CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2006**, *5*, 431–446 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

August 2006, Volume 5, Number 8, Pages 431–446

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday

Application of a Fragment–based Model to the Prediction of the Genotoxicity of Aromatic Amines

Mose' Casalegno,[1] Emilio Benfenati,[1] and Guido Sello [2]

[1] IRFMN, Istituto di Ricerche Farmacologiche Mario Negri, via Eritrea 62, 20157 Milano, Italy
[2] Universita' degli Studi di Milano, Dipartimento di Chimica Organica e Industriale, via Venezian 21, 20133 Milano, Italy

Received: February 28, 2006; Revised: June 5, 2006; Accepted: June 9, 2006; Published: August 31, 2006

Citation of the article:
M. Casalegno, E. Benfenati, and G. Sello, Application of a Fragment–based Model to the Prediction of the Genotoxicity of Aromatic Amines, *Internet Electron. J. Mol. Des.* **2006**, *5*, 431–446, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Application of a Fragment–based Model to the Prediction of the Genotoxicity of Aromatic Amines

Mose' Casalegno,[1] Emilio Benfenati,[1] and Guido Sello [2,]*

[1] IRFMN, Istituto di Ricerche Farmacologiche Mario Negri, via Eritrea 62, 20157 Milano, Italy
[2] Universita' degli Studi di Milano, Dipartimento di Chimica Organica e Industriale, via Venezian 21, 20133 Milano, Italy

Received: February 28, 2006; Revised: June 5, 2006; Accepted: June 9, 2006; Published: August 31, 2006

Internet Electron. J. Mol. Des. **2006**, *5* (8), 431–446

Abstract

Motivation. Aromatic amines are well known mutagenic compounds, and their toxic effects have been thoroughly studied, thus making them a good test dataset for computational models. We developed a general approach to model compound toxicity, particularly to aquatic organisms. The model uses only diatomic fragments to estimate the compound biological response. We are now going to apply this model to a 95 compound dataset of aromatic amines to predict their genotoxicity. In particular, the computed activities are compared to the *Salmonella* mutagenicity tests that are a sufficiently homogeneous set of experimental data.

Method. The method used is very straightforward. Each molecule is dissected into diatomic fragments; each of them is represented by atom type, interatomic bond type, and neighboring bond type. In this way, the variability of atomic fragments is well represented and guarantees a good representation of the molecular differences. Statistical analysis uses both standard multilinear regression and neural network analyses.

Results. The initial dataset has been analyzed several times, using both the complete dataset and four partitions of this dataset in order to: (*a*) validate the model; (*b*) discuss the meaning of the result from a chemical viewpoint. The results are interesting because they demonstrate that our modeling approach is of general application and that the statistical analysis allows for the identification of some hints on the molecular characteristics that differentiate the biological activity of each compound.

Conclusions. This new application of our modeling approach demonstrates that it could be used in several different models and applications.

Keywords. Genotoxicity; aromatic amines; QSAR; diatomic fragments; MLR; neural network.

Abbreviations and notations	
QSAR, quantitative structure–activity relationships	MLR, multi linear regression
DFG, diatomic fragment	ANN, artificial neural network

1 INTRODUCTION

In the last three decades mathematical modeling has attained a special position in many chemical topics, from synthesis planning to structure activity–prediction. The use and manipulation of

Dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday.

* Correspondence author; phone: +39–02–50314107; fax: +39–02–50314106; E–mail: guido.sello@unimi.it.

theoretical molecular description, of interaction simulation, and of qualitative/quantitative correlation with experiments, have allowed the conception, realization, and application, of several original models that showed their usefulness and potency both in predicting unknown experimental results and in contributing to the better understanding of chemical data. On this line, toxicity prediction has moved its first steps more recently, as a consequence of the great interest shown by many parts towards the development of alternatives to experimental measures in order to reduce both costs and time [1–3]. Here, the number and types of different end points are particularly challenging, together with the variety and uncertainty of experimental data. In addition, the difficulty of the analyses at the molecular level due to the complexity of the modes of action made intricate the interpretation of the model reliability and applicability. Nevertheless, more and more efforts are put in the development of new models in the search of better hints to the solution of the toxicity modeling problem. The need to improve knowledge imposes the requirement of both restricting the compound diversity and making the effort of separating the diverse elements that determine the toxic effect at the molecular level.

