

Prediction of the ability of hydrofluorocarbons to dissolve hydrocarbons by means of neural networks

A.A. Dowman, A.A. Woolf *

Faculty of Applied Sciences, University of the West of England, Bristol BS16 1QY, UK

Received 21 November 1994; accepted 11 February 1995

Abstract

An alternative approach to predicting alkane solubilities in hydrofluorocarbons using artificial neural networks is described. The networks are trained with a minimum set of compounds above and below a set solubility limit. Each potential solvent is given descriptors, such as the number of atoms or their ratios as deduced from the structural formula. Further descriptors, one summing the number of adjacent dipoles of opposite polarity and another ratioing hydrocarbon content of solvent relative to solute, are also needed. The differentiation of solvents from non-solvents is better than that found previously with a predominant area diagram drawn on polar and non-polar contributing axes. The smallest networks provide inequalities which are linear in the descriptors and become greater or lesser than zero for non-solvents and solvents, respectively. The order of solubility is predictable from the inequalities.

Keywords: Neural networks; Hydrofluorocarbons; Alkanes; Solubilities; Fluorocarbons

1. Introduction

It is difficult to predict the properties of hydrofluorocarbons (HFCs) from those of the corresponding hydrocarbons (HCs) and perfluorocarbons (FCs). Thus the boiling points of two-carbon compounds do not increase monotonically with molecular mass. Instead they rise from minima at C_2H_6 and C_2F_6 to maxima at CH_2FCH_2F and CHF_2CH_2F with intermediate values around CH_3CF_3 [1]. Hence HFCs cannot be modelled in the same manner as HCs or FCs because HFCs are polar. They combine opposite polarities ($H-C-F$) unlike HCs or FCs which are non-polar overall.

HFCs have been proposed for replacing chlorinated solvents for cleaning or degreasing. Van Der Puy and coworkers have attempted to estimate the solubilities of HCs in possible replacement solvents [2]. The HFCs were modelled with a solubility parameter SP given by Eq. (1), i.e.

$$SP = 1.175 \ln(np) + [0.025H - 0.063F - 0.028\alpha - 0.018\beta] \quad (1)$$

where the non-polar (np) term was equal to $R_m/V_m^{-1/3}$ (where R_m is the molar refractivity and V_m the molar volume, both of which can be summed from group contributions). This term can be identified with the London dispersion forces

in the liquid and had previously been related linearly to the boiling points of HCs and FCs as far as the octanes [3]. (However the relationship overestimates the boiling points increasingly as the chain lengthens. For $C_{16}H_{34}$ the estimate is 594.5 K compared with the actual value of 560 K and for $C_{16}F_{34}$ the estimate is 567.0 K compared with 512.2 K.) The rest of the terms sum to a much smaller polar contribution. The number of hydrogens (H) and fluorines (F) can be taken to represent $C-H$ and $C-F$ dipoles and α and β their adjacency in $H-C-F$ and $H-C-C-F$, respectively. Thus the evaluation of the five constants required at least five known solubilities, and to be realistic at least 10 to allow partial optimization. By plotting the polar versus the non-polar terms, Van Der Puy and coworkers produced a predominant area diagram for a particular alkane ($C_{16}H_{34}$) where the solubility was taken as $>10\%$ by volume. (In their Table 1 solubilities were expressed as solute/solution volumes as opposed to solute/solvent ones in their Fig. 1. Conversion using fractional volumes may be achieved using $[\{v/(V+v)\}^{-1} - 1]^{-1} = v/V$, where v and V are the solute and solvent volumes, respectively.) Solvents were separated from non-solvents, albeit with about six of them on, or close to, the boundary line. For convenience, the diagram is reproduced as Fig. 1 with the compounds numbered (see Appendix 1 for the formulae). Corrections are also indicated as crosses (x) where the β terms were underestimated. The additional

* Corresponding author.

