Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 23 (2005) 503–523

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

# Assessing the reliability of a QSAR model's predictions

Linnan He, Peter C. Jurs *

*Department of Chemistry, The Pennsylvania State University, 104 Chemistry Building, University Park, PA 16802, USA*

Received 11 November 2004; received in revised form 1 March 2005; accepted 1 March 2005
Available online 17 May 2005

**Abstract**

Quantitative structure activity relationships (QSAR) are one of the well-developed areas in computational chemistry. In this field, many successful predictive models have been developed for various property, activity or toxicity predictions. However, the predictive power of models for new query compounds is often not well characterized. The breadth of applicability of models is often not characterized. In other words, with a given QSAR model and a specific query compound to be predicted, can the model be used reliably for the desired prediction? In this study, we assessed the reliability of QSAR models' prediction on query compounds. Our approach, employing hierarchical clustering, was developed and tested using a test dataset containing 322 organic compounds with fathead minnow acute aquatic toxicity as the activity of interest. The hypothesis of the approach was that if a query compound is more similar to the compounds used to generate the QSAR model, it should be predicted more accurately. Thus, the core of the approach is to determine the relationship between the similarity of query compounds to the training set compounds of the QSAR model and the prediction accuracy given by that model. This relationship determination was achieved by comparing the results given by the two major components of the approach: objects clustering and activity prediction. With the resultant information from the two steps, a direct relationship was shown.

*Keywords:* QSAR model; Hierarchical clustering; Fathead minnow acute aquatic toxicity

## 1. Introduction

Similar molecules with just a slight variation in their structures can have quite different biological activities. This kind of relationship between molecular structure and changes in biological activity is the center of focus for the field of quantitative structure activity relationships (QSAR). In the field of QSAR, the main objective is to investigate these relationships by building mathematical models that explain the relationship in a statistical way. QSAR first dates back to the 19th century, with A.F.A. Cros' discovery of the inverse relationship between the water solubility of alcohols and the toxicity of alcohols to mammals [1]. Today, QSARs are being applied in many disciplines with much emphasis in drug design. Over the years of development, many methods, algorithms and techniques have been discovered and applied in QSAR studies. With the success of their applications, QSAR has becoming one of the well-developed areas in computational chemistry.

Clearly, QSAR has matured, however, there are still many interesting questions raised in the field. One of them is the ability of a previously developed QSAR model to predict a query compound's activity. For instance, many successful QSAR models with high predictability have been reported in the literature for the prediction of fathead minnow acute toxicity [2–5]. However, the good predictability was only applicable to the group of compounds that were excluded from the original dataset prior to model building. Thus, it would be interesting to determine if the good predictability also applies to a new query compound, a compound not included in the original dataset.

In general, the question is: given a QSAR model and a query compound for prediction, can the developed QSAR model be reliably used to provide an accurate and reliable prediction? To address this question, we have developed a new approach using hierarchical clustering combined with the usual QSAR model generation methods. The main task of this new approach involves testing the following

* Corresponding author. Tel.: +1 814 865 3739; fax: +1 814 865 3314.
*E-mail address:* pcj@psu.edu (P.C. Jurs).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.03.003

hypothesis: if a query compound is found to be more similar in structure to the compounds used to generate the QSAR model, and then it should be predicted more reliably and more accurately by that model. If the hypothesis is shown to be true, then we can answer the question and conclude that a QSAR model can reliably predict a query compound's activity if the query compound is sufficiently similar to the compounds used to generate the QSAR model. Thus, finding a correlation between the similarity of a query compound to the compounds used to build the QSAR model and the prediction accuracy of the query compound given by that QSAR model was the core of this approach.

In this study, our main objective is to assess QSAR models' reliability for activity prediction of new compounds. The designed approach was tested on 322 organic compounds with fathead minnow acute toxicity as the activity of interest. Even though only one particular dataset was studied, we believe our approach is quite general, and it can assist in the assessment of the reliability of a QSAR model's prediction in general.

## 2. Experimental methods

Our approach consisted of two major steps, object clustering and activity prediction. The clustering step involved grouping dataset compounds into clusters using hierarchical clustering. The main purpose, here, was to form dissimilar clusters of objects, to which the query compounds would be compared to for determination of degree of similarity. We choose hierarchical clustering for this similarity comparison step over other more sophisticated similarity determination methods because of its simplicity and speed. The results showed that its similarity determination ability was sufficient for this application. In the second step, a QSAR model was developed with each of the formed clusters to predict the query compounds' activity of interest. This step gave us an idea about the prediction accuracy of each QSAR model for each query compound's activity prediction. Four different trials were designed with this approach. In these trials, the number of query compounds, the number of clusters and the method of clustering varied. The overall approach is illustrated in Fig. 1.

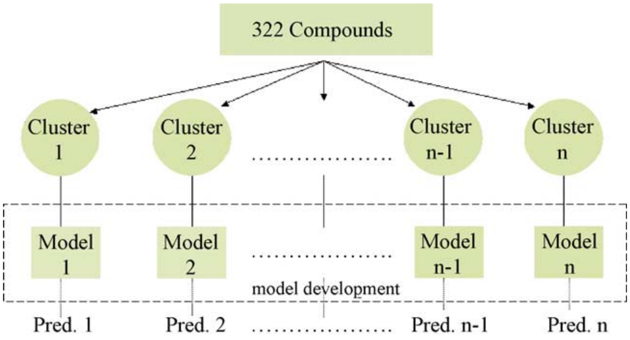

Fig. 1. Study schema.

Table 1
Dataset information

| No. | CAS# | Name | Exp. |
|---|---|---|---|
| 1 | 110-82-7 | Cyclohexane | 2.96 |
| 2 | 67-64-1 | Acetone | 0.85 |
| 3 | 78-93-3 | 2-Butanone | 1.35 |
| 4 | 107-87-9 | 2-Pentanone | 1.84 |
| 5 | 591-78-6 | 2-Hexanone | 2.37 |
| 6 | 110-43-0 | 2-Heptanone | 2.94 |
| 7 | 111-13-7 | 2-Octanone | 3.45 |
| 8 | 821-55-6 | 2-Nonanone | 3.97 |
| 9 | 693-54-9 | 2-Decanone | 4.5 |
| 10 | 6175-49-1 | 2-Dodecanone | 5.19 |
| 11 | 563-80-4 | 3-Methyl-2-butanone | 2 |
| 12 | 108-10-1 | 4-Methyl-2-pentanone | 2.27 |
| 13 | 75-97-8 | 3,3-Dimethyl-2-butanone | 3.06 |
| 14 | 96-22-0 | 3-Pentanone | 1.75 |
| 15 | 110-12-3 | 5-Methyl-2-hexanone | 2.86 |
| 16 | 502-56-7 | 5-Nonanone | 3.66 |
| 17 | 108-94-1 | Cyclohexanone | 2.19 |
| 18 | 76-22-2 | Camphor | 3.14 |
| 19 | 67-56-1 | Methanol | 0.05 |
| 20 | 64-17-5 | Ethanol | 0.52 |
| 21 | 71-23-8 | 1-Propanol | 1.12 |
| 22 | 71-36-3 | 1-Butanol | 1.59 |
| 23 | 71-41-0 | 1-Pentanol | 2.21 |
| 24 | 111-27-3 | 1-Hexanol | 2.94 |
| 25 | 111-70-6 | 1-Heptanol | 3.51 |
| 26 | 111-87-5 | 1-Octanol | 3.98 |
| 27 | 143-08-8 | 1-Nonanol | 4.41 |
| 28 | 112-30-1 | 1-Decanol | 4.82 |
| 29 | 112-42-5 | 1-Undecanol | 5.22 |
| 30 | 112-53-8 | 1-Dodecanol | 5.27 |
| 31 | 67-63-0 | Isopropanol | 0.78 |
| 32 | 78-92-2 | 2-Butanol | 1.31 |
| 33 | 104-76-7 | 2-Ethyl-1-hexanol | 3.66 |
| 34 | 78-83-1 | 2-Methyl-1-propanol | 1.69 |
| 35 | 75-65-0 | 2-Methyl-2-propanol | 1.06 |
| 36 | 600-36-2 | 2,4-Dimethyl-3-pentanol | 2.85 |
| 37 | 77-74-7 | 3-Methyl-3-pentanol | 2.18 |
| 38 | 108-93-0 | Cyclohexanol | 2.06 |
| 39 | 111-46-6 | Diethyleneglycol | 0.15 |
| 40 | 107-21-1 | Ethyleneglycol | 0.04 |
| 41 | 107-41-5 | Hexyleneglycol | 1.13 |
| 42 | 57-55-6 | 1,2-Propyleneglycol | 0.13 |
| 43 | 2216-51-5 | *L*-Menthol | 3.92 |
| 44 | 75-09-2 | Dichloromethane | 2.42 |
| 45 | 67-66-3 | Chloroform | 3.06 |
| 46 | 56-23-5 | Carbontetrachloride | 3.56 |
| 47 | 107-06-2 | 1,2-Dichloroethane | 2.9 |
| 48 | 71-55-6 | 1,1,1-Trichloroethane | 3.4 |
| 49 | 79-00-5 | 1,1,2-Trichloroethane | 3.21 |
| 50 | 79-34-5 | 1,1,2,2-Tetrachloroethane | 3.92 |
| 51 | 76-01-7 | Pentachloroethane | 4.43 |
| 52 | 67-72-1 | Hexachloroethane | 5.23 |
| 53 | 58-89-9 | Hexachlorocyclohexane | 6.52 |
| 54 | 78-87-5 | 1,2-Dichloropropane | 2.93 |
| 55 | 142-28-9 | 1,3-Dichloropropane | 2.94 |
| 56 | 96-18-4 | 1,2,3-Trichloropropane | 3.41 |
| 57 | 110-56-5 | 1,4-Dichlorobutane | 3.39 |
| 58 | 628-76-2 | 1,5-Dichloropentane | 3.75 |
| 59 | 627-30-5 | 3-Chloro-1-propanol | 2.07 |
| 60 | 115-20-8 | 2,2,2-Trichloroethanol | 2.8 |
| 61 | 57-15-8 | 1,1,1-Trichloro-2-methyl-2-propanol | 3.12 |
| 62 | 106-94-5 | 1-Bromopropane | 3.26 |
| 63 | 109-65-9 | 1-Bromobutane | 3.57 |

Table 1 (*Continued*)

