

Theoretical prediction of the polarity/polarizability parameter π_2^H

Olivier Lamarche,^a James A. Platts*^a and Anne Hersey^b

^a Department of Chemistry, Cardiff University, P.O. Box 912, Cardiff, UK CF10 3TB.
E-mail: platts@cf.ac.uk

^b Mechanism and Extrapolation Technologies, GlaxoSmithKline, Park Road, Ware,
UK SG12 0DP

Received 23rd March 2001, Accepted 16th May 2001

First published as an Advance Article on the web 19th June 2001

Ab initio and DFT calculations on the structure and properties of over 98 molecules are reported. Properties calculated for the molecules are assessed for their ability to correlate and predict experimentally derived values of the polarity/polarizability parameter, π . Using multivariate linear regression and partial least squares methods, four properties stand out as predictors of π : the molecular dipole moment, the polarizability, the CHelpG atomic charges and the frontier molecular orbital energies. These properties correlate π to close to the standard deviation in a previously published fragmental approach. The partial least squares method is shown to result in significantly better predictions for an external validation set.

Introduction

It is generally accepted that the main factors affecting the solubility, partition, and passive biological distribution of molecules are size, hydrogen bonding, and polarity. Hydrogen bonding describes the specific attractions between acidic molecule and basic solvent, and *vice versa*, while molecular size is related both to energetically unfavourable formation of cavities in the solvent and favourable dispersion interactions.¹ Polarity and polarizability then account for electrostatic, inductive and dispersive interactions between solvent and solute, particularly those not associated with specific hydrogen bonds. Thus, knowledge of these molecular properties allows understanding and prediction of the solvation properties of molecules.

One particularly successful method of relating such properties to the solvation of molecules is Abraham's linear free energy relation (LFER) approach.² This method uses five properties, or descriptors, in linear combinations to describe solvation. Scales of hydrogen bond acidity and basicity, $\sum \alpha_2^H$ and $\sum \beta_2^H$, were established from complexation equilibrium constants,³ while size is described by McGowan's molecular volume, V_x .⁴ The excess electron density associated with n- and π -electron pairs is modelled by the 'excess molar refraction', R_2 . R_2 , R_2 and V_x are easily calculated by summation of atomic contributions, while previous studies⁵ show how $\sum \alpha_2^H$ and $\sum \beta_2^H$ may be estimated from theoretical calculations.

Finally, a general model of solvation must take account of the polarity and polarizability of a solute. Initial attempts to use the dipole moment were unsatisfactory, and Kamlet and Taft's solvatochromic parameter π_1^* has a number of experimental difficulties, particularly with regard to measurement for new solutes. A new descriptor, the solute dipole/polarizability parameter π_2^H , was therefore developed from a large set of gas-liquid chromatography data,⁶ using an iterative self-consistent method. For clarity we will use the simplified notation π , α and β throughout this paper. π has the dimensions of a free energy, but was scaled in a similar manner to α and β , with n-alkanes defined as having zero π , and hence cannot be expressed in conventional units. Unlike α and β , it is possible for molecules to have negative π values,

since species such as fluorocarbons are less polar/polarizable than simple alkanes. In this work, we seek theoretical models of π similar to those already established for α and β .

π is used as a descriptor in the correlation and prediction of solvation and related properties³ by expressing these as linear combinations of π along with the size, hydrogen bond capacity and refraction descriptors discussed above to yield the solvation eqn. (1)

$$\log SP = c + rR_2 + s\pi + a\alpha + b\beta + vV_x \quad (1)$$

where the dependent variable SP is a property of a series of solutes in a given solute system.

Regression coefficients c , r , s , a , b and v are obtained *via* multiple linear regression against known log SP values and characterise the interaction properties of the solvent system in question. The r -coefficient corresponds to the ability of the phase to interact with π - and n-electron pairs of the solutes, the s -coefficient gives the tendency of the phase to interact with dipolar/polarizable solutes, the a - and b -coefficient are measures of, respectively, the hydrogen bond basicity and hydrogen bond acidity of the phase and the v -coefficient describes the dispersion interactions and cavitation forces. In this manner, it has proved possible to model a wide range of solvation-related properties, including gas-solvent and water-solvent partition coefficients,⁷ many combinations of chromatographic stationary and mobile phases⁸ and biological properties including blood-brain distribution,⁹ skin permeation,¹⁰ anaesthesia,¹¹ membrane irritation,¹² and uptake into plants.¹³

The descriptors R_2 and V_x in eqn. (1) are easily calculated from structure, but traditionally the polarity and hydrogen bonding descriptors had to be determined experimentally. This could be either directly from complexation measurements or indirectly *via* back calculations from partition measurements, which can be difficult and time consuming.⁶ Allied to this, it has been shown⁹ that these descriptors can be accurately estimated from fragment values if the fragments are well chosen. However, this manual fragmentation approach is slow, limiting the use of eqn. (1) for large datasets. Recently, a fast, automated method has been developed for the estimation of these descriptors based on fragmental contributions.¹⁴ This method is capable of reproducing experimental data, including

partition coefficients,¹⁵ blood–brain distribution values,¹⁶ and skin and cell permeabilities.¹⁷

The generality of this method is limited by the lack of experimental data for important fragments. The fragmental contributions to descriptors were taken from an experimental database of descriptors and clearly, if a given fragment is not present in the database then no values can be assigned. There is therefore a pressing need for a general method to predict descriptors for fragments, in order to feed values into the predictive model. Perhaps the most accurate method of achieving this would be to “back-calculate” fragment values from experimental data. However, experimental limitations mean that this cannot be truly general—for example, highly reactive compounds could not be measured easily, and actual samples of the material in question would be required, preventing use in the screening of virtual libraries. We seek methods for the prediction of π using quantum chemistry, with the ultimate aim of predicting values for any molecule purely from structure. The fragmental model¹⁴ resulted in an estimated error of 0.16 units of π : similar errors from our quantum chemical model would be acceptable.

