Prediction of vapour pressure using descriptors derived from molecular dynamics

Prem KC Paul

Unilever R & D, Port Sunlight, Quarry Road East, Wirral, UK CH63 3JW.
E-mail: prem.paul@unilever.com; Fax: +44 0151 6411812; Tel: +44 0151 6413905

OBC
www.rsc.org/obc

Vapour pressures of organic materials can be predicted to high levels of accuracy using cohesive energies and solubility parameters derived from molecular dynamics simulations that use good forcefields. It is found that over 90% of the correlation with vapour pressure is accounted for by a single cross term involving the product of either the molecular weight or molar volume of a molecule and its cohesive energy density.

Introduction

Concerns regarding the environmental impact of chemicals like volatile organics have meant that prediction of important physico-chemical properties like vapour pressure continues to receive wide attention. While some earlier methods depend on one or more experimental parameters like boiling point (BP) as descriptors, the more recent ones predict solely from molecular structure.[1] The field has advanced from predicting vapour pressure from purely thermodynamic considerations to using structure–property relationships and concomitantly from simple equations to neural network models. Some recent developments[2] have been to predict vapour pressures from quantum mechanical descriptors and to predict the temperature dependence of vapour pressure starting from molecular structure. Amongst molecular simulation methods,[3] Monte Carlo techniques have been used in the study of liquid–vapour equilibria and prediction of physico-chemical properties.[4]

Here a QSAR/QSPR approach to predicting vapour pressures has been pursued, wherein some of the descriptors in the models are extracted from molecular dynamics (MD) simulations. The descriptors chosen from MD are such that they are chemically intuitive to have a bearing on vapour pressure. Additionally, model building has been restricted to finding simple equations with fewer terms such that the understanding from the model to physical–chemical principles is easy and immediate.

In physical chemistry, it is well known that within a Clausius–Clapeyron framework at a given temperature, the vapour pressure[5] of a substance (assuming ideality) is related to its enthalpy of vapourisation (ΔH_{vap}). This enthalpy describes the cohesive energy density (E_{coh}) of the material, which in turn is the square of the solubility parameter (SP) of a material (see eqns. (1–3)).

$$\ln(vp) = B - \Delta H_{vap}/RT \quad (1)$$

$$E_{coh} = (\Delta H_{vap} - RT)/V_m \quad (2)$$

$$SP = SQRT(E_{coh}) \quad (3)$$

In this paper we have tried to use the solubility parameter of a material as a descriptor in the prediction of vapour pressure.

The literature regarding solubility parameters is considerable. Foremost is an extensive review of this area provided by Barton's Handbook.[6] Solubility parameters were first defined to explain the relative miscibility of solvents by Joel H. Hildebrand[7] who proposed that the solvency behaviour of non-electrolytes is best

described by the square root of the internal pressure or the cohesive energy density of the pure substances. Over the years SPs began to be described in terms of various contributions; one example is for the cohesive energy to be described by polar, dispersive and hydrogen bonding effects, as given in Hansen's parameters.[8]

Solubility parameters of small molecules and polymers can be approximately calculated from the molecular structure of the molecule and the monomer, respectively. These methods[8,9] use contributions from various groups in the molecule to arrive at the value. However they are approximate and are not always accurate for novel combinations of different functional groups.

In this paper, solubility parameters are calculated using molecular dynamics simulations and modern forcefields. The advantages of this method are that because the description of the system is fully atomistic, specific interactions, such as hydrogen bonding, molecule and chain structure effects such as chirality and tacticity respectively are easily included. Furthermore, since SPs are calculated for the temperature at which the simulation is performed, the temperature dependence of SPs can be studied. It has recently been found that simulations on bulk systems of moderate size using systematically-derived class II forcefields such as COMPASS[10] are capable of making predictions of solubility parameters, with an accuracy that compares favourably with experimentally derived values.

Methods

Solubility parameters of the molecules under consideration were calculated using molecular dynamics simulations and COMPASS forcefields using the software DISCOVER by Accelrys Inc.[11] All structure–activity and structure–property relationships were obtained using Cerius2 software.[11]

Cohesive energy is defined as the increase in energy per mole of a material if all intermolecular forces are eliminated. The cohesive energy density (CED) corresponds to the cohesive energy per unit volume. CED is simply the difference in energy between a fully periodic, bulk (b) system and one where each of the components is in isolation *i.e.* in the gas (g) phase ($E_{coh} = E_b - E_g$) which is easily calculated from molecular dynamics simulations. The solubility parameter is given by the square root of the cohesive energy density (eqn. (3)). The calculation details given next are very similar to the ones used to predict bulk density.[12]

Initially an amorphous cell containing many copies of the molecule (approximately 1500 atoms in the box) is built to a target density. After some equilibration dynamics, a large (100 picosecond) *NPT* simulation ($P = 0.0$) is run where the number of particles (N), the pressure (P) and temperature (T) are kept constant, using Berendsen's pressure control algorithm[13] and applying Andersen's stochastic collision temperature control method.[14] This enables the simulated system to 'find' its own density, and from the average volume a further large *NVT* ($V =$ volume) simulation is performed. After sufficient equilibration steps, trajectories are collected at appropriate time intervals.