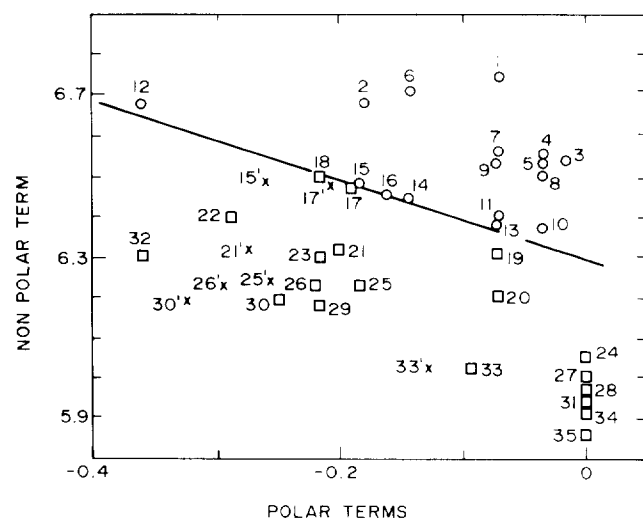


Fig. 1. Solvents (○) and non-solvents (□) for hexadecane with corrected values (x). Terms are given in Eq. (1) and formulae are listed in Appendix 1.

β terms are those bracketed after the following numbered compounds: 15(4); 17(1); 21(2); 25(4); 26(4); 30(4); and 33(2).

2. Application of neural networks

The classification of HFCs is essentially a pattern recognition problem which can be treated by such methods as principal, or discriminant, component analysis [4] without having to postulate any mechanism for solubility at the molecular level. The method followed here uses artificial neural networks (NNs) [5] which input numerical descriptors for each compound and transform them via intermediate layers to outputs set at 1 and 0 for solvents and non-solvents, respectively. The connections (synapses) between inputs and each member (neuron) of the intermediate (hidden) layer are weighted and then summed at that neuron. The sums are scaled between 0 and 1 with a sigmoidal function $[1/(1+e^{-x})]$. Each column of descriptors for the set of compounds is also scaled, usually relative to the highest value as unity. The hidden neurons are fully connected to one or more outputs in similar fashion. The network is trained with known solvents and non-solvents to give unit and zero outputs by systematically adjusting weights and biases in a back-propagation mode. This collection of weights and biases enables other HFCs to be classified rapidly as solvents or non-solvents, under the conditions pertaining to the training set, by a forward propagation.

A relevant application of NNs was in the prediction of CFC and halomethane boiling points from their structural formulae [6]. The number of carbons and number of each halogen were input as descriptors together with some topological indices expressing molecular shape in numerical form. For example the simplest such index due to Wiener [7] is half the sum of the interatomic distances in the molecular graph. Because

both boiling points and solubilities depend on intermolecular forces, it seemed reasonable to first input C , H , F , H/F and C/F counts as descriptors. However topological descriptors were omitted because FC and HFC boiling points (in °C) are barely structure-dependent as amply illustrated by the values provided by Van Der Puy et al. [2]: Et_t^iPr , 36–37; $^i\text{Pr}_t^i\text{Et}$, 37–39; Et_t^iBu , 61–62; $^i\text{Pr}_t^i\text{Pr}$, 62; Pr_t^iPr , 59–61; Bu_t^iPr , 86–88; Bu_t^iPr , 87.

