Quantitative structure–sublimation enthalpy relationship studied by neural networks, theoretical crystal packing calculations and multilinear regression analysis

Michael H. Charlton,* Robert Docherty and Michael G. Hutchings
Zeneca Specialities Research Centre, Hexagon House, PO Box 42, Blackley, Manchester, UK M9 8ZS

Three different techniques have been used to analyse the relationship between the structure of 62 organic compounds and their sublimation enthalpies. Using a neural network based on molecular structure descriptors (molecular formula, hydrogen bonding and π-characteristics), sublimation enthalpies can be modelled. The best of the neural network models yielded an average error of 2.5 kcal mol^{-1} in a series of 'leave-one-out experiments'. The same sublimation enthalpy data have been studied using theoretical techniques based upon crystal packing calculations, and also with a simple three parameter multilinear regression model. The latter two methods produced results that were superior to the neural network in this particular study (mean errors of 1.4 and 1.8 kcal mol^{-1}, respectively), although in the case of MLRA, this is the result of the model fitting exercise, and not a predictive run. It was surprising to find such a simple linear relationship between characteristics describing the molecular formula and the sublimation enthalpy. Nevertheless, the results here have highlighted the potential of neural networks and MLRA as useful tools for the approximate prediction of physical properties, as demonstrated for a series of compounds not included in the training set.

Introduction

The sublimation enthalpy ($\Delta_{sub}H$) of a compound is the energy required to break up the solid state and convert the system into the gas phase. It is a property of particular interest in the crystal chemistry of molecular materials. Pigments and disperse dyes are two typical examples in which the $\Delta_{sub}H$ plays an important role, since many characteristic properties of these compounds are governed by solid state interactions. Colour, light fastness, thermal stability and bleeding characteristics are often dependent upon the intermolecular interactions within the solid state. The prediction of thermochemical data are also of considerable importance in the identification of commercially viable processes, in the formulation of many products such as dyes, drugs and agrochemicals, and in plant design and operation.

Clearly, the $\Delta_{sub}H$ and other thermochemical data can be measured, but this may be a costly and time-consuming process. The prediction of solid state properties (such as sublimation enthalpy) before a compound is synthesised would obviously be of considerable use. Other multilinear regression analysis (MLRA) studies of the $\Delta_{sub}H$ have been performed, but these have been upon restricted classes of compounds.[1,2] They also tend to use considerably more complex input parameters which include some experimental information, meaning that *a priori* predictions cannot be made.

This work outlines the use of various quantitative techniques in the modelling and prediction of the $\Delta_{sub}H$ of a series of organic molecules, ranging from simple hydrocarbons, polycyclic aromatic compounds, carboxylic acids, amides and amino acids to heterocycles and dyestuffs. The compounds under study in this work are limited to those containing only carbon, hydrogen, oxygen and nitrogen, although the system could, in principle, be extended to other elements.

The first technique used is that of Neural Networks (NN), and this is then compared both with calculated values from traditional molecular modelling techniques and with a MLRA model. The input to the NN has been kept as simple as possible to facilitate the prediction of values for novel molecules. The MLRA uses the same input data as the NN, although some covariance was found between the input parameters, indicating that the number of inputs could be reduced. In addition, some other simple parameters have also been investigated in the MLRA study.

Neural networks

NNs are so called because they attempt to model some of the processing functions of the human brain. The idea of neural processing was originally developed in the 1940s,[3] but subsequently fell out of favour until the 1980s. This renewed interest followed the publication of work which discussed the limitations of these earlier systems,[4] as well as the dramatic growth of traditional digital computing. In 1982, Hopfield published work that led to the re-emergence of neural nets as computational tools.[5]

NNs are able to perform highly non-linear pattern recognition, classification and regression tasks, the results of which are often superior to traditional approaches. Recent applications of NNs include the determination of structure–activity relationships in drug design,[6–10] the prediction of protein structure[10] and the classification of spectra.[10,11] Thermochemical data such as solubilities[12] and boiling points of organic heterocycles[13] have also been the subject of investigation. There are also a number of reviews of the use of NNs in chemistry.[10,14,15]

The network used in this study is a program called PSDD[16] (Perceptron Simulation for Drug Design), which is freely available. In general, NNs can have any number of hidden layers, but the PSDD program is restricted to only one, although this can contain any number of nodes. The authors of the program consider a three layer network (with input, output and one hidden layer) to be sufficient for practical structure–activity studies.[16,17] The PSDD simulation is a feed forward network, in which each layer is only connected to the outputs of all the nodes in the preceding layer. This is relatively simple to construct and easier to train than networks with more layers.[6] A number of different NN architectures have been studied, but to date, the feed-forward network has achieved greatest popularity in chemical studies.

Network training

PSDD uses a supervised training scheme. This involves showing the network a set of sample data, for which the desired physical or chemical property is known. The main learning scheme used by the program is Back Propagation Learning.[15] This starts at the output layer, and systematically alters the weights of the preceding layer to yield the required output. The program passes through all the data repeatedly, until the difference between predicted and experimental values falls below a preset limit, *i.e.* if we have a training pattern t, then the learning process is repeated until the error function, E, becomes small enough, eqn. (1). Further details of the learning process are given elsewhere.[15]

$$E = \sum_{j} (\text{Output}_j - t_j)^2 \tag{1}$$

Overfitting of the data

In networks with a large number of hidden nodes, there is a danger of overfitting the data. If the number of weights controlled by the NN exceeds the number of items in the training set, then the net will reproduce these items extremely well by memorising each of the values. In such a situation, it is acting rather like a look-up table, and is likely to be unable to interpolate its knowledge to systems that it has not yet encountered.[6] With a small number of hidden neurons, the node efficiency increases, and the network is able to infer relationships between the data more readily. To determine a suitable network architecture, a ratio, ρ, has been defined,[18] eqn. (2):

$$\rho = \frac{N_t}{N_w} \tag{2}$$

where N_t is the number of items in the training set, and N_w is the number of weights in the network. If training bias is included in the net, then N_w can be calculated from the number of input (I), hidden (H) and output (O) nodes, eqn. (3):

$$N_w = (I + 1)H + (H + 1)O \tag{3}$$

It has been suggested that to maintain the balance, allowing generalisation and preventing the NN from simply memorising the data, the network should be designed so that ρ lies between 1.8 and 2.2.[4] Other workers propose that a value of $\rho > 1$ is sufficient.[15]