Ševčík and co-workers reported a neural network approach to estimating π .¹⁸ They took a number of topological and quantum mechanical properties as input, combining them nonlinearly *via* a feed-forward neural network (NN). In order to get an acceptable model, they restricted the diversity of compounds to benzene and phenol types. The calculations performed using the NN method resulted in $R^2 = 0.979$ for the training set (62 compounds) and $R^2 = 0.932$ for the test set (after removal of statistical outliers) for a model with 16 descriptors. However, this method is unsuitable for our purposes, since this impressive accuracy is destroyed with a more general dataset, including outliers and a wider chemical diversity of compounds. They were able to train this larger NN to $R^2 = 0.908$ using 7 descriptors, but reported very poor prediction of π from this model, with $R^2 = 0.537$ for the test set. In addition to this lack of generality, the authors did not report sufficient information on their NN to allow us to calculate values of π for new compounds. Our aim is to find a robust, and ideally quick, method for the prediction of *any* organic compound, not simply benzenes and phenols, and to set out exactly how other researchers may apply this method.

Calculation methods

A set of 58 compounds was selected from an experimental database of π values,¹⁹ values being obtained *via* the experimental methods outlined above. Molecules were selected to cover both the numerical spread and the chemical diversity of molecules with known π values. The molecules are tabulated in Table 1, along with a sequential number and their experimental π value. Another set of 40 compounds was selected to test the models. Molecules were selected to cover known or suspected weaknesses of the model. The molecules are tabulated in Table 2 and 7 (see later), along with a sequential number and their experimental and observed π values.

All *ab initio* and DFT calculations were performed using GAUSSIAN98²⁰ running on a Compaq workstation. Initially, the geometries of the molecules were optimised at the HF/3-21G²¹ level and the resulting structures were confirmed as minima through harmonic frequency calculations. Where it was deemed necessary, the conformational space of the molecules was explored at the same level to ensure that the final optimised structure corresponded to the global minimum. A number of properties outlined below were then computed as possible descriptors of π . Starting from these HF/3-21G structures, the geometries of these molecules were then re-optimised at the HF/6-31G(d)²² level and the properties were recomputed at this and B3LYP/6-31++G(d,p)^{23,24} levels.

Table 1 Values of the dipole/polarizability parameter for the training set

Number	Name	π	Ref.
1	Tetrafluoromethane	−0.25	^a
2	Chloropentafluoroethane	−0.12	^a
3	Chlorotrifluoromethane	−0.05	^a
4	Trifluoromethane	0.04	^f
5	1,2-Dichlorotetrafluoroethane	0.05	^a
6	Propene	0.08	^b
7	Ethene	0.10	^b
8	Fluorotrichloromethane	0.18	^c
9	Trimethylamine	0.20	^c
10	Methyl isobutyl ether	0.22	^b
11	Propyne	0.25	^c
12	Dimethyl ether	0.27	^b
13	Dimethylamine	0.30	^b
14	Ethylthiol	0.35	^b
15	Methylamine	0.35	^b
16	Propan-2-ol	0.36	^b
17	Trichloroethene	0.37	^c
18	Dimethyl sulfide	0.38	^b
19	1-Chloropropane	0.40	^b
20	Ethanol	0.42	^b
21	Tetrahydropyran	0.49	^c
22	3-Chloropropyne	0.50	^c
23	Dimethoxymethane	0.52	^c
24	Benzene	0.52	^d
25	Acrylonitrile	0.54	^c
26	Ethyl acetate	0.62	^b
27	Acetic acid	0.65	^b
28	Methyl formate	0.68	^b
29	Acetone	0.70	^b
30	Thiazole	0.74	^c
31	2-Chloroethanol	0.77	^c
32	4-Fluoropyridine	0.77	^c
33	Methyl urethan	0.82	^f
34	3-Chloropyridine	0.83	^c
35	Pyridine	0.84	^c
36	Imidazole	0.85	^c
37	Benzyl alcohol	0.87	^c
38	Phenol	0.89	^c
39	Acetonitrile	0.90	^b
40	Pyrrole	0.91	^c
41	3-Vinylpyridine	0.93	^e
42	Pyrrolidine	0.95	^c
43	Aniline	0.96	^c
44	3-Fluorophenol	0.98	^c
45	2,2,2-Trichloroethanol	1.01	^e
46	1,2,4-Triazole	1.04	^c
47	3-Chloropheno	1.06	^c
48	3-Nitrobenzoic acid	1.08	^c
49	Trimethyl phosphate	1.10	^c
50	4-Chloroaniline	1.13	^c
51	2-Furaldehyde	1.13	^c
52	3-Cyanopyridine	1.26	^c
53	Formamide	1.30	^e
54	Trichloroacetic acid	1.33	^d
55	Acetanilide	1.36	^c
56	N-Methylbenzamide	1.49	^c
57	N-Methylbenzenesulfonamide	1.50	^c
58	3-Cyanophenol	1.55	^c

^a M. H. Abraham and J. R. Gola, unpublished work. ^b M. H. Abraham, G. S. Whiting, R. M. Doherty and W. J. Shuely, *J. Chromatogr.*, 1991, **587**, 213. ^c Various back calculations from partition and chromatographic retention data. ^d M. H. Abraham and G. S. Whiting, *J. Chromatogr.*, 1992, **594**, 229. ^e Estimated by comparison to closely related compounds. ^f calculated using UNIX descriptor calculation program.