| No. | CAS# | Name | Exp. |
|---|---|---|---|
| 64 | 111-25-1 | 1-Bromohexane | 4.68 |
| 65 | 629-04-9 | 1-Bromoheptane | 5.09 |
| 66 | 111-83-1 | 1-Bromoocatne | 5.36 |
| 67 | 109-64-8 | 1,3-Dibromopropane | 5.05 |
| 68 | 107-10-8 | *n*-Propylamine | 2.28 |
| 69 | 109-73-9 | *n*-Butylamine | 2.44 |
| 70 | 33966-50-6 | Sec-butylamine | 2.42 |
| 71 | 110-58-7 | *n*-Pentylamine | 2.69 |
| 72 | 111-26-2 | *n*-Hexylamine | 3.25 |
| 73 | 111-68-2 | *N*-Heptylamine | 3.72 |
| 74 | 111-86-4 | *n*-Octylamine | 4.4 |
| 75 | 112-20-9 | *n*-Nonylamine | 4.82 |
| 76 | 2016-57-1 | *n*-Decylamine | 5.18 |
| 77 | 7307-55-3 | *n*-Undecylamine | 2.91 |
| 78 | 124-22-1 | *n*-Dodecylamine | 6.26 |
| 79 | 2869-34-3 | *n*-Tridecylamine | 6.45 |
| 80 | 13952-84-6 | 2-Butanamine | 2.42 |
| 81 | 598-74-3 | 1,2-Dimethylpropylamine | 2.49 |
| 82 | 5813-64-9 | 2,2-Dimethyl-1-propylamine | 2.26 |
| 83 | 15673-00-4 | 3,3-Dimethylbutylamine | 2.23 |
| 84 | 107-45-9 | *tert*-Octylamine | 3.72 |
| 85 | 693-16-3 | 1-Methylheptylamine | 4.4 |
| 86 | 141-43-5 | 2-Aminoethanol | 1.47 |
| 87 | 78-96-6 | 1-Amino-2-propanol | 1.47 |
| 88 | 109-85-3 | 2-Methoxyethylamine | 2.16 |
| 89 | 109-89-7 | Diethylamine | 1.93 |
| 90 | 143-16-8 | Di-*n*-hexylamine | 5.38 |
| 91 | 110-73-6 | 2-(Ethylamine)ethanol | 1.78 |
| 92 | 100-37-8 | *N*,*N*-Diethylethanolamine | 1.82 |
| 93 | 96-80-0 | 2-(Diisopropylamino)-ethanol | 2.86 |
| 94 | 105-14-6 | 5-(Diethylamino)-2-pentanone | 2.67 |
| 95 | 102-69-2 | Tripropylamine | 3.45 |
| 96 | 91-65-6 | *N*,*N*-Diethylcyclohexylamine | 3.86 |
| 97 | 103-76-4 | 1-(2-Hydroxyethyl)piperazine | 1.31 |
| 98 | 140-31-8 | *N*-Aminoethylpiperazine | 1.82 |
| 99 | 78-90-0 | 1,2-Propanediamine | 1.81 |
| 100 | 107-15-3 | Ethylenediamine | 2.55 |
| 101 | 96-29-7 | Methylethylketoxime | 2.01 |
| 102 | 127-06-0 | Acetoneoxime | 2.12 |
| 103 | 100-64-1 | Cyclohexanoneoxime | 2.74 |
| 104 | 761-65-9 | *N*,*N*-Dibutylformamide | 3.25 |
| 105 | 68-12-2 | *N*,*N*-Dimethylformamide | 0.84 |
| 106 | 1634-04-4 | Methyltert-butylether | 2.12 |
| 107 | 142-96-1 | Di-*n*-butylether | 3.61 |
| 108 | 60-29-7 | Diethylether | 1.46 |
| 109 | 108-20-3 | Diisopropylether | 2.11 |
| 110 | 693-65-2 | Di-*n*-pentylether | 4.71 |
| 111 | 109-87-5 | Dimethoxymethane | 1.04 |
| 112 | 123-91-1 | 1,4-Dioxane | 0.93 |
| 113 | 110-88-3 | Trioxane | 1.18 |
| 114 | 109-99-9 | Tetrahydrofuran | 1.52 |
| 115 | 470-82-6 | 1,8-Epoxy-*p*-menthane | 3.18 |
| 116 | 64-19-7 | Aceticacid | 2.85 |
| 117 | 109-52-4 | *n*-Pentanoicacid | 3.12 |
| 118 | 142-62-1 | *n*-Hexanoicacid | 2.76 |
| 119 | 112-05-0 | *n*-Nonanoicacid | 3.18 |
| 120 | 124-04-9 | Adipicacid | 3.18 |
| 121 | 75-07-0 | Acetaldehyde | 3.11 |
| 122 | 123-72-8 | Butanal | 3.65 |
| 123 | 110-62-3 | Pentanal | 3.82 |
| 124 | 66-25-1 | Hexanal | 3.66 |
| 125 | 590-86-3 | 3-Methyl-butanal | 4.42 |
| 126 | 111-30-8 | Glutaraldehyde | 3.94 |
| 127 | 107-22-2 | Glyoxal | 2.43 |
| 128 | 141-78-6 | Ethylacetate | 2.58 |
| 129 | 79-20-9 | Methylacetate | 2.27 |
| 130 | 109-21-7 | *n*-Butyln-butyrate | 4.09 |
| 131 | 123-86-4 | *n*-Butylacetate | 3.81 |
| 132 | 540-88-5 | *tert*-Butylacetate | 2.55 |
| 133 | 142-92-7 | *n*-Hexylacetate | 4.56 |
| 134 | 123-66-0 | Ethylhexanoate | 4.21 |
| 135 | 109-60-4 | *n*-Propylacetate | 3.23 |
| 136 | 111-15-9 | 2-Ethoxyethylacetate | 3.5 |
| 137 | 108-59-8 | Dimethylmalonate | 4.03 |
| 138 | 87-91-2 | Diethyll-(+)-tartrate | 2.5 |
| 139 | 105-53-3 | Diethylmalonate | 4.01 |
| 140 | 123-25-1 | Diethylsuccinate | 3.09 |
| 141 | 141-28-6 | Diethyladipate | 4.05 |
| 142 | 110-40-7 | Diethylsebacate | 4.98 |
| 143 | 105-99-7 | Dibutyladipate | 4.85 |
| 144 | 818-61-1 | 2-Hydroxyethylacrylate | 4.38 |
| 145 | 140-88-5 | Ethylacrylate | 4.6 |
| 146 | 106-63-8 | Isobutylacrylate | 4.79 |
| 147 | 999-61-1 | 2-Hydroxypropylacrylate | 4.59 |
| 148 | 868-77-9 | 2-Hydroxyethylmethacrylate | 2.76 |
| 149 | 80-62-6 | Methylmethacrylate | 2.5 |
| 150 | 75-05-8 | Acetonitrile | 1.49 |
| 151 | 107-12-0 | Propionitrile | 1.56 |
| 152 | 2243-27-8 | *n*-Octylcyanide | 4.3 |
| 153 | 764-13-6 | 2,5-Dimethyl-2,4-hexadiene | 4.46 |
| 154 | 513-81-5 | 2,3-Dimethyl-1,3-butadiene | 4.08 |
| 155 | 5194-50-3 | 2,4-Hexadiene | 3.61 |
| 156 | 5989-27-5 | *d*-Limonene | 5.29 |
| 157 | 77-73-6 | Dicyclopentadiene | 3.63 |
| 158 | 1647-16-1 | 1,9-Decadiene | 5.68 |
| 159 | 78-79-5 | Isoprene | 2.95 |
| 160 | 75-35-4 | 1,1-Dichloroethylene | 2.84 |
| 161 | 79-01-6 | Trichloroethylene | 3.47 |
| 162 | 127-18-4 | Tetrachloroethylene | 3.91 |
| 163 | 107-19-7 | Propargylalcohol | 4.56 |
| 164 | 818-72-4 | 1-Octyn-3-ol | 5.49 |
| 165 | 110-65-6 | 2-Butyne-1,4-diol | 3.21 |
| 166 | 764-01-2 | 2-Butyn-1-ol | 3.84 |
| 167 | 95-63-6 | 1,2,4-Trimethylbenzene | 4.19 |
| 168 | 2416-94-6 | 2,3,6-Trimethylphenol | 4.22 |
| 169 | 527-60-6 | 2,4,6-Trimethylphenol | 4.02 |
| 170 | 25167-83-3 | Tetrachlorophenol | 6.13 |
| 171 | 98-86-2 | Acetophenone | 2.87 |
| 172 | 100-51-6 | Benzylalcohol | 2.37 |
| 173 | 623-25-6 | 1,4-Bis(chloromethyl)benzene | 6.65 |
| 174 | 100-44-7 | Benzylchloride | 4.4 |
| 175 | 100-46-9 | Benzylamine | 3.02 |
| 176 | 100-52-7 | Benzaldehyde | 3.93 |
| 177 | 100-10-7 | 4-(Dimethylamino)-benzaldehyde | 3.51 |
| 178 | 122-03-2 | 4-Isopropylbenzaldehyde | 4.35 |
| 179 | 446-52-6 | 2-Fluorobenzaldehyde | 4.96 |
| 180 | 104-88-1 | 4-Chlorobenzaldehyde | 4.81 |
| 181 | 613-45-6 | 2,4-Demethoxybenzaldehyde | 3.92 |
| 182 | 1761-61-1 | 2-Hydroxy-5-bromobenzaldehyde | 5.19 |
| 183 | 635-93-8 | 2-Hydroxy-5-chlorobenzaldehyde | 5.31 |
| 184 | 90-02-8 | 2-Hydroxybenzaldehyde | 4.73 |
| 185 | 121-33-5 | 3-Methoxy-4-hyroxybenzaldehyde | 3.12 |
| 186 | 708-76-9 | 2-Hydroxy-4,6-dimethoxybenzaldehye | 4.83 |
| 187 | 653-37-2 | Pentafluorobenzaldehyde | 5.25 |
| 188 | 387-45-1 | 2-Chloro-6-fluorobenzaldehyde | 4.23 |
| 189 | 874-42-0 | 2,4-Dichlorobenzaldehyde | 4.99 |
| 190 | 58-90-2 | 2,3,4,6-Tetrachlorophenol | 5.35 |
| 191 | 4901-51-3 | 2,3,4,5-Tetrachlorophenol | 5.74 |
| 192 | 3481-20-7 | 2,3,5,6-Tetrachloroaniline | 5.93 |
| 193 | 732-26-3 | 2,4,6-*tri*-*tert*-Butylphenol | 6.39 |

Table 1 (*Continued*)

| No. | CAS# | Name | Exp. |
|---|---|---|---|
| 194 | 150-76-5 | 4-Methoxyphenol | 3.27 |
| 195 | 120-07-0 | *N*-Phenyldiethanolamine | 2.39 |
| 196 | 103-83-3 | *N*,*N*-Dimethylbenzylamine | 3.55 |
| 197 | 150-19-6 | 1-Hydroxy-3-methoxybenzene | 3.22 |
| 198 | 150-78-7 | 1,4-Dimethoxybenzene | 3.07 |
| 199 | 5673-07-4 | 2,6-Dimethoxytoluene | 3.88 |
| 200 | 13608-87-2 | 2,3,4-Trichloroacetophenone | 5.05 |
| 201 | 95-95-4 | 2,4,5-Trichlorophenol | 5.34 |
| 202 | 88-06-2 | 2,4,6-Trichlorophenol | 4.64 |
| 203 | 937-20-2 | 2,4-Dichloroacetophenone | 4.21 |
| 204 | 70-69-9 | 4-Aminopropiophenone | 3.01 |
| 205 | 102-27-2 | *N*-Ethyl-*m*-toluidine | 3.44 |
| 206 | 100-61-8 | *N*-Methylaniline | 3.03 |
| 207 | 121-69-7 | *N*,*N*-Dimethylaniline | 3.27 |
| 208 | 91-66-7 | *N*,*N*-Diethylaniline | 3.96 |
| 209 | 59-50-7 | *P*-Chloro-*m*-cresol | 4.4 |
| 210 | 24544-04-5 | 2,6-Diisopropylaniline | 4.1 |
| 211 | 10031-82-0 | 4-Ethoxybenzaldehyde | 3.74 |
| 212 | 634-93-5 | 2,4,6-Trichloroaniline | 4.59 |
| 213 | 634-67-3 | 2,3,4-Trichloroaniline | 4.74 |
| 214 | 615-65-6 | 2-Chloro-4-methylaniline | 3.59 |
| 215 | 71-43-2 | Benzene | 3.5 |
| 216 | 1746-23-2 | *p*-*t*-Butylstyrene | 3.51 |
| 217 | 100-42-5 | Styrene | 3.51 |
| 218 | 1745-81-9 | 2-Allylphenol | 3.95 |
| 219 | 97-53-0 | Eugenol | 3.84 |
| 220 | 108-86-1 | Bromobenzene | 4.45 |
| 221 | 608-71-9 | Pentabromophenol | 6.72 |
| 222 | 106-37-6 | 1,4-Dibromobenzene | 5.28 |
| 223 | 118-79-6 | 1,3,5-Tribromo2-hydroxybenzene | 4.71 |
| 224 | 106-40-1 | 4-Bromoaniline | 3.56 |
| 225 | 108-88-3 | Toluene | 3.42 |
| 226 | 108-90-7 | Chlorobenzene | 3.7 |
| 227 | 108-41-8 | 1-Chloro-3-methylbenzene | 3.84 |
| 228 | 106-43-4 | 1-Chloro-4-methylbenzene | 4.33 |
| 229 | 95-50-1 | *o*-Dichlorobenzene | 4.19 |
| 230 | 541-73-1 | *m*-Dichlorobenzene | 4.27 |
| 231 | 106-46-7 | *p*-Dichlorobenzene | 4.27 |
| 232 | 95-49-8 | 2-Chlorotoluene | 4.23 |
| 233 | 95-73-8 | 2,4-Dichlorotoluene | 4.54 |
| 234 | 120-82-1 | 1,2,4-Trichlorobenzene | 4.8 |
| 235 | 87-61-6 | 1,2,3-Trichlorobenzene | 4.89 |
| 236 | 108-70-3 | 1,3,5-Trichlorobenzene | 4.74 |
| 237 | 95-94-3 | 1,2,4,5-Tetrachlorobenzene | 5.83 |
| 238 | 634-66-2 | 1,2,3,4-Tetrachlorobenzene | 5.29 |
| 239 | 87-86-5 | Pentachlorophenol | 6.04 |
| 240 | 771-60-8 | Pentafluoroaniline | 3.69 |
| 241 | 371-40-4 | 1-Amino-4-fluorobenzene | 3.82 |
| 242 | 128-37-0 | 2,6-Di-*tert*-butyl-4-methylphenol | 5.78 |
| 243 | 108-95-2 | Phenol | 3.5 |
| 244 | 105-67-9 | 2,4-Xylenol | 3.86 |
| 245 | 95-65-8 | 3,4-Xylenol | 3.94 |
| 246 | 95-75-0 | 3,4-Dichlorotoluene | 4.74 |
| 247 | 1126-79-0 | *n*-Butylphenylether | 4.42 |
| 248 | 39905-57-2 | 4-Hexyloxyaniline | 4.78 |
| 249 | 122-99-6 | 2-Phenoxyethanol | 2.6 |
| 250 | 95-48-7 | *o*-Cresol | 3.9 |
| 251 | 108-39-4 | *m*-Cresol | 3.29 |
| 252 | 106-44-5 | *p*-Cresol | 3.76 |
| 253 | 95-57-8 | *o*-Chlorophenol | 4 |
| 254 | 106-48-9 | *p*-Chlorophenol | 4.46 |
| 255 | 120-83-2 | 2,4-Dichlorophenol | 4.3 |
| 256 | 62-53-3 | Aniline | 3.03 |
| 257 | 95-51-2 | *o*-Chloroaniline | 4.35 |
| 258 | 106-47-8 | *p*-Chloroaniline | 3.62 |
| 259 | 95-76-1 | 3,4-Dichloroaniline | 4.32 |
| 260 | 554-00-7 | 2,4-Dichloroaniline | 4.07 |
| 261 | 106-49-0 | 4-Methylaniline | 2.83 |
| 262 | 95-47-6 | *o*-Xylene | 3.81 |
| 263 | 108-38-3 | *m*-Xylene | 3.82 |
| 264 | 106-42-3 | *p*-Xylene | 4.21 |
| 265 | 1689-84-5 | 1-Cyano-3,5-dibromo-4-hydroxybenzene | 4.3 |
| 266 | 5922-60-1 | 1-Cyano-2-amino-5-chlorobenzene | 3.73 |
| 267 | 6575-09-3 | 2-Cyano-6-methyl-benzonitrile | 4 |
| 268 | 529-19-1 | 1-Cyano-2-methylbenzene | 3.42 |
| 269 | 100-47-0 | Benzonitrile | 2.98 |
| 270 | 100-41-4 | Ethylbenzene | 3.59 |
| 271 | 141-93-5 | *m*-Diethylbenzene | 4.51 |
| 272 | 123-07-9 | *p*-Ethylphenol | 4.07 |
| 273 | 589-16-2 | 4-Ethylaniline | 3.22 |
| 274 | 104-13-2 | 4-Butylaniline | 4.17 |
| 275 | 16245-79-7 | 4-Octylaniline | 6.24 |
| 276 | 37529-30-9 | 4-Decylaniline | 6.57 |
| 277 | 80-46-6 | *p*-*tert*-Amylphenol | 4.8 |
| 278 | 98-54-4 | *p*-*tert*-Butylphenol | 4.47 |
| 279 | 89-83-8 | Thymol | 4.67 |
| 280 | 98-82-8 | Cumene | 4.23 |
| 281 | 538-68-1 | *n*-Pentylbenzene | 4.94 |
| 282 | 25154-52-3 | Nonylphenol | 6.22 |
| 283 | 131-11-3 | Dimethylphthalate | 3.7 |
| 284 | 84-66-2 | Diethylphthalate | 4.12 |
| 285 | 84-69-5 | Diisobutylphthalate | 5.49 |
| 286 | 84-74-2 | Dibutylphthalate | 5.33 |
| 287 | 84-62-8 | Diphenylphthalate | 6.6 |
| 288 | 93-89-0 | Ethylbenzoate | 4.23 |
| 289 | 1129-35-7 | Methyl4-cyanobenzoate | 3.54 |
| 290 | 94-09-7 | Ethyl4-aminobenzoate | 3.67 |
| 291 | 1126-46-1 | Methyl4-chlorobenzoate | 4.15 |
| 292 | 368-77-4 | 3-(Trifluoromethyl)benzonitrile | 3.56 |
| 293 | 831-82-3 | 4-Phenoxyphenol | 4.58 |
| 294 | 119-61-9 | Benzophenone | 4.11 |
| 295 | 101-84-8 | Diphenylether | 4.63 |
| 296 | 14548-46-0 | 4-Benzoylpyridine | 3.25 |
| 297 | 122-39-4 | Diphenylamine | 4.65 |
| 298 | 118-55-8 | Phenylsalicylate | 5.27 |
| 299 | 97-23-4 | 2,2′-Methylenebis(4-chloro)phenol | 5.94 |
| 300 | 91-94-1 | 3,3′-Dichlorobenzidine | 5.09 |
| 301 | 92-52-4 | Biphenyl | 4.8 |
| 302 | 90-43-7 | *o*-Phenylphenol | 4.5 |
| 303 | 109-06-8 | 2-Methylpyridine | 2.02 |
| 304 | 108-99-6 | 3-Methylpyridine | 2.81 |
| 305 | 108-89-4 | 4-Methylpyridine | 2.36 |
| 306 | 100-70-9 | 2-Pyridinecarbonitrile | 2.16 |
| 307 | 500-22-1 | 3-Pyridinecarboxaldehyde | 3.81 |
| 308 | 104-90-5 | 5-Ethyl-2-methylpyridine | 3.17 |
| 309 | 1122-54-9 | 4-Acetylpyridine | 2.86 |
| 310 | 5683-33-0 | 2-Dimethylaminopyridine | 2.98 |
| 311 | 110-86-1 | Pyridine | 2.9 |
| 312 | 2859-67-8 | 3-Pyridinepropanol | 2.96 |
| 313 | 2176-62-7 | Pentachloropyridine | 5.73 |
| 314 | 939-23-1 | 4-Phenylpyridine | 3.98 |
| 315 | 91-20-3 | Naphthalene | 4.32 |
| 316 | 90-15-3 | 1-Naphthalenol | 4.54 |
| 317 | 135-19-3 | 2-Naphthol | 4.62 |
| 318 | 90-12-0 | 1-Methylnaphthalene | 4.2 |
| 319 | 90-13-1 | 1-Chloronaphthalene | 4.85 |
| 320 | 91-22-5 | Quinoline | 3.45 |
| 321 | 260-94-6 | Acridine | 4.89 |
| 322 | 253-52-1 | Phthalazine | 3.11 |