Sublimation enthalpy prediction using the neural network

It was decided that the ability to predict the $\Delta_{sub}H$ of a compound would be of most use if the input to the network was as simple as possible. The absence of any quantum mechanically derived parameters (*e.g.* dipole moments, charges) from the model would enable much faster prediction. With this in mind, seven parameters were selected that could be important factors in determining the $\Delta_{sub}H$. The parameters are the number of carbon atoms (C), the number of hydrogen atoms (H), the number of nitrogen atoms (N), the number of oxygen atoms (O), the number of π-atoms (PI), the number of hydrogen bond donors (HBD) and the number of hydrogen bond acceptors (HBA). Initial examination of $\Delta_{sub}H$ trends indicates that molecular size, the π–π interactions and hydrogen bonding are of greatest importance. These features are represented by this series of parameters. It should be noted that in this context, HBD has been defined as the number of hydrogens attached to either N or O. Weak C–H ··· acceptor interactions have been ignored in this analysis. The HBA term is defined by the total number of oxygen atoms plus all nitrogen atoms except those in -NO_2 and aryl-NR_2 groups. The nitrogen atoms in N_2 are also excluded. It should also be stressed that all types of acceptors are treated equally, with no differentiation between either atom type or environment. This is also true for hydrogen-bond donors.

These values were tabulated for each of a list of 62 molecules which formed the training set. This initial dataset, although not exhaustive, was originally selected to represent a wide range of crystal chemistry for molecules containing only C, H, O and N atoms, for the purpose of evaluating theoretical crystal packing techniques. The dataset includes aliphatic and aromatic hydrocarbons, oxohydrocarbons, azahydrocarbons, carboxylic acids, amides and amino acids. One constraint on the dataset was that only compounds for which both $\Delta_{sub}H$ and accurate cystal structures are known were selected.

The information was entered into the PSDD program, along with the experimental values for the SE, most of which were obtained from Cox and Pilcher.[19] The experimental values are given in Table 1. A number of predictive runs termed 'Leave-One-Out Experiments' were performed to assess the performance of the network. For K compounds, this involves the network learning for $(K - 1)$ compounds, and then predicting the remaining one. This process is repeated for each of the molecules in the training set, giving a predicted value for each compound, involving no prior knowledge of the experimental value. Although such a run involves a series of quitc lengthy calculations, a 'production' version of the network could calculate an unknown $\Delta_{sub}H$ extremely rapidly, having been trained upon all the molecules in the training set.

Eight different runs were attempted to test the effect of altering values for the number of hidden neurons and the error cutoff function [eqn. (1)]. The runs were limited to a maximum of 8×10^5 training cycles, except for run 8, for which the limit was 1×10^6. The results for each run were analysed and compared to the experimental values. The statistical results of these runs are shown in Table 2, and enable the optimum parameters for the network to be determined, and show the modelling power of the network. The optimum NN architecture selected from these 8 was NN7, based upon best values for the square of the multiple correlation coefficient, r^2, maximum and average errors, and SD. The predictive results for this leave-one-out experiment are presented in Table 1.

Effect of *ρ* and *E* on network predictions

By varying the learning termination threshold E_{max} {the maximum allowable value of E [eqn. (1)]}, the learning behaviour of the net can be modified. For different values of ρ, the effect of altering E_{max} has been examined in network runs 1 to 8. Whilst large values lead to insufficient training, smaller values mean that the network takes longer to train, because it is harder to obtain predictions with very small errors during training. Low values can also lead to the overtraining phenomenon previously described when the value of ρ is also small. In this situation, the network starts to reproduce the noise in the input data, rather than the overall trends, and this leads to a poor ability to predict values for unknown compounds. The effect can be seen by comparing the results for Runs 1, 7 and 8. Run 7 has the best results, and an intermediate value of E_{max}. This implies that not enough training has occurred in NN1, and the network has overtrained in NN8. Reducing the number of hidden neurons (runs NN7, NN3 and NN6) leads to a gradual deterioration in results. For intermediate numbers of hidden neurons ($\rho = 2.21$), the amount of training seems to make little difference (NN4 and NN6). Despite this, even the best network architecture still produces a maximum error of 10.1 kcal mol^{-1}, and two negative results (nitrogen and cyanogen), which are chemically meaningless. Both of these molecules have no hydrogens and a high nitrogen to carbon ratio, which could be

Table 1 Experimental and predicted values of the $\Delta_{sub}H$ of selected organic compounds (kcal mol^{-1})

