Studies of Imidazole and Pyrazole Protonation using Electrostatically Trained Neural Networks

Howard B. Broughton,[a] Stuart M. Green[*,†,b] and Henry S. Rzepa[b]
[a] *Neuroscience Research Centre, Merck Sharp and Dohme, Terlings Park, Harlow, Essex, UK CM20 2QR*
[b] *Department of Chemistry, Imperial College, London, UK SW7 2AY*

A backpropagation neural network was trained with the molecular electrostatic potentials (MEPs) of a series of substituted imidazoles to predict their corresponding pK_a. Using MEPs determined with a variety of semiempirical and *ab initio* methods, the predictive power of the trained network was found to be sensitive to the quality of the basis set. The network was also trained to predict the proton affinity (E_{pa}) and pK_a, both individually and combined, for a series of pyrazole MNDO MEPs.

Neural networks are now finding regular application to a variety of chemical problems: spectroscopy,[1,2] prediction of protein structure[3–6] and water binding sites,[7] synthetic analysis,[8–10] and solvation studies.[11] This rapidly growing field has already spawned two reviews, by Burns and Whitesides,[12] and Gasteiger and Zupan.[13] The realm of quantitative structure–activity relationships (QSAR) has seen a great deal application of such neural paradigms[14–18] as they make an obvious alternative to the multivariate statistics traditionally used in Hansch analysis.

In preliminary studies we successfully trained a neural network with a series of imidazole MEPs (AM1 derived) in an attempt to predict histidine pK_a perturbation in triose phosphate isomerase.[19] Such a regime uses the neural network to derive a three-dimensional quantitative structure–activity relationship (3D-QSAR); inputs to the network are sourced directly from grid points of the MEP (freeing us from the restrictions imposed by substituent parameters). The network is then trained to map these points to the molecular property (pK_a). Statistical techniques such as comparative molecular field analysis (CoMFA[20]) have also been developed to perform 3D-QSAR, but these methods presuppose some form of non-parametric[21] or more restrictive parametric[20] relationship between individual field points and the observable quantity. Neural networks offer an alternative where no assumption is made about the relationship between property and field; instead this is derived during the process of training the network.

To test the 3D-QSAR abilities of the neural network further, this paper details the results of using different basis sets or Hamiltonians for deriving the MEP in the imidazole pK_a case. The result of training pyrazole pK_a and/or proton affinity (E_{pa}) against MNDO MEPs are also discussed.

Computational Methods

All structures were optimised using either MOPAC[22] or Gaussian[23] for the respective semiempirical method (AM1, PM3 and MNDO, using EF and PRECISE) or basis set (STO-3G, 3-21G and 6-31G*). Each optimised molecule was then aligned such that atom 1 (Figs. 1 and 2) was placed at the origin, with atom 2 extending out along the positive X-axis and atom 3 in the XY-plane. Atoms 1, 2 and 3 are consistently defined for all. The MEP was then evaluated, with each respective method, on a 1.5 Å grid excluding points that were within the van der Waals radii and those which were more than 6 Å from the closest atom.‡

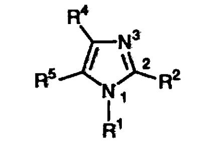

Fig. 1 Imidazole atom numbering

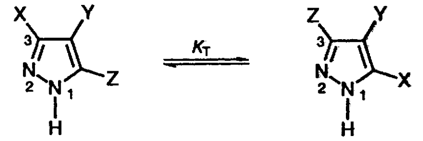

Fig. 2 Pyrazole tautomerism and atom numbering

For suitable input to the neural network only grid points common to all molecules in either imidazole or pyrazole set were used. Each MEP point was then scaled such that the value presented to the network lay in the range 0.1 to 0.9 (in order to prevent network saturation): eqn. (1) where x_{pi} is the scaled

$$x_{pi} = 0.8\left(\frac{C_{pi} - V_i^{\min}}{V_i^{\max} - V_i^{\min}}\right) + 0.1 \qquad (1)$$

input for molecule p at grid location i, V_{pi} is the MEP value, $V_i^{\max}$ and $V_i^{\min}$ are the respective maxima and minima for all molecules at the MEP point i. The pK_a values to which the inputs would be trained against were also similarly normalised.