### 2.1. Compounds

The majority of the data for this study came from the AQUIRE database [4]. Additional data were collected from the literature [5–7]. The final dataset included 322 organic compounds with fathead minnow acute aquatic toxicity, expressed as: −log (mmol/L), as the activity of interest. A list of the 322 compounds is presented in Table 1. The most toxicologically active compound in this dataset was pentabromophenol with a toxicity of 6.72 log units, and the least toxicologically active compound was ethylene glycol with a toxicity of 0.04 log units. The compounds included in the dataset were selected based on the requirements of the ADAPT software package [8–11]. The system handles only neutral organic compounds, but not charged species, and it supports atom types C, H, O, N, P halogens and sulfur compounds. All computations in this study were performed on a DEC 300 AXP Model 500 workstation with the ADAPT software package.

### 2.2. Objects clustering

In the cluster analysis step, hierarchical cluster analysis was carried out using MATLAB for windows [12] and using a set of descriptors generated by the ADAPT software package. Descriptor generation will be described in detail later in this paper. The next few paragraphs describe the cluster algorithm used in this study.

#### 2.2.1. Summary of hierarchical clustering

The aim of the clustering was the recognition of groups of objects based on their similarity. It involves grouping a collection of objects into clusters (subsets), such that objects within each cluster are more closely related to one another than objects in different clusters. In this study, hierarchical clustering was used [13]. In hierarchical clustering, the objects are clustered into subgroups in a series of partitions. The basic process of hierarchical clustering starts by assigning each object as a cluster, so that *N* objects will form *N* clusters with each cluster containing just one item. Then the similarities between the clusters are determined by the distances (similarities) between the items they contain. The next step is to find the most similar (closest) pair of clusters and merge them into a new cluster, so that now the number of clusters becomes *N*–1. Then the distances (similarities) between the newly formed cluster and each of the old clusters are computed. The previous two steps are iterated until all items are clustered into a single cluster containing *N* objects. Within hierarchical clustering, there are two subdivisions: agglomerative methods and divisive methods [13]. The algorithm just described is an agglomerative method, and it is the method implemented in this study. The divisive method works in the opposite way. It separates *N* objects into finer groups. Every object is initially in a huge cluster. Then the cluster is divided continuously until the desired number of clusters is formed. In brief, agglomerative method is a bottom–up clustering method, whereas divisive method uses a top–down approach. The graphical representation of the hierarchical clustering and agglomerative methods are presented in Figs. 2 and 3.

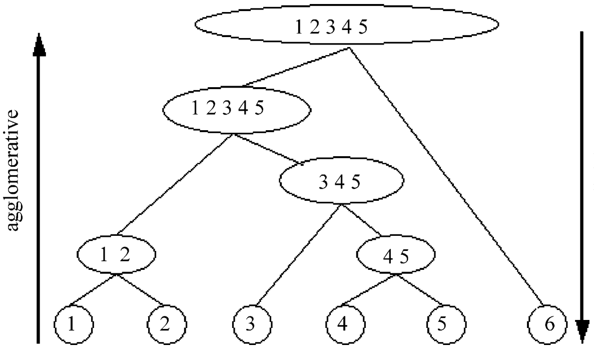

Fig. 2. Hierarchical clustering algorithm.

*2.2.1.1. Linkage methods.* In hierarchical clustering, the distances (similarities) computation (distance calculation between clusters) can be done in different ways. This cluster fusion step is often called the linkage method, which is what distinguishes single linkage from average linkage and complete linkage clustering [13]. In single linkage clustering, the shortest distance from any object in one cluster to any object in the other cluster is considered as the distance between one cluster and another cluster. In average linkage clustering, the distance between two clusters is the average distance from any object in one cluster to any object in the other cluster. Complete linkage clustering considers the longest distance from any object in one cluster to any object in the other cluster. In this study, single linkage and average linkage clustering were used in different parts of the study.

Single linkage is one of the simplest agglomerative hierarchical clustering methods. The distance between clusters is given by the value of the shortest link between the clusters, illustrated in Fig. 4. The distance $D(x, y)$ is computed as

$$D(x, y) = \mathrm{Min}(\mathrm{d}(i, j))$$

where object $i$ is in cluster $x$ and object $j$ is in cluster $y$.

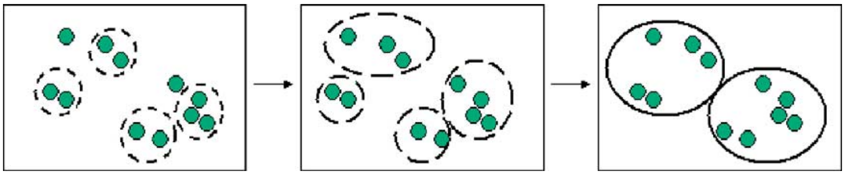

Fig. 3. Agglomerative method.

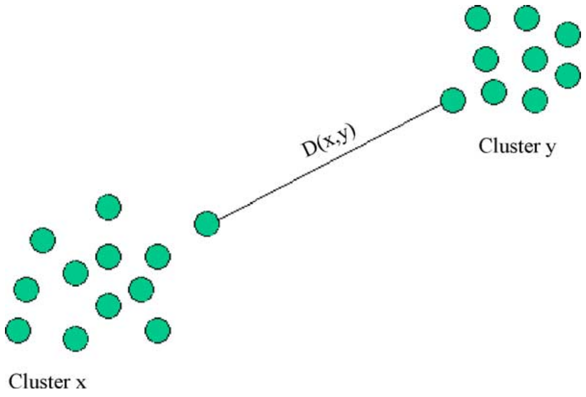

Fig. 4. Single linkage clustering.

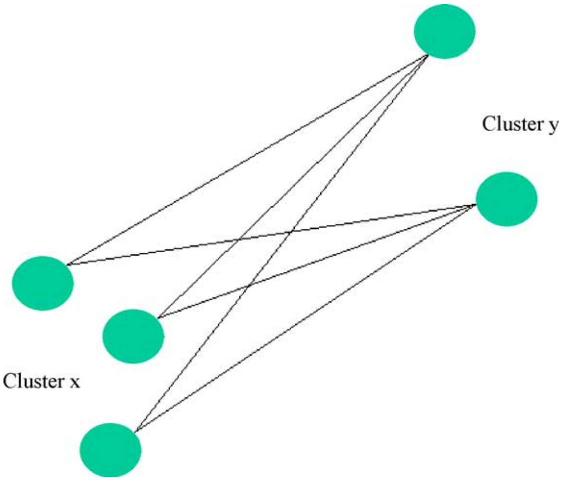

Fig. 5. Average linkage clustering.

At each stage of hierarchical clustering, the cluster *x* and *y*, for which $D(x, y)$ is minimum, are merged. Average linkage clustering is illustrated in Fig. 5. In average linkage method, the distance $D(x, y)$ is computed as,

$$D(x,y) = \frac{T_{xy}}{N_x N_y}$$

where $T_{xy}$ is the sum of all pairwise distances between cluster *x* and *y*. $N_x$ and $N_y$ are the sizes of the clusters *x* and *y*, respectively.

At each stage of hierarchical clustering, the cluster *x* and *y*, for which $D(x, y)$ is minimum, are merged. Briefly, the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.

*2.2.1.2. Similarity metrics.* Various types of hierarchical clustering methods differ not only in the linkage method used to merge objects, but also in the distance metric used to calculate the distance matrix between objects. Frequently used similarity metrics include Euclidean distance, standardized Euclidean distance, city-block distance, Minkowski distance and Mahalanobis distance. In this study, standardized Euclidean distance and city-block distance [13] were selected as they provided the best clustering. Standardized Euclidean distance was calculated by the following equation,

$$d_{xy}^2 = (T_x - T_y)D^{-1}(T_x - T_y)'$$

where $T$ is a $m \times n$ matrix, $d_{xy}$ is the distance between the vector $T_x$ and $T_y$. $D$ is the diagonal matrix with diagonal elements given by the variance of the variable $T_i$ over the $m$ objects.

City-block distance is simply the summed difference across dimensions and was calculated as

$$d_{xy} = \sum_{i=1}^{n} |T_{xi} - T_{yi}|$$

With the formation of clusters, the similarity was calculated between the query compounds and the clusters. The distance measure used to determine this similarity was the distance from the query compound to the centroid of the cluster in the feature space.

## 3. QSAR model development

After clusters were formed, QSAR models were developed with each cluster. Here, four major steps were included: structure entry-optimization, structure representation (descriptor generation), feature selection and mapping.

### 3.1. Structure entry-optimization

Structure entry started with drawing the 322 organic compounds using a commercial software product, HyperChem (HyperCube Inc., Waterloo, ON) on a PC. The two-dimensional sketched structures were represented by connection tables, which contain atom types and the connections between atoms. The structure information was then transferred and stored in an ADAPT work area. Then the geometry of each compound was optimized with the PM3 Hamiltonian [14] using MOPAC [15]. A single-point energy calculation was also performed using the AM1 Hamiltonian [16] on the PM3 geometry optimized structures using MOPAC in order to generate accurate charge information. Information for both two-dimensional and three-dimensional representations were saved and used later for descriptor generation.

### 3.2. Descriptor generation

In developing predictive models, in order to mathematically link the structure and activity of interest, organic compounds' chemical structures must be numerically encoded by descriptors. Descriptors are simply numerical values that describe aspects of the molecular structures.

These numerical representations were calculated using ADAPT descriptor generation routines. Based on the information content, these descriptors can be classified into four categories: topological descriptors, geometric descriptors, electronic descriptors and hybrids of the other three. Topological descriptors [17–19] are calculated from connection tables only, and they do not require accurate 3D optimized structural information. These descriptors reflect the molecular connectivity and branching of the molecule. They include atom, bond, functional group, path and substructure counts and connectivity indices [20]. Geometric descriptors [21] are calculated from the PM3 geometry optimized 3D structures. They encode information on the molecules' 3D structure and shape of the molecule. Examples of these descriptors include principal moments of inertia [22], solvent accessible surface area [23], molecular volume, length to breadth ratio [21], shadow projection areas [24] and gravitational index G [25]. Electronic descriptors include information, such as partial atomic charges [26], dipole moment, HOMO and LUMO energy levels [27] that quantify charge information of the molecule. In many cases, they are also related to the molecular topology and composition. The calculation of these descriptors requires not only accurate 3D geometry optimized structural information, but also the single-point AM1 charges calculated using MOPAC. Hybrid descriptors are calculated by combining two or more of the above descriptor information sources. Charged partial surface areas (CPSAs) [28] are examples of these descriptors. They combine geometric and electronic properties of the molecules, e.g. surface areas and fractional partial charges. These descriptors characterize a molecule's ability to engage in polar interactions with the information on partial positive/negative charges relative to surface areas, total weighted partial charges relative to surface areas and fractional partial charges relative to surface areas. Hydrogen bonding descriptors [29] are also calculated to encode those molecular features that are important for intermolecular hydrogen bonding.

The structure entry-optimization and structure representation steps were performed at the earliest stage due to the approach designed in this study. The clustering step requires the input of the set of descriptors encoding the compounds.

### 3.3. Feature selection

Once approximately 230 descriptors had been generated by the descriptor generation routines, feature selection was used to reduce the number of descriptors per compound. Feature selection is used to choose a subset of descriptors that are best in encoding the property of interest, since many of the calculated descriptors carry redundant and highly correlated information or very little useful information. Feature selection includes objective methods and subjective methods. Objective feature selection uses the independent variables alone to filter out non-useful descriptors without using the dependent variables. It employs an identical test, pairwise correlation test and vector space descriptor analysis (VSDA) [30] to eliminate descriptors which contain little information or which have a high correlation with another descriptor. Descriptors which had identical or zero values for greater than 90% of the compounds were eliminated. One descriptor in any pair of descriptors whose pairwise correlation exceeding 0.85 was also eliminated. The remaining descriptors were then further reduced by subjective feature selection, which searches for an information-rich subset of descriptors. Here, the dependent variables, acute aquatic toxicity values, were considered in descriptor selection. The selected subset of descriptors usually ranging from 3 to 10 descriptors per model was used to map the set of molecular structure to the toxicity. An evolutionary optimization technique, simulated annealing [31], was employed in subjective feature selection to search the space for optimal subsets of descriptors.