Molecule	Expt.	Theory[a]	NN[b]	MLRA[c]
Pentane	9.9	10.3[d]	12.1	10.5
Hexane	12.6	12.6[d]	11.4	11.9
Octane	15.9	16.5[d]	15.9	14.8
Octadecane	37.8	35.2[e]	30.7	28.9
Benzene	12.5	12.5[d]	9.6	11.9
Biphenyl	20.7	21.6[e]	18.9	20.4
Naphthalene	17.3	19.7[e]	17.9	17.6
Anthracene	24.4	26.9[d]	22.6	23.2
Phenanthrene	20.7	23.3[f]	23.4	23.2
Chrysene	28.4	31.9[f]	28.7	28.9
Triphenylene	27.4	30.7[f]	27.5	28.9
Perylene	31.0	32.5[f]	29.7	31.7
Ovalene	50.6	52.7[d]	49.7	48.6
Benzoquinone	14.9	14.9[e]	18.9	16.5
Anthraquinone	26.1	26.8[e]	25.8	27.7
Benzophenone	23.9	24.5[e]	19.3	24.1
4-Methylphenol	17.7	15.4[g]	22.0	20.2
Succinic anhydride	19.6	17.8[h]	16.0	15.9
Maleic anhydride	16.4	15.2[h]	17.8	15.9
Phthalic anhydride	21.1	22.5[h]	22.9	21.5
Naphthaquinone	21.7	22.7[h]	19.4	22.1
9,10-Phenanthrenequinone	25.8	24.4[h]	24.8	27.7
Cyclohexane-1,4-dione	20.2	19.5[h]	16.2	16.5
2,2,4,4-Tetramethylcyclobutane-1,3-dione	17.3	16.1[h]	19.7	19.3
s-Trioxane	13.8	14.5[h]	18.7	14.5
1,3,5,7-Tetraoxocane	19.0	18.0[h]	15.3	18.2
Phenyl Benzoate	23.7	23.9[h]	26.9	26.3
Nitrogen	2.0	2.0[d]	−0.9	3.5
Cyanogen	8.7	8.4[d]	−1.4	10.8
Dicyanoacetylene	10.6	10.6[i]	16.7	13.6
Tetracyanoethylene	20.6	21.8[d]	17.8	21.0
Pyrimidine	11.7	13.5[d]	15.2	13.6
Pyrazine	14.5	14.3[j]	13.0	13.6
Trinitrotoluene	24.4	25.1[d]	21.0	26.9
N,*N*-Dimethyl-*p*-nitroaniline	23.8	26.3[k]	23.2	19.3
Formic acid	15.2	13.3[l]	15.6	14.0
Acetic acid	16.3	15.2[l]	16.7	15.4
Propanoic acid	17.7	17.6[l]	18.0	16.8
Butyric acid	19.2	19.1[l]	19.2	18.2
Valeric acid	20.2	21.3[l]	20.6	19.6
Oxalic acid	24.8	25.9[m]	28.5	24.4
Succinic acid	29.3	32.0[l]	28.6	27.3
Glutaric acid	29.0	31.0[l]	31.1	28.7
Adipic acid	32.1	34.5[l]	32.0	30.1
Suberic acid	35.4	31.5[l]	36.7	32.9
Sebacic acid	39.6	41.9[l]	37.7	35.7
Benzoic acid	23.0	20.4[e]	23.9	22.4
Oxamide	28.2	25.4[l]	25.7	29.0
Malonamide	28.8	31.0[l]	30.9	30.4
Succinamide	32.3	34.3[l]	28.7	31.8
Urea	22.2	23.4[l]	21.4	25.3
Formamide	17.5	15.7[l]	11.4	16.2
Diketopiperazine	26.0	27.0[l]	28.4	22.7
7,7,8,8-Tetracyanoquinodimethane	26.5	27.0[n]	27.6	29.5
Indigo	31.8	32.5[o]	40.4	39.7
Acridine	22.0	24.0[j]	23.6	24.1
Azobenzene	22.0	23.4[j]	28.1	24.9
Mesitonitrile (2,4,6-trimethylcyanobenzene)	18.6	17.0[j]	20.9	19.8
N-Methylcarbazole	22.8	21.6[j]	21.8	21.8
p-Dicyanobenzene	21.2	18.0[j]	15.5	19.3
s-Triazine	13.4	13.4[j]	13.3	14.5
Stearic Acid	39.8	41.3[p]	41.0	37.9
Max. error		3.5	10.1	8.9
Mean error		1.4	2.5	1.8
r^2		0.97	0.87	0.92

[a] Theoretical prediction using crystal packing. [b] Neural network prediction (Run 3, see Table 2). [c] Multilinear Regression Analysis prediction from eqn. (5). References: *d*[29], *e*[24], *f*[30], *g*[31], *h*[1], *i*[32], *j*[2], *k*[33], *l*[34], *m*[35], *n*[36], *o*[37] and *p*[38].

leading to confusion in the network. The inclusion of both compounds simultaneously during training may be enough to make the results more stable during prediction.

Theoretical crystal packing calculations

In order to understand the principles which govern the wide variety of solid state properties and structures of organic

Table 2 Analysis of predictions of $\Delta_{sub}H$

Run	Network parameters[a]			Max. error	Av. error[b]	r^2	SD[c]
	H	ρ	E_{max}				
NN1	7	1.13	0.001	14.472	2.501	0.803	2.396
NN2	5	1.68	0.001	16.194	2.374	0.797	2.751
NN3	5	1.68	0.0005	15.072	2.509	0.826	2.529
NN4	4	2.21	0.001	19.324	3.565	0.701	3.089
NN5	3	3.26	0.001	16.934	2.881	0.757	2.714
NN6	4	2.21	0.0005	16.771	3.205	0.706	3.309
NN7	7	1.13	0.0005	10.126	2.489	0.865	2.174
NN8	7	1.13	0.00001	24.247	3.672	0.599	4.231
Theory				3.5	1.385	0.971	0.939
MLRA[d]				8.9	1.777	0.917	1.623

[a] H = number of hidden neurons including a bias neuron in the hidden layer, ρ defined in eqn. (2). E_{max} = optimisation error threshold (eqn. 1). [b] Unsigned mean error. [c] Standard deviation based upon unsigned mean of errors in prediction. [d] Multilinear regression analysis.