From these trajectories, the cohesive energy density and solubility parameter are calculated.

Other descriptors[15] that have been used in building the QSAR/QSPR are AlogP (calculated octanol–water partition coefficient), molecular weight (MW), molecular refractivity (MolRef), atomic polarisabilities (Apol), number of rotational bonds (Rotlbonds), number of h-bond donors or HBD (*e.g.* R–NH, R–OH *etc.*), number of h-bond acceptors or HBA (*e.g.* C=O, R–O–R, R–N–R *etc.*), free energies of solvation in water (*F*h2o) and octanol (*F*oct).

Results and discussion

A starting list of 22 molecules was used in the training set for the prediction of vapour pressure. The vapour pressure data at 25 °C comes mainly from the CRC Handbook.[16] Table 1 gives a list of these molecules and some of the descriptors used in the model building.

As can be seen the table covers a broad spread of vapour pressures (∼5 orders of magnitude) and represents a wide spectrum of molecules having ring structures and straight chains that include aldehydes, alcohols *etc.*

Before model building using a genetic function algorithm[11] (GFA) was performed, simple linear regression models relating vapour pressure to one or more of the descriptors were looked at. This is shown in Table 2.

Table 2 shows that while molecular weight, MolRef and the solubility parameter have some bearing on the vapour

Table 1 Molecules used in the training set. Vapour pressure (vp) in µm

Molecule	log(vp)	MW	AlogP	SP
α-Ionone	1.20	192.30	3.68	17.20
Anethole	1.76	148.20	2.79	19.86
Benzyl salicylate	−0.82	228.25	3.27	21.19
Benzophenone	0.00	182.22	3.27	21.32
Cinnamic alcohol	0.62	134.18	2.04	23.97
Coumarin	0.48	146.15	1.94	22.64
Hexyl cinnamic aldehyde	−0.15	216.32	4.00	18.94
Indole	1.07	117.15	1.77	25.11
Limonene	3.15	136.24	2.94	17.44
Benzyl acetate	2.08	136.15	1.55	20.99
Menthol	1.73	156.27	2.78	18.97
Benzaldehyde	3.04	106.12	1.72	22.12
α-Pinine	3.64	136.24	2.80	15.83
C7-aldehyde	3.53	114.19	1.63	18.06
C9-aldehyde	2.41	142.24	2.43	17.83
Anisole	3.52	108.14	1.79	19.77
Citronellal	2.36	154.25	2.25	17.90
Citronellol	1.18	156.27	2.75	20.02
Eucalyptol	3.22	154.25	1.83	17.33
Methyl amyl ketone	3.59	114.19	2.20	17.94
Methyl hexyl ketone	2.88	128.21	2.59	17.77
Vanillin	−0.77	152.15	1.19	24.72

[a] MW = molecular weight, AlogP from Cerius2 software[11], SP from molecular dynamics simulations in MPa$^{0.5}$ or ((J cm^{-3})$^{0.5}$)

Table 2 Simple equations of log(vp) with parameters and cross terms

x = ?	y = mx + C	R^2
MW	y = −0.03x + 6.50	0.50
AlogP	y = −0.13x + 2.19	0.08
SP	y = −0.35x + 8.78	0.39
MolRef	y = −0.83x + 5.60	0.35
MW*SP	y = −0.0019x + 7.39	0.90
SP*SP	y = −0.0084x + 5.17	0.39
MW*SP*SP	y = −0.00007x + 6.13	0.93

[a] y = log(vp)

pressure, cross terms have a much larger impact. Indeed the term MW*SP*SP shows very good correlation (R^2 ∼0.93).

This finding was made use of in building more complex and complete models using a GFA. Here, apart from descriptors given in Table 1, cross terms like MW*SP*SP and other parameters such as MolRef, *F*h2o and *F*oct were used in the model building. Many models were obtained. Given below is a simple model with an R^2 of 0.96 for the 22 molecules in Table 1.

$$\log(\text{vp}) = a1 - a2*MW*SP*SP - a3*MolRef \qquad (4)$$

(where a1 = 6.9, a2 = 0.000067 and a3 = 0.025).