## 4. Model building

Prior to QSAR model development, the dataset was split into three sets for model building, validation and prediction purposes. They are the training set (TSET), cross-validation set (CVSET) and external prediction set (PSET). The TSET contained 80% of the compounds and was used for model building. For model validation, 10% of the compounds were assigned as the CVSET. The remaining 10% of the compounds were assigned as the PSET. In multiple linear

Table 2
Summary of four studied trials

| | No. of clusters | No. of query compounds | Similarity measures | Linkage methods |
|---|---|---|---|---|
| Trial I | 5 | 4 | City-block | Single linkage |
| Trial II | 5 | 9 | City-block | Single linkage |
| Trial III | 4 | 15 | City-block | Single linkage |
| Trial IV | 3 | 18 | Standardized. Euclidean distance | Average linkage |

| | No. of query compounds | Identities of query compounds |
|---|---|---|
| Trial I | 4 | **150**, **197**, **217**, **221** |
| Trial II | 9 | **23**, **65**, **111**, **150**, **197**, **217**, **221**, **228**, **292** |
| Trial III | 15 | **16**, **23**, **32**, **65**, **111**, **150**, **161**, **197**, **217**, **221**, **228**, **235**, **261**, **268**, **292** |
| Trial IV | 18 | **16**, **23**, **32**, **65**, **111**, **150**, **161**, **187**, **197**, **211**, **217**, **228**, **235**, **261**, **268**, **292**, **298**, **316** |

regression modeling, the CVSET and PSET were combined and treated as an external prediction set. The compounds were randomly selected for inclusion in the three sets. In addition, a set of query compounds was randomly withdrawn from the dataset prior to the sets formation step. For the purpose in this study, only query compounds were included in the external prediction set, thus the number of compounds in the PSET varies from trial to trial.

Three types of models were developed, distinguished by the linearity of feature selection and model formation. Models developed with linear feature selection and linear model formation method were classified as type I models. Models with linear feature selection and non-linear model formation method were classified as type II models. Type III models were developed with non-linear feature selection and non-linear model formation methods.

### 4.1. Type I models

Type I models were developed from the reduced pool of descriptors established after objective feature selection using the simulated annealing algorithm to select subsets of descriptors [32]. The quality of these potential models was assessed based on the root-mean-square (rms) error of the TSET compounds. Other statistical factors, such as correlation coefficients ($R$), P-statistic and T-statistic were also examined. The algorithm generated 10 models for each subset of descriptors. The optimal model size was decided after comparing the rms errors of the smaller models to those of the models with more descriptors. Once the optimal model size was determined, the 10 models' multicollinearities among the descriptors were then determined by a variance inflation factor (VIF). VIF was calculated for each descriptor in the model as $1/(1 - R^2)$, where $R$ is a multiple correlation coefficient [33]. Multicollinearities were considered to exist when the VIF was greater than 10. Models with excessive multicollinearities were eliminated. The resulting models were checked for outliers in the TSET. Six statistical tests were used to check for the presence of outliers [34]. They are residuals, standardized residuals, studentized residuals, leverage points, DFFITS values and Cook's distances. Any member flagged as an outlier by four or more of the tests were investigated for its effects on the model. Outliers were removed from the model and the coefficients were recalculated. The suspect compounds were excluded from the dataset if the outlier removal did improve

Table 3
Degree of similarity between query compounds and clusters in trial I

| Query compounds | Most similar → Least similar | | | | |
|---|---|---|---|---|---|
| | Cluster no. | | | | |
| **150** | 2 | 3 | 1 | 4 | 5 |
| **197** | 1 | 3 | 4 | 2 | 5 |
| **217** | 3 | 1 | 2 | 4 | 5 |
| **221** | 4 | 5 | 1 | 3 | 2 |

Table 4
Descriptors chosen for the QSAR model for the prediction of fathead minnow acute toxicity in trial I

| Label | Description |
|---|---|
| Cluster 1 | |
| KAPA 4 | Kier shape descriptor, atom type corrected first-order Kappa index [37] |
| S4PC | Fourth-order simple path cluster [38] |
| S6CH | Simple molecular connectivity index descriptor of paths of length six [38] |
| MOLC 8 | Weighted 4th-order path/clusters [39] |
| NBr | Number of basis rings |
| MDE 33 | Molecular distance edge between 3°/3° carbons [40] |
| V4P 5 | Valence-corrected fourth-order path molecular connectivity index [38] |
| SYMM15 | Geometric symmetry index |
| Cluster 2 | |
| S6CH | Simple molecular connectivity index descriptor of paths of length six [38] |
| N6PC | Number of sixth-order path clusters [38] |
| MOLC 8 | Weighted 4th-order path/clusters [39] |
| NO 3 | Number of oxygen atoms |
| NDB | Number of double bond |
| NLP 19 | Number of lone pairs |
| WTPT 2 | Average molecular ID (molecular ID/no. atoms) [41] |
| MDE 12 | Molecular distance edge between 1° and 2° carbons [40] |
| Cluster 3 | |
| ALLP 4 | (Total number of weighted paths)/(number of atoms) [42,43] |
| V6CH | Atom valence-corrected sixth-order chain values [38] |
| MDE 11 | Molecular distance edge between primary carbons [40] |
| MDE 13 | Molecular distance edge between 1° and 3° carbons [40] |
| MDE 44 | Distance edge for 4° to 4° carbons [40] |
| Cluster 4 | |
| MOLC 5 | Path three molecular connectivity [39] |
| NC | Number of carbons |
| NDB | Number of double bond |
| MDE 24 | Molecular distance edge between 2° and 4° carbons [40] |
| MOLC 9 | Heteroatom and aromatic ring corrected average distance sum connectivity [39] |
| NCl 7 | Number of chlorine atoms |
| MDE 14 | Molecular distance edge between 1° and 4° carbons [40] |
| Cluster 5 | |
| KAPPA 2 | Kappa 2 index [37] |
| S4PC | Fourth-order simple path cluster [38] |
| NAB | The count of aromatic bonds |
| 2SP3-1 | No. $sp^3$ carbons attached to two carbon atoms |
| V2 | Second order valence path molecular connectivity [38] |
| MOLC 8 | Weighted 4th-order path/clusters [39] |
| WTPT 2 | Average molecular ID (molecular ID/no. atoms) [41] |
| MDE 12 | Molecular distance edge between 1° and 2° carbons [40] |

the rms error greatly. The query compounds in the PSET were untouched until model development was finished. The acute aquatic toxicity values of the query compounds were then predicted with the final type I QSAR model.

### 4.2. Type II models

Type II models were developed in the belief that they should provide better toxicity prediction if the relationship between the descriptors and the acute toxicity is nonlinear. The descriptors from the best type I model were used to build a computational neural network (CNN) model [35], a type II model. CNN is a computer program designed to simulate the way the human brain processes information. The network gathers knowledge by detecting patterns and relationships in the input data. Its goal is to predict an output value (property of interest) from input variables (descriptor values) for each compound in the

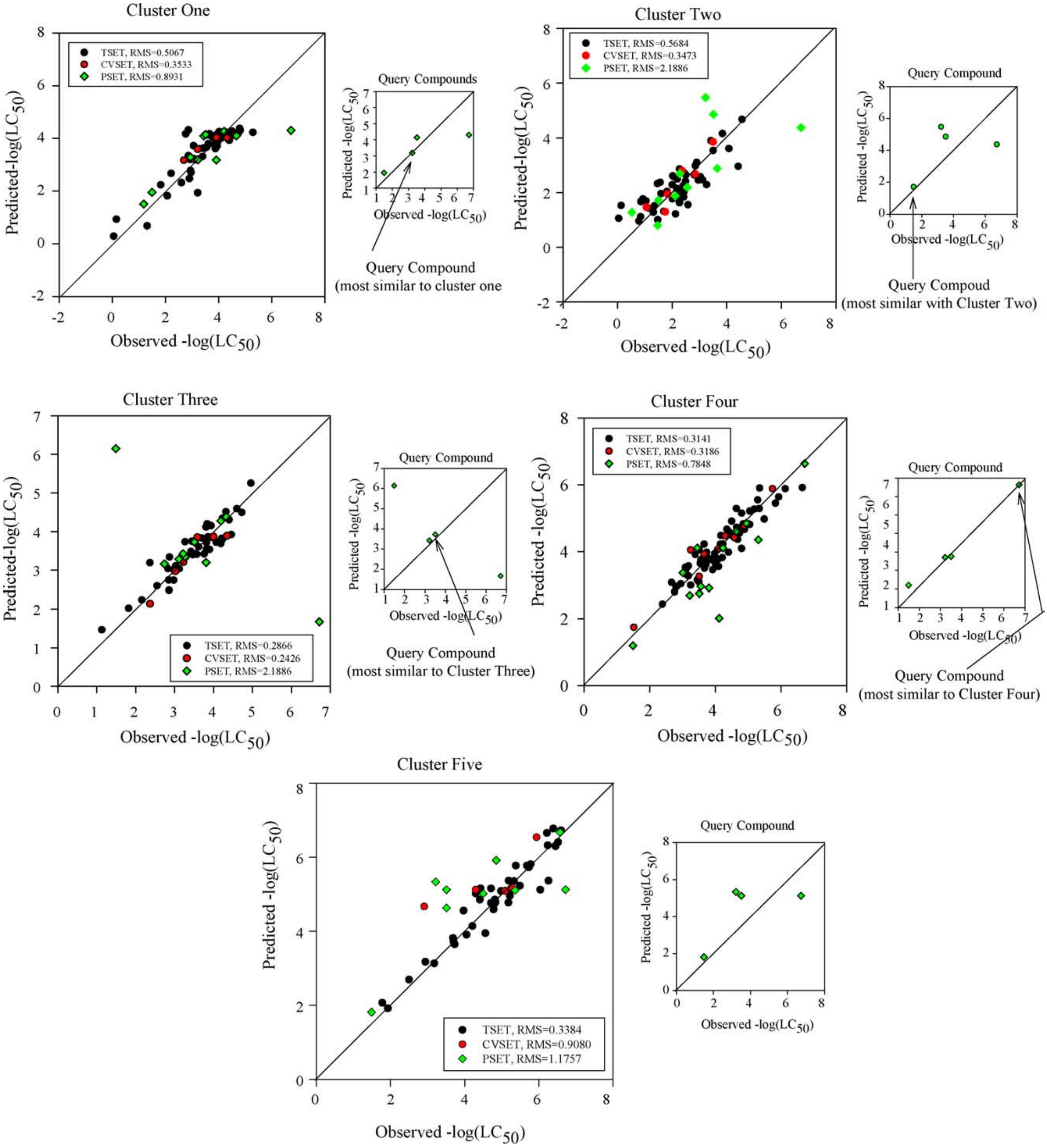

Fig. 6. Predicted –log(LC$_{50}$) vs. observed –log(LC$_{50}$) of the five developed QSAR model in trial I.

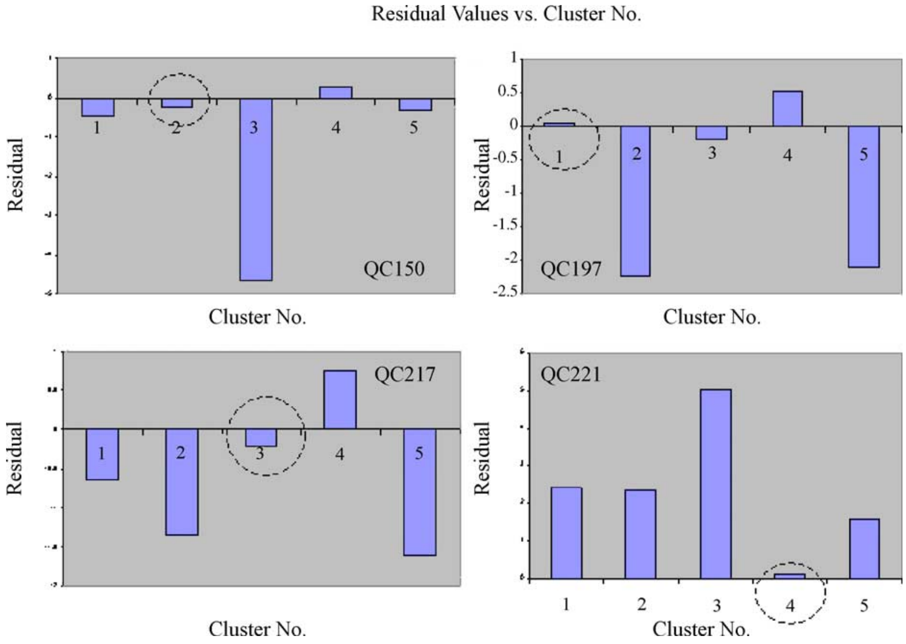

Fig. 7. Residual value plots of the query compounds in trial I.

dataset. Three fully connected, feed-forward layers of neurons comprise the basic structure of the networks used in this study. These layers are called the input, hidden and output layers. The input layer accepts descriptor values and the hidden layer processes these values and passes them to the output layer, which produces a value of the acute toxicity. The number of input neurons was equal to the number of descriptors in the type I model. The network was trained with the Broyden–Fletcher–Goldfarb–Shanno optimization method [36] and used a simulated annealing optimization algorithm to select optimal starting weights and biases. During the information processing in the hidden or output layer neurons, a sigmoid transfer function was applied. Therefore, the CNN produced non-linear models. The network architecture with the lowest rms errors for the TSET and CVSET was selected as the best type II model.

### 4.3. Type III models

CNN was also used in type III model development. Here, CNN was implemented for both descriptor selection and model building. Thus, a type III model is a fully non-linear model. The descriptors composing a type III model were also chosen by simulated annealing, but with a CNN fitness function to determine the optimal set of descriptors from the reduced pool. The training was performed as described above with the selected best subsets of descriptors. The best type III model was chosen based on low rms errors of TSET and CVSET and its generalization ability.