Aromatic and heteroaromatic amines are well known mutagenic and carcinogenic compounds [4–5]. Many experimental data concerning their bio–activity have been collected, thus they can be appropriately used as a test dataset in the elaboration of new models. Moreover, the existence of many good models of their toxicity, mainly of their mutagenicity, stimulates a productive discussion [6–17]. In this paper, we are going to present our models for the prediction of amine genotoxicity

2 MATERIALS AND METHODS

2.1 Chemical Data

We use a dataset containing 95 aromatic amines that are quite representative of diverse structural features, as ring number, functional group type and position. In addition, this dataset has been used by several authors [6–10,12–24] for toxicity prediction. The compound structures are reported in the Supplementary Material.

2.2 Biological Data

The biological activity chosen to test the model of toxicity prediction is the compound mutagenicity as measured in the Ames test (*Salmonella typhimurium*) [4]. For the present work, we chose a dataset of experimental mutagenic potencies of aromatic amines (Table 1) towards *Salmonella typhimurium* TA98 + S9 microsomial preparation (expressed by the logarithm of the number of revertants per nanomole) previously utilized in the development of genotoxicity QSAR models [27].

2.3 Statistical Analyses

The statistical analyses are performed using MLR and neural network as present in the Statistica software package [25]. The parameters used in modeling are:

(*a*) MLR forward insertion: p1 = 0.05; F1 = 1.0; sweep delta = 10^{-7}; inverse delta = 10^{-12}; steps = 8.

(*b*) MLR Backward elimination: p2 = 0.05; F2 = 1.0; sweep delta = 10^{-7}; inverse delta = 10^{-12}; steps = 26.

(*c*) Neural network (multilayer perceptron): 1 input layer, 1 hidden layer, 1 output layer; hidden nodes: 16 when using 28 variables, 8 when using 8 or 9 variables; logistic interval of the regression output function = 0.9; uniform casual initialization; learning coefficient = 0.01; momentum = 0.3; mixed at every epoch; resampling with cross validation; 30 resamplings; first phase uses back propagation algorithm for 100 epochs, second phase uses conjugate gradient descent for 500 epochs. The partition of cases depends on the total case number and are: for 95 cases: 58 training cases, 29 selection cases, 3 test cases; for 77 cases: 38 training cases, 20 selection cases, 2 test cases.

2.4 Calculation of Diatomic Fragments

Structure descriptors are determined following the approach we previously developed and that was applied to the modeling of the aquatic toxicity of several compounds (96–h LC_{50}) for the fathead minnow (*Pimephales promelas*) [26].

Molecular structures are analyzed by extracting atomic reference indexes, atom types, connectivity matrices, and bond orders. All hydrogen atoms are deleted from the original structures, then saturation level is calculated; for example, for the carbon atom, C, we can get the following atom types: C, CH, CH_2, CH_3, with 4, 3, 2, and 1 bonding positions, respectively. The program detects aromatic rings in the molecules. To all bonds belonging to an aromatic ring, a bond order of 4 is assigned instead of the original alternating 1 – 2 values. Condensed aromatic systems are also considered, by increasing the bond order of common bonds by 1. For example, a bond order of 5 is obtained for the bond common to both naphthalene aromatic rings. The molecular representation described so far was used as the starting point to generate DFGs (Diatomic FraGments). A simple rule was adopted to break down each molecule. After the selection of each atom, each of its nearest neighbors was considered, and the resulting atomic pair was isolated from the surrounding atoms by cutting all bonds connecting it to other atoms, providing a DFG.

DFGs from breakdown of the dataset molecules were afterwards collected and compared to eliminate redundancies and to account for multiple fragments occurrences. Three features were thus used to univocally identify each DFG:

(*a*) The main bond order, of the bond joining the DFG constituent atoms.

(*b*) The elements of the constituent atoms, such as C, N, P, O, etc.