Properties calculated were atomic charges, which were collected into the single parameter Q , defined as the sum of the absolute values of the atomic charges (by definition, the sum is zero) over the entire molecule (eqn. (2)).

$$Q = \sum_i |q_i| \quad \text{and} \quad \sum_i q_i = 0 \quad (2)$$

Table 2 Values of the dipole/polarizability parameter for the validation set

Number	Name	π		Ref.
		Obs.	Calc.	
1	1,1,1,2,3,3,3-Heptafluoropropane	0.01	0.21	^a
2	<i>trans</i> -But-2-ene	0.08	0.24	^c
3	Cyclohexane	0.10	0.22	^b
4	Diffuorodichloromethane	0.13	0.06	^a
5	Cyclohexene	0.20	0.48	^b
6	Carbon dioxide	0.28	0.15	^e
7	1,1,1,2-Tetrafluoroethane	0.34	0.19	^a
8	Bromomethane	0.43	0.41	^b
9	Tetrahydrofuran	0.52	0.55	^b
10	Thiophene	0.57	0.48	^c
11	Ethylene oxide	0.59	0.34	^c
12	Styrene	0.65	0.79	^d
13	Acetaldehyde	0.67	0.70	^b
14	1,3-Dichlorobenzene	0.73	0.85	^c
15	Methoxybenzene	0.75	0.84	^c
16	1,4-Dichlorobenzene	0.75	0.64	^c
17	1,2-Dichlorobenzene	0.78	0.99	^c
18	Morpholine	0.79	0.68	^e
19	3,4-Lutidine	0.85	1.09	^c
20	Tribromoethene	0.86	0.63	^f
21	<i>N</i> -Methylaniline	0.90	1.04	^c
22	Urea	1.00	1.24	^c
23	Pyrimidine	1.00	1.11	^c
24	Pyrazole	1.00	0.66	^c
25	1,2,4-Tribromobenzene	1.07	1.06	^g
26	2-Fluoroacetophenone	1.07	1.19	^g
27	3-Chloroacetophenone	1.07	1.22	^g
28	Catechol	1.10	1.10	^c
29	2,3-Dichloroaniline	1.24	1.32	^f
30	Acetamide	1.30	1.09	^e
31	2-Cyanoaniline	1.37	1.52	^c
32	2-Cyanopyridine	1.44	1.58	^c

^a M. H. Abraham and J. R. M. Gola, unpublished work. ^b M. H. Abraham, G. S. Whiting, R. M. Doherty and W. J. Shuely, *J. Chromatogr.*, 1991, **587**, 213. ^c Various back calculations from partition and chromatographic retention data. ^d M. H. Abraham and G. S. Whiting, *J. Chromatogr.*, 1992, **594**, 229. ^e Estimated by comparison to closely related compounds. ^f Calculated using UNIX descriptor calculation program. ^g M. H. Abraham, *J. Chromatogr.*, 1993, **644**, 95.

Atomic charges were computed using three different methods: Mulliken analysis,²⁵ natural population analysis²⁶ (NPA) and electrostatic potential-derived charges using the CHelpG scheme.²⁷ A second atomic charge descriptor, Q/N was constructed by dividing Q by the number of atoms in the molecule N , thereby normalising this polarity descriptor to the size of the molecule. The first two non-zero molecular multipole moments, *i.e.* the dipole μ and quadrupole Θ were calculated. In regression against π , the magnitude of the dipole and the trace of the quadrupole tensor were employed. These properties describe the charge distribution and its deviation from sphericity over a body consisting of more than an atom. Higher multipole moments were not used as they are rarely encountered in chemistry.

Another descriptor used was the molecular polarizability, defined as the change in dipole moment in response to an applied electromagnetic field. Like the quadrupole moment, the trace of the polarizability tensor was used as a descriptor of π . The first (β) and second (γ) hyperpolarizabilities were not used as they only apply for a strong electromagnetic field. Polarizability was calculated using the coupled-perturbed Hartree–Fock (CPHF) method, except in the case of the largest molecules in Table 7, for which CPHF failed due to memory requirements, and finite-field methods were employed. It is worth noting that the methods employed here are unlikely to give accurate polarizabilities, since it is well known²⁸ that this requires larger basis sets with diffuse basis functions. However, our requirement is to reproduce the

trends in polarizability, such that they can be used in regression analysis, rather than their absolute values for any given compound.

Other descriptors used were the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}) and the energy gap between these two molecular orbitals (E_{GAP}). Finally, the molecular volume (V_{M}) as delineated by the $0.001 e a_0^{-3}$ iso-surface in the electron density was calculated.

To begin with, the computed properties were used as independent variables in a standard least-squares regression. All simple and multivariate linear regression analysis employed the JMP discovery software.²⁹ Subsequently, we used a second multivariate data analytical tool, namely partial least squares projections to latent structures, PLS. All PLS regression employed the SIMCA-P 8.0 software.³⁰

Results and discussion

(i) The multivariate linear regression analysis (MLRA) approach

A. Hartree–Fock models. The different properties computed at the HF/3-21G level were correlated against π ; the results are displayed in Table 3. Some properties, such as the dipole moment and polarizability, show some correlation. However, none of these simple correlations show sufficient correlation to be used independently. Changes in π will be better explained by using more than one independent variable: a multivariate linear regression analysis.

Using a stepwise regression control, the best correlation was obtained by using a combination of CHelpG charges (Q), dipole moment (μ) and the GAP energy (E_{GAP}). They correlate to π according to:

$$\pi = 0.871 + 0.133Q + 0.115\mu - 1.399E_{\text{GAP}}$$

$$n = 58, R^2 = 0.764, R_{\text{CV}}^2 = 0.735,$$

$$\text{rms} = 0.219, F = 58.2 \quad (3)$$

where n is the number of data in the regression, R^2 is the overall correlation coefficient, R_{CV}^2 is the cross-validated or “leave-one-out” correlation coefficient, rms is the root mean square error and F is Fischer’s F -statistic.