This model was tested against 40 other molecules[17] (open triangles in the graph of Fig. 1) which include many simple organic molecules. The results are shown in Fig. 1.
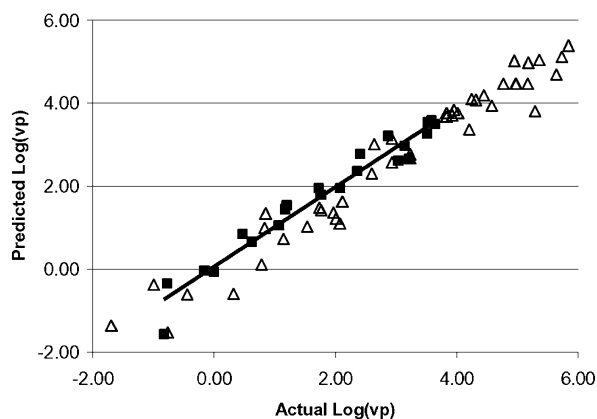

Fig. 1 Predicted *vs.* actual log(vp) of training set (squares) and test set (open triangles). R^2 ∼0.95 using eqn. (4).

From the figure it is clear that eqn. (4) is an excellent model for vapour pressure. In general the ranges for the three constants a1, a2 and a3 of eqn. (4) are: a1 ∼6.5–7.5, a2 ∼0.000055–0.000065 and a3 ∼0.020–0.040. For example the R^2 for all the data in Fig. 1 goes up to 0.97 if a1 = 7.16, a2 = 0.000057 and a3 = 0.036.

Other models for log(vp) obtained for all data given in Fig. 1 using a GFA are shown below. Adding other terms improves the R^2, but might also result in overfitting the model.

$$\log(\text{vp}) = 12.61 - 0.11*MolRef - 0.27*SP \ (R^2 \sim 0.94) \quad (5)$$

$$\log(\text{vp}) = 7.16 - 0.000057*MW*SP*SP \\ - 0.036*MolRef \ (R^2 \sim 0.97) \qquad (6)$$

$$\log(\text{vp}) = 8.81 + 0.2*HBA - 0.00005*MW*SP*SP \\ - 0.05*MolRef - 0.08*SP \ (R^2 \sim 0.98) \qquad (7)$$

The dependence of the model on MW*SP*SP is looked at next. Intuitively, the dependence of molecular weight is not surprising since 'lighter' (*i.e.* low molecular weight) molecules are expected to evaporate readily and consequently have a high vapour pressure. However, if the molecules have many interacting groups they would have a greater cohesive energy and this would slow down the evaporation and manifest as a low vapour pressure. Thus the cross term MW*SP*SP captures not only the size of the molecule but also its interacting ability and is therefore intimately related to the vapour pressure.

More rigorously, we can see from eqns. (1–3) that by substituting for ΔH_{vap} from eqn. (2) into eqn. (1) we get ln(vp) proportional to $V_{\text{m}}*E_{\text{coh}}$ (or $V_{\text{m}}*SP*SP$). Therefore the term $V_{\text{m}}*SP*SP$ is expected to be as well or better related to log(vp) than MW*SP*SP. In fact for all the points in Fig. 1, a simple equation with an R^2 of 0.94 relates log(vp) and $V_{\text{m}}*SP*SP$ and is given by

$$\log(\text{vp}) = 8.133 - 0.00006*V_{\text{m}}*SP*SP \ (R^2 \sim 0.94) \qquad (8)$$

V_{m} of course is the molar volume and is related to the molecular weight (MW) through the density of the material. It is easy

to see that the term MW*SP*SP was an approximation of V_m*SP*SP. This is reasonable since the density of most organic materials does not vary a lot. The disadvantage of using V_m in correlations is that molar volumes are a lot more difficult to compute accurately than molecular weights. However, using MD simulations[12] and an *NPT* protocol (at $P = 0$) because the molecule is allowed to explore different conformational states one can quite accurately obtain V_m.[18]

Apart from MW*SP*SP or V_m*SP*SP, other terms in the models (eqns. (4–7)) are probably correction terms and add only marginal improvements. The corrections could occur because (a) the forcefield used may not be well parameterised for the molecules under consideration, (b) the Clausius–Clapeyron equation may not work well for real (non-ideal) systems. A compressibility factor ($Z = PV/nRT$) may be involved in eqn. (1) and (c) the experimental data of vapour pressure may have been extrapolated and may not be accurate for 25 °C.