Table 3 MLRA results

Entry	Parameters	r^2
1	*C H N O PI HBD HBA*	0.93
2	*C PI HBD HBA*	0.93
3	*C PI HBD HBA HBX*	0.93
4	*C PI MW HBD HBA*	0.93
5	*C HBD HBA*	0.92
6	*C HBD HBA AR ARH S*	0.94
7	*C HBD HBA ARH CSQ*	0.94
8	*C HBD HBA ARH*	0.94

materials it is important to describe the interactions of molecules in specific orientations and directions. As a result of the pioneering work of Williams[20] and Kitaigordsky[21] on the use of atom–atom potentials, and in more recent times, the elegant work of Gavezzotti and co-workers,[22] it is now possible to interpret packing effects in organic crystals in terms of interaction energies. The basic assumption of the atom–atom method is that the interaction between two molecules can be considered to consist simply of the sum of the interactions between the constituent atom pairs.

The lattice energy $\Delta_{latt}H$ often referred to as the crystal binding or cohesive energy, can, for molecular materials, be calculated by summing all the interactions between a central molecule and all the surrounding molecules. The lattice energy can be compared to the experimental sublimation enthalpy, having the same magnitude but the opposite sign. Each intermolecular interaction can be considered to consist of the sum of the constituent atom–atom interactions. If there are n atoms in the central molecule and n' atoms in each of the M surrounding molecules then lattice energy can be calculated by eqn. (4), shown below. In most cases n and n' will be equal, but in the case of molecular complexes they may differ. M is simply the total number of molecules in the crystal.

$$E_{latt} = 1/2 \sum_{k=1}^{M} \sum_{i=1}^{n} \sum_{j=1}^{n'} V_{kij} \tag{4}$$

V_{kij} is the interaction between atom i in the central molecule and atom j in the k^{th} surrounding molecule. It includes the electrostatic and Van der Waals' terms found in typical force field methodologies.[23]

The lattice energy ($-\Delta_{sub}H$) is a crucial parameter to be determined in the study of molecular materials. The calculated value has the advantage that the value can be broken down into the specific interactions along particular directions and further partitioned into the constituent atom–atom contributions. This is the key link between molecular structure and crystal packing arrangement. This allows a profile of the important intermolecular interactions to be built up within families of compounds and an understanding of the interactions which contribute to particular packing motifs.

The calculated lattice energies for the 62 compounds are reported in Table 1. The sources of the calculated data are clearly referenced. The details of those calculations are described within those references. Calculations carried out by the authors will be outlined in a future publication. They essentially involve the use of the HABIT program,[24] with a summation limit of 30 Å. Partial atomic charges were assigned using the AM1[25] method within the MOPAC program.[26] No minimisation of the experimental structure was permitted during either the molecular orbital or the crystal packing calculations.

Multinuclear regression analysis of sublimation enthalpy data

In order to check if the NN model for sublimation enthalpy of the data set already discussed was superior to conventional statistical modelling of the data, a MLRA was undertaken.[27] The independent variables and abbreviations used were exactly those used in the NN analysis above, with the following additional ones: the square of the number of carbon atoms (*CSQ* = *C***C*), product of *HBD* and *HBA* (*HBX*), number of aromatic rings in the molecule (*AR*), number of aromatic hydrogen atoms (*ARH*), the symmetry number of the molecule in its most symmetrical conformation (*S*), the molecular weight (*MW*) and an indicator variable to reflect if the molecule was dipolar (*DIP* = 1) or not (*DIP* = 0).

As a pre-evaluation of the independent variables, the Pearson correlation matrix was determined. This revealed strong colinearities amongst some variables. In particular, it was found that *C* correlates strongly with *PI* (0.789), *AR* (0.828), *ARH* (0.768) and *MW* (0.939). Thus, it is statistically unacceptable to include any of these other parameters as well as *C* in the MLRA. Next, various parameter combinations were checked for multilinearities. Of special relevance is the correlation between *C* and *HBD* and *HBA*. The square of the regression coefficient was found to be 0.21, and it was concluded that this combination would not result in coefficient bias. The original 7 NN input data are shown in Appendix 1. The additional parameters discussed above are not included because of their high correlation with the number of carbon atoms.

The $\Delta_{sub}H$ values were then regressed against various parameter combinations. Statistical results are recorded in Table 3. It should be noted that it was not the intention of this part of the study to develop an optimal model based upon MLRA. Rather, the aim was to determine how the molecular parameters used in the NN development performed in MLRA. Admittedly, a few extra parameters were checked in the MLRA, but a good additive model would probably need more subtle features of molecular structure than reflected by the list of parameters used here.

Results and discussion

The calculated $\Delta_{sub}H$ from the best of the NN predictions (Run 7, from Table 2), the theoretical crystal packing calculations and the best MLRA model are presented in Table 1, alongside the experimental values. Statistical results for all of the NN runs are shown in Table 2. Regression results from the MLRA appear in Table 3. Deviations are generally given as absolute values, however, the *relative* deviations in these predictions show a slightly different perspective on the results. For example, in the MLRA case of N_2, the error is 1.4 kcal mol^{-1} on an absolute value of 2.0, which is an error of 70%, whilst for