It is somewhat encouraging that such a small basis set can give a reasonable correlation. However, a bigger basis set should increase the flexibility and obtain more accurate calculations of properties. Therefore, the above stepwise procedure was repeated using properties calculated at the HF/6-31G(d) level. The best model found is a combination of CHelpG charges (Q/N), dipole moment (μ), polarizability (α) and the HOMO energy (E_{HOMO}):

$$\pi = 0.071 + 1.419Q/N + 0.115\mu + 0.0036\alpha + 1.052E_{\text{HOMO}}$$

$$n = 58, R^2 = 0.812, R_{\text{CV}}^2 = 0.772,$$

$$\text{rms} = 0.197, F = 57.1 \quad (4)$$

The improvement of this model over eqn. (3) is impressive, indicating that improved calculations do indeed lead to better

Table 3 Correlations with individual HF/3-21G properties

Property	R^2	rms error	Units
Q_{Mulliken}	0.347	0.357	au
Q_{CHelpG}	0.429	0.334	au
Q_{NPA}	0.239	0.386	au
μ	0.479	0.319	D
Θ	0.256	0.381	D Å
α	0.460	0.325	au
E_{HOMO}	0.265	0.379	au
E_{LUMO}	0.220	0.391	au
E_{GAP}	0.379	0.348	au
V_{M}	0.335	0.361	au

correlations with π . We therefore carried out the same process with a still larger basis set, and also accounted for the effects of electron correlation by use of density functional methods.

B. Density functional models. In order to get a good compromise between time and accuracy, we computed the properties using density functional theory; at the B3LYP/6-31++G(d,p) level. The choice of angular functions is straightforward³¹ for the calculation of polarizability and dipole moment (the more flexible, the better for the dipole moment). It is established that diffuse functions are essential for the calculation of polarizability.³² It will also be interesting to see the impact of electron correlation on the model. Furthermore, DFT was already used with success in the calculation of descriptors for α and β .⁵ The HF/6-31G(d) geometries were used to calculate the DFT properties for two reasons: the optimisation step is the time-consuming step and, according to our calculations on a small set,³³ there are no major differences between properties if we use B3LYP/6-31++G(d,p) geometries or HF/6-31G(d) geometries.

Correlations of individual molecular properties with π showed no great improvement over those found with Hartree–Fock properties, as in Table 3. Both the overall correlation coefficients of the dipole moment and the polarizability were slightly better; $R_\mu^2 = 0.495$ (*vs.* 0.479 from Table 3) and $R_\alpha^2 = 0.479$ (*vs.* 0.467 from Table 3). Using a stepwise regression control, the best least squares regression was obtained by using a combination of CHelpG charges (Q/N and Q), dipole moment (μ), polarizability (α) and the HOMO energy (E_{HOMO}). They correlate to π according to:

$$\begin{aligned} \pi = & -0.019 + 2.365Q/N - 0.101Q + 0.122\mu \\ & + 0.0043\alpha + 1.859E_{\text{HOMO}} \\ n = & 58, R^2 = 0.831, R_{\text{CV}}^2 = 0.781, \\ \text{rms} = & 0.188, F = 51.1 \end{aligned} \quad (5)$$

This method to compute the properties is a good compromise between time and good prediction of experimentally derived values of π . This model correlates and predicts π close to the fragmental approach root mean square error (rms = 0.19 *vs.* 0.16). The use of bigger basis sets or higher levels of theory may improve this model somewhat, but would involve significantly greater time and computational resources, and was not explored further.

Our efforts to improve the model now focused on the molecules with high residual π values ($|\Delta\pi| > 0.25$). Nine molecules have such errors from eqn. (5): these are listed in Table 4. Firstly, the structures of all the set were confirmed as minima through harmonic frequency calculations, which revealed that 4-chloroaniline had a negative frequency, despite being a minimum at the HF/3-21G level. Its structure was then reoptimised in lower symmetry at HF/6-31G(d) level and its properties were recomputed at B3LYP/6-31++G(d,p) level. Secondly, the conformational space of the molecules was

Table 4 Molecules with high $|\Delta\pi|$

Name	$ \Delta\pi $
Trimethyl phosphate	0.47
Pyrrolidine	0.42
4-chloroaniline	0.38
Methyl isobutyl ether	0.32
Acrylonitrile	0.30
Trichloroacetic acid	0.29
2,2,2-Trichloroethanol	0.29
Acetonitrile	0.29
Dimethoxymethane	0.27

explored to ensure the final optimised structure corresponded to the global minimum. Four molecules of Table 4 (trimethyl-phosphate, pyrrolidine, methyl isobutyl ether and dimethoxymethane) have more than one low energy minimum.

The different conformers of trimethyl phosphate, pyrrolidine and dimethoxymethane have been studied extensively.³⁴ The reported low energy conformers are; C_2 and C_1 for dimethoxymethane; C_3 , C_1 and C_5 for trimethyl phosphate and C_5 (axial) and C_5 (equatorial) for pyrrolidine. According to these publications, the properties of each conformer contribute, to some extent, to the total properties of the molecule. In order to better represent these molecules in the model, we optimised the structures of conformations having energy within 2.5 kcal mol⁻¹ of the global minimum. The properties of these molecules were calculated by taking the average value of all the different conformations. This is not truly rigorous as the conformer populations depend on parameters such as the degeneracy of the conformation, the shape of the conformational energy well and the energy difference between conformations. However, we can justify this simplified hypothesis, as the purpose of this method is not to find the most accurate properties of the molecules but the ones that fit the model best. Furthermore, the conformer energy profiles of molecules can differ greatly with the computational level used.^{34c}

For consistency, this method has been applied to every molecule of our set that has alternative conformers (but with $\Delta E < 2.5$ kcal mol⁻¹ and with sensibly different values of properties between conformers). In our set this means that 2-chloroethanol, 1,2-dichlorotetrafluoroethane, 3-chlorophenol, 3-fluorophenol and 3-cyanophenol required such treatment. Some molecules, such as ethylthiol, 2,2,2-chloroethanol or ethanol, have alternative conformers but their properties do not greatly differ from one to another (less than 2% variation). In this case, only the properties of the global minimum have been taken into account.