Next, a comparison is made of vapour pressure relationships with the term MW*SP*SP using SPs calculated by the molecular dynamics method (as in this paper) and Hansen's SPs. The comparison could not be made for all materials in this paper, as Hansens's parameters for such substances are not widely reported so instead the comparison was done for a list[19] that includes common, well known substances whose vapour pressures differ by almost 9 orders of magnitude.

The following points emerged from this study:

1. The correlation of MW*SP*SP with log(vp) using SPs obtained by MD ($R^2 \sim 0.89$) is better than that obtained using Hansen's SP ($R^2 \sim 0.8$).

2. The correlation is better at high vapour pressures for well known common small molecules for both sets of SPs as forcefield and group contribution terms are well parameterised for common small systems (and may have used vapour pressure data in the parameterisation).

3. As the molecules get bigger and more flexible, the correlation using Hansen's parameters gets worse. In fact for the 7 biggest molecules in the list, the correlation (R^2) with Hansen's SP is 0.37 while it is 0.81 for SPs from molecular dynamics.

Thus, predictions of vapour pressure using solubility parameters calculated from group contribution methods are not consistently good. It was also found from additional studies made here that using other group contribution methods[9] (not Hansen's method) and for the kind of molecules looked at here the best R^2 value obtained did not exceed 0.7.

This highlights the importance of the method by which solubility parameters are calculated. Molecular dynamics methods being more rigorous offer a much better 'quality' of parameter. This quality is of course fully dependent on the forcefield used. Conversely, forcefield parameters may be improved by checking how well they predict vapour pressures.

Other types of molecular simulations especially Monte Carlo methods using specialised protocols and methodologies[3] (like Gibbs ensemble or Widom particle insertion) have been used to study liquid–vapour equilibrium and predict important bulk properties.[4] While good results have been obtained for model systems and specific classes of compounds, the accuracy of the predictions have largely been dependent on the size and density of the system studied, and more importantly on the potential functions describing the molecular interactions of the systems. Though a promising route for calculating bulk properties, considerable work may need to be invested in (a) tailoring the implementation of these methodologies and protocols according to the type of system studied (b) improving the potential functions describing interactions; before such simulations can be routinely and universally applied to bigger and more complex molecules.

The downside of MD methods used here is the computational time required for each calculation (roughly 2 h on a R12000 processor for a small molecule). However, since chirality, tacticity and conformational variability can be fully accounted for in

MD, bulk parameters like density[12] and solubility parameters can be calculated to a high degree of accuracy. Moreover, since molecular dynamics is performed at a given temperature, the temperature dependence of parameters is also expected to be reproduced. Indeed initial work in this area suggests that this a promising route for calculating quite accurately vapour pressures at different temperatures and the boiling point.

Conclusions

In this paper, it has been shown that the vapour pressure of a substance at 25 °C can be accurately predicted from solubility parameters obtained from molecular dynamics simulations. It is found that a single cross term that is the product of either the molar volume or molecular weight of the substance and the square of the solubility parameters (V_m*SP*SP or MW*SP*SP), dominates the equation for vapour pressure. The solubility parameter method to predict vapour pressure is especially appealing as it is sensible in physical chemical terms and can be understood by the underlying theory for liquid–vapour equilibria within the Clausius–Clapeyron framework, lending confidence in the models. The models that have been built being representative of the data set from which they were derived will probably work best for small molecules, *i.e.* small organic molecules (MW <500) containing mostly carbon and oxygen atoms. However, the fact that the models can be explained in physical–chemical terms suggests a wider applicability, though caution may need to be exercised. Finally, given the slight uncertainties of experimental vapour pressure data, a correlation consistently over 0.9, and predictions as good as those shown in eqns. (4–8), the models for vapour pressure from parameters derived from molecular dynamics simulations can be considered to be very good indeed.

Acknowledgements

References

1 J. C. Dearden, *Environ. Toxicol. Chem.*, 2003, **22**, 1696–1709 and references cited therein;; B. E. Poling, J. M. Prausnitz and J. P. O'Connell, *The properties of gases and liquids*, Mcgraw-Hill, New York, 5th edn., 2000.

2 S. T. Lin, J. Chang, S. Wang, W. A. Goddard and S. I. Sander, *J. Phys. Chem. A*, 2004, **108**, 7429–7439; A. Habibi-Yangjeh, *Indian J. Chem., Sect. B*, 2004, **43**, 1504–1526; J. W. Chang, P. Yang, S. Chen, X. Quan, X. Yuan, K. W. Schraam and A. Kettrup, *SAR QSAR Environ. Res.*, 2003, **14**, 97–111; A. J. Chalk, B. Beck and T. Clark, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1053–1059; B. Beck, A. Briendl and T. Clark, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1046–1051.