Initially, the choice of training set, guided by the distribution shown in Fig. 1, avoided compounds near the boundary line. Compounds 1–6 and 8 were trained to unit output and 26, 29, 30, 32, 33 and 34 to zero using H , C , F , H/F and C/F numbers as descriptors in a 5–3–1 network, where 3 is a three-neuron hidden layer. Five of the remaining 23 compounds were wrongly predicted. Hence an additional descriptor was introduced to improve prediction, namely the aforementioned α and β factors which allowed for polarity. They were assumed to be of equal importance, and since there was no priori reason to favour either, were inputted as simple sums. They compactly store structural information such as could be provided at greater length by listing all CH_nF_m groups with n and m ranging from 1 to 3. They also serve to indicate group position. For example, $\text{CF}_3\text{CF}_2\text{CH}(\text{CH}_3)_2$ gives a 2β value while the isomer $\text{F}_2\text{CHCF}_2\text{CF}(\text{CH}_3)_2$ gives a $2\alpha + 6\beta$ sum. The new 6–3–1 network only left compound 25 ($\text{FCH}_2\text{CH}_2\text{F}$) as the outstanding exception, classifying it clearly as a solvent. (Compound 13 also became a marginal non-solvent.) Although 25 is the only possible two-carbon HFC suitable for a cleaning solvent — all others boil below room temperature — another descriptor was devised to remove the anomaly because it would also be applicable when shorter chain alkanes were considered as solutes. Arguing that 'like dissolves like', the HC content of a potential solvent relative to a solute, expressed as the ratio R of the number of carbons, would be low for non-solvents and higher for solvents. This ratio was down-weighted for molecules with non-adjacent CH_n groups because the 'clustered' compounds R_tR have been shown to be better solvents than their mixed isomers. The down-weighting factor was varied and a value of 0.75 was found to be satisfactory. Rather than expand to a 7–3–1 network, it was possible to omit H and C/F as redundant information and return to a 5–3–1 network with the original training set. This led to an all-correct prediction of solubility as judged by a boundary at greater than 0.5. Contracting the hidden layer to give a 5–2–1 architecture gave as good a partition.

However, because no further improvement (i.e. moving all outputs closer to 1 or 0) was possible at this stage, the training set was re-examined. In Fig. 1 there is a poor correspondence between a molecule's position and solubility, especially near the boundary line. After several trials, the final choice for the solvent set was the most soluble compounds 2 and 5 and the reasonably soluble compounds 13 and 16, although these were close to the line. For the non-solvent set, all the zero polarity FCs were excluded, as well as the monohydrogen HFC 33, as obvious non-solvents.

Instead, compounds 19, 21, 23, 25 and 29 were taken. Compound 25, $F(CH_2)_2F$, was trained as a non-solvent because it was a good test for predicting that $F(CH_2)_3F$, with an extra CH_2 group, would be a solvent. In all, a quarter of the compounds constituted the training set.

The neurons in the hidden layer were systematically varied and the descriptors reduced to H , C and F numbers together with $(\alpha + \beta)$ and R values. The resulting 5–1–1 network, using the above training set, was compared with the 7–1–1 network which included the C/F and H/F descriptors (Appendix 2). The solvency of the remaining 28 compounds was correctly predicted with all outputs within 0.02 of the required unit or zero values. As the hidden layer has been reduced to one neuron, these networks can be expressed as linear inequalities. For example, with the 5–1–1 network

$$9.99F/20 - 7.40H/20 + 3.25C/10 + 9.42(\alpha + \beta)/20 - 9.60R/0.25 - 0.66 \geq 0 \quad (2)$$

where the numerator constants are the weights connecting the inputs to the hidden neuron, and the last constant is the bias weight. The denominators are the maximum descriptor values used to normalize the column values listed in Appendix 2. Inequalities of this type are not unique but will vary with the choice of training set. Overall they should all produce the same sequence of solvency.

For the 7–1–1 network, inequality (2) is:

$$9.65F/20 - 7.13H/10 + 2.48C/20 - (1.54C/F)/2 - (2.25H/F)3 + 9.46(\alpha + \beta)/20 - 10.36R/0.25 + 07 \geq 0 \quad (3)$$

Interestingly the available experimental solubility values [2] parallel the inequality values (Table 1). There is at least a

Table 1
Comparison of experimental with predicted order of solubilities in hydrofluorocarbons

Solubility (solute/solvent) by volume [2]	Compound ^a	Predicted order of solubility from 5–1–1 network
Miscible ^b	1	1
17.6 ^b	4	4
6.4 ^b	3	3
Miscible	2, 5, 7, 10	5 2 7 10
49.3	11	11
25.0	15	15
12.4	13	14
9.9	12, 14	13 12

^a No values were reported for 6, 8 and 9 and these have not been included in the predicted order.