The prediction over the last four molecules in Table 4 (acrylonitrile, acetonitrile, 2,2,2-trichloroethanol and trichloroacetic acid) could not be improved with this method. Through the attempt to improve the π model, our goal was not to compute accurate properties but properties with a consistent error with respect to the true values. According to the literature, properties of molecules containing CN-bonds do not follow the same pattern as the others and have therefore an error value different from the set (*e.g.* SCF polarizabilities are below experimental values, whereas the CN-containing molecules are slightly overestimated²⁸). In the case of trichloromethyl compounds, the values of the CHelpG atomic charges are different from what one could expect ($Q_{\text{Cl}} \approx 0$ and $Q_{\text{C}} \approx -0.2$). It is expected that calculation of properties could generate errors from certain molecules due to their chemical diversity. Even the introduction of electron correlation by DFT did not change their behaviour. It seems, therefore, that care must be taken in using models such as eqn. (5) to predict π values for molecules containing CN-bonds and/or trichloromethyl groups.

Using properties corrected for multiple low energy conformations, the best correlation was similar to, but much superior to eqn. (5).

$$\begin{aligned} \pi = & 0.077 + 1.333Q/N + 0.150\mu + 0.003\alpha + 1.697E_{\text{HOMO}} \\ n = & 58, R^2 = 0.857, R_{\text{CV}}^2 = 0.827, \\ \text{rms} = & 0.171, F = 79.4 \end{aligned} \quad (6)$$

In contrast to eqn. (5), the sum of the CHelpG atomic charges, Q , was not included in this model, since it displays large pairwise correlation with Q/N . Despite this, all the statistical measures show improvement over the previous model, with a particularly noticeable increase in the cross-validated R_{CV}^2 .

(ii) Partial least squares (PLS) methods

Our best MLR model (eqn. (6)) correlates and predicts experimentally derived values, which are very close to fragmental approach rms error (0.17 vs. 0.16). Unfortunately, some molecules still have high residual values. Multivariate linear regression analysis is a method which is only suitable when the variables are devoid of noise.³⁵ We therefore apply PLS regression to the same data set, with the aim of improving on these shortcomings. PLS derives its usefulness from its ability to analyse data with many, noisy, collinear, and even incomplete variables. The basic conceptual model is such that the variable correlations are modelled as arising from a small set of 'latent variables', where all the measured variables are modelled as linear combinations of these latent variables.³⁶

Three PLS models of π , differing in the properties used and the number of PLS components selected, are reported in Table 5, e.g. model 1 contracts six descriptors to two latent variables. R^2 is a quantitative measure of the fit between observed and calculated data, while Q^2 measures the predictive ability of the model. In a similar manner to R_{CV}^2 used in the MLR models, this is achieved *via* cross-validation techniques. Unlike R_{CV}^2 , this is done by omitting groups of around 10% of the data and predicting values for this 10%. For this reason, it is more realistic than the so-called "leave-one-out" approach.

The first model has the highest R^2 and Q^2 parameters of all we have found using PLS methods, where six calculated properties are reduced to two components. However, this model contains two very highly correlated properties, namely E_{GAP} and E_{HOMO} —removing one of these has little effect on the model, as does omitting the molecular quadrupole. The second model in Table 5 removes these properties, and reduces the remaining four properties to one PLS component with no real loss of accuracy. The third model is directly comparable to the best MLRA mode, eqn. (6) (in fact, if model 3 used all four components the PLS model is identical to the MLRA model).

Usually, the R^2 and Q^2 parameters vary differently with increasing model complexity. R^2 is inflationary and rapidly approaches unity as model complexity increases (number of free model parameters; terms, latent variables, etc). However, Q^2 will typically reach a plateau and can even head downwards (indicating that prediction ability is degrading). Thus, PLS analysis finds a trade-off between fit and prediction ability although MLRA analysis will give the best R^2 parameter but not obviously the best Q^2 . Our preferred PLS model on these grounds is model 2 of Table 5, shown below as eqn. (7).

$$\pi = 0.488 + 0.962Q/N + 0.145\mu + 0.0023\alpha - 2.685E_{GAP}$$
$$n = 58, R^2 = 0.840, Q^2 = 0.822, \text{rms} = 0.176. \quad (7)$$

This model has been chosen as it gives the best R^2/Q^2 fraction, the best balance between predictive power and reasonable fit. Moreover, the descriptors used are not cross-correlated. Fig. 1 is a plot of the observed *vs.* calculated π values using eqn. (7). This model was constructed using raw (*i.e.* unscaled and uncentred) calculated properties, such that the coefficients of eqn. (7) do not reflect each property's importance in the model. To see the weight of each descriptor, the

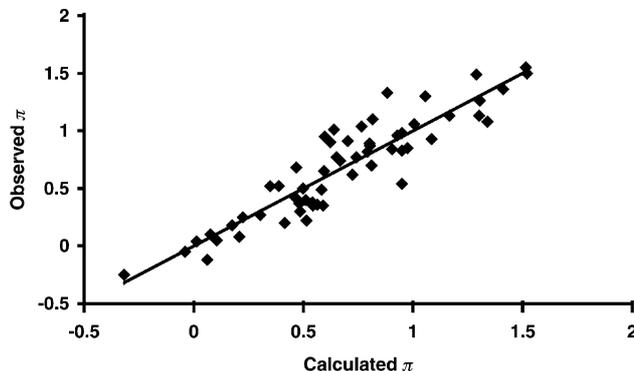


Fig. 1 Observed *vs.* calculated π values from eqn. (7) for the training set of 58 molecules.

model was recomputed with scaled and centred variables, and the relative values of the regression coefficients displayed in Fig. 2. As the solvatochromic polarity/polarizability parameter takes care of electrostatic and induction/dispersion interactions, it is not surprising to see that the dipole moment and the normalised overall sum of absolute CHelpG atomic charges act as descriptors for the electrostatic term and the trace of the polarizability tensor and the energy gap between higher and lower occupied molecular orbitals for the induced part.