The neural network used in this study was of the backpropagation type[24,25] with one hidden layer of 10 neurons‡ and a final layer with one neuron, for single property prediction, or two, when combined properties are used. All neurons had a sigmoidal activation function and a single bias term (fixed input of 1.0). Initially, all neuronal weights were randomly set (uniform distribution) to values between ± 0.3. The scaled MEPs were trained to map to their respective properties in 'batch mode', *i.e.*, all patterns were presented and the network updated with the sum of errors. Network updates were performed using a fast adaptive learning algorithm (superSAB[26]). This method was found to be much

† Current address: Center for Molecular Design, Washington University, Box 1099, St. Louis, MO 63130-4899, USA.

‡ Alternative grid definitions and numbers of neurons in the hidden layer have been tested but this prescription was found to be best in a trade-off between performance and computation time.[33]

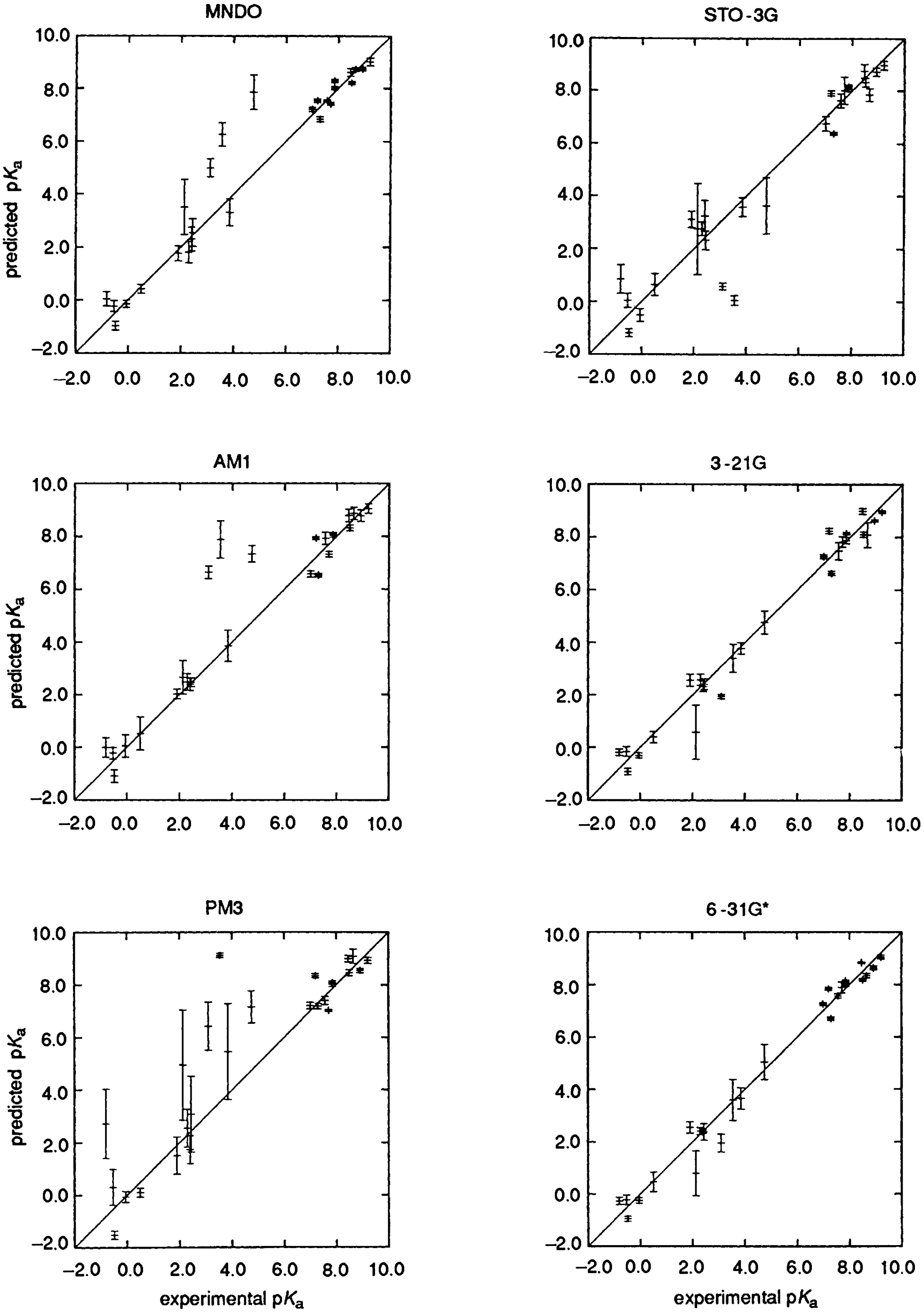

Fig. 3 Leave-one-out cross-validation results for mapping of MEP to pK_a for 26 imidazoles. The predicted pK_a is the mean of five independent trainings with error bars as the standard deviation.

more stable and to require far fewer iterations to achieve the required degree of training than the traditional generalised delta rule.[24]

Owing to the initial random weights the network cannot be guaranteed to produce the same results for every random seed, hence many differently seeded trainings were used for each cross-validation step with all predictions included in the subsequent analysis.

Results and Discussion