^b These values were for a mineral oil as the solute with a longer chain hydrocarbon than hexadecane.

^c Obtained from the inequality values listed in Table 2.

semi-quantitative relation quite unlike that which could be anticipated from Fig. 1.

3. Conclusions

The advantage of the NN approach is the ease with which networks can be retrained for different conditions. Thus different solubility limits could be chosen or the alkane solute changed, provided appropriate training sets were employed. It also seems possible to make quantitative solubility predictions if sufficient solubility data exist; perhaps with a more extended network. All the descriptors were derivable from the structural formulae without recourse to experimental data or preconceived notions of mechanism.

Appendix 1

List of FCs and HFCs used in Fig. 1 and in neural networks

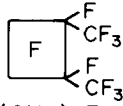
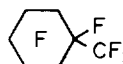
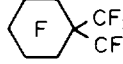

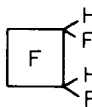

- | | |
|---|---|
| 1. Et _f Bu | 20. Oct _f Et |
| 2. (EtC ₂ F ₄) ₂ CF ₂ | 21. H(CF ₂) ₄ Me |
| 3. isoPr _f isoPr | 22. CF ₃ CH ₂ CF ₂ Me |
| 4. Et _f isoPr | 23. CF ₃ (CH ₂) ₂ CF ₃ |
| 5. Pr _f isoPr | 24.  |
| 6. Me(CF ₂) ₂ Et | 25. F(CH ₂) ₂ F |
| 7. Pr _f Pr | 26. H(CF ₂)H |
| 8. Bu _f isoPr | 27.  |
| 9. Bu _f Pr | 28.  |
| 10. isoPr _f Et | 29. CF ₃ CH ₂ CH(CF ₃) ₂ |
| 11. Pr _f Et | 30. HCF ₂ CFHCF ₂ H |
| 12. MeCF ₂ CH ₂ CF ₂ Me | 31. C ₅ F ₁₂ |
| 13. Pr _f Et | 32. CF ₃ CH ₂ CF ₂ CH ₂ CF ₃ |
| 14.  | 33. CF ₃ (CF ₂) ₅ H |
| 15. F(CH ₂) ₃ F | 34. C ₆ F ₁₄ |
| 16. CF ₃ CH(Me)CH ₂ CF ₃ | 35. C ₈ F ₁₈ |
| 17. Pr _f CH ₂ CHFMe | 36. cis  |
| 18. Me(CF ₂) ₃ Me | 38. cis  |
| 19. Hexyl _f Et | |