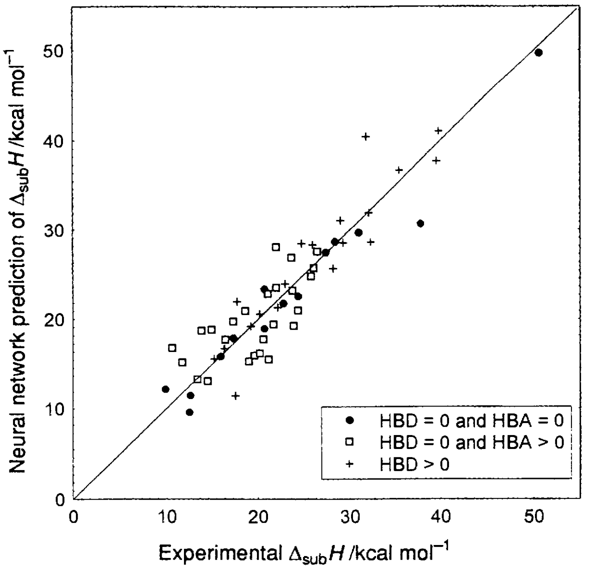


Fig. 1 Plot of NN prediction *vs*. experimental $\Delta_{sub}H$

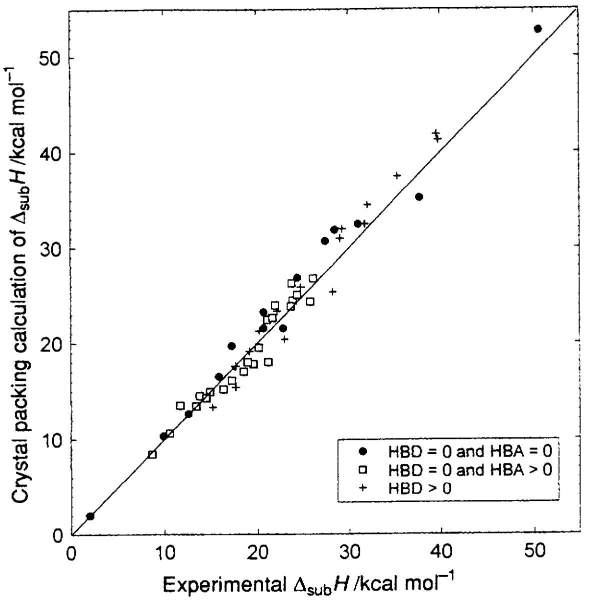


Fig. 2 Plot of crystal packing calculation *vs*. experimental $\Delta_{sub}H$

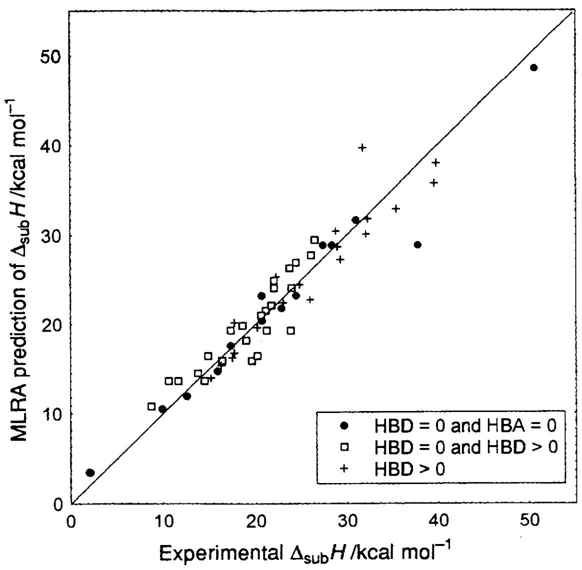


Fig. 3 Plot of MLRA prediction *vs*. experimental $\Delta_{sub}H$

ovalene, the error is 2.0 in 50.6 kcal mol^{-1}, which is only 4%. Nevertheless, we feel that quoting such errors unfairly biases the results against the smaller molecules, whereas it is often the largest molecules that have the greatest experimental error.

The results have been plotted in Figs. 1, 2 and 3. In all cases, the plots are of predicted *vs*. experimental sublimation enthalpy, and the straight line represents the linear regression of the data, from which the gradient, intercept and r^2 values have been determined.

It is clear that the NN results are less statistically accurate than the crystal packing calculations, as would be expected from such a simplifed model. Nevertheless, Fig. 1 shows that with the exception of a few outliers, the NN reproduces the experimental trends reasonably well. In general, the NN predictions are worse for larger values of $\Delta_{sub}H$, which is in line with the fact that there are fewer large molecules in the training set. The larger molecules, which tend to have the largest values of $\Delta_{sub}H$, are also inclined to have the largest values for the 7 input parameters. It is documented that neural nets, although good at interpolation, are relatively poor at extrapolating beyond the maximum values within the training set. Indeed, the PSDD program issues warnings if any parameters in the test set are greater than the maximum found in the training set.[16] This could partly explain the poor performance for these molecules.

The MLRA results are presented in Table 3, and it is clear that a 7-parameter combination (entry 1)—which in any case is statistically unreliable—is hardly any better than a 3-parameter model based on *C*, *HBA* and *HBD* (entry 5). Inclusion of the alternative extra parameters *CSQ*, *HBX*, *AR* and *S* caused no improvement in the model. While *ARH* was *statistically* significant, it was *chemically* insignificant, in that its coefficient was negative, implying that aromatic hydrogen atoms should lead to a lowering of the $\Delta_{sub}H$, contrary to chemical experience. The fact that it correlates so strongly with *C* probably causes the magnitude of its coefficient in the first place, and re-emphasises the need for careful pre-assessment of independent variable colinearity prior to MLRA.

It is therefore concluded that a simple 3-parameter MLRA model [entry 5; eqn. (5)] is the best, both statistically and

$$\Delta_{sub}H = (3.47 \pm 0.88) + (1.41 \pm 0.06)\,C + (4.55 \pm 0.30)\,HBD + (2.27 \pm 0.24)\,HBA \quad (5)$$

$$(n = 62, r^2 = 0.92, s = 1.6 \text{ kcal mol}^{-1})$$

chemically, that can be derived with the data used. A plot of calculated values of $\Delta_{sub}H$, based on the model following, and experimental $\Delta_{sub}H$ is shown above in Fig. 3.