(iii) Validation of the proposed model: Prediction of π values of an external set

In order to validate this goodness of prediction parameter, it is necessary to test the model with an external set of molecules. A test set of 32 compounds (see Table 2) was selected from the experimental database of π values, with molecules selected to cover both the numerical spread and the chemical diversity of molecules of the proposed model. Some molecules were selected to test possible weaknesses of the model (*e.g.* three dichlorobenzenes: three different isomers with almost identical π values; morpholine and tetrahydrofuran: analogues of pyrrolidine; heptafluoropropane: two trihalomethyl groups).

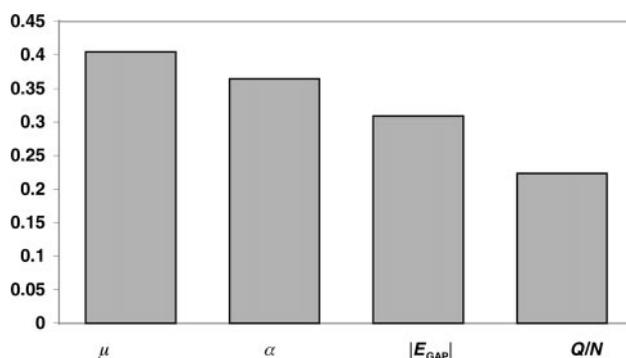


Fig. 2 Weight of each descriptor of eqn. (7) using scaled and centred variables.

Table 5 Summary of PLS models

	E_{HOMO}	E_{GAP}	α	μ	Q/N	Θ	R^2	Q^2	rms	Components
Model 1	Yes	Yes	Yes	Yes	Yes	Yes	0.853	0.829	0.168	2
Model 2	No	Yes	Yes	Yes	Yes	No	0.840	0.822	0.176	1
Model 3	Yes	No	Yes	Yes	Yes	No	0.851	0.819	0.171	1

Table 6 Prediction statistics for the 32 validation molecules

	Analysis	Components	R^2	sd	ae	aae	Max – Min
Model 1	MLRA	≡6	0.787	0.178	–0.032	0.143	0.767
	PLS	2	0.796	0.175	–0.031	0.139	0.773
Model 2	MLRA	≡4	0.787	0.179	–0.031	0.149	0.748
	PLS	1	0.825	0.162	–0.031	0.141	0.619
Model 3	MLRA	≡4	0.776	0.183	–0.032	0.153	0.723
	PLS	1	0.805	0.171	–0.026	0.145	0.645

The predictive ability of the different models was tested using (i) R^2 of correlation between predicted and observed π values; (ii) the standard deviation of the residual π values (sd); (iii) the average error (ae); (iv) the average absolute error (aae); and (v) the difference between the maximum and minimum residual π value (Max – Min). Statistics for each model are displayed in Table 6.

It is encouraging to note that the PLS models predict better π values than their MLRA analogues. This must be principally due to the ability of PLS to analyse data with noisy variables. The best PLS predictive equation is model 2 (eqn. (7)) as the standard deviation of the residual π values and its R^2 value are the lowest (standard deviation less than the rms of eqn. (8)), its Max – Min value shows no eccentric residual values and the average of the absolute residual π value is less than the rms of the fragmental approach. Fig. 3 is a plot of the observed π values *vs.* calculated π values (eqn. (7)) from Table 2.

From the previous results and those displayed in Table 2, we can conclude that the main strength of the proposed model is its ability to predict π values for a wide diversity of compounds and a large range of π values. Even new chemical structures not present in the training set (*e.g.* ethylene oxide, bromomethane) or molecules with different properties, but similar π values, are predicted with similar accuracy. However, it is difficult to classify and obtain a clear picture of the strengths and weaknesses of the model as the properties calculated (dipole moment, polarizability and CHelpG atomic charges) can vary greatly from conformers, isomers or molecules of the same class. For example, the previous discussion on CN-bonds seems not to apply for aromatic compounds (*e.g.* 3-cyanopyridine or 2-cyanoaniline). From some results, *e.g.* methyl isobutyl ether and pyrrolidine, we could conclude that the introduction of saturated alkyl groups degrades the π prediction as these have significant polarizability, but should not contribute to the π value. However, this conclusion appears not to hold for tetrahydrofuran or even morpholine, a close analogue of pyrrolidine. It is even difficult to draw some conclusions from structurally similar molecules; imidazole is predicted well (residual = –0.12) unlike pyrazole (0.34).

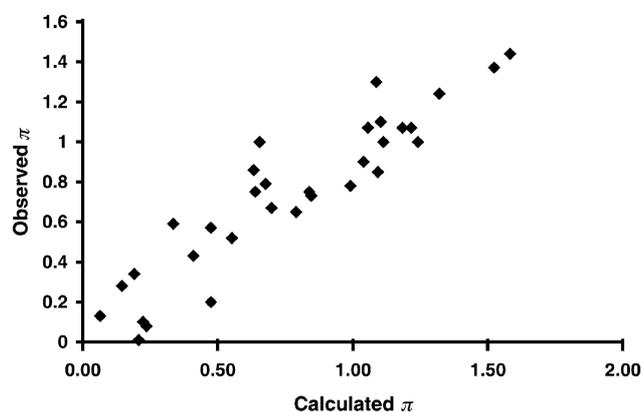


Fig. 3 Observed *vs.* calculated π values from eqn. (7) for the validation set of 32 molecules.