Table 2
Final neural networks used to differentiate solvents from non-solvents

Compound No.	Descriptor inputs							Taught to value	Network output		Values of inequalities	
	F	H	C	C/F	H/F	$\alpha + \beta$	R		7-1-1	5-1-1	Eq. (2)	Eq. (3)
1	5	9	6	1.20	1.80	4	0.250		1.00	1.00	-12.41	-10.46
2	10	10	9	0.90	1.00	8	0.250	1	1.00	1.00	-7.27	-5.83
3	7	7	6	0.86	1.00	1	0.188		1.00	1.00	-8.04	-7.03
4	5	7	5	1.00	1.40	2	0.188		1.00	1.00	-9.19	-7.89
5	7	7	6	0.86	1.00	2	0.188	1	1.00	1.00	-7.57	-6.56
6	4	8	5	1.25	2.00	10	0.125		1.00	1.00	-4.67	-2.98
7	7	7	6	0.86	1.00	4	0.188		1.00	1.00	-6.63	-5.62
8	9	7	7	0.78	0.78	2	0.188		1.00	1.00	-6.14	-5.23
9	9	7	7	0.78	0.78	4	0.188		1.00	1.00	-5.20	-4.29
10	7	5	5	0.71	0.71	2	0.125		1.00	1.00	-3.49	-3.02
11	7	5	5	0.71	0.71	4	0.125		1.00	1.00	-2.55	-2.08
12	4	8	5	1.25	2.00	20	0.188		0.99	0.96	-2.56	-0.67
13	9	5	6	0.67	0.56	4	0.125	1	0.98	0.98	-1.20	-0.75
14	4	4	4	1.00	1.00	8	0.125		0.99	0.98	-2.09	-1.29
15	2	6	3	1.50	3.00	8	0.063		1.00	0.99	-4.05	-1.74
16	6	6	5	0.83	1.00	9	0.188	1	1.00	1.00	-4.28	-3.36
17	8	6	6	0.75	0.75	10	0.094		0.00	0.00	1.50	2.01
18	6	6	5	0.83	1.00	12	0.094		0.01	0.01	0.98	1.62
19	13	5	8	0.62	0.38	4	0.125	0	0.02	0.02	1.40	1.91
20	17	5	10	0.59	0.30	4	0.125		0.00	0.00	3.90	4.57
21	8	4	5	0.63	0.50	10	0.063	0	0.00	0.00	4.22	4.34
22	5	5	4	0.80	1.00	16	0.094		0.00	0.00	2.87	3.41
23	6	4	4	0.67	0.67	12	0.125	0	0.00	0.01	1.25	1.59
24	12	0	6	0.50	0.00	0	0.000		0.00	0.00	7.59	7.32
25	2	4	2	1.00	2.00	8	0.000	0	0.00	0.00	0.83	1.80
26	9	3	5	0.56	0.33	10	0.000		0.00	0.00	8.17	7.97
27	14	0	7	0.50	0.00	0	0.000		0.00	0.00	8.80	8.65
28	16	0	8	0.50	0.00	0	0.000		0.00	0.00	10.01	9.98
29	9	3	5	0.56	0.33	12	0.063	0	0.00	0.00	6.53	6.52
30	5	3	3	0.60	0.60	10	0.000		0.00	0.00	5.52	5.31
31	12	0	5	0.42	0.00	0	0.000		0.00	0.00	7.40	6.99
32	8	4	5	0.63	0.50	20	0.094		0.00	0.00	7.65	7.86
33	13	1	6	0.46	0.08	4	0.000		0.00	0.00	9.21	8.96
34	14	0	6	0.43	0.00	0	0.000		0.00	0.00	8.60	8.32
35	18	0	8	0.44	0.00	0	0.000		0.00	0.00	11.01	10.98
36	6	2	4	0.67	0.33	4	0.000		0.00	0.00	4.29	4.06
38	8	2	5	0.63	0.25	4	0.000		0.00	0.00	5.59	5.39

Appendix 2

The neural networks were investigated using Neural Desk neural software supplied by Neural Computer Science (Southampton, UK) run on a 486 DX processor. The learning time was <1 min for ca. 1500 epochs converging with an average error of 0.01.

Table 2 lists details of the final neural network inputs and outputs.

References

- [1] A.A. Woolf, *J. Fluorine Chem.*, 50 (1990) 89.
- [2] M. Van Der Puy, A.J. Poss, P.J. Persichini and L.A.S. Ellis, *J. Fluorine Chem.*, 67 (1994) 215.
- [3] M. Van Der Puy, *J. Fluorine Chem.*, 63 (1993) 165.
- [4] R.G. Brereton, *Chemometrics*, Ellis Horwood, Chichester, UK, 1990.
- [5] J. Zupan and J. Gasteiger, *Anal. Chim. Acta*, 248 (1991) 1; M.T. Spining, J.A. Darsey, B.G. Sumpter and D.W. Nold, *J. Chem. Educ.*, 71 (1994) 406.
- [6] A.T. Balaban, S.C. Basak, T. Colburn and G.D. Grunwald, *J. Chem. Inf. Comput. Sci.*, 34 (1994) 1, 118.
- [7] H. Weiner, *J. Am. Chem. Soc.*, 69 (1947) 17.