(*a*) *Imidazole* pK_a *QSAR with MEPs*.—A collection of small, conformationally invariant imidazoles of varied substitution were selected (Table 1), these covered a range of 10 pK_a units. The geometry of each of these 26 imidazoles was optimised and MEPs calculated for all methods described (Table 2). A leave-one-out cross-validation analysis was then performed for all 26 imidazoles training five differently seeded networks to map MEP to pK_a for 25 of the imidazoles, then predicting the pK_a of the excluded imidazole (Fig. 3).

Two molecules (5-NO_2, 1-Me and 5-F, 1-Me) are expressed as outliers independent of the MEP derivation methods, this is due to the lack of 5-substituted moieties and the network is relatively uneducated about these types of species. With the semiempirical methods and STO-3G the common outliers are all molecules containing chlorine suggesting a poor electrostatic description of chlorine by these methods due to an incorrect definition of the atomic core.[27]

Overall, 6-31G* > 3-21G ≫ MNDO > STO-3G > AM1 ≫ PM3. The best performing networks were those trained with the high level *ab initio* derived MEPs, although there is little improvement (but considerable effort) in going from 3-21G to 6-31G*. The poor performance of PM3 is to be expected due to the poor parametrisation of nitrogen.[28]

Table 1 Substituted imidazole pK_a[34]

Imidazole	pK_a	Imidazole	pK_a
2-NH_2, 4,5-Me_2	9.21	5-F, 1-Me	3.85
2,4,5-Me_3	8.92	2-Cl	3.50
2-NH_2, 1-Me	8.65	4-Cl, 1-Me	3.10
2,4-Me_2	8.50	4-F	2.44
2-NH_2	8.46	2-F	2.40
1,2-Me_2	7.85	2-F, 1-Me	2.30
2-Me	7.85	5-NO_2, 1-Me	2.13
1,5-Me_2	7.70	4-F, 1-Me	1.90
4-Me	7.56	4-NO_2, 2-Me	0.50
1-Me	7.30	4-NO_2	−0.05
1,4-Me_2	7.20	2-NO_2, 1-Me	−0.48
Unsubstituted	7.00	4-NO_2, 1-Me	−0.53
5-Cl, 1-Me	4.75	2-NO_2	−0.81

Table 2 Cross-validation statistics for mapping of MEP to pK_a for 26 imidazoles; each analysis is based on all five different trainings[a]

MEP	r^2	r_s	*rv*
MNDO	0.91	0.96	1.06
AM1	0.89	0.94	1.69
PM3	0.70	0.86	3.43
STO-3G	0.89	0.90	1.32
3-21G	0.97	0.95	0.37
6-31G*	0.97	0.97	0.30

[a] r^2 and r_s are the correlation coefficients of Pearson and Spearman, respectively; *rv* is the residual variance.

