A Comparison of Linear and Non-linear QSAR Modelling to Predict Toxicity of Phenols to *Tetrahymena pyriformis*

S. J. Enoch[1], T. W. Schultz[2], J. C. Madden[1] and M. T. D. Cronin[1]

[1] School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, England
[2] The University of Tennessee, College of Veterinary Medicine, 2407 River Drive, Knoxville, TN 379961-4543, USA

Contact author: s.j.enoch@ljmu.ac.uk

INTRODUCTION

•The REACH legislation has led to considerable interest in QSAR modelling for the prediction of environmental toxicity and fate.

•Toxicity datasets frequently consist of groups of compounds acting via a number of differing mechanisms of action.

•Complex modelling techniques such as neural network analysis have been suggested to model such data better than simple regression methods.

•However, for regulators, simpler, more transparent, models e.g. those based on regression analysis, are of benefit due to their transparency.

AIMS

•The aim of this study was to investigate if increasing the complexity of statistical technique improved the modelling of the individual toxic mechanisms in a dataset of phenol toxicity to *Tetrahymena pyriformis*.[2]

METHODS

Data sets

•250 phenol compounds with five mechanisms of action were sourced from the literature.

•Data were split into 200 training and 50 validation compounds.

•The five mechanisms of action were identified in a previous study as (training:validation):

- •weak acid respiratory uncouplers (15:4)
- •soft electrophiles (22:5)
- •pre-electrophiles (22:5)
- •pro-redox cyclers (3:1)
- •polar narcotics (138:35)
- •soft electrophiles (22:5)

•Training and validation sets were constructed so that each mechanism of action was equally represented

Statistical Analysis

•Two QSAR models utilised the same 200 training and 50 validation compounds in their construction.

•Descriptors listed below used in the construction of both models

•Seven parameter stepwise regression model used as previously reported

•Multi-layer perceptron neural network (MLP 11:11-9-1:1) model developed

•Model quality was assessed in terms of statistical fit and predictivity using:

- •the square root of the mean error in the training or validation data (RMSE).
- •leave-one-out cross-validated coefficient of determination (r^2_{cv})
- •coefficient of determination for the validation data (q^2_{ext}).

Descriptors

•The descriptors used in the two models were:

- •LogD: logarithm of the octanol-water dissociation constant
- •LUMO: energy of the lowest unoccupied molecular orbital
- •MW: molecular weight
- •P_{NEG}: negatively charged surface area
- •SsOH: electrotopological state index for the hydroxyl group
- •ABSQon: sum of the absolute charges on nitrogen and oxygen
- •MaxHp: largest positive charge on a hydrogen atom

REFERENCES

1. J. Devillers, *SAR and QSAR in Environmental Research,* **15**, 237 (2004)
2. M. T. D. Cronin, A. O. Aptula, J. C. Duffy, T. I. Netzeva, P. H. Rowe, I. V. Valkova, and T. W. Schultz, *Chemosphere,* **49**, 1201 (2002).
3. T. W. Schultz, Sinks, G. D., and Cronin, M. T. D., *in* "Quantitative Structure-Activity Relationships in Environmental Sciences VII" (Chen, F. Schuurmann, G., ed.), p. 329. SETAC, Florida 1997.

RESULTS

•Results of the statistical analysis for the regression and neural network models are shown below

Training							
	Mechanism	polar-narcosis	pro-electrophiles	pro-redox	resp-uncouplers	soft-electrophiles	global
	No. Cmpds.	138	22	3	15	22	200
SR	RMSE	0.36	0.92	1.36	0.48	0.42	0.48
NN		0.10	0.55	0.16	0.12	0.19	0.16
SR	r^2_{cv}	0.82	-0.22	-0.66	0.95	0.50	0.66
NN		0.83	0.26	0.95	0.90	0.85	0.77

Validation							
	No. Cmpds	35	5	1	4	5	50
SR	RMSE	0.44	1.35	-	0.32	0.42	0.55
NN		0.23	1.22	-	0.08	0.11	0.28
SR	q^2_{ext}	0.74	-2.50	-	0.93	0.45	0.59
NN		0.69	-1.34	-	0.95	0.66	0.60

Table 1: Statistical results for stepwise regression (SR) and neural network analysis (NN) (- indicates not calculated)