(iv) Limitation of the proposed model: Prediction of π values of an out of range set

The 32 molecules in Table 2 were chosen to be broadly similar to those used to construct the model of π , thereby confirming that models such as eqn. (7) are not simply a result of over-fitting. A much more stringent test of the model would be to predict π values for molecules well outside the training sets, including larger compounds and new functional groups. We therefore selected a second test set of 8 molecules, which are tabulated in Table 7, along with a sequential number, their experimental π value, their calculated π value (eqn. (7)) and their π value from the fragmental approach. This set gives us an insight into the limit of the model.

From Table 7, we can see that the model can well predict π values for molecules of similar size to the one in the training and test set ($N \leq 25$ for 1, 4 and 6), even if their π value is outside the model range or the chemical diversity is pushed further. Interestingly, the proposed model shows its limit for larger molecules ($N \geq 40$ for 2, 3, 7 and 8). These molecules have, in general, many functional groups and, whatever is the observed π value, a cumulative effect of the individual error of each group is likely to occur (even observed on simple long alkane chains). For example, morphine has certain groups such as an alcohol and a tertiary amine, which are slightly overestimated individually. It is therefore not surprising to predict an overestimated value for this drug. Allied to this, it has been found impossible to compute their properties in a reasonable time and the level of theory had to be switched to the less constrained B3LYP/6-311G(d,p) level.

The most interesting conclusion comes from the study of the last drug: cimetidine. The predicted π value of this molecule was indeed very bad and only one of its functional groups, the cyano-guanidine group, was responsible for this value: *N,N'*-dimethylcyanoguanidine, was also predicted very badly (1.89 *vs.* 0.6, $N = 16$). We had already noticed that molecules with important charge separation give poor predictions (*e.g.* 3-nitrobenzoic acid and urea, residual π values of –0.24).

Table 7 Predicted π values for larger molecules

Number	Name	N	π			Ref.
			Obs.	Calc.	Fragment.	
1	Ferrocene	21	0.85	1.01	0.53	^a
2	Morphine ^e	40	1.25	1.95	1.68	^b
3	Mepyramine ^e	44	1.25	1.94	1.89	^c
4	Caffeine	24	1.63	1.64	1.81	^c
5	Cimetidine	33	1.73	2.97	2.11	^d
6	Clonidine	23	1.83	1.78	1.92	^d
7	Estrone ^e	42	2.05	1.96	2.59	^c
8	Testosterone ^e	49	2.59	2.00	2.32	^d

^a M. H. Abraham and J. R. M. Gola, unpublished work. ^b M. H. Abraham, K. Takacs-Novak, and R. C. Mitchell, calculated from partition of the neutral form. ^c Various back calculations from partition and chromatographic retention data. ^d Calculated by H. Chadha. ^e Due to the size of this molecule (40 atoms), the properties have been computed at the less constrained B3LYP/6-311G(d,p) level instead of the usual B3LYP/6-31++G(d,p) level; the polarizability was computed numerically.

Molecules such as cyano-guanidine are an extreme case as there is a high degree of charge separation, perhaps even a zwitterionic form,³⁷ giving a very high value for the dipole moment (7.8 D) and CHelpG atomic charges, not easily compatible with our model and a low π value.

Conclusions

We have shown that density functional calculations using moderately sized basis sets can yield accurate models of Abraham's polarity/polarizability descriptor π . Four properties (CHelpG atomic charges, dipole moment, polarizability and frontier molecular orbital energies) computed at the B3LYP/6-31++G(d,p) level using HF/6-31G(d) optimised geometries are found to be significant in several models. The best model developed here used PLS regression methods, and allowed us to predict experimental values of π with almost the same errors as a previously published fragmental model, *i.e.* rms errors of 0.17 over a range of 1.7 units. Particular attention must be paid to molecules with energetically close conformers ($\Delta E < 2.5 \text{ kcal mol}^{-1}$), as the properties can differ greatly from one to another.

The models developed were tested against two separate validation sets of molecules: one larger set of molecules similar to the training set confirmed the predictive ability of the model for smaller molecules. Thus, we have satisfied the main aim of the current study, namely to develop methods for predicting π for as yet uncharacterised fragments. The model was also tested against several larger molecules, including drug-like compounds, with encouraging results.

Accurate π values for a wide diversity of medium-sized molecules ($N \leq 25$) can be predicted without, apparently, π value restrictions. The π values from certain molecules such as molecules containing CN-bonds and/or trichloromethyl groups should be used with caution, as these molecules could not fit the model well due to possible inconsistency in the calculations. Attention must also be drawn to molecules with important charge separation as high dipole moment and CHelpG atomic charges are incompatible with low π values (according to the model). For large-sized molecules, the fragmental approach seems to be a better calculation alternative as the computational resources required to calculate π in this manner are faster and the predicted π values more reliable due to non-cumulative effects.

Acknowledgements

OL thanks GlaxoSmithKline for a Ph.D. studentship. The authors are also grateful to Mr. Chris Luscombe for help in the use of PLS models.