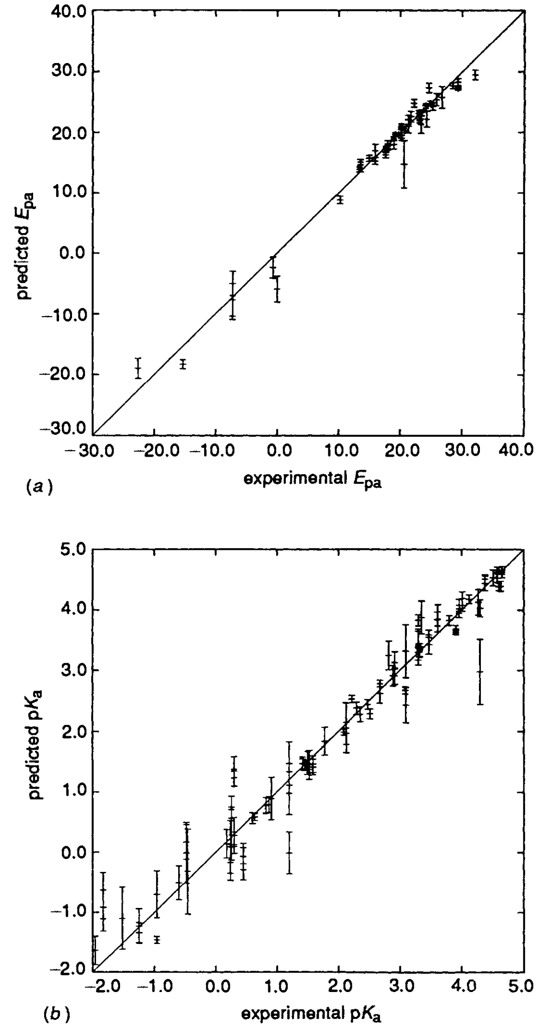

Fig. 4 Leave-one-out cross-validation results for mapping of MEP to (*a*) E_{pa} for 49 pyrazoles and (*b*) pK_a for 91 pyrazoles. The predicted property is the mean of five independent trainings with error bars as the standard deviation.

Table 3 Experimental[31] pyrazole E_{pa}

Pyrazole	E_{pa}/kcal mol^{-1}	Pyrazole	E_{pa}/kcal mol^{-1}
3,5-$(NO_2)_2$	−22.6	1-Me, 5-Ph	20.6
1-Me, 3,5-$(NO_2)_2$	−15.3	1,5-Me_2	20.7
3-NO_2	−7.2	3-Et, 5-Ph	21.3
5-NO_2	−7.2	3-Ph, 5-Et	21.3
1-Me, 3-NO_2	−0.7	1-Me, 3-NH_2	21.7
1-Me, 5-NO_2	0.0	1-Me, 5-Bu^t	22.2
H	10.2	1-Me, 3-Bu^t	23.0
4-Me	13.3	3-Me, 5-Bu^t	23.1
3-Me	13.5	3-Bu^t, 5-Me	23.1
5-Me	13.5	3,5-Ph_2	23.3
1-Me	14.9	3,4,5-Me_3	23.4
3-Ph	15.9	1,3,4-Me_3	24.1
5-Ph	15.9	1,4,5-Me_3	24.1
3-NH_2	17.6	1-Me, 5-NH_2	24.2
5-NH_2	17.6	3,5-Et_2, 4-Me	24.6
3-Bu^t	17.9	3,5-$(Bu^t)_2$	24.8
5-Bu^t	17.9	1,3,5-Me_3	24.9
3,4-Me_2	18.9	1,5-Me_2, 3-Ph	25.2
4,5-Me_2	18.9	1,3-Me_2, 5-Ph	25.8
1,4-Me_2	19.2	1-Me, 3,5-Ph_2	26.7
1,3-Me_2	19.7	3,5-$(Bu^t)_2$, 4-Me	28.5
3-Me, 5-Ph	20.1	1,3,4,5-Me_4	29.3
3-Ph, 5-Me	20.1	1-Me, 3,5-$(Bu^t)_2$	29.3
1-Me, 3-Ph	20.2	1,4-Me_2, 3,5-$(Bu^t)_2$	32.1
3,5-Me_2	20.2		