## 5. Relationship determination

After model building, only one model among the three types of models was chosen for each cluster, and the prediction accuracy of that model was taken into consideration when determining the relationships between the degree of similarity and prediction accuracy. With the developed QSAR models, the query compounds' acute aquatic toxicity were predicted. In the four trials, 4, 9, 15 and 18 query compounds' activities were predicted, respectively, with the QSAR models generated with each cluster. With the resultant activity predictions of the query compounds, the residual values of each query compound were then determined by calculating the difference between the observed value and the predicted value given by each model. For each single query compound, the residual values given by all the clusters' model prediction were compared with the degree of similarity of that query

Table 5
Degree of similarity between query compounds and clusters in trial II

| Query compounds | Most similar → Least similar | | | | |
|---|---|---|---|---|---|
| | Cluster no. | | | | |
| **23** | 2 | 3 | 1 | 4 | 5 |
| **65** | 5 | 4 | 1 | 3 | 2 |
| **111** | 2 | 3 | 1 | 4 | 5 |
| **150** | 2 | 3 | 1 | 4 | 5 |
| **197** | 1 | 3 | 4 | 2 | 5 |
| **217** | 3 | 1 | 2 | 4 | 5 |
| **221** | 4 | 5 | 1 | 3 | 2 |
| **228** | 1 | 3 | 2 | 4 | 5 |
| **292** | 4 | 1 | 3 | 2 | 5 |

compound to all the clusters. QSAR plots and residual plots were generated in determining this relationship.

## 6. Results and discussion

There is a direct relationship between the residual values and the degree of similarity between the query compound and the cluster of compounds used to generate the model. In other words, the higher the degree of similarity between the query compound and the cluster, the lower the residual value of its prediction. This trend has been observed in all four trials involving testing with different numbers of clusters and query compounds with different distance metrics and linkage methods.

In this study, four trials were designed to test the hypothesis. The trials were designed to test the effects of varying cluster numbers, query compound numbers and different clustering methods. Information about how each trial was done is presented in Table 2. The results of each trial are presented and discussed in the following sections.

### 6.1. *Trial I*

In trial I, five clusters were formed with agglomerative hierarchical clustering. The clusters contained 61, 63, 55, 85 and 54 compounds, respectively. City-block was used as the distance metric and the single linkage method was used for objects fusion. Prior to clustering, four query compounds were randomly selected from the dataset. They were acetonitrile (**150**), 1-hydroxy-3-methoxybenzene (**197**), styrene (**217**) and pentabromophenol (**221**). The distance between each query compound and the centroid of each cluster in the feature space was computed. Acetonitrile was found to be closest to cluster 2, 1-hydroxy-3-methoxybenzene was most similar to cluster 1, styrene was most similar to cluster 3 and pentabromophenol to cluster 4. None of the four compounds was found to be most similar to cluster 5. The rank order of similarity between the four query compounds and each cluster is summarized in Table 3 with the cluster number that the query compound was most similar to listed first.

Table 6
Descriptors chosen for the QSAR model for the prediction of fathead minnow acute toxicity in trial II

| Label | Description |
|---|---|
| Cluster 1 | |
| KAPPA 2 | Kappa 2 index [37] |
| V5C | Atom valence-corrected fifth-order cluster [38] |
| S4PC | Fourth-order simple path cluster [38] |
| NC | Number of carbons |
| NAB | The count of aromatic bonds |
| 3SP2 1 | Count of $sp^2$-hybridized carbons bonded to three other heteroatoms |
| MDE 24 | Molecular distance edge between 2° and 4° carbons [40] |
| EMAX 1 | Maximum atomic e-state value [45] |
| Cluster 2 | |
| KAPPA 4 | Kappa 1 index-atom corrected [37] |
| V6P | Valence-corrected sixth-order path $\chi$index [38] |
| V7CH | Seventh-order valence chain [38] |
| MOLC 5 | Path three molecular connectivity [39] |
| NCl 7 | Number of chlorine atoms |
| MDE 13 | Molecular distance edge between 1° and 3° carbons [40] |
| MDE 24 | Molecular distance edge between 2° and 4° carbons [40] |
| Cluster 3 | |
| ALLP 4 | (Total number of weighted paths)/(number of atoms) [42,43] |
| V6P | Valence-corrected sixth-order path $\chi$index [38] |
| S6P | Simple sixth-order path $\chi$index [38] |
| S5C | Fifth-order mole connectivity [38] |
| NC | Number of carbons |
| 3SP2 | Count of $sp^2$-hybridized carbons bonded to three other heteroatoms |
| MDE 23 | Molecular distance edge between 2° and 3° carbons [40] |
| MDE 44 | Distance edge for 4° to 4° carbons [40] |
| Cluster 4 | |
| KAPPA 3 | Kappa 3 index [37] |
| ALLP 1 | Total number of paths [42,43] |
| V6PC 14 | Valence-corrected sixth-order path cluster $\chi$index [38] |
| NC | Number of carbons |
| NAB | The count of aromatic bonds |
| WTPT 2 | Average molecular ID (molecular ID/no. atoms) [41] |
| MDE 34 | Molecular distance edge between 3° and 4° carbons [40] |
| EMAX 1 | Maximum atomic e-state value [45] |
| Cluster 5 | |
| KAPPA 3 | Kappa 3 index [37] |
| ALLP 1 | Total number of paths [42,43] |
| S4PC | Fourth-order simple path cluster [38] |
| N5C | Number of fifth-order paths |
| MOLC 5 | Path three molecular connectivity [39] |
| NDB | Number of double bond |
| WTPT 2 | Average molecular ID (molecular ID/no. atoms) [41] |
| 3SP2 | Count of $sp^2$-hybridized carbons bonded to three other heteroatoms |

Following the QSAR model development procedures described in the previous section, one single best QSAR model was chosen from the pool of developed types I–III models for each cluster. Thus, five QSAR models were selected, one for each cluster, for the final query compounds' toxicity prediction. The descriptors included in these models are listed in Table 4. The model's QSAR plots, predicted $-\log(LC_{50})$ versus observed $-\log(LC_{50})$, are shown in Fig. 6. The rms errors are included in each plot. They ranged from 0.29 to 0.57 for TSETs, 0.24 to 0.91 for CVSETs and 0.89 to 2.19 for PSETs. The side plots show the query compounds' toxicity prediction accuracy. In each of the side plots, the query compound that was most similar to the presented cluster is pointed out with an arrow. In these plots, the differences in prediction accuracy between the query compound that was similar to the corresponding cluster and the rest of the query compounds are revealed. These plots show visually that the query compound that was most similar to the presented cluster was predicted more accurately than other less similar query compounds. The same conclusion was also drawn from the residual plots of Fig. 7. These four graphs plot the residual values of the four query compounds for each cluster's QSAR model prediction. In these graphs, the circled number indicates the cluster that the query compounds was most similar to. For example, query compound **150** had the greatest similarity with cluster 2. As can be seen from these graphs, the shortest bars (lowest residual values) also correspond to the cluster numbers that the query compound shared with the greatest similarity. From the results given in trial I, it was concluded that the QSAR models had higher prediction accuracy on compounds that are similar in structure with the compounds used to build that QSAR model.

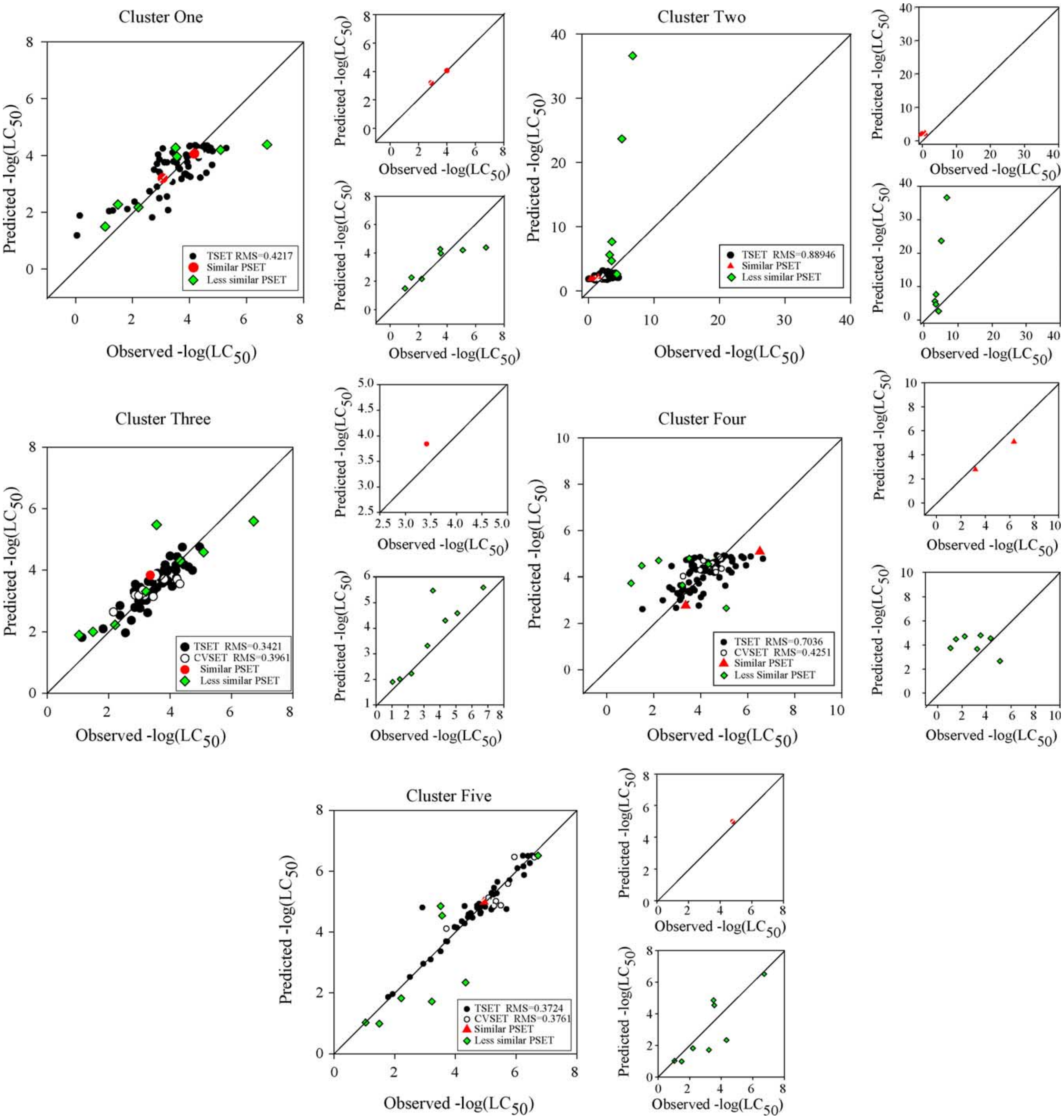

Fig. 8. Predicted –log(LC$_{50}$) vs. observed –log(LC$_{50}$) of the five developed QSAR model in trial II.

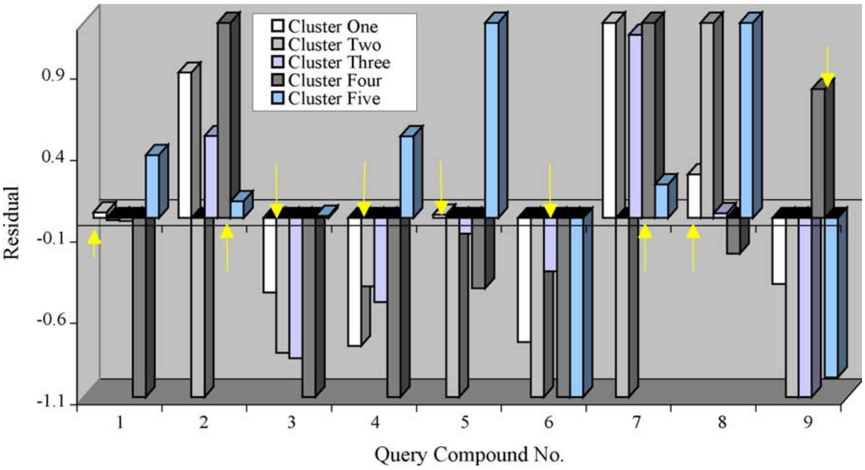

Fig. 9. Overall view of the residual values of all query compounds in trial II.

### 6.2. Trial II

Were the four query compounds predicted well due to chance? Further experiments were carried out with five additional query compounds withheld from the dataset. Five clusters were formed with agglomerative hierarchical clustering, and they contained 60, 62, 54, 84 and 53 compounds, respectively. Nine query compounds were excluded, including the four tested in trial I. The five additional compounds were 1-pentanol (**23**), 1-bromoheptane (**65**), dimethoxy methane (**111**), 1-chloro-4-methylbenzene (**228**) and 3-(trifluoromethyl) benzonitrile (**292**). Of the nine compounds, two compounds (**197** and **228**) were found to be most similar to cluster 1, three compounds (**23**, **111** and **150**) were most similar to cluster 2, one compound (**217**) was most similar to cluster 3, two compounds (**221** and **292**) were most similar to cluster 4 and one compound (**65**) was most similar to cluster 5. The rank order of similarity between query compounds and clusters is presented in Table 5.