A crude physical interpretation of this model can be attempted as follows. It is believed that the *C* parameter is simply reflecting the size of the molecule. It is apparent that larger molecules have higher SE, other factors being equal. Parameter *C* correlates highly with molecular weight, *MW*, but substitution of the latter parameter to reflect size leads to no improvement in the model (r^2 = 0.91), and leads to TNT becoming a serious outlier. The *HBD* parameter reflects potential for intermolecular H-bonding, which inevitably increases the $\Delta_{sub}H$. The magnitude of the regression coefficient suggests that each H-bond is worth about 4.5 kcal mol^{-1} to the $\Delta_{sub}H$. The *HBA* parameter must complement *HBD*, but we believe that its role is mainly to reflect the polar nature of functional groups in the molecules studied, since it is possible to

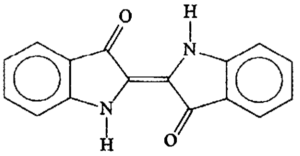
Table 4 Predictions of $\Delta_{sub}H$ of a test by neural networks and MLRA model (all values in kcal mol^{-1})

Molecule	Expt.[a] $\Delta_{sub}H$	MLRA Pred.	MLRA Dev.	Neural network[b] Pred.	Neural network[b] Dev.
Pentadecane	25.80	24.80	−1.00	27.6	1.8
Adamantane	14.30	17.50	3.20	16.0	1.7
trans-Stilbene	23.70	23.40	−0.35	22.5	−1.2
Tetraphenylmethane	36.00	39.40	3.40	41.4	5.4
2-Isopropyl-5-methylphenol	21.80	24.50	2.70	22.5	0.7
Furan-1-carboxylic acid	25.90	21.85	−4.10	25.3	−0.6
Azabicyclononane	13.85	21.60	7.80	17.8	4.0
Dimethyl glyoxime	23.20	22.70	−0.50	32.0	8.8
Pentaerythritol tetranitrate	36.30	37.90	1.60	21.7	−14.6
DMCTCN[c]	26.70	29.70	3.00	27.4	0.7
Mean unsigned deviation			2.80		3.95
Standard unsigned deviation			2.50		4.58

[a] Ref. 40. [b] NN prediction using all 7 parameters, 5 hidden nodes. [c] 4,5-Dimethyl-4-cyclohexene-1,1,2,2-tetracarbonitrile.

have hydrogen bond acceptors in a molecule, without having donors. Indeed, both dipole–dipole and dipole–induced dipole interactions will lead to an increased $\Delta_{sub}H$.

The standard deviation of the model is 1.6 kcal mol^{-1}, which is, in general, within experimental error. The main outliers are indigo **1**, octadecane, *N*,*N*-dimethyl-*p*-nitroaniline, and succinic anhydride. The first of these is widely overestimated by the model, whilst the other three compounds are underestimated.



1 Indigo

Indigo has two potential N–H hydrogen-bond donor sites, but each of these is strongly involved in intramolecular H-bonds. They are therefore less available for intermolecular H-bonding, and would be expected to contribute less to the sublimation enthalpy. In fact, the crystal structure of indigo [28] indicates only weak intermolecular H-bonds, with long H-bond distances of 2.1 Å. If the *HBD* parameter for indigo is given a value of zero to reflect the absence of intermolecular H-bonding, and the MLRA re-run, the model improves ($r^2 = 0.94$) and indigo falls nicely onto the regression line. Unfortunately, there are no other comparable molecules in the current dataset where potential for intramolecular H-bonding weakens intermolecular H-bonding.

The deviations of the other three molecules are harder to rationalise. Perhaps conformational effects in octadecane are significant, and these are omitted from the NN and MLRA models. There may also be multipolar effects in succinic anhydride and dimethyl-*p*-nitroaniline which go beyond the crude additive model of eqn. (5).

The MLRA model is far from optimal. There are chemical features of these molecules which contribute to the intermolecular forces in the solid state which are not included in the parameter values used. On the other hand, it is interesting and very surprising that a model which is so crude should reproduce the trend in the sublimation enthalpy as closely as shown in Fig. 3. Furthermore, the derivation by MLRA of the model is much simpler and faster than the derivation of the NN, although prediction of the $\Delta_{sub}H$ of a new molecule would be equally quick by either method. In this study using these parameters at least, the NN approach has nothing to add to the results of conventional MLRA.

The work of Gavezzotti [1,2] has shown that the lattice energy (and consequently the $\Delta_{sub}H$) can be predicted based upon linear regression studies on molecular crystals. The heat of sublimation can be estimated from the packing potential energy (PPE) which correlates with molecular descriptors such as molecular weight, Van der Waals' surface, volume and molecular outer surface. Others have shown that the $\Delta_{sub}H$ can be modelled by linear regression, but in a much more restricted series of molecules containing only hydrocarbons.[39]

Following the success of the three parameter MLRA study, the same three inputs were used in a NN containing 8 hidden neurons (ρ = 1.72, E_{max} = 0.001) in a comparative study. This yielded (in kcal mol^{-1}) a mean error of 3.1, maximum error = 27.4 (Ovalene), SD of 4.7 and r^2 = 0.60. These results are worse than Run 7 in Table 2. The maximum error here is significantly greater than that obtained using theoretical crystal packing methods for which the value is 3.5 kcal mol^{-1}.

Although neither neural network (mean error 2.5, max. error 10.1 kcal mol ^{1}) nor regression analysis (mean error 1.8, max. error 8.9 kcal mol^{-1}) reproduced the experimental results nearly as accurately as the theoretical crystal packing calculations (mean error 1.4, max. error 3.5 kcal mol^{-1}), both have yielded surprisingly good models. It is particularly interesting that the $\Delta_{sub}H$ could be reproduced from the simple three parameter MLRA, and this in itself could be a useful tool. We are currently investigating much larger data sets to determine the scope and limitations of MLRA models for predicting $\Delta_{sub}H$.

