to the calculation of nonproprietary data, log P_{ow}, in both explanatory and predictive modes. It will be followed by other publications illustrating its application to a range of data derived from biological systems.

Introduction

A major goal within chemistry is to discover chemicals with specific physical, chemical or biochemical properties in order to use them as ‘effect’ or ‘performance’ chemicals such as pharmaceuticals, crop-protection agents, catalysts, etc. In general, these properties tend to be very complex, and are often difficult to summarise numerically and conceptually. Although modern computational methods based on quantum or molecular mechanics allow the reliable simulation of simple molecular properties (such as 3D structure, or even bulk properties such as viscosity), it is impractical to attempt to simulate more complex properties, such as the response of a biological system to interaction with the chemical, from first principles.

A practical solution is to exploit observed empirical correlations between measured chemical and physical properties, an approach stated most clearly by Cramer [1]: "A fundamental objective in scientific research is the discovery of unifying relationships among a body of data. Ideally, such relationships in chemistry are derived from a theoretical model of molecular behaviour, such as the ideal gas law or the Schrödinger equation. Some useful concepts, for example the Hammett equation, have a primarily empirical rationale."

The Hammett equation and its many derivatives [2] are examples of so-called ‘linear free energy relationships’ and are the origins of the widely applied quantitative structure–property and –activity relationships (QSPRs and QSARs, respectively) such as those of Hansch and Leo [3]. These equations are, however, better described as quantitative property–property relationships. For example, a typical approach would be to derive an equation that expresses a performance property, *P* (e.g. a chemical’s activity in a biological test), as a function of a set of physicochemical parameters, such as the octanol–water

*Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, U.S.A.
**Shell Research Ltd., Thornton Research Centre, Thornton-le-Moors, Cheshire, U.K.
***Procter and Gamble Pharmaceuticals, Staines, Middlesex TW18 3AZ, U.K.
****To whom correspondence should be addressed at: Centre for Molecular Design, University of Portsmouth, Halpern House, 1-2 Hampshire Terrace, Portsmouth PO1 2QF, U.K.
*****Shell International Petroleum Company, Shell Centre, London, U.K.
******Shell Français, Paris, France.

partition coefficient, molar polarisability, dipole moment, etc. The equation is derived from a training set of compounds, for each of which the value of P is known from experiment. The collection of physicochemical parameters, which may be measured experimentally or calculated empirically or theoretically, is used as a proxy for molecular structure (i.e. as a 'molecular descriptor', M).

Generally, P is expressed as:

$$P = \mathrm{f(M)} + \mathrm{E} + \mathrm{e}$$

where f is typically a regression function (e.g. derived from multiple linear, principal components or continuum regression models), a binary discriminant function which permits classification (for example into classes of 'active' and 'inactive' molecules), or some other form of linear or nonlinear model; E and e are the errors associated with the molecular descriptors M and the performance properties P.

Typically, such relationships apply only to a specific region of chemistry (e.g. to a homologous series, or one in which there are comparatively simple structural differences between molecules which have a common framework). At best they are explanatory, but overinterpretation is common due to a tendency to assume that correlation implies 'cause and effect'. Frequently, attempts are made to use the equations to predict the properties of interest for some hypothetical structures, to guide say an analogue synthesis programme in which there may be many thousands of alternative possibilities to explore.

One approach may be to maximise P in terms of M, and make a chemical that possesses the implied optimal values of the physicochemical parameters within M. A major limitation lies, however, not only in the inadequacy of the set of physicochemical parameters to act as a comprehensive molecular descriptor, but also in the fact that the parameters themselves may be highly intercorrelated, i.e. they are not independent variables. The predictions can only be interpolative, since it is impossible to extrapolate to structures whose features are not explicitly within the training set. In addition, there may well be many hypothetical structures which would fit the specification implied by the optimal set of values for the physicochemical parameters, but the majority would lie outside the structural space bounded by the model; the resulting conclusions would be ill-founded.

Cramer et al. [4] have devised an alternative molecular descriptor, for use with biological response data, which has the acronym CoMFA (Comparative Molecular Field Analysis). As the descriptor, this uses a string of n numbers (n typically greater than 2000) representing the values of the van der Waals and electrostatic energies at each of n/2 points sampled in a uniform cubic array centred on a molecule. Again the approach is to use a training set of structures corresponding to chemicals whose properties of interest (in this case response in a biological system) have been measured. Within the SYBYL molecular modelling package, a multivariate statistics method, in the form of partial least squares (PLS) [5], is used to derive a calibrative function, f. A well-recognised limitation of CoMFA [6] lies in the need to superimpose *all* the molecular structures in the training set, and with so many degrees of freedom (internal motions as well as rigid body rotations) this is difficult and arbitrary unless there are sufficient reasons to assume a particular conformation and common centroid for the overlap. With care, however, the method is useful for molecules of closely related structure and, despite having many more variables than observations, may perform significantly better than the older methods based on a linear combination of univariate descriptors. Since 1988, we have been developing a different approach in which we sought to utilise information from a vibrational normal coordinate analysis as the molecular descriptor, together with a variety of multivariate statistical methods. This has proved successful initially with performance data derived from both in vitro and in vivo measurements in biological systems relevant to the discovery of new crop-protection chemicals and subsequently in broader areas of molecular design. The method is versatile, overcomes alignment-related limitations of the CoMFA approach and seems able to encompass greater changes in molecular structure than are implicit in the descriptor rather than explicit. Our experience has shown that it implicitly covers a large volume of structural space, and as a consequence its predictive powers extend outside the structural features which are explicit within the molecules chosen for the training set. (In principle, CoMFA also does this.) Subsequent papers will examine the process of transformation of the property of interest into the new vibrational descriptor EVA (vide infra), and show a comparison of EVA with CoMFA on the same data sets of structure and biological property, using the same relational analysis.