As in trial I, five best QSAR models were selected for the final prediction step. The descriptors used in the five models

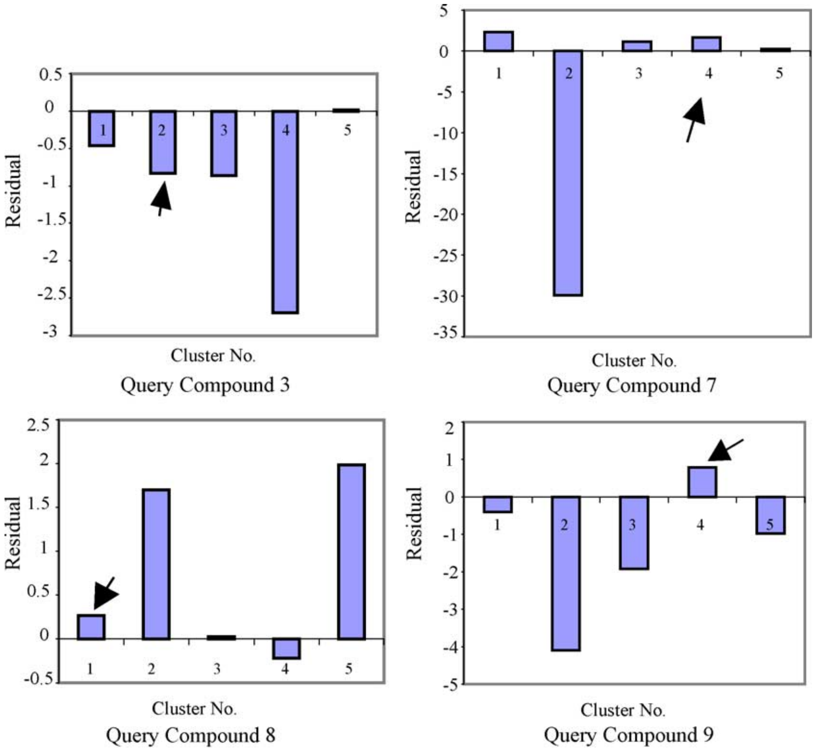

Fig. 10. Residual values plots of the not-well predicted query compounds in trial II.

are listed in Table 6. For each model, Fig. 8 shows the predicted –log($LC_{50}$) versus the observed –log($LC_{50}$). Two small side plots are also shown next to each cluster's main QSAR plot. The upper plot shows the predictions of query compounds that were similar to that cluster, and the bottom plot shows the predictions of query compounds, which were less similar to that cluster. The rms errors of TSETs are shown with each QSAR plot. They ranged from 0.34 to 0.90 log units. Some models presented might not be excellent models; the purpose here was to compare the prediction accuracy with the degree of similarity between query compounds and clusters. Through close examination of each cluster's two side plots, it was easy to discover the trend that occurred with the query compounds' predictions. With the more similar query compounds, the QSAR models were able to provide better predictions. The more similar query compounds appear closer to the diagonal line, whereas less similar query compounds are randomly scattered in the plots. Therefore, it was concluded that of the nine query compounds, the ones that were more similar to a particular cluster were predicted more accurately by the corresponding QSAR model. This conclusion was drawn from a general view of the QSAR plots. Would the same conclusion hold true if the prediction of compounds were examined individually?

An overview of the residual values of all the query compounds is presented in Fig. 9. In this figure, each query compound's residual values given by all five models are represented by bars with arrows pointing to the clusters that they were most similar to. Careful study of the graph shows that query compounds most similar with the corresponding cluster were predicted the best by that model, except for four query compounds. For these exceptions, the compounds were predicted well by another cluster's model that they were not most similar to. Query compound 3, dimethoxymethane (**111**), which was most similar to cluster 2 was predicted more accurately by cluster 5's QSAR model. Query compound 7, pentabromophenol (**221**), was better predicted by cluster 5, not by its most similar cluster. The same behavior was observed with 1-chloro-4-methylbenzene (**228**) and 3-(trifluoromethyl) benzonitrile (**292**). Enlarged residual plots for these four compounds are shown in Fig. 10. To further investigate this behavior, the similarity between clusters was determined. We suspected that the correlation between clusters might be playing a role here. The cluster similarity was determined by calculating the distances between the centroids of each pair of clusters. The similarity between the query compound and the clusters was also determined by an atom pair similarity calculation in addition to the previous determination done by distance similarity measure. The earlier determination showed that QC 3 was most similar to cluster 2, followed by cluster 3, 1, 4 and 5. The atom pair similarity indicated the following similarity order: cluster 2, 1, 5, 4 and 3 (from most to least similar). Cluster 2 was found to be most similar to cluster 3. QC 3 was predicted best by cluster 5's QSAR model, but none of the similarity determinations could explain it. Cluster 5 was neither similar with QC 3 nor close to cluster 2 in the feature space. The poor prediction with QC 3 was left

Table 7
Degree of similarity between query compounds and clusters in trial III

| Query compounds | Most similar → Least similar | | | |
|---|---|---|---|---|
| | Cluster no. | | | |
| **16** | 3 | 1 | 2 | 4 |
| **23** | 2 | 1 | 3 | 4 |
| **32** | 2 | 1 | 3 | 4 |
| **65** | 4 | 3 | 1 | 2 |
| **111** | 2 | 1 | 3 | 4 |
| **150** | 2 | 1 | 3 | 4 |
| **161** | 1 | 2 | 3 | 4 |
| **197** | 1 | 3 | 2 | 4 |
| **217** | 1 | 2 | 3 | 4 |
| **221** | 4 | 3 | 1 | 2 |
| **228** | 1 | 2 | 3 | 4 |
| **235** | 3 | 1 | 2 | 4 |
| **261** | 1 | 2 | 3 | 4 |
| **268** | 3 | 1 | 2 | 4 |
| **292** | 1 | 2 | 3 | 4 |

Table 8
Descriptors chosen for the QSAR model for the prediction of fathead minnow acute toxicity in trial III

| Label | Description |
|---|---|
| Cluster 1 | |
| S6CH | Index of chains of length six [38] |
| N3P | Number of third order paths [38] |
| 2SP3 1 | No. $sp^3$ carbons attached to two carbon atoms |
| EMAX 1 | Maximum atomic e-state value [45] |
| Cluster 2 | |
| V6P | Valence-corrected sixth-order path $\chi$ index [38] |
| V7CH | Seventh-order valence chain [38] |
| MOLC 8 | Weighted fourth-order path/clusters [39] |
| NAB | The count of aromatic bonds |
| EMAX 1 | Maximum atomic e-state value [45] |
| PND 1 | All pendant vertices [44] |
| Cluster 3 | |
| ALLP 1 | Total number of paths [42,43] |
| NLP 19 | Number of lone pairs |
| MDE 13 | Molecular distance edge between 1° and 3° carbons [40] |
| GEOM 3 | Z-principal geometric moment |
| Cluster 4 | |
| V4P | Fourth-order valence path molecular connectivity [38] |
| V6P | Sixth-order valence path molecular connectivity [38] |
| V6CH | Sixth-order valence chain [38] |
| V7CH | Seventh-order valence chain [38] |
| NLP | Number of tone pairs |
| MDE 11 | Molecular distance edge between primary carbons [40] |
| MDE 33 | Molecular distance edge between tertiary carbons [40] |
| ELOW 1 | Through space distance between EMIN and EMAX [45] |

unexplained. The second questionable query compound, QC 7, was predicted best by cluster 5's QSAR model. Our previous similarity calculation showed that it was most similar to cluster 4, followed by cluster 5, 1, 3 and 2. However, the atom pair similarity showed this compound was closest to cluster 5, which was also the cluster that provided the best prediction. Therefore, QC 7 was predicted best by the QSAR model of the cluster it was most similar to, according to atom pair similarity. The same reasoning applied with QC 8. Cluster 1, determined initially, was most similar to QC 8. Later determination by atom pair similarity showed cluster 3 was the most similar cluster. This aligned with the fact that QC 8 was predicted best by cluster 3. The last questionable compound, QC 9, was studied as above. QC 9 was most similar to cluster 4 but was predicted best by cluster 1. This behavior might be explained by the fact that cluster 1 was closest to cluster 4. Therefore, QC 9 was also very similar to cluster 1.

In this trial, five query compounds were well predicted by the QSAR model developed from the cluster they were most similar to. Only four questionable query compounds were not predicted best by the cluster's QSAR model that they should be. The four questionable compounds' similarity with each cluster was determined again with atom pair similarity. The atom pair similarity showed the cluster that provided the best prediction was in fact the one they were most similar to. Only one query compound was left unexplained. Overall, the majority of the query compounds were predicted well by the QSAR model of the cluster they were most similar to. The atom pair similarity was also tested on all nine-query compounds. The results coincided with those determined using the distance similarity measure.

### 6.3. Trial III

In trial III, we experimented with fewer clusters and more query compounds. Four clusters were formed containing 111, 61, 82 and 53 compounds, respectively. Fifteen query compounds were excluded for the study. They were 5-nonanone (**16**), 1-pentanol (**23**), 2-butanol (**32**),

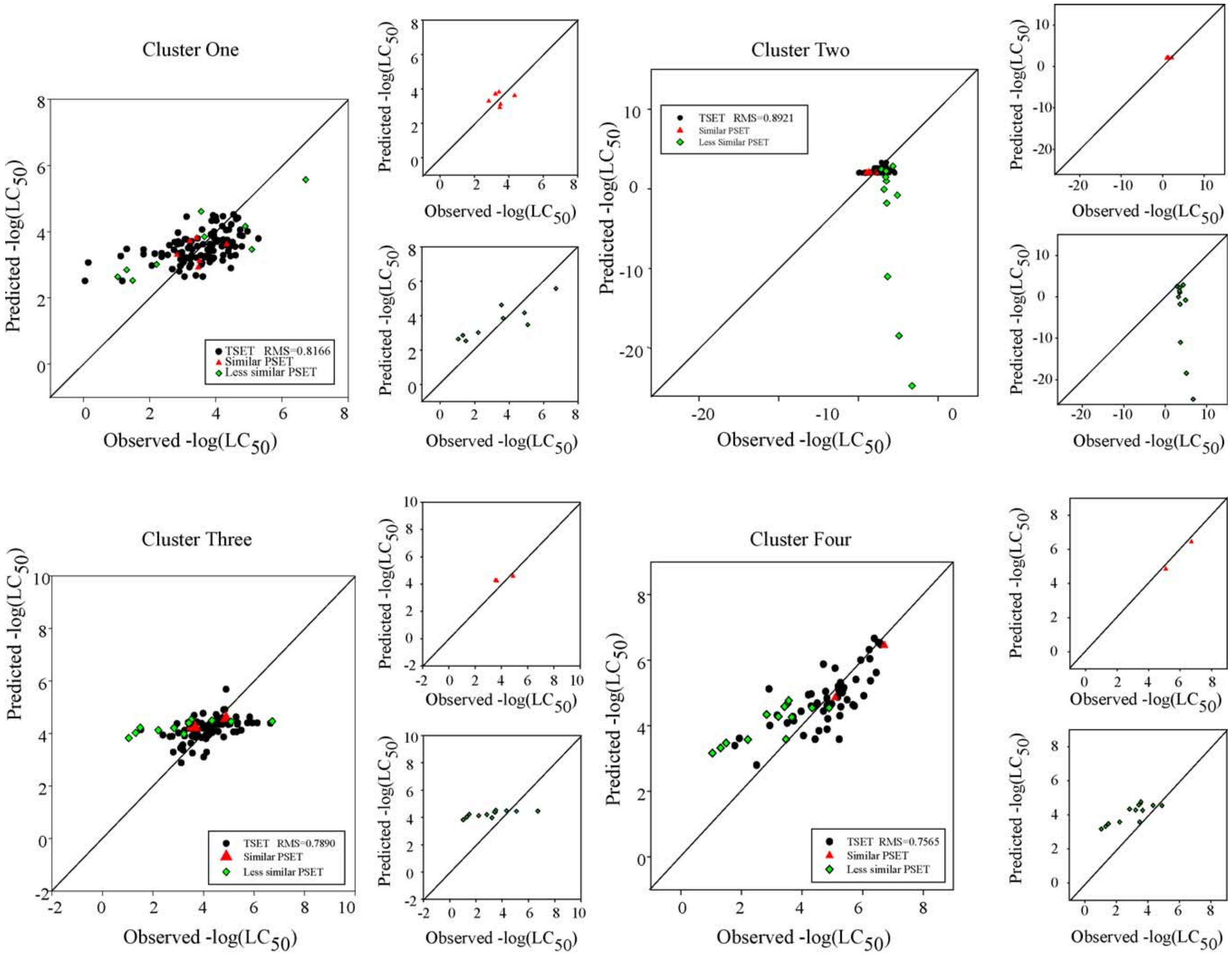

Fig. 11. Predicted –log(LC$_{50}$) vs. observed –log(LC$_{50}$) of the five developed QSAR model in trial III.

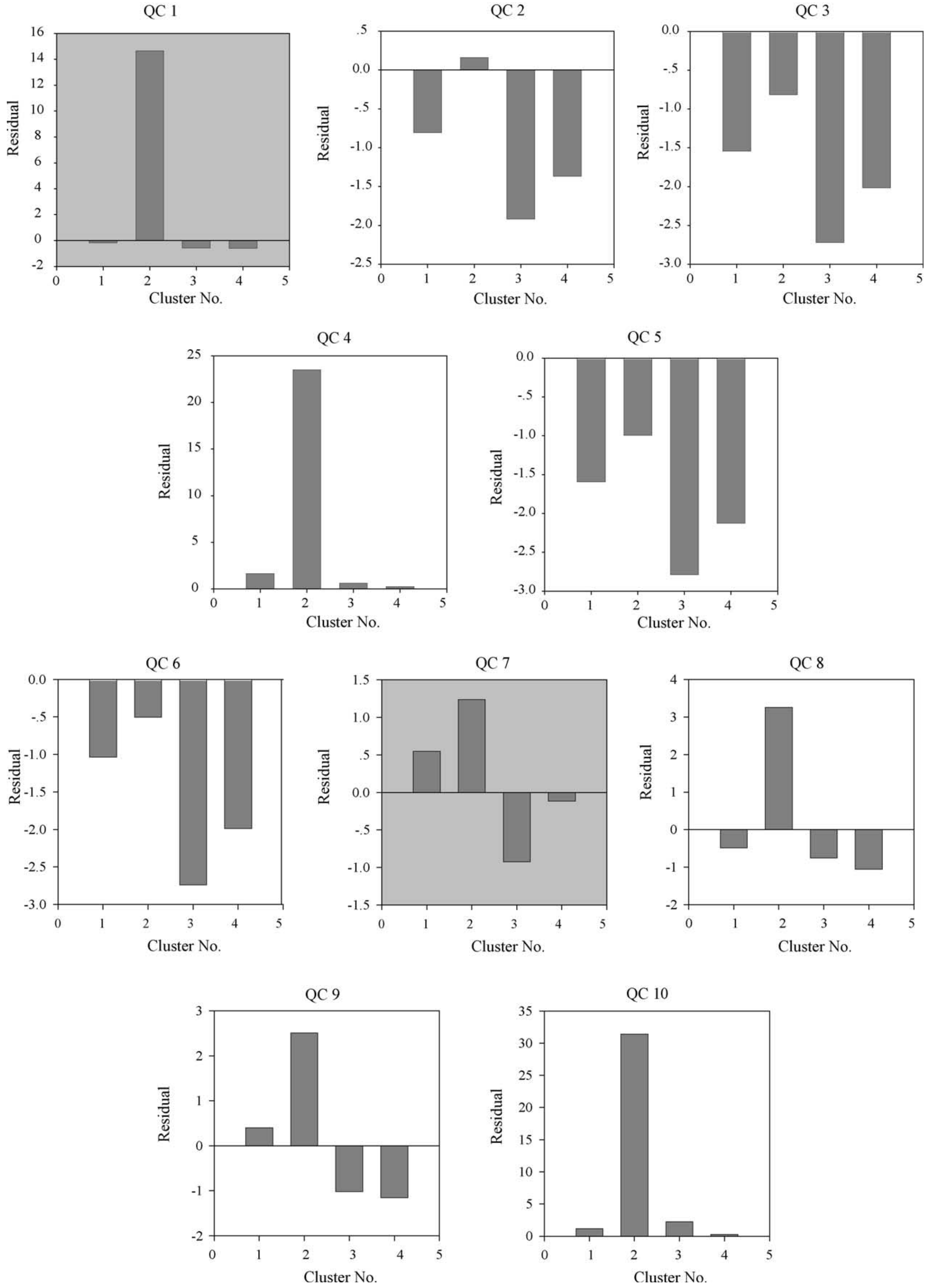

Fig. 12. Residual values plots of the not-well predicted query compounds in trial III.