In the meantime, the two NN and the MLRA models have been tested by using them to predict the $\Delta_{sub}H$ of a series of molecules not used in the model development. These were selected to be roughly the same size of those of Table 1, and include molecules from different chemical classes to those used in the training set. The experimental and predicted $\Delta_{sub}H$ values and deviations are recorded in Table 4.

The predictions derived from the MLRA model, eqn. (5), are reasonably good, giving a mean unsigned error of 2.8 kcal mol^{-1}. However, they are inferior to the model fit for the data in Table 1. The major deviant is 3-azabicyclo[3.2.2]nonane by 7.8 kcal mol^{-1}. If the *HBA* and *HBD* terms are artifically set to zero for this molecule, implying no intermolecular dipole–dipole or H-bonding interactions between molecules in the crystal, the deviation reduces to 0.8 kcal mol^{-1}. However, this may be pure serendipity. Preliminary application of the simple MLRA model to larger molecules indicates that extrapolation beyond the scope of the data set in Table 1 is not warranted. Interpolation appears to be satisfactory.

The NN predictions are less satisfactory but in most cases still reasonable. The unsigned mean deviation is 3.9 kcal mol^{-1} (standard deviation 4.6 kcal mol^{-1}), with a maximum deviation of 14.6 kcal mol ^{1} for pentaerythritol tetranitrate. The molecule has 12 oxygens and 12 donors, both figures being significantly

Appendix 1 Input parameters for neural network and MLRA analyses

Column	Meaning
1	Code for molecular name
2	Number of carbons
3	Number of hydrogens
4	Number of nitrogens
5	Number of oxygens
6	Number of π-atoms
7	Number of hydrogen-bond donors
8	Number of hydrogen-bond acceptors
9	Experimental value (used in NN training)

PENTAN	5	12	0	0	0	0	0	9.9
HEXANE	6	14	0	0	0	0	0	12.6
OCTANE	8	18	0	0	0	0	0	15.9
OCTADE	18	38	0	0	0	0	0	37.8
BENZEN	6	6	0	0	6	0	0	12.5
BIPHEN	12	10	0	0	12	0	0	20.7
NAPHTH	10	8	0	0	10	0	0	17.3
ANTHRA	14	10	0	0	14	0	0	24.4
PHENAN	14	10	0	0	14	0	0	20.7
CHRYSE	18	12	0	0	18	0	0	28.4
TRIPHE	18	12	0	0	18	0	0	27.4
PERYLE	20	12	0	0	20	0	0	31.0
OVALEN	32	14	0	0	32	0	0	50.6
BENZOQ	6	4	0	2	8	0	2	14.9
ANTHRO	14	8	0	2	16	0	2	26.1
BENZOP	13	10	0	1	14	0	1	23.9
MEPHEN	7	8	0	1	7	1	1	17.7
SUCCIN	4	4	0	3	4	0	3	19.6
MALEIC	4	2	0	3	6	0	3	16.4
PHTHAL	8	4	0	3	10	0	3	21.1
NAPHAQ	10	6	0	2	12	0	2	21.7
PHENAN	14	8	0	2	16	0	2	25.8
CYCLOH	6	8	0	2	4	0	2	20.2
TETRAM	8	12	0	2	4	0	2	17.3
TRIOXA	3	6	0	3	0	0	3	13.8
TETRAO	4	8	0	4	0	0	4	19.0
PHENYL	13	10	0	2	14	0	2	23.7
NITROG	0	0	2	0	2	0	0	2.0
CYANOG	2	0	2	0	4	0	2	8.7
DICYAN	4	0	2	0	6	0	2	10.6
ETHYLE	6	0	4	0	10	0	4	20.6
PYRIMI	4	4	2	0	6	0	2	11.7
PYRAZI	4	4	2	0	6	0	2	14.5
TNTXXX	7	5	3	6	15	0	6	24.4
DIMETH	8	10	2	2	9	0	2	23.8
FORMIC	1	2	0	2	2	1	2	15.2
ACETIC	2	4	0	2	2	1	2	16.3
PROPAN	3	6	0	2	2	1	2	17.7
BUTYRI	4	8	0	2	2	1	2	19.2
VALERI	5	10	0	2	2	1	2	20.2
OXALIC	2	2	0	4	4	2	4	24.8
SUCCIN	4	6	0	4	4	2	4	29.3
GLUTAR	5	8	0	4	4	2	4	29.0
ADIPIC	6	10	0	4	4	2	4	32.1
SUBERI	8	14	0	4	4	2	4	35.4
SEBACI	10	18	0	4	4	2	4	39.6
BENZOI	7	6	0	2	8	1	2	23.0
OXAMID	2	4	2	2	4	4	2	28.2
MALONA	3	6	2	2	4	4	2	28.8
SUCCIN	4	8	2	2	4	4	2	32.3
UREAXX	1	4	2	1	2	4	1	22.2
FORMAM	1	3	1	1	2	2	1	17.5
DIKETO	4	6	2	2	4	2	2	26.0
TCNQXX	12	4	4	0	16	0	4	26.5
INDIGO	16	10	2	2	20	2	2	31.8
ACRIDI	13	9	1	0	14	0	1	22.0
AZOBEN	12	10	2	0	14	0	2	22.0
NITRIL	10	11	1	0	8	0	1	18.6
N-METH	13	11	1	0	13	0	0	22.8
P-DICY	8	4	2	0	10	0	2	21.2
TRIAZI	3	3	3	0	6	0	3	13.4
STERIC	18	36	0	2	2	1	2	39.8

greater than the values in the training set. In fact, the PSDD software gives a warning that this molecule is outside the limits defined by the training set.

Overall, we conclude that both the neural networks and the MLRA model are able to give reasonable estimates for unknown sublimation enthalpies, especially if the test molecules are restricted to interpolations. It should also be noted that there is at times a large degree of uncertainty in the experimental data. For example, two values quoted for biphenylene in Pedley's collection of physical data[40] differ by over 10 kcal mol^{-1}.