The purpose of the present paper is to introduce and define the methodology, and to illustrate its application by analysing, inter alia, the octanol–water partition coefficients of a set of compounds of very diverse structure. This set of chemicals was used by Cramer [1] in an earlier factor analysis of physicochemical properties via principal components designated as 'BCDEF'.

The method

An ideal molecular descriptor should, in numerical form, encode all the features of structure which determine chemical and physical properties, conventionally the constituent atoms, their bonding and spatial relationships, molecular shape and size, and electronics allowing for modifications due to interaction of the molecule with its environment. In terms of quantum mechanics, the most complete descriptor of molecular structure and properties is the molecular wave function, which characterises nuclear and electronic properties fully. In principle, the interaction of any molecule with any electromagnetic field can be predicted precisely by solving the appropriate Schrödinger equation. Therefore, all molecular properties (including interactions with biological systems) can be quantified if we know the molecular wave function.

Within the Born–Oppenheimer approximation, the electronic potential energy function is the prime tool used by theoreticians to rationalise and predict chemical phenomena. It can be regarded as an encoding of the molecular structure which contains sufficient information to make it a comprehensive molecular descriptor; the difficulty remains, however, in trying to extract the information pertinent to QSPR from the potential energy function. In principle, it should be possible to do so by using the modern methods of multivariate statistics in what is essentially a 'chemometric' analysis.

Following preliminary studies with experimentally derived mid- and near-infrared spectroscopy (to be published elsewhere), our approach is to use, as a multivariate descriptor, a fundamental molecular property which may be characterised reliably and easily from the potential energy function, namely vibration. By means of a classical analysis [7], the normal coordinate eigenvalues and eigenvectors (corresponding to vibrational frequencies and atomic displacements) can be calculated using standard quantum or molecular mechanical methods. Given their implicit dependence on the molecular wave function, it is reasonable to assume that these properties should well-characterise molecular structure and offer a unique molecular description containing adequate information on shape, size and electronic properties.

This paper introduces a molecular descriptor, based on the normal coordinate EigenVAlues, which we refer to as 'EVA'. Together with recently developed multivariate statistical techniques, we have used this descriptor successfully for some 5 years to analyse biological performance data and to produce usefully predictive QSPRs where other methods (such as those discussed in Ref. 3 or Ref. 4) have been less effective. The objective of the present paper is to describe the process and illustrate its wider application.

A major limitation of current statistical methods is the need for the descriptor for each molecule in the training set to comprise the same number of well-defined components (i.e. the molecular descriptor must be a vector of fixed dimensionality). 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 will generally be of different dimensionality. Data transformation is therefore required to convert, without loss of information, a varying number of vibrational eigenvalues into the descriptor EVA which has fixed dimensionality.

We do this by projecting the eigenvalues (vibrational frequencies) onto a bounded frequency scale with individual vibrations represented by points along this axis. The frequency range is chosen to be 0–4000 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 10–20 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 detect vibrational 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 which inevitably causes 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, we sample the resultant spectrum across its whole width in fixed increments, typically of width 2 or 5 cm^{-1}. The choice of increment determines the number of variables in the EVA descriptor; thus, for example, a 2 cm^{-1} increment would result 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. We have, however, introduced three new variables into the model, namely the frequency range, the width of the sampling increment and pseudo-'intensities' of vibrations. (These should not be confused with the intensities associated with lines in experimentally derived vibrational spectra such as infrared or Raman; we do not calculate transition moments, and each normal mode is of unit intensity.) The choice of these new variables can have a marked effect on the quality of a subsequently derived QSAR/QSPR model, and this will be discussed in a subsequent paper.

Two types of descriptor transformation have been considered, namely normalisation and scaling. The influence of these on the analysis will also be discussed elsewhere, but briefly.

(1) Data normalisation applies to the EVA descriptor separately for each molecule. It corresponds to normalising the total descriptor intensity to unity (i.e. 'total absorption' normalisation), thereby preserving relative peak intensities. However, since the total number of peaks is a function of the number of atoms present in the molecule, normalisation effectively reduces the descriptor's sensitivity to molecular size.