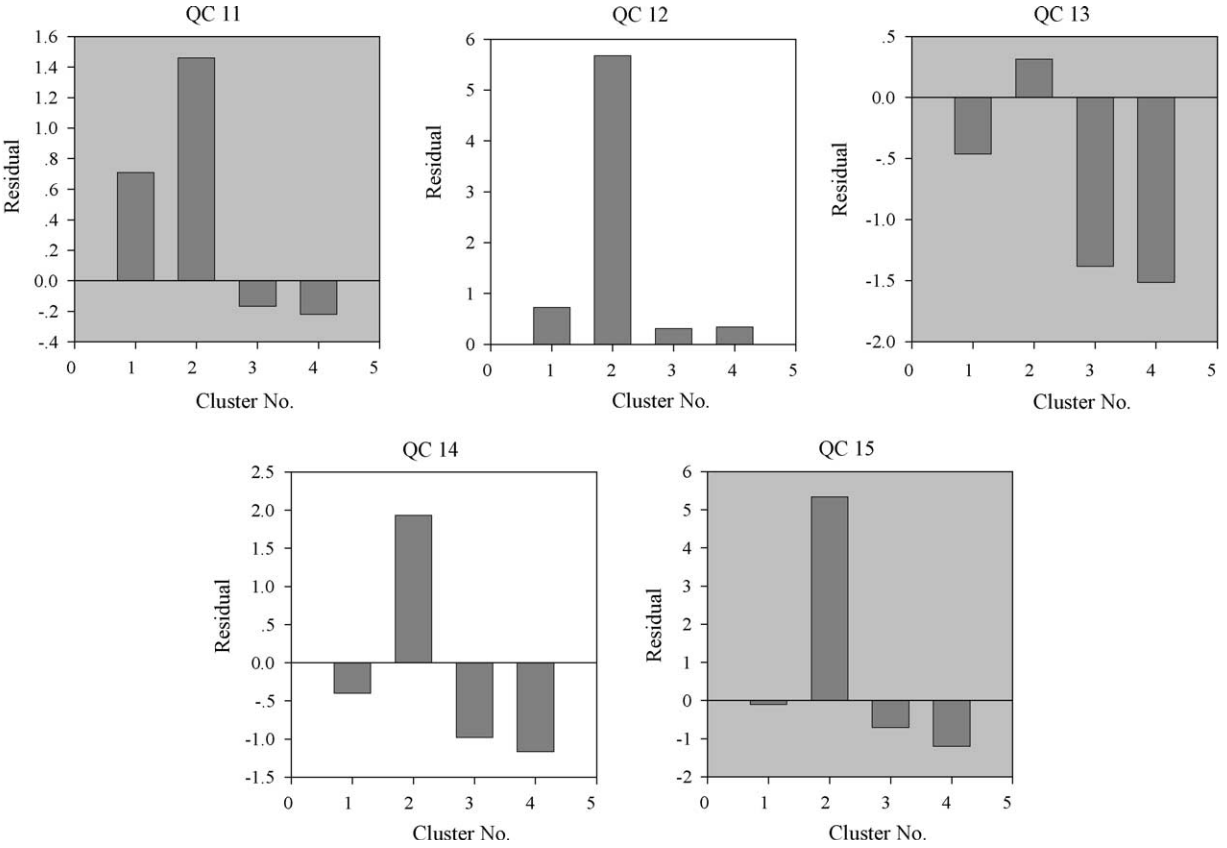

Fig. 12. (*Continued*).

1-bromoheptane (**65**), dimethoxymethane (**111**), acetonitrile (**150**), trichloroethylene (**161**), 1-hydroxy-3-methoxybenzene (**197**), styrene (**217**), pentabromophenol (**221**), 1-chloro-4-methylbenzene (**228**), 1,2,3-trichlorobenzene (**235**), 4-methylaniline (**261**), 1-cyano-2-methylbenzene (**268**) and 3-(trifluoromethyl)benzonitrile (**292**). City-block

Table 9
Degree of similarity between query compounds and clusters in trial IV

| Query compounds | Most similar → Least similar | | |
|---|---|---|---|
| | Cluster no. | | |
| **16** | 2 | 1 | 3 |
| **23** | 1 | 2 | 3 |
| **32** | 1 | 2 | 3 |
| **65** | 3 | 2 | 1 |
| **111** | 1 | 2 | 3 |
| **150** | 1 | 2 | 3 |
| **161** | 1 | 2 | 3 |
| **187** | 2 | 1 | 3 |
| **197** | 1 | 2 | 3 |
| **211** | 2 | 1 | 3 |
| **217** | 1 | 2 | 3 |
| **228** | 1 | 2 | 3 |
| **235** | 2 | 1 | 3 |
| **261** | 1 | 2 | 3 |
| **268** | 1 | 2 | 3 |
| **292** | 2 | 1 | 3 |
| **298** | 2 | 3 | 1 |
| **316** | 2 | 1 | 3 |

Table 10
Descriptors chosen for the qsar model for the prediction of fathead minnow acute toxicity in trial IV

| Label | Description |
|---|---|
| Cluster 1 | |
| S6P | Sixth-order simple path molecular connectivity [38] |
| NO 3 | Number of oxygen atoms |
| MDE 12 | Molecular distance edge between 1° and 2° carbons [40] |
| ESUM 1 | Sum of E-state values over all heteroatoms [45] |
| PND 3 | Superpendentic index over all pendant nitrogen atoms [44] |
| PND 5 | Superpendentic index from pendant O atoms [44] |
| Cluster 2 | |
| KAPPA 2 | Kappa 2 index [37] |
| V6P | Sixth-order valence path molecular connectivity [38] |
| V6CH | Sixth-order valence chain [38] |
| NO 3 | Number of oxygen atoms |
| NSB 12 | Number of single bonds |
| ESUM 2 | Sum of E-state values over all heteroatoms [45] |
| Cluster 3 | |
| KAPPA 2 | Kappa 2 index [37] |
| V6P | Sixth-order valence path molecular connectivity [38] |
| V6CH | Sixth-order valence chain [38] |
| V7CH | Seventh-order valence chain [38] |
| S7CH | Seventh-order simple chain [38] |
| N3P | Number of third-order paths [38] |
| NSB 12 | Number of single bonds |
| NBr | Number of basis rings |
| ESUM 2 | Sum of E-state values over all heteroatoms [45] |

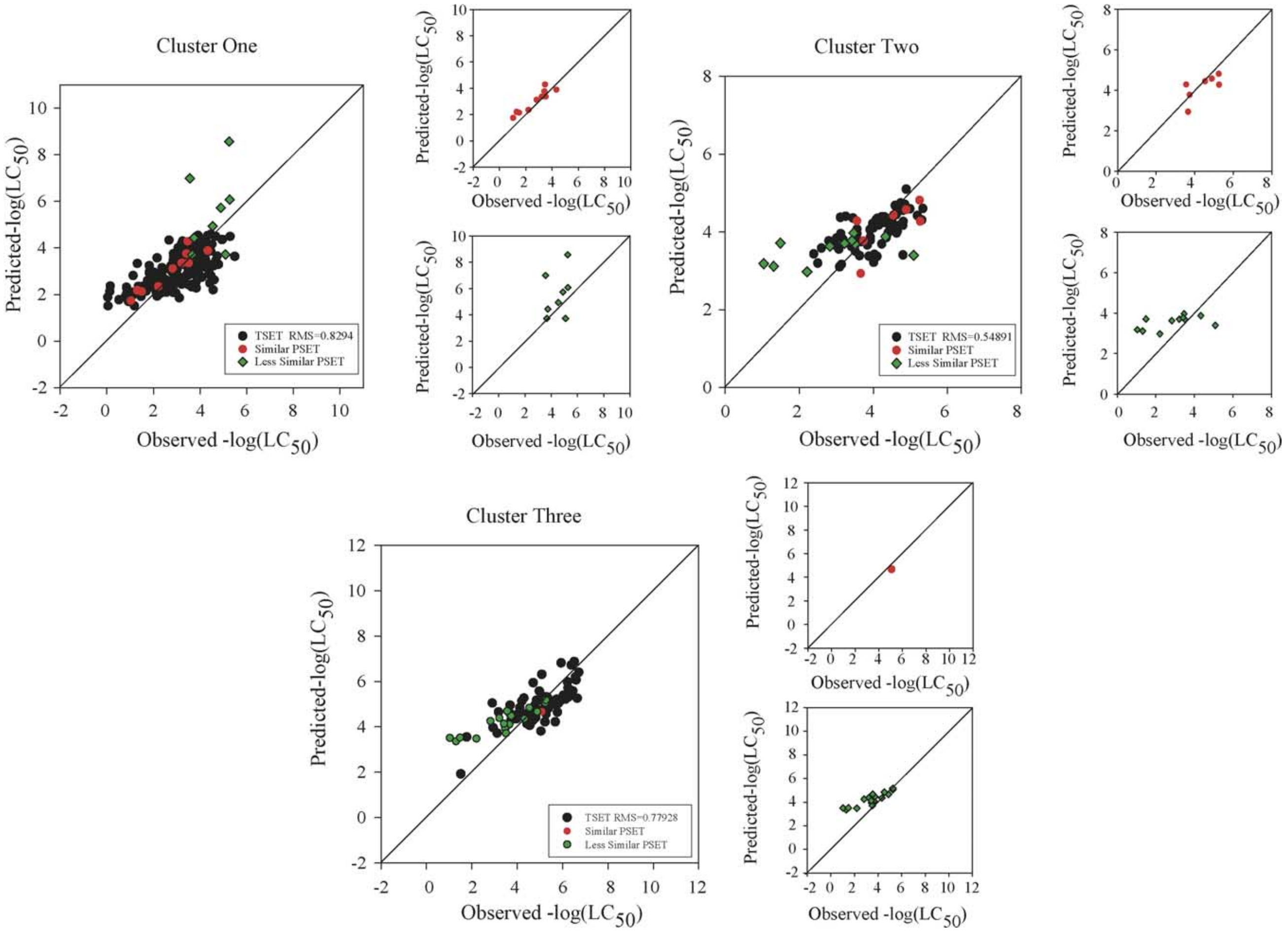

Fig. 13. Predicted –log($LC_{50}$) vs. observed –log($LC_{50}$) of the five developed QSAR model in trial IV.

and single linkage were used as the similarity measure and fusion methods. Prior to model development and toxicity prediction, the similarity between clusters and the similarity between query compounds and clusters were determined. Table 7 lists the rank order of the similarity of the 15 query compounds with each cluster. Six compounds were closest to cluster 1. Cluster 2 had four most similar compounds. Three compounds were most similar to cluster 3 and two were most similar to cluster 4. QSAR models were again developed with each cluster. Four best QSAR models were chosen for the final prediction. The descriptors in these models are shown in Table 8. With these four models, the less similar query compounds were poorly predicted. ''Less similar compounds'' means they were further away from the clusters in feature space. The QSAR plots in Fig. 11 showed the comparison of the prediction accuracy between the ''more similar'' query compounds and ''less similar'' query compounds given by each cluster's QSAR model. Again, predicted –log($LC_{50}$) versus observed –log($LC_{50}$) was plotted with two side small plots with the top one representing the ''more similar'' compounds and the bottom one representing the ''less similar'' compounds. For each QSAR plot, the more similar query compounds with that cluster were predicted more accurately than the less similar ones. The ''less similar'' query compounds were poorly predicted, as the diamonds representing them were located far away from the diagonal line. The residual plots in Fig. 12 also indicated that the higher the similarity between the query compound and the cluster, the better the prediction on the query compounds. The individual residual plots shown with shaded background represent the query compounds that fell out of the trend. They were predicted well by clusters other than the ones they were most similar to. Even though the predictions were not what we expected, the similarity between clusters showed QC 1, QC 11 and QC 15 were best predicted by a cluster that was most similar to the cluster that they should be predicted the best. In other words, in these plots the cluster that showed the lowest residual values were in fact the ones that were most similar to the cluster which the query compound was most similar to. The only query compound that was unexplainable was QC 7. QC 1 was predicted best by cluster 1 but not cluster 3, which could be due to the fact that cluster 1 is most similar to cluster 3. QC 11 was most similar to cluster 1, but was predicted most accurately by cluster 3, the most similar cluster-to-cluster 1. For QC 13, cluster 2 had the lowest residual value, and it is