Table 4 Experimental[32] pyrazole pK_a

Pyrazole	pK_a	Pyrazole	pK_a	Pyrazole	pK_a
4-NO_2	−1.96	3-Me, 4-Cl	1.42	5-Et	3.30
3,4-Br_2	−1.83	4-Cl, 5-Me	1.42	5-Bu^t	3.30
4,5-Br_2	−1.83	3-Me, 4-Br	1.46	3-Me	3.32
1,5-Me_2, 3,4-Br_2	−1.52	4-Br, 5-Me	1.46	5-Me	3.32
3-Me, 4-NO_2	−1.25	3-Et, 4-Cl	1.50	3,4-[-camphyl-]	3.35
4-NO_2, 5-Me	−1.25	4-Cl, 5-Et	1.50	3,4-Me_2, 5-Ph	3.47
3,4-Br_2, 5-Me	−0.96	3-Et, 4-Br	1.53	3-Ph, 4,5-Me_2	3.47
3-Me, 4,5-Br_2	−0.96	4-Br, 5-Et	1.53	3,4-[-$(CH_2)_3$-]	3.61
1,3-Me_2, 4,5-Br_2	−0.60	3-Me, 4-I	1.59	4,5-[-$(CH_2)_3$-]	3.61
3-Cl	−0.48	4-I, 5-Me	1.59	1,3,5-Me_3	3.80
5-Cl	−0.48	1,3,5-Me_3, 4-Br	1.78	3,4-Me_2	3.91
3,5-Me_2, 4-NO_2	−0.46	1-Me	2.09	4,5-Me_2	3.91
1-Me, 4-Br	0.18	3-Ph	2.13	3,4-[$(CH_2)_5$-]	3.96
3-Br, 4-Me	0.24	5-Ph	2.13	4,5-[-$(CH_2)_5$-]	3.96
4-Me, 5-Br	0.24	3,5-Me_2, 4-Cl	2.22	3,4-[-$(CH_2)_4$-]	4.01
3-Ph, 4-Cl	0.26	3,5-Me_2, 4-Br	2.30	4,5-[-$(CH_2)_4$-]	4.01
4-Cl, 5-Ph	0.26	3,5-Me_2, 4-I	2.36	3,5-Me_2	4.12
3-Cl, 5-Me	0.30	1,4-Me_2	2.48	1,3,4,5-Me_4	4.27
3-Me, 5-Cl	0.30	H	2.52	3-C_3H_5, 4-Me, 5-Et	4.30
3-Ph, 4-Br	0.30	3-Ph, 4-Me	2.68	3-Et, 4-Me, 5-C_3H_5	4.30
4-Br, 5-Ph	0.30	4-Me, 5-Ph	2.68	3,4-[-$(CH_2)_3$-], 5-Me	4.38
3-Br, 5-Me	0.45	1,3-Me_2	2.82	3-Me, 4,5-[-$(CH_2)_3$-]	4.38
3-Me, 5-Br	0.45	1,3-Me_2	2.89	3,5-Et_2, 4-Me	4.51
4-Cl	0.60	3-Me, 5-Ph	2.92	3,4-[-$(CH_2)_5$-], 5-Me	4.57
4-Br	0.64	3-Ph, 5-Me	2.92	3-Me, 4,5([-$(CH_2)_5$-]	4.57
4-I	0.82	4-Me	3.09	3,4-Me_2, 5-Et	4.60
1,3-Me_2, 4-Br	0.87	3-C_3H_5	3.10	3-Et, 4,5-Me_2	4.60
1,5-Me_2, 4-Br	0.91	5-C_3H_5	3.10	3,4,5-Me_3	4.63
1,3-Me_2, 5-Br	1.20	3-Et	3.30	3,4-[-$(CH_2)_4$-], 5-Me	4.65
3-Me, 4-Br, 5-Ph	1.20	3-Bu^t	3.30	3-Me-4,5-[-$(CH_2)_4$-]	4.65
3-Ph, 4-Br, 5-Me	1.20				