TABLE 1

CALCULATED AND EXPERIMENTAL log P_{ow} VALUES FOR A TRAINING SET OF MOLECULES

Compound name	log P_{ow} (exp.)	log P_{ow} (calc.)	log P_{ow} (exp. – calc.)	log P_{ow} (pred.)	log P_{ow} (exp. – pred.)
1,1,1-Trichloroethane	2.49	2.07	0.42	1.05	1.44
1,1-Dichloroethane	1.79	1.88	–0.09	1.42	0.37
1,1-Dichlorotetrafluoroethane	2.82	2.79	0.03	0.78	2.04
1,1-Difluoroethane	0.75	0.52	0.23	0.55	0.2
1,2-Dichlorobenzene	3.38	3.58	–0.02	2.32	1.06
1,2-Dichloroethane	1.48	1.99	–0.51	1.85	–0.37
1,3-Butadiene	1.99	1.81	0.18	1.29	0.7
1,3-Dichlorobenzene	3.38	3.14	0.24	2.43	0.95
1,3-Dichloropropane	2.01	2.03	–0.03	2.18	–0.17
1,4-Dichlorobenzene	3.39	3.28	0.11	3.21	0.18
1,4-Dioxane	–0.42	0.19	–0.61	0.51	–0.94
1,4-Pentadiene	1.48	2.03	–0.55	1.75	–0.27
1-Bromopropane	2.11	1.94	0.17	2.71	–0.61
1-Butanol	0.88	0.89	–0.01	1.22	–0.34
1-Butene	2.41	2.09	0.32	1.88	0.53
1-Chlorobutane	2.64	2.63	0.01	2.83	–0.19
1-Chloropropane	2.04	2.33	–0.29	2.36	–0.32
1-Hexanol	2.03	1.98	0.05	1.99	0.04
1-Nitropropane	0.87	0.81	0.06	1.15	–0.28
1-Octanol	3.15	2.67	0.48	1.81	1.34
1-Pentanol	1.39	1.55	–0.16	1.37	0.02
1-Pentyne	1.98	1.94	0.04	1.89	0.09
1-Propanol	0.31	0.32	–0.01	0.63	–0.32
2,2-Dimethylbutane	3.82	3.84	–0.02	3.21	0.61
2,2–Dimethylpropane	3.11	3.04	0.67	2.98	0.13
2,6-Dimethylpyridine	1.68	1.65	0.03	1.15	0.53
2-Butanol	0.61	0.72	–0.11	1.02	–0.41
2-Butanone	0.29	0.36	–0.07	0.14	0.15
2-Chloropropane	1.91	1.85	0.46	1.99	–0.08
2-Cresol	1.95	2.31	–0.36	2.46	–0.51
2-Ethylpyridine	1.69	1.88	–0.19	1.56	0.13
2-Methyl-2-butanol	0.89	1.02	–0.13	1.71	–0.83
2-Methylpropane	2.76	2.36	0.4	2.37	0.39
2-Methylpropene	2.34	2.4	–0.06	1.08	1.26
2-Methylpyrazine	0.23	0.89	–0.66	2.29	–2.06
2-Methylpyridine	1.11	1.26	–0.15	1.38	–0.27
2-Nitrotoluene	2.31	2.09	0.22	1.42	0.89
2-Pentanone	0.91	0.99	–0.08	1.32	–0.41
2-Propanol	0.05	–0.17	0.22	–0.28	0.33
2-Propylbenzene	3.66	3.34	0.32	3.09	0.57
3-Methylpyridine	1.24	1.63	–0.39	1.79	–0.55
3-Nitrotoluene	2.45	1.96	0.49	1.62	0.83
4-Cresol	1.95	1.64	0.31	2.71	–0.76
4-Methylpyridine	1.22	1.02	0.2	0.72	0.5
4-*tert*-Butylphenol	2.94	2.86	0.08	3.98	–1.04
Acetamide	–1.15	–1.11	–0.04	0.33	–1.48
Acetic acid	–0.17	–0.35	0.18	–0.09	–0.08
Acetone	–0.24	–0.43	0.19	–0.42	0.18
Acetonitrile	–0.34	–0.38	0.04	0.55	–0.89
Acetophenone	1.73	1.69	0.04	2.85	–1.12
Acetylene	0.37	0.57	–0.2	0.99	–0.62
Allyl alcohol	0.17	0.16	0.01	0.33	–0.16
Anisole	2.09	2.13	–0.04	1.05	1.04
Anthracene	4.45	4.43	0.02	5.06	–0.61
Benzaldehyde	1.48	1.91	–0.43	2.59	–1.11
Benzene	2.13	2.36	–0.23	3.24	–1.11
Bromobenzene	2.99	2.62	0.37	2.74	0.25
Bromoethane	1.61	1.87	–0.26	1.71	–0.1
Bromomethane	1.19	0.97	0.22	0.09	1.1

TABLE 1 (continued)

Compound name	log P_{ow} (exp.)	log P_{ow} (calc.)	log P_{ow} (exp. – calc.)	log P_{ow} (pred.)	log P_{ow} (exp. – pred.)
Bromotrifluoromethane	1.86	1.8	0.06	1.27	0.59
Butylamine	0.88	0.77	0.11	0.78	0.1
Butyric acid	0.79	0.83	–0.04	0.87	–0.08
Chlorobenzene	2.84	3.19	–0.35	2.98	–0.14
Chlorodifluoromethane	1.08	1.44	–0.36	1.36	–0.28
Chloroethane	1.43	1.8	–0.37	1.43	0
Chloromethane	0.91	0.85	0.06	0.26	0.65
Chlorotrifluoromethane	1.65	1.69	–0.04	1.51	0.14
Cyclohexane	3.44	3.49	–0.05	3.01	0.43
Cyclohexanol	1.23	1.29	–0.06	1.46	–0.23
Cyclohexene	2.86	2.82	0.04	1.89	0.97
Cyclopentane	2.99	2.22	0.77	0.69	2.3
Dibutylamine	2.68	2.84	–0.16	2.45	0.23
Dichlorofluoromethane	2.16	1.82	0.34	1.63	0.53
Dichloromethane	1.25	1.55	–0.3	1.15	0.1
Diethylamine	0.57	0.79	–0.22	0.57	0
Diethyl ether	0.77	0.89	–0.12	1.07	–0.3
Diethyl sulphide	1.95	1.94	0.01	2.64	–0.69
Dimethylacetamide	–0.77	–0.95	0.18	–0.46	–0.31
Dimethyl ether	0.1	–0.55	0.65	–1.02	1.12
Di-*n*-propylamine	1.73	1.8	–0.07	1.74	–0.01
Di-*n*-propyl ether	2.03	2.35	–0.32	2.46	–0.43
Ethane	1.81	1.84	–0.03	1.71	0.1
Ethanol	–0.31	–0.31	0	–0.06	–0.25
Ethyl acetate	0.73	0.79	–0.06	0.27	0.46
Ethyl propionate	1.21	1.26	–0.05	1.28	–0.07
Ethylamine	–0.13	0.09	–0.22	0.48	–0.61
Ethylbenzene	3.15	2.99	0.16	3.35	–0.2
Ethylene	1.13	1.14	–0.01	1.42	–0.29
Ethylene glycol	–1.93	–1.45	–0.48	0.59	–2.52
Fluoromethane	0.51	0.67	–0.16	0.63	–0.12
Hexylamine	1.98	2.19	–0.21	2.31	–0.33
Iodoethane	2.01	1.73	0.28	1.36	0.65
Iodomethane	1.69	1.41	0.28	1.29	0.4
m-Xylene	3.19	2.78	0.41	2.16	1.03
Methane	0.79	0.65	0.14	0.93	–0.14
Methanol	–0.64	–1.04	0.4	–0.29	–0.35
Methyl acetate	0.18	0.44	–0.26	–0.04	0.22
Methyl benzoate	2.12	2.16	–0.04	2.28	–0.16
Methylacetamide	–1.1	–1.1	0	–0.09	–1.01
n-Butane	2.89	2.48	0.41	2.76	0.13
n-Butylbenzene	4.26	4.21	0.05	3.88	0.38
n-Pentane	3.23	3.27	–0.04	2.42	0.81
Naphthalene	3.59	3.69	–0.1	3.99	–0.4
Nitrobenzene	1.85	2.07	–0.22	2.53	–0.68
Nitroethane	0.18	–0.09	0.27	1.09	–0.91
o-Xylene	2.77	2.63	0.14	1.89	0.88
p-Xylene	3.15	3.26	–0.11	3.02	0.13
Pentylamine	1.49	1.52	–0.03	1.67	–0.18
Phenanthrene	4.46	4.73	–0.27	4.92	–0.46
Phenol	1.48	1.34	0.14	1.33	0.15
Piperidine	0.85	0.81	0.04	1.39	–0.54
Propane	2.36	2.34	0.02	2.61	–0.25
Propionic acid	0.33	–0.04	0.37	0.1	0.23
Propionitrile	0.16	0.27	–0.11	1.23	–1.07
n-Propyl acetate	0.83	0.86	–0.03	1.03	–0.2
n-Propylamine	0.48	0.39	0.09	0.82	–0.34
n-Propylbenzene	3.68	3.7	–0.02	3.68	0
Propene	1.77	1.41	0.36	1.07	0.7