It should also be stressed that the theoretical and predicted methods are not in direct competition. Whilst the former requires prior knowledge of the crystal structure, it yields detailed analysis of the important forces involved in the solid state structure formed. This is vital information relating the molecular structure to solid state properties. The latter methods yield no such information, but are quick and simple, requiring no experimental data in the input, and therefore have potential in novel molecule design.

This study has also shown that in this case, neural network predictions are inferior to those from the regression analysis. This is probably because of the unexpectedly highly linear dependency of the sublimation enthalpy upon the input data, and the NN would be anticipated to perform better in other, non-linear cases. Furthermore, it is possible that the neural network would benefit further from a different choice of molecules in the training set. Pre-selection of molecules that have a more even distribution of both experimental $\Delta_{sub}H$ and of input data values would be likely to enhance the network performance.

References

1 A. Gavezzotti, *J. Phys. Chem.*, 1991, **95**, 8948.

2 A. Gavezzotti and G. Filippini, *Acta Cryst., Sect. B*, 1992, **48**, 537.

3 W. S. McCulloch and W. Pitts, *Bull. Math. Biophysics*, 1943, **5**, 115.

4 M. L. Minsky and S. S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1943.

5 J. J. Hopfield, *Proc. Natl. Acad. Sci. USA*, 1982, **79**, 2554.

6 S.-S. So and W. G. Richards, *J. Med. Chem.*, 1992, **35**, 3201.

7 H. Oinuma, K. Miyako, M. Yamanaka, K.-I. Nomoto, H. Katoh, K. Sawada, M. Shino and S. Hamano, *J. Med. Chem.*, 1990, **33**, 905.

8 T. Aoyama, Y. Suzuki and H. Ichikawa, *J. Med. Chem.*, 1990, **33**, 2583.

9 T. Aoyama and H. Ichikawa, *Chem. Pharm. Bull.*, 1991, **39**, 372.

10 M. E. Lacy, *Tetrahedron Computer Methodology*, 1990, **3**, 119.

11 T. Aoyama, Y. Suzuki and H. Ichikawa, *Chem. Pharm. Bull.*, 1989, **37**, 2558.

12 N. Bodor, A. Harget and M.-J. Huang, *J. Am. Chem. Soc.*, 1991, **113**, 9480.

13 L. M. Egolf and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 616.

14 J. Zupan and J. Gasteiger, *Anal. Chim. Acta*, 1991, **248**, 1–30; *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 503.

15 J. Zupan and J. Gasteiger, *Neural Networks for Chemists*, VCH, Weinheim, Germany, 1993.

16 QCPE Program 615, Quantum Chemical Program Exchange, Creative Arts Building 181, Indiana University, Bloomington, Indiana 47405.

17 T. Aoyama and H. Ichikawa, *Chem. Pharm. Bull.*, 1991, **39**, 358.

18 D. T. Manallack and D. J. Livingstone, *Med. Chem. Res.*, 1992, **2**, 181.

19 J. D. Cox and G. Pilcher, *Thermochemistry of Organic and Organometallic Compounds*, Academic Press, New York, 1970.

20 D. E. Williams, *J. Chem. Phys.*, 1965, **43**, 4424.

21 A. I. Kitaigordsky, *Molecular Crystals and Molecules*, Academic Press, New York, 1973.

22 G. Filippini and A. Gavezzotti, *Acta Crystallogr., Sect. B*, 1993, **49**, 868.

23 A. T. Hagler, S. Lifson and P. Dauber, *J. Am. Chem. Soc.*, 1979, **101**, 5122.

24 R. Docherty, G. Clydesdale, K. J. Roberts and P. Bennema, *J. Phys. D. Appl. Phys.*, 1991, **24**, 89.

25 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.

26 QCPE Program 455, Quantum Chemical Program Exchange,

Creactive Arts Building 181, Indiana University, Bloomington, Indiana 47405.
27 SAS Institute Inc., *SAS/STAT User's Guide*, Version 6, 4th Ed., Vols. 1 and 2, Cary, NC, SAS Institute Inc., 1989.
28 P. Susse, M. Steins and V. Kupcik, *Z. Kristallogr.*, 1988, **184**, 269.
29 D. E. Williams and S. R. Cox, *Acta Cryst., Sect. B*, 1989, **40**, 404.
30 R. Docherty, in preparation.
31 J. Royer, C. Decoret, B. Tinland, M. Perrin and R. Perrin, *J. Phys. Chem.*, 1989, **93**, 3393.
32 H. A. J. Govers, *Acta Cryst., Sect. A*, 1975, **31**, 380.
33 N. Higgins, personal communication.
34 S. Lifson, A. T. Hagler and P. Dauber, *J. Am. Chem. Soc.*, 1979, **101**, 5122.
35 F. A. Momany, L. M. Carruthers, R. F. McGuire and H. A. Scheraga, *J. Phys. Chem.*, 1974, **78**, 1595.
36 H. A. J. Govers, *Acta Crystallogr., Sect. A*, 1978, **34**, 960.
37 R. Docherty and A. P. Chorlton, in preparation.
38 Y. Michopoulos and C. Adam, presented at British Crystallographic Assoc. Spring Meeting, Oxford, 1989.
39 J. S. Chickos and R. Annunziata, L. H. Ladon, A. S. Hyman and J. F. Liebman, *J. Org. Chem.*, 1986, **51**, 4311; K. Nass, D. Lenoir and A. Keltrup, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 1735.
40 *Thermochemical Data of Organic Compounds*, J. B. Pedley, R. D. Naylor and S. P. Kirby, Chapman and Hall, 1986.

Paper 5/02045J
Received 31*st May* 1995
Accepted 12*th June* 1995