Table 5 Cross-validation statistics for pyrazole MEP training; all five different leave-one-out cross-validations are included in each analysis[a]

	E_{pa}/kcal mol^{-1}			pK_a		
Pyrazoles	r^2	r_s	rv	r^2	r_s	rv
49	0.97	0.96	4.17	—	—	—
91	—	—	—	0.95	0.98	0.17
21	0.75	0.87	5.56	0.86	0.92	0.09
21	0.72	0.87	6.80	—	—	—
21	—	—	—	0.87	0.93	0.08

[a] r^2 and r_s are the correlation coefficients of Pearson and Spearman, respectively; *rv* is the residual variance.

In previous comparisons of MEP derivation methods,[29,30] workers have assumed those determined for the highest feasible level of theory were the most 'realistic' and used this level as a standard. From this work with imidazoles one can suggest an order of how realistic the techniques are in producing MEPs by correlating them with real experimental data, rather than *via* comparison with a presumed standard.

(*b*) *Pyrazole* pK_a *and* E_{pa} *QSAR with MEPs*.—Neural networks were also trained to map the MEP to the pK_a and E_{pa} of a series of pyrazoles. Recent experimental (Fourier transform ion cyclotron resonance spectroscopy) and theoretical (6-31G geometries and energies) work has demonstrated that for the tautomerism of NH pyrazoles (Fig. 2) the equilibrium constant[31] $K_T = 1$, *i.e.*, 3- and 5-substituted moieties are inseparable, hence their pK_a and E_{pa} values are identical and both tautomers should be considered equally.

Proton affinity data[31] for 39 pyrazoles were used; 10 of these had equivalent tautomers, so a total of 49 pyrazoles were considered (Table 3). The geometries of these were optimised and the MEP evaluated as before for the imidazoles but only at the MNDO level. A complete leave-one-out cross-validation study for each pyrazole was carried out for five differently seeded networks. The results of this study are shown in Fig. 4(*a*). A much larger dataset of 58 pyrazole pK_a values was available.[32] Of these 33 had equivalent tautomers making an overall set of 91 pyrazoles (Table 4). Exactly the same approach was used as in the proton affinity study, results in Fig. 4(*b*).

Training is not restricted to the mapping of only one property at a time, any number of outputs can be included and the network trained to predict many things at once. Of the pyrazoles considered 21 are coincident between the pK_a and proton affinity sets. Thus, these were used in a cross-validation study but with training based on two outputs. For comparison purposes E_{pa} and pK_a were also trained independently for the 21 pyrazoles. The results of this combined training (Table 5) are rather poor, the cross-validations are nowhere near as good for those observed previously. This is not a failure of the network to cope with two outputs, but is just due to a lack of training data. Independent trainings on just the 21 pyrazoles against proton affinity or pK_a yield almost identical results, hence the training of multiple properties is not detrimental to predictive performance.

Conclusions

We have demonstrated the utility of neural networks in the correlation of complex quantum mechanically defined features (MEPs) with experimentally determined parameter(s) for a given molecular series. The neural network, far from being the blind pattern classification method, shows the ability to generalise as it remains sensitive to the relative level of theory. Further studies based on trainings using alternative grid-based molecular descriptors and other properties (*e.g.*, biological activities) will hopefully expand this technique fully within the realm of 3D-QSAR.

Acknowledgements

We thank Merck Sharp and Dohme Research Laboratories and the SERC for financial support (to S. M. G.).