TABLE 1 (continued)

Compound name	log P_{ow} (exp.)	log P_{ow} (calc.)	log P_{ow} (exp. – calc.)	log P_{ow} (pred.)	log P_{ow} (exp. – pred.)
Propyne	0.94	1.02	–0.08	0.74	0.2
Pyridine	0.62	0.52	0.1	0.59	0.03
Pyrrolidine	0.46	0.42	0.04	1.07	–0.61
t-Butanol	0.37	0.51	–0.14	0.83	–0.46
t-Butylbenzene	4.11	3.92	0.19	4.03	0.08
Tetrachloromethane	2.83	3.1	–0.27	2.57	0.26
Tetrafluoromethane	1.18	1.61	–0.43	–0.06	1.24
Tetrahydrofuran	0.46	0.65	–0.19	0.74	–0.28
Thioanisole	2.74	2.74	0	2.19	0.55
Thiophenol	2.52	2.41	0.11	2.53	–0.01
Toluene	2.69	2.64	0.05	2.81	–0.12
Trichloroethene	2.29	2.18	0.11	1.24	1.05
Trichloromethane	1.97	1.97	0	1.51	0.46
Triethylamine	1.44	1.38	0.06	1.84	–0.4
Trifluoromethane	0.64	1.34	–0.7	1.19	–0.55
Trimethylamine	0.27	0.26	0.01	–0.55	0.82
Water	–1.38	–1.33	–0.05	0.48	–1.86

log P_{ow} (exp.) = experimentally determined log(octanol–water partition coefficient).
log P_{ow} (calc.) = log P_{ow} calculated from the full regression equation.
log P_{ow} (exp. – calc.) = log P_{ow} (exp.) – log P_{ow} (calc.).
log P_{ow} (pred.) = log P_{ow} 'predicted' by leave-one-out cross-validation.
log P_{ow} (exp. – pred.) = log P_{ow} (exp.) – log P_{ow} (pred.).

(2) Data scaling is performed on the descriptors for a set of molecules as a whole. By setting the variance of each frequency increment to unity, equal importance is placed upon them in subsequent statistical analysis. A consequence of this is that the original descriptor profiles are not preserved.

In summary, the EVA descriptor involves projecting molecular vibrational frequencies that have been transformed into Gaussian form onto a bounded frequency axis which is sampled in fixed increments along the axis. The similarities and differences between the EVA descriptors of a set of molecules are then compared statistically in terms of the overlap of projected intensities.

It should be noted that the 'EVA transformation' applied to vibrational frequencies can in essence be applied to a whole range of properties. Preliminary studies have