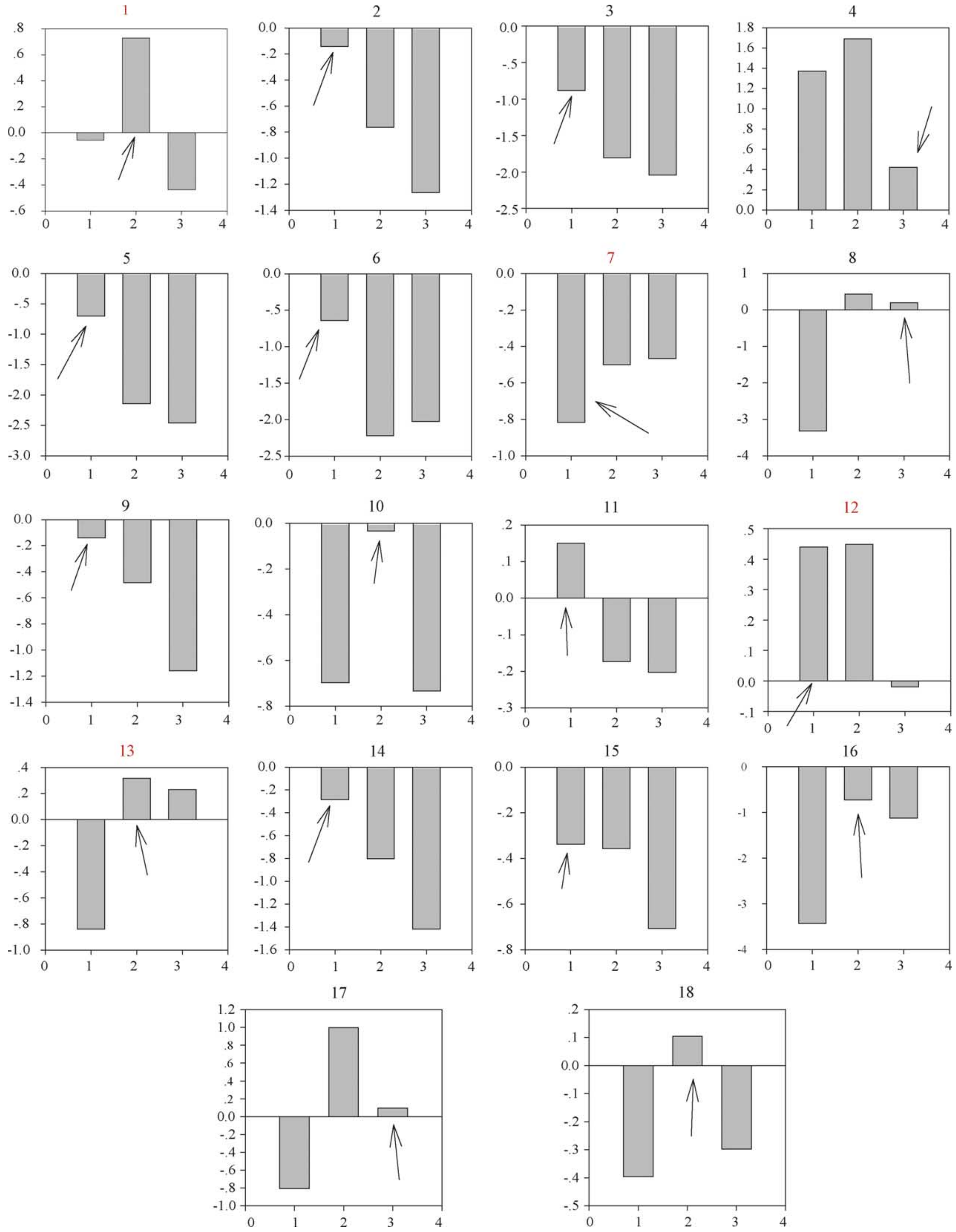

Fig. 14. Residual values plots of the not-well predicted query compounds in trial IV.

also most similar to cluster 1, which QC 13 was most similar to. The same reasoning applied with QC 15. Cluster 3 is most similar to cluster 1, cluster 1 had the lowest residual value. In summary, the questionable query compounds were all predicted accurately by a QSAR model of their second most similar cluster. These clusters were also the ones that were most similar to the query compounds' first most similar cluster.

### 6.4. Trial IV

In the last trial, a completely different similarity measure and fusion method was used from previous trials. They were standardized Euclidean distance and the average linkage method. Here, only three clusters were formed, which contained 175, 67 and 62 compounds. Eighteen query compounds (**16**, **23**, **32**, **65**, **111**, **150**, **161**, **187**, **197**, **211**, **217**, **228**, **235**, **261**, **268**, **292**, **298** and **316**) were used for toxicity prediction. They included some compounds that were tested in the previous trials. Four new compounds were pentafluorobenzaldehyde (**187**), 4-ethoxybenzaldehyde (**211**), phenylsalicylate (**298**) and 1-naphthalenol (**316**). Ten compounds were closest to cluster 1 (QC 2, 3, 5, 6, 7, 9, 11, 12, 14 and 15), seven compounds were closest to cluster 2 (QC 1, 8, 10, 13, 16–18) and one compound was most similar to cluster 3 (QC 4). The rank order of the similarity with each cluster is shown in Table 9. The three best QSAR models, selected as described in the previous section, contained 21 descriptors. The descriptors used are shown in Table 10. Next, the QSAR plots and residual plots were generated as in previous trials. They are shown in Figs. 13 and 14. The rms errors of the three QSAR models ranged from 0.55 to 0.83 log units. They all provided better predictions on ‘‘closer’’ query compounds than the further ones, relatively to each cluster. The two side plots show this observation clearly. In the individual residual plots, arrows are drawn to the cluster that the query compound was most similar to. In the ideal situation, we would like to see the arrows pointing to the lowest residual values, the shortest bars. That would indicate that the query compound was predicted best by the QSAR model generated from the cluster it was most similar to. Here, four cases deviated from expectations. They were QC 1, 7, 12 and 13. For QC 1, 7 and 13, even though they were not predicted well by their most similar cluster, they were predicted accurately with a cluster that was very similar to their most similar cluster. The only unexplainable prediction was QC 12.

Although a different similarity measure and fusion method was implemented in this trial, the same conclusion was drawn as in the other three trials. Query compounds that were more similar to the training set compounds used to generate the QSAR model would be generally predicted well by that model. Table 11 summarizes the studies of the four trials. The effects of varying cluster number, query compound number and different clustering methods on the final conclusions were studied through the four trials. From trial I to IV, cluster numbers from five to three were examined with increasing number of query compounds. The overall results showed that the majority of the query compounds were predicted more accurately by the QSAR model developed from the cluster they were most similar to.

Table 11
Final summary of the four trials

| | No. of clusters | No. of query compounds | No. of QC (well-predicted) | No. of QC (not well-predicted) |
|---|---|---|---|---|
| Trial I | 5 | 4 | 4 | 0 |
| Trial II | 5 | 9 | 5 | 4 |
| Trial III | 4 | 15 | 10 | 5 |
| Trial IV | 3 | 18 | 14 | 4 |

## 7. Conclusion

In this study, we have studied the relationship between the similarity of query compounds with the compounds used to generate the QSAR models, and the acute toxicity prediction accuracy given by the QSAR models. This relationship determination process with the fathead minnow dataset served as a means to examine the feasibility and effectiveness of the approach designed in this study for assessing the reliability of QSAR models' prediction in general. Our approach also implies a general methodology for activity prediction of a query compound. The activity could be predicted through first developing QSAR models for each distinguishable cluster found in a large dataset, and then predicting the query compound's activity using the most reliable model from the most similar cluster.

For the dataset employed in this study, we found that a QSAR model can reliably predict a query compound's activity if the query compound is sufficiently similar in structure to the group of compounds used to generate that QSAR model. Therefore, the two-step approach involving object clustering and activity prediction provided an efficient way to assess the reliability of a QSAR model's prediction. In conclusion, the new, non-computationally intensive and easily understood approach designed in this study not only solved the problem, but also should have practical applications in the QSAR field. Future work may include examining the approach with more datasets with different activities.

## Reference

[1] S. Borman, New QSAR techniques eyed for environmental assessments, Chem. Eng. News 68 (1990) 20–23.

[2] M. Nendza, C.L. Russom, QSAR modeling of the ERL-D fathead minnow acute toxicity database, Xenobiotica 21 (1991) 147–170.

[3] C.L. Russom, S.P. Bradbury, S.J. Borderius, D.E. Hammermeister, R.A. Drummond, Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales Promelas*), Environ. Toxicol. Chem. 16 (1997) 948–967.

[4] D.V. Eldred, C.L. Weikel, P.C. Jurs, K.L.E. Kaiser, Prediction of fathead minnow acute toxicity of organic compounds from molecular structure, Chem. Res. Toxicol. 12 (1999) 670–678.

[5] K.L.E. Kaiser, S.P. Niculescu, Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales Promelas*): a study based on 965 compounds, Chemosphere 38 (1999) 3237–3245.

[6] U.S. Environmental Protection Agency, AQUIRE database http://www.epa.gov/ecotox/, 2001.

[7] L.D. Newsome, D.E. Johnson, R.L. Lipnick, S.J. Broderius, C.L. Russom, A QSAR study of the toxicity of amines to the fathead minnow, Sci. Total Environ. 109/110 (1991) 537–551.

[8] L.H. Hall, E.L. Maynard, L.B. Kier, Structure–activity relationship studies on the toxicity of benzene derivatives: III. Predictions and extensions to new substituents, Environ. Toxicol. Chem. 8 (1989) 431–436.

[9] D.L. Geiger, C.E. Northcott, D.J. Call, L.T. Brooke (Eds.), Acute Toxicities of Organic Chemicals to Fathead Minnows (*Pimephales promelas*), 1–5, Center for Lake Superior Environmental Studies, University of Wisconsin, Superior, WI, 1984–1990.

[10] P.C. Jurs, J.T. Chou, M. Yuan, in: E.C. Olson, R.E. Christofferson (Eds.), Computer-Assisted Drug Design, American Chemical Society, Washington, DC, 1979.

[11] A.J. Tuper, W.E. Brugger, P.C. Jurs, Computer-Assisted Studies of Chemical Structure and Biological Function, Wiley, New York, 1979.

[12] MATLAB, http://www.mathworks.com.

[13] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2000(Chapter 8).

[14] J.P.P. Stewart, MOPAC: a semiempirical molecular orbital program, J. Comput. Aided Mol. Des. 4 (1990) 1–105.

[15] J.P.P., Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, Indiana University, Bloomington, 1990, Program 455.

[16] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.P.P. Steward, AM1: a new general purpose quantum mechanical molecular model, J. Am. Chem. Soc. 107 (1985) 3902–3909.

[17] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Analysis, Research Studies Press Ltd., John Wiley & Sons, New York, 1986.

[18] A.T. Balaban, Highly discriminating distance-based topological index, Chem. Phys. Lett. 89 (1982) 399–404.

[19] A.K. Madan, S. Gupta, M. Singh, Superpendentic index: a novel highly discriminating topological descriptor for predicting biological activity, J. Chem. Inf. Comput. Sci. 39 (1999) 9.

[20] L.B. Kier, L.H. Hall, Intermolecular accessibility: the meaning of molecular connectivity, J. Chem. Inf. Comput. Sci. 40 (2000) 792–795.

[21] R.H. Rohrbaugh, P.C. Jurs, Molecular shape and the prediction of high-performance liquid chromatographic retention indexes of polycyclic aromatic hydrocarbons, Anal. Chem. 59 (1987) 1048–1054.

[22] H. Goldstein, Classical Mechanics, Addison-Wesley, Reading, MA, 1950, pp. 144–156.

[23] R.S. Pearlman, in: S.H. Yalkowsky, A.A. Sinkula, S.C. Valvani (Eds.), Molecular Surface Area and Volumes and their Use in Structure/Activity Relationships, Marcel Dekker, New York, 1980.

[24] T.R. Stouch, P.C. Jurs, A simple method for the representation, quantification, and comparison of the volumes and shapes of chemical compounds, J. Chem. Inf. Comput. Sci. 26 (1986) 4–12.

[25] A.R. Katritzky, L. Mu, V.S. Lobanov, M. Karelson, Correlation of boiling pints with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple organics, J. Phys. Chem. 100 (1996) 10400–10407.

[26] S.L. Dixo, P.C. Jurs, Atomic charge calculations for quantitative structure–property relationships, J. Comput. Chem. 18 (1992) 492–504.

[27] R.J. Abraham, P.E. Smith, Charge calculations in molecular mechanics IV: a general method for conjugated systems, J. Comput. Chem. 13 (1987) 288–297.

[28] D.T. Stanton, P.C. Jurs, Development, Use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies, Anal. Chem. 62 (1990) 2323–2329.

[29] S.N. Vinogradov, R.H. Linnell, Hydrogen Bonding, Van Nostrand Reinhold, New York, 1971.

[30] C.J. Russell, S.L. Dixon, P.C. Jurs, Computer assisted study of the relationship between molecular structure and Henry's law constant, Anal. Chem. 64 (1992) 1350.

[31] J.M. Sutter, S.L. Dixon, P.C. Jurs, Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing, J. Chem. Inf. Comput. Sci. 35 (1995) 77–84.

[32] J.M. Sutter, P.C. Jurs, Selection of molecular descriptors for quantitative structure–activity relationships, in: J. Kalivas (Ed.), Adaption of Simulated Annealing to Chemical Optimization Problems, Elsevier Science Publishing Co., Amsterdam, 1995.

[33] D.T. Stanton, P.C. Jurs, M.G. Hicks, Computer-assisted prediction of normal boiling points of furans, tetrahydrofurans, and thiophenes, J. Chem. Inf. Comput. Sci. 31 (1991) 301–310.

[34] D.A. Belsey, E. Kuh, R.E. Welsch, Regression Diagnostics, Wiley, New York, 1980.

[35] M.D. Wessel, P.C. Jurs, Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks, Anal. Chem. 66 (1994) 2480–2487.

[36] L. Xu, J.W. Ball, S.L. Dixon, P.C. Jurs, Quantitative structure–property relationships for toxicity of phenols using regression analysis and computational neural networks, Environ. Toxicol. Chem. 13 (1994) 841–851.

[37] L.B. Kier, A shape index from molecular graphs, Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol. 4 (*3*) (1985) 109–116.

[38] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure Activity Analysis, Research Studies Press Ltd., John Wiley & Sons, 1986.

[39] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.

[40] S. Liu, C. Cao, Z. Li, Approach to estimation and prediction for normal boiling point (nbp) of alkanes based on a novel molecular distance edge (mde) vector, lambda, J. Chem. Inf. Comput. Sci. 38 (1998) 387–394.

[41] M. Randic, On molecular idenitification numbers, J. Chem. Inf. Comput. Sci. 24 (1984) 164–175.

[42] M. Randic, Comput. Chem. 3 (1979) 5.

[43] H. Wiener, J. Am. Chem. Soc. 69 (1947) 17.

[44] S. Gupta, M. Singh, A.K. Madan, Superpendentic index: a novel topological descriptor for predicting biological activity, J. Chem. Inf. Comput. Sci. 39 (1999) 272–277.

[45] L.B. Kier, L.H. Hall, An electrotopological-state index for atoms in molecules, Pharm. Res. 7 (1990) 801–807.