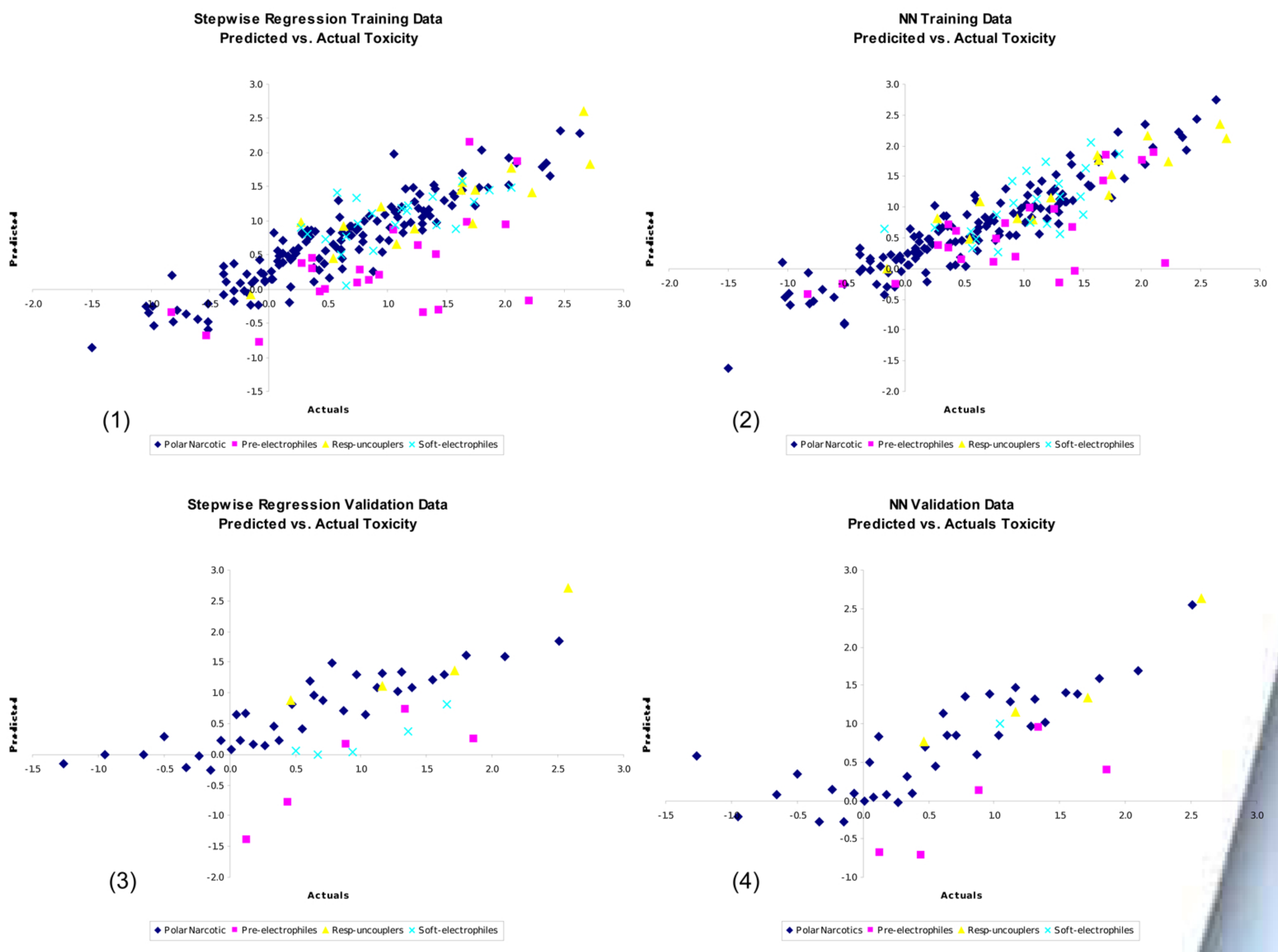

Figure 1: Plots of predicted versus actual toxicity for training and validation data for SR and NN models. Four of the mechanisms are as illustrated (pro-redox omitted due to the lack of test compounds).

DISCUSSION

•Fit (as measured by RMSE and r cv for the training data in Table 1) generally increases upon increasing model complexity for all mechanisms.

•Predictivity (as measured by RMSE and q ext for the validation data in Table 1) does not increase with model complexity.

•Polar narcotic and respiratory uncouplers are equally well predicted by both modelling methods.

•Electrophilic mechanisms are predicted poorly no matter what the modelling method.

•The poor predictivity of electrophiles is probably due to the lack of adequate molecular descriptors capable of describing these chemical reactions.

CONCLUSIONS

•This study has demonstrated that for a typical high quality toxicological database containing multiple mechanisms of toxic action there appears to be no benefit in increasing model complexity.

•Instead, it appears that modelling within mechanism based domains is more successful no matter what the modelling technique.

•The results also suggest that the poor modelling of the electrophilic compounds is likely to be due to inadequate descriptors capable of describing such processes.

•This suggests that the most useful modelling approaches in terms of REACH and for regulators are likely to be simple mechanism based regression models.

AKNOWLEDGMENT

The funding of the European Union 6th Framework CAESAR Specific Targeted Project (SSPI-022674-CAESAR) is gratefully acknowledged.