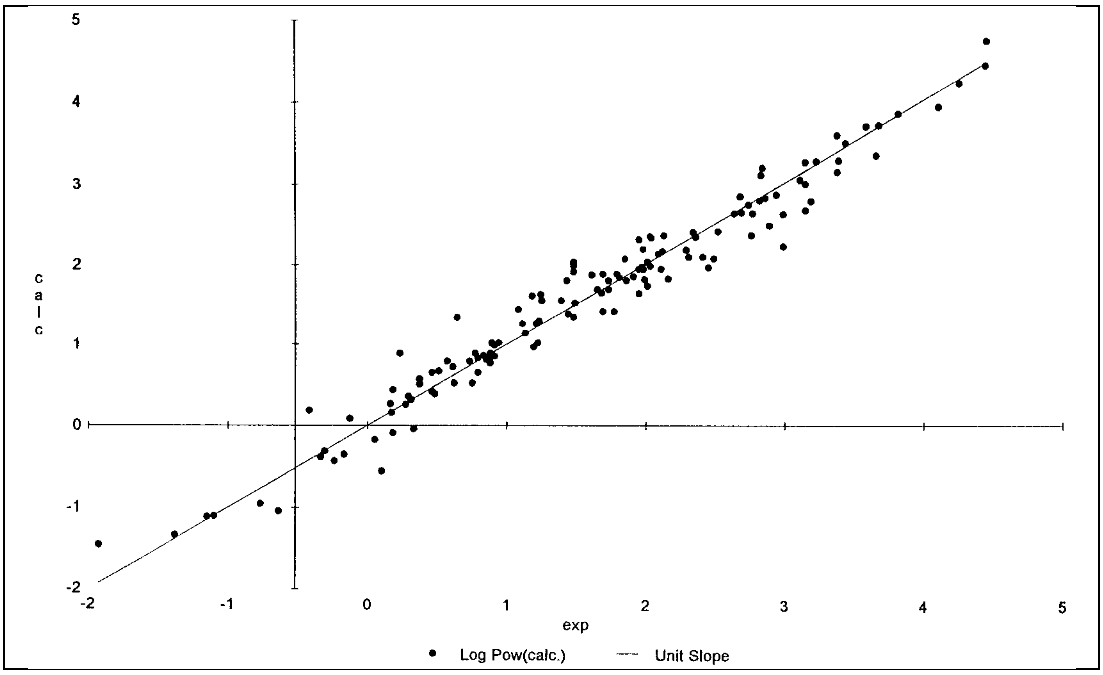

Fig. 1. log P_{ow} (training set; full-fit regression).

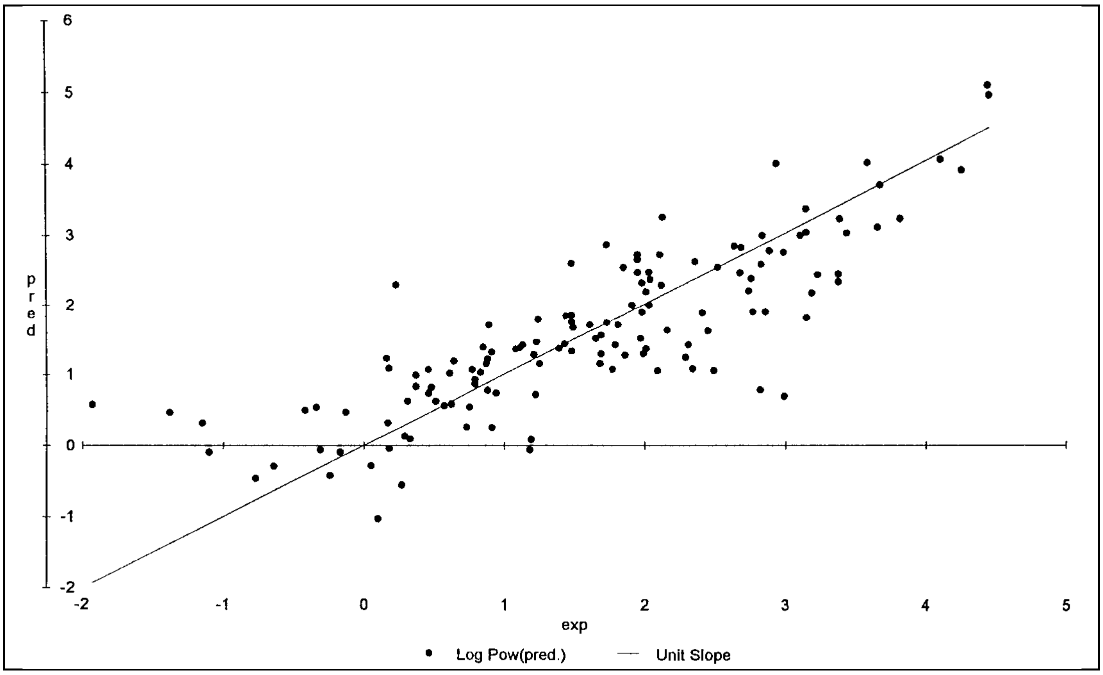

Fig. 2. log P_{ow} (training set; cross-validated prediction).

demonstrated that robust QSAR/QSPR models can result from application of the same principles to other properties derived from the molecular electronic wave function, and also to experimental properties (A.M. Ferguson, unpublished results proprietary to Shell Research Ltd.).

Experimental details and results

Data sets which were used to develop and validate the method described above remain proprietary and confidential to Shell. Consequently, we give here an illustration of the use of EVA with a data set available in the open literature; the present study was based on the 142 chemicals in the training set used by Cramer [1] in his wholly empirical factor analysis of physical properties. Of this set, only 135 were considered here because of limitations in the semiempirical molecular orbital method used to derive the normal modes. The structures are listed in Table 1, together with reported values of their octanol–water partition coefficients, log P_{ow} [8].

All structures were created initially in SMILES notation, and the SMILES strings were processed using CONCORD v. 2.9.3 to yield 3D coordinates for all the structures. These were read into SYBYL 6.0 and the conformation of minimum energy was derived using the BFGS optimiser. The normal coordinate frequencies were generated using the MOPAC 6.0 AM1 Hamiltonian (keywords: FORCE, SCFCRT=1.D-12, GNORM=0.05). They were then converted into EVA descriptors, as described above, using a Gaussian width of 10 cm^{-1} and a sampling increment of 5 cm^{-1}. (For any given molecule, the EVA descriptor varies with conformation; the consequences of this will be discussed in a subsequent paper, but in the present study the variations in EVA between different molecules are likely to be much larger than the variations in EVA between conformations of any individual. The adoption of minimum-energy conformation therefore provides a convenient standard state, with little effect on the subsequent analysis.) For each of the 135 chemicals listed in Table 1, there corresponds an EVA descriptor consisting of a string of 800 numbers representing the summed intensities within the increments of 5 cm^{-1}. Effectively we have a matrix of 135 rows of 800 variables, and to each row there corresponds an observable – an experimental value of log P_{ow}.

The matrix of variables was then regressed onto the log P_{ow} values using the method of partial least squares (PLS) [5] encoded within SAS. A regression equation using only six PLS factors was produced which explained 96% of the variation in the data ($R^2=0.96$). This is illustrated in Fig. 1 and the individual values of log P_{ow} calculated by this equation are listed in Table 1 alongside the experimental values. In this analysis, data transformation (normalisation or scaling) was not found to be beneficial.