References

- (a) M. J. Kamlet, R. M. Doherty, J.-L. M. Abboud and M. H. Abraham, *CHEMTECH*, 1988, **16**, 566; (b) M. H. Abraham and J. Liszi, *J. Chem. Soc., Faraday Trans. 1*, 1978, **74**, 1604.
- M. J. Kamlet, R. M. Doherty, M. H. Abraham and R. W. Taft, *J. Am. Chem. Soc.*, 1984, **106**, 464.
- (a) M. H. Abraham, P. L. Grellier, D. V. Prior, P. P. Duce, J. J. Morris and P. J. Taylor, *J. Chem. Soc., Perkin Trans. 2*, 1989, 699; (b) M. H. Abraham, P. L. Grellier, D. V. Prior, J. J. Morris and P. J. Taylor, *J. Chem. Soc., Perkin Trans. 2*, 1990, 521.
- M. H. Abraham and J. C. McGowan, *Chromatographia*, 1987, **23**, 243.
- (a) J. A. Platts, *Phys. Chem. Chem. Phys.*, 2000, **2**, 3115; (b) J. A. Platts, *Phys. Chem. Chem. Phys.*, 2000, **2**, 973.
- M. H. Abraham, G. S. Whiting, R. M. Doherty and W. J. Shuely, *J. Chromatogr.*, 1991, **587**, 213.
- (a) M. H. Abraham, J. A. Platts, A. Hersey, A. J. Leo and R. W. Taft, *J. Pharm. Sci.*, 1999, **88**, 670; (b) M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. J. Leo and R. W. Taft, *J. Chem. Soc., Perkin Trans. 2*, 1994, 1777.
- (a) M. Reta, P. W. Carr, P. C. Sadek and S. C. Rutan, *Anal. Chem.*, 1999, **71**, 3484; (b) M. H. Abraham, C. F. Poole and S. K. Poole, *J. Chromatogr. A*, 1999, **842**, 79; (c) F. Z. Oumada, M. Roses, E. Bosch and M. H. Abraham, *Anal. Chim. Acta*, 1999, **382**, 301.
- (a) J. A. Gratton, M. H. Abraham, M. W. Bradbury and H. S. Chadha, *J. Pharm. Pharmacol.*, 1997, **49**, 1211; (b) M. H. Abraham, H. S. Chadha and R. C. Mitchell, *J. Pharm. Sci.*, 1994, **83**, 1257.
- M. H. Abraham, F. Martins and R. C. Mitchell, *J. Pharm. Pharmacol.*, 1997, **49**, 858.
- M. H. Abraham and C. Rafols, *J. Chem. Soc., Perkin Trans. 2*, 1995, 1843.
- M. H. Abraham, R. Kumarsingh, J. E. Cometto-Muniz and W. S. Cain, *Toxicol. in Vitro*, 1998, **12**, 403.
- J. A. Platts and M. H. Abraham, *Environ. Sci. Technol.*, 2000, **34**, 318.
- J. A. Platts, M. H. Abraham and A. Hersey, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 835.
- J. A. Platts, M. H. Abraham, A. Hersey and D. Butina, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 71.
- J. A. Platts, M. H. Abraham, Y. H. Zhao, A. Hersey, L. Ijaz and D. Butina, *J. Pharm. Pharmacol.*, submitted.
- J. A. Platts, M. H. Abraham, A. Hersey and D. Butina, *Pharm. Res.*, 2000, **17**, 1013.
- D. Svozil, V. Kvasnička and J. G. Ševčík, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 338.
- M. H. Abraham, personal communication.
- M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle and J. A. Pople, *GAUSSIAN98 Rev. A.6*, Gaussian Inc., Pittsburgh, PA, 1998.
- J. S. Binkley, J. A. Pople and W. J. Hehre, *J. Am. Chem. Soc.*, 1980, **102**, 939.
- (a) R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724; (b) M. S. Gordon and Chem., *Phys. Lett.*, 1980, **76**, 163.
- (a) A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648; (b) C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785.
- T. Clark, J. Chandrasekhar and G. W. Spitznagel, *J. Comput. Chem.*, 1983, **4**, 294.
- R. S. Mulliken, *J. Chem. Phys.*, 1959, **23**, 1833.
- For a comprehensive review see A. E. Reed, L. A. Curtiss and F. Weinhold, *Chem. Rev.*, 1988, **88**, 899.
- C. M. Breneman and K. B. Wiberg, *J. Comput. Chem.*, 1990, **11**, 361.
- (a) M. A. Spackman, *J. Phys. Chem.*, 1989, **93**, 7594; (b) G. Deniau, G. Lécayon, P. Viel, G. Hennico and J. Delhalle, *Langmuir*, 1992, **8**, 267.
- J. Sall *et al.*, JMP Rev. 4, SAS Institute Inc., Cary, NC, 2000.
- SIMCA-P Version 8.0*, Umetrics AB, Umeå, Sweden, 2000.
- P. A. Christiansen and E. A. McCullough, *Chem. Phys. Lett.*, 1978, **55**, 439.
- P. A. Christiansen and E. A. McCullough, *Chem. Phys. Lett.*, 1977, **51**, 468.
- O. Lamarche and J. Platts, unpublished work.
- (a) G. D. Smith, R. L. Jaffe and D. Y. Yoon, *J. Phys. Chem.*, 1994, **98**, 9072; (b) L. George, K. S. Viswanathan and S. Singh, *J. Phys. Chem. A*, 1997, **101**, 2459; (c) L. Carballera and I. Pérez-Juste, *J. Chem. Soc., Perkin Trans. 2*, 1998, 1339.
- L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, *Introduction to Multi- and Megavariate Data Analysis using Projection Methods*, Umetrics AB, Umeå, Sweden, 1999.
- S. Wold, C. Albano, W. J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström, in *Chemometrics: Mathematics and Statistics in Chemistry*, ed. B. R. Kowalski, Reidel, Dordrecht, 1984.
- R. C. Young, C. R. Ganellin and M. J. Graham, *Tetrahedron*, 1982, **38**, 1493.