3 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, Oxford, 1987; D. Frenkel and B. Smit, *Understanding Molecular Simulations*, Academic Press Inc., San Diego, 1996.

4 Example references are: M. G. AhunBay, S. Kranias, V. Lachet and P. Ungerer, *Fluid Phase Equilib.*, 2004, **224**, 73–81; A. Z. Panagiotopoulos, *J. Phys.: Condens. Matter*, 2000, **12**, R25–R52; G. C. Boulougouris, J. R. Errington, I. G. Economou, A. Z. Panagiotopoulos and D. N. Theodorou, *J. Phys. Chem. B*, 2000, **104**, 4958–4963; R. J. Sadus, *J. Phys. Chem. B*, 1997, **101**, 3834–3838; J. I. Siepmann, *Adv. Chem. Phys.*, 1999, **105**, 443–460.

5 Here it is assumed that ΔH_{vap} does not depend on temperature (*T*). R is the universal gas constant and B is a constant. V_m is the molar volume and ln(vp) refers to natural logarithm.

6 A. F. M. Barton, *CRC Handbook of Solubility Parameters and Other Cohesion Parameters*, CRC Press Inc., Boca Raton, 2nd edn., 1991.

7 J. H. Hildebrand and R. L. Scott, *The Solubility of Nonelectrolytes*, Dover Publications Inc., NY, 1964 (or Reinhold, NY, 1950).

8 C. M. Hansen, *Hansen's Solubility Parameters–a users handbook*, CRC Press Inc, Boca Raton, 2000.

9 S. A. Siddiqui and H. L. Needles, *Text. Res. J.*, 1982, 570–579; B. C. Hancock, P. York and R. C. Rowe, *Int. J. Pharm.*, 1997, **148**, 1–21; D. W. Van Krevelen, P. J. Hoftyzer, *Properties of Polymers: Their estimation and correlation with chemical structure*, Elsevier,

Amsterdam, 1976; R. F. Fedors, *Polym. Eng. Sci.*, 1974, **14**, 147–154; (this is by no means a complete list of such references).

10 D. Rigby, H. Sun and B. E. Eichinger, *Polym. Int.*, 1997, **44**, 311–330; H. Sun and D. Rigby, *Spectrochim. Acta, Part A*, 1997, **A53**, 1301–1323.

11 All software used in this report like Insight II, Cerius2 and DISCOVER are products of Accelrys Inc., 9685 Scranton Road, San Diego 92121, CA, USA.

12 P. K. C. Paul, S. Fraser and D. Rigby, *J. Chem. Soc., Perkin Trans. 2*, 2002, 199–200.

13 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

14 H. C. Andersen, *J. Chem. Phys.*, 1980, **72**, 2384–2393; T. A. Andrea, W. C. Swope and H. C. Andersen, *J. Chem. Phys.*, 1983, **79**, 4576–2393.

15 For a parameter description see the Cerius2–QSAR+ software manual or the Accelrys web page; Todeschini *et al.*, *Handbook of molecular descriptors*, Wiley-VCH, Weinheim, 2000.

16 *CRC Handbook of Chemistry and Physics*, 82nd edn., 2001–2002.

17 The 40 molecules in the test set were: acetic acid, acetone, acetonitrile, acetophenone, C8-aldehyde, amyl butyrate, benzene, benzyl benzoate, n-butanol, iso-butanol, sec-butanol, chloroform, dichloromethane, diethyl phthalate, diethyl ether, dimethyl benzyl carbinyl acetate (dmbca), fixolide, florhydral, dioxane, ethanol, ethyl acetate, phenethyl acetate, methyl salicylate, γ-ionone, hedione, phenethyl alcohol, thujone, eugenol, linalyl acetate, musk ambrette, glycerol, hexane, isopentane, octene, propanol, propylene, propylene glycol, *m*-xylene, *p*-xylene and toluene.

18 V_m is calculated by dividing the average volume of an amorphous box of molecules after a large MD (with *NPT* protocol and $P = 0$), by the number of molecules in the simulation box.

19 The list of molecules with Hansen's solubility parameters[6] used in this paper: methanol, acetonitrile, ethanol, acetone, acetamide, acetic acid, isopentane, butanol, diethyl ether, propylene glycol, benzene, crotonic acid, hexane, dioxane, ethyl acetate, glycerol, toluene, phenol, benzaldehyde, benzyl alcohol, anisole, octene, indole, acetophenone, benzoic acid, acetanilide, methyl benzoate, anethole, benzophenone, diethyl phthalate and stearic acid.