Discussion

The above analysis shows clearly that the EVA descriptor – derived solely from theoretical methods – contains sufficient information to explain the way in which an important physicochemical parameter, log P_{ow}, varies across a remarkably wide range of chemicals. These compounds cover the most commonly encountered structural types, varying from the large polycyclic aromatic (and highly

TABLE 2
CALCULATED (PREDICTED) AND EXPERIMENTAL log P_{ow} VALUES FOR A PREDICTION SET OF MOLECULES

Compound name	log P_{ow}	log P_{ow}	log P_{ow}	Compound name	log P_{ow}	log P_{ow}	log P_{ow}
	(exp.)	(pred.)	(exp. – pred.)		(exp.)	(pred.)	(exp. – pred.)
1,2,3,5-Tetramethylbenzene	4.17	2.91	1.26	*N*,*N*-Dimethylacetamide	–0.77	–0.58	–0.19
1,2,3-Trimethylbenzene	3.66	3.4	0.26	*N*,*N*-Dimethylaniline	2.31	2.19	0.12
1,3,5-Trimethylbenzene	3.42	2.71	0.71	Diphenylamine	3.5	5	–1.5
1,2-Diaminoethane	–2.04	–0.47	–1.57	Di-*n*-propylamine	1.67	1.57	0.1
1,2-Dichloroethene	2.09	1.51	0.58	Durene	4	3.3	0.7
1,3-Dichloropropane	2	2.28	–0.28	*N*-Ethylaniline	2.16	1.96	0.2
1,4-Pentadiene	2.48	1.21	1.27	Ethylene oxide	–0.3	0.76	–1.06
1-Heptanol	2.72	2.44	0.28	Ethyl phenyl ether	2.51	1.71	0.8
1-Hexene	3.39	2.45	0.94	Fluorobenzene	2.27	1.76	0.51
1-Pentyne	1.98	2.04	–0.06	Furan	1.34	0.31	1.03
2,3-Dimethylbutane	3.85	3.42	0.43	Glycerol	–1.76	–0.05	–1.71
2,6-Dimethylpyridine	1.68	1.28	0.4	Heptene	3.99	3.01	0.98
2-Aminopropane	0.26	0.8	–0.54	Hexafluoroethane	2	1.4	0.6
cis-2-Butene	2.33	0.58	1.75	Hexanal	1.79	2.45	–0.66
2-Ethylpyridine	1.69	1.56	0.13	Hexanoic acid	1.92	2.38	–0.46
2-Hexanone	1.38	1.31	0.07	Hexylamine	2.06	2.38	–0.32
2-Methylnaphthalene	3.86	3.61	0.25	Iodobenzene	3.25	2.5	0.75
2-Methylpyrazine	0.23	1.27	–1.04	*N*-Methylacetamide	–1.05	–0.37	–0.68
2-Methylpyridine	1.11	1.1	0.01	*N*-Methylaniline	1.66	1.16	0.5
3-Methylpyridine	1.2	1.38	–0.18	Methylcyclopentane	3.37	2.15	1.22
4-Methylpyridine	1.22	0.69	0.53	Nitromethane	–0.35	0.21	–0.56
2-Naphthol	2.7	3.23	–0.53	Pentanoic acid	1.39	2.46	–1.07
2-Pentanone	0.91	1.21	–0.3	Pentylamine	1.49	0.87	0.62
Acrolein	–0.01	1.01	–1.02	Phenanthrene	4.46	4.98	–0.52
Allene	1.45	1.29	0.16	Propionic acid	0.33	0.52	–0.19
Aniline	0.9	2.21	–1.31	Pyrrolidine	0.46	1.13	–0.67
Anthracene	4.45	4.65	–0.2	Quinoline	2.03	3.28	–1.25
n-Butylamine	0.97	1.47	–0.5	Styrene	2.95	3.12	–0.17
Butyronitrile	0.56	1.1	–0.54	Tetrachloroethane	2.39	2.02	0.37
Chloroacetic acid	0.22	0.39	–0.17	Tetrachloroethene	3.4	2.2	1.2
m-Cresol	1.96	1.48	0.48	Thiazole	0.44	0.91	–0.47
Cyclohexene	2.86	2.46	0.4	Thiomethylbenzene	2.74	0.5	2.24
Cyclopropane	1.72	0.19	1.53	Thiophene	1.81	0.91	0.9
Di-*n*-butylamine	2.83	2.99	–0.16	Thymol	3.3	2.59	0.71
Dichloroacetic acid	0.92	0.98	–0.06	*p*-Toluidine	1.39	3.16	–1.77
Dichlorofluoromethane	1.55	1.87	–0.32	Triethylamine	1.45	0.9	0.55
Diisopropylether	1.52	0.68	0.84	Trifluoromethane	0.64	1.07	–0.43
Dimethylamine	–0.38	–0.08	–0.3	Tri-*n*-propylamine	2.79	2.24	0.55

log P_{ow} (exp.) = experimentally determined log(octanol–water partition coefficient).
log P_{ow} (pred.) = log P_{ow} calculated from the regression equation derived using data in Table 1.
log P_{ow} (exp. – pred.) = log P_{ow} (exp.) – log P_{ow} (pred.).

lipophilic) phenanthrene (log P_{ow} = 4.46) to small hydrophilic chemicals such as methanol (log P_{ow} = –0.64). To be really useful, however, the analysis should yield a regression equation which is capable of predicting values of log P_{ow} for chemicals outside the data set used to derive the relationship; the quality of fit displayed in Fig. 1 can be misleading in this context.