References

1 B. Curry and D. E. Rumelhart, *Tetrahedron Comput. Methodol.*, 1990, **3**, 213.

2 R. J. Fessenden and L. Gyorgyi, *J. Chem. Soc., Perkin Trans. 2*, 1991, 1755.

3 N. Qian and T. J. Sejnowski, *J. Mol. Biol.*, 1988, **202**, 865.

4 M. J. McGregor, T. P. Flores and M. J. E. Sternberg, *Protein Eng.*, 1989, **2**, 521.

5 D. G. Kneller, F. E. Cohen and R. Langridge, *J. Mol. Biol.*, 1990, **214**, 171.

6 B. A. Metfessel, P. N. Saurugger, D. P. Connelly and S. S. Rich, *Protein Sci.*, 1993, **2**, 1171.

7 R. C. Wade, H. Bohr and P. G. Wolynes, *J. Am. Chem. Soc.*, 1992, **114**, 8284.

8 V. Simon, J. Gasteiger and J. Zupan, *J. Am. Chem. Soc.*, 1993, **115**, 9148.

9 H. H. Luce and R. Govind, *Tetrahedron Comput. Methodol.*, 1990, **3**, 143.

10 D. W. Elrod, G. M. Maggiora and R. G. Trenary, *Tetrahedron Comput. Methodol.*, 1990, **3**, 163.

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

12 J. A. Burns and G. M. Whitesides, *Chem. Rev.*, 1993, **93**, 2583.

13 J. Gasteiger and J. Zupan, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 503.

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

15 T. A. Andrea and H. Kalayeh, *J. Med. Chem.*, 1991, **34**, 2824.

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

17 J. H. Wikel and E. R. Dow, *Bioorg. Med. Chem. Lett.*, 1993, **3**, 645.

18 Ajay, *J. Med. Chem.*, 1993, **36**, 3565.

19 H. B. Broughton, S. M. Green and H. S. Rzepa, *J. Chem. Soc., Chem. Commun.*, 1992, 1178.

20 R. D. Cramer III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.

21 H. B. Broughton, S. M. Green and H. S. Rzepa, *J. Chem. Soc., Chem. Commun.*, 1992, 37.

22 J. J. P. Stewart, MOPAC version 6.0. *QCPE*, Program No. 455, local modifications for grid-based MEP calculation by S. M. Green.

23 M. J. Frisch, G. W. Trucks, M. Head-Gordon, P. M. W. Gill, M. W. Wong, J. B. Foresman, B. G. Johnson, H. B. Schlegle, M. A. Robb, E. S. Replogle, R. Gomperts, J. L. Andres, K. Raghavachari, J. S. Binkley, C. Gonzalez, R. L. Martin, D. J. Fox, D. J. Defrees, J. Baker, J. J. P. Stewart and J. A. Pople. GAUSSIAN 92, Revision B, Gaussian, Inc., Pittsburgh PA, 1992.

24 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature* (*London*), 1986, 323, 533.

25 J. McClelland and D. Rumelhart, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, MA, 1986.

26 T. Tollernaere, *Neural Networks*, 1990, **3**, 561.

27 L. Bonati, U. Cosentino, E. Fraschini, G. Moro and D. Pitea, *J. Comput. Chem.*, 1992, **13**, 842.

28 I. Juranic, H. S. Rzepa and M. Yi, *J. Chem. Soc., Perkin Trans. 2*, 1990, 877.

29 K. M. Merz Jr., *J. Comput. Chem.*, 1992, **13**, 749.

30 I. Alkorta, H. O. Villar and G. A. Arteca, *J. Comput. Chem.*, 1993, **14**, 530.

31 J. L. M. Abboud, P. Cabildo, T. Cañada, J. Catálan, R. M. Claramunt, J. L. G. de Paz, J. Elguero, H. Homan, R. Notario, C. Toiron and G. I. Yranzo, *J. Org. Chem.*, 1992, **57**, 3938.

32 J. Elguero, E. Gonzalez and R. Jacquier, *Bull. Chim. Soc. Fr.*, 1968, **12**, 5009.

33 S. M. Green, *Development of Neural Networks of Structure–Activity Relationships*, Ph.D. thesis, University of London, 1993.

34 M. R. Grimmett, in *Comprehensive Heterocyclic Chemistry*, Pergamon, 1984, vol. 5, p. 384.

Paper 4/05093B
Received 19*th August* 1994
Accepted 20*th October* 1994