In order to get a more meaningful measure of the predictive capability of the relationship, we apply the 'leave-one-out' cross-validation or 'jackknife' technique [9], whereby each molecular descriptor and its corresponding observable (experimental log P_{ow} value) are removed in turn from the data set, and a new regression equation is calculated at each iteration. This equation is then used to calculate a value of log P_{ow} for the chemical whose descriptor and observable were omitted. This builds up a set of log P_{ow} values which are 'real predictions' in the sense that the regression equation from which they are derived contains no prior information about them. The results of this operation on the 'training' set of 135 compounds are also shown in Table 1, together with the corresponding experimental data and the values calculated from the full regression equation. The cross-validated 'predictive' R^2, a robust indicator of predictive power, is 0.68 and the results are shown graphically in Fig. 2. Given the approximations inherent in the use of a semiempirical quantum mechanical method applied to isolated molecules, it is gratifying that the value of R^2 is so high. Not surprisingly, the poorest predictions are for compounds such as ethylene glycol with its strong intermolecular associations. As defined by Cramer [1], $R^2 = 1 - (1/ns^2)(\text{pred.} - \text{exp.})^2$, where n is the number of predictions and s is the standard deviation of a 'typical' set of property values (in this case the training data set values); pred. is the predicted value corresponding to an experimental value, exp.

Perhaps a more severe and more convincing test is to use the full regression equation to predict the values of log P_{ow} for an entirely new set of compounds, and to examine how well these predictions compare with experimental values. For this purpose we took a subset of those used by Cramer as the 'prediction set' upon which he tested his BCDEF analysis [10]. The 76 compounds we chose were those for which we could locate reliable values for log P_{ow} from the MEDCHEM database [8]. These are listed in Table 2, together with the values calculated for log P_{ow}; Fig. 3 illustrates these results graphically. The value of R^2 for these predictions is 0.65, sufficiently similar to the 'cross-validated' R^2 of 0.68 obtained for the training set to confirm the conclusion that the use of EVA as a molecular descriptor, together with PLS regression, provides a usefully predictive, as well as explanatory, tool for an important physicochemical property of a set of chemicals of remarkably diverse structures. This is further illustrated in Fig. 4, which shows the results of plotting the cross-validated results derived from the training set together with the results of unsupervised prediction; they clearly lie within the same space.

It is instructive to apply the criteria recently developed elsewhere in order to evaluate the quality and significance of these results. Wold [11], for example, makes use of the ratio of the PRESS (Predictive REsidual Sum of Squares) statistic to SSY, the sum of squares of the response values (in this case the experimental values of log P_{ow}). In the present example, PRESS = 69.06 and SSY = 556.59, giving a PRESS/SSY ratio of 0.12. According to Wold, a value of 0.4 indicates a 'reasonable' QSPR, and anything less than 0.1 is 'excellent'; our result of 0.12 with such a broad set of structures speaks for itself. Using the criteria laid down by Cramer, the probability of a PLS correlation of this quality arising by chance is vanishingly small [12], covering as it does 135 response data with a cross-validated R^2 of 0.68.

To the best of our knowledge, this is the first attempt to describe the variation of a fundamental physical property of a broad range of chemicals by means of a chemometric analysis of descriptions of molecular structure so closely derived from the molecular potential energy function. The only previous work with which the results may be compared is that of Cramer [1,10], from which the examples for this study were taken. Cramer's approach is a wholly empirical correlation analysis, using factor analysis to derive a relationship between certain physical properties of a training set of 114 chemicals and a linear combination of six physicochemical parameters (aqueous activity coefficient, partition coefficient, molar refractivity, boiling point, molar volume and heat of vaporisation). Inevitably, such an approach is internally self-consistent. Thus, the five principal components BCDEF derived from the set of six properties mentioned above can be decomposed to 'predict' values of each of these

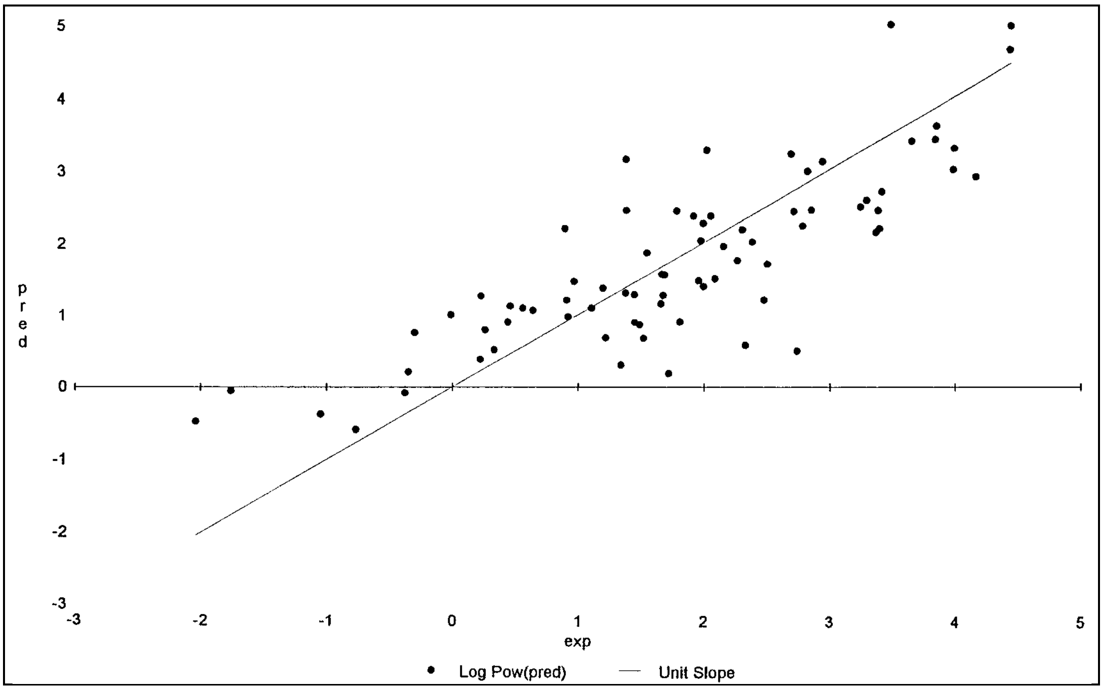

Fig. 3. log P_{ow} (prediction set; cross-validated prediction).

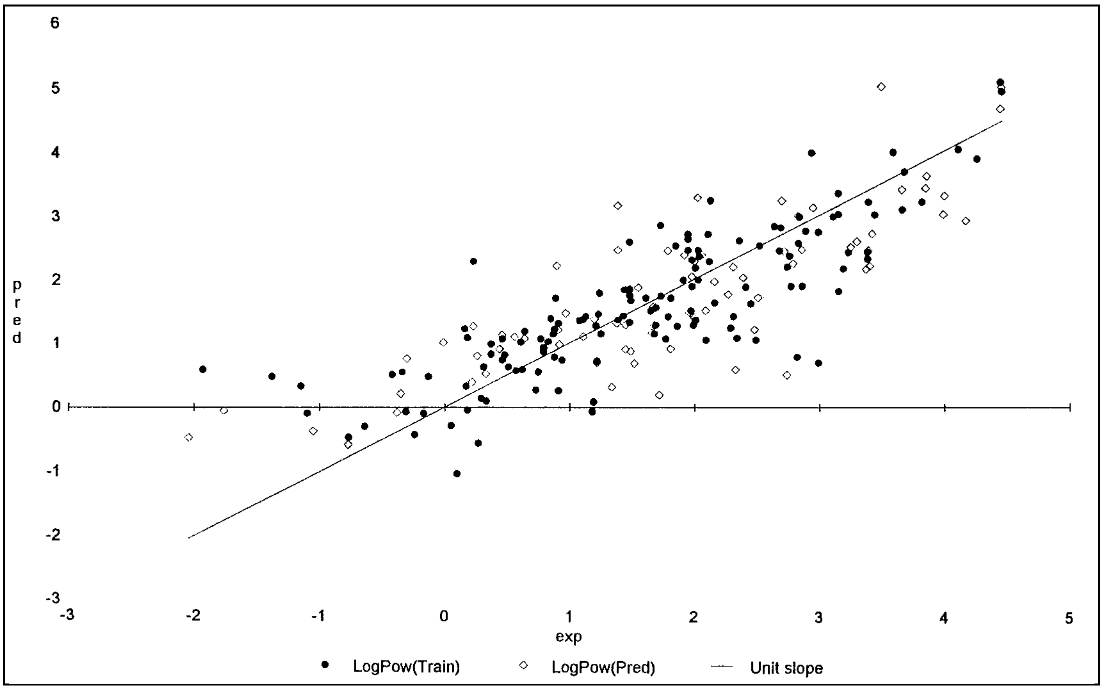

Fig. 4. log P_{ow} (predictions; training set and prediction data set).

properties for every compound in the training set as an inevitable consequence of the mathematics involved.

Regression of BCDEF onto sets of observables corresponding to measured values of other physicochemical properties, such as viscosity, boiling point, dielectric constant, etc., produces equations which explain the variation in these parameters in terms of their (nonobvious) underlying correlations with the original set of six properties. Although there is no attempt at cross-validation, application of the regression equations to a 'prediction set' of chemicals [10] does demonstrate that the method has predictive power. The derivation of BCDEF values for molecules outside the training set is, however, very complex and relies on an additive-constitutive approach based on molecular fragments present in the training set; it is inevitable that physicochemical parameters predicted in this way will be doubly confounded with the data of the training set, and it is difficult to assess the true potential of the method as a predictive tool.

In contrast, the method described in the present work is free from such limitations. The PLS factors are derived from a quantum mechanical picture of structure, free from contamination by any measured physical property (with the exception, arguably, of those used in the semi-empirical parameterisation of the MOPAC AM1 Hamiltonian). The calculation of the EVA descriptor for any structure, either in or out of the training set, is a trivial and independent operation. Subsequent papers will describe the application of this methodology to the more familiar area of drug design, using examples taken from research into crop-protection chemicals.

Acknowledgements

We take pleasure in acknowledging the contributions made to the development and subsequent use of the methods described herein by the following former members of Staff of the Sittingbourne Research Centre: Dr. Andrew Dearing, Dr. Michael King, Mr. 'Mac' McCarthy, Mr. Adrian Roberts, Dr. Mervyn Thomas and the late Dr. James Cox.

References

1 Cramer III, R.D., J. Am. Chem. Soc., 102 (1980) 1837.
2 Johnson, C.D., The Hammett Equation, Cambridge University Press, Cambridge, U.K., 1973.
3 Hansch, C., Sammes, P.G., Taylor, J.B. and Ramsden, C.A. (Eds.) Comprehensive Medicinal Chemistry, Vol. 4, Pergamon, Oxford, U.K., 1990.
4 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
5 Stahle, L. and Wold, S., Prog. Med. Chem., 25 (1988) 292.
6 Sun Jin Cho and Tropsha, A., J. Med. Chem., 38 (1995) 1060.
7 Golton, A.V., In Straughan, B.P. and Walker, S. (Eds.), Spectroscopy, Vol. 2, Chapman and Hall, London, U.K., 1976, p. 9.
8 Leo, A.J., MEDCHEM database, Pomona College Medicinal Chemistry Project, Claremont College, CA, U.S.A.
9 Cramer III, R.D., Bunce, J.D., Patterson, D.E. and Frank, I., Quant. Struct.–Act. Relatsh., 7 (1988) 18.
10 Cramer III, R.D., J. Am. Chem. Soc., 102 (1980) 1849.
11 Wold, S., Quant. Struct.–Act. Relatsh., 10 (1991) 191.
12 Clark, M. and Cramer III, R.D., Quant. Struct.–Act. Relatsh., 12 (1993) 137.