(*c*) The nearest neighbor bond orders, *i.e.*, the orders of all bonds connecting the constituent atoms with their nearest neighbors.

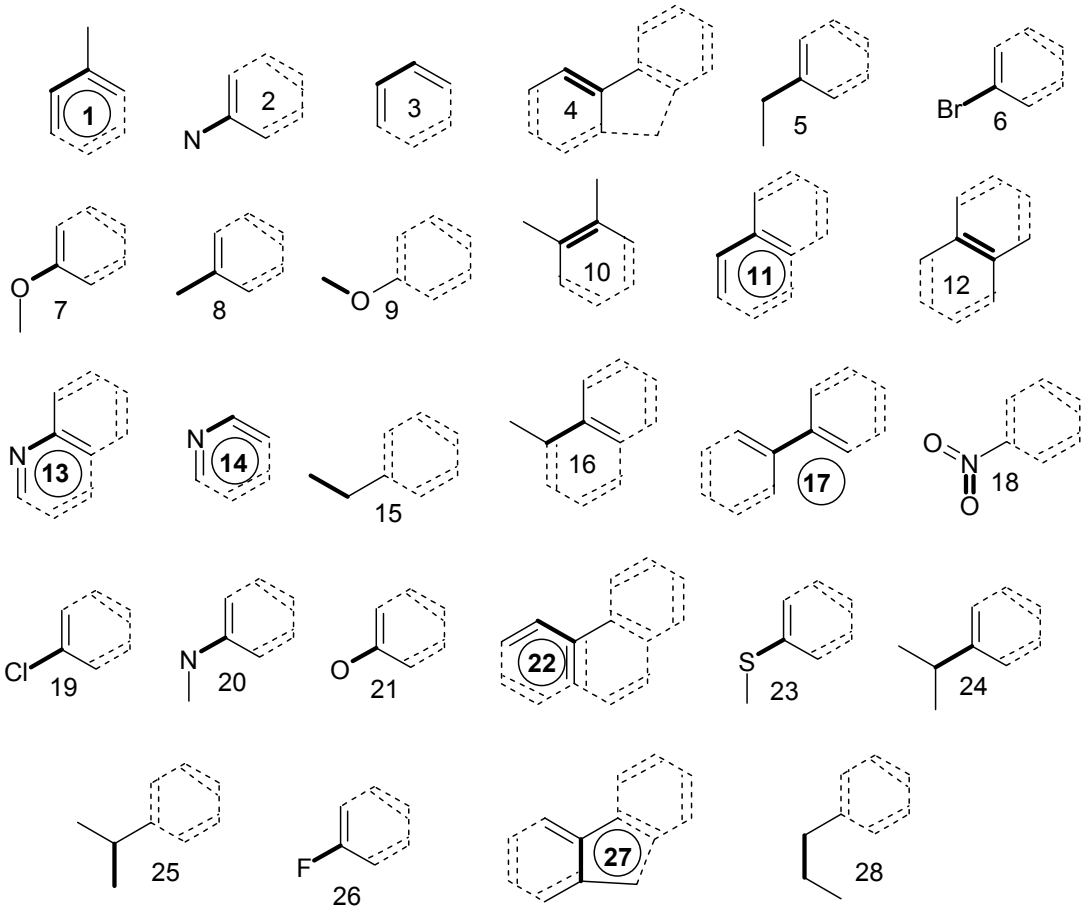

Figure 1. DFGs present in the dataset. Circled DFGs are those selected in almost all set partitioning.

This scheme gave N_{tot} = 28 not unique different DFGs, describing the entire dataset (see Figure 1). To each DFG we assigned an integer from 1 to N_{tot}. For each molecule a description array was then filled with integers accounting for fragments occurrences. For QSAR purposes, we created a matrix containing all descriptors and genotoxicity values. The matrix listing all molecules and their DFG representations was then partitioned four times. Each time a training set and a test set were obtained, with a ratio of 3:1. All 95 toxicity values were preliminary ordered in monotonic decreasing fashion. Then, beginning from the highest value, one out of four molecules was picked out and placed in the test set, leaving the other three in the training set. This procedure was repeated four times, picking out a different molecule each time. This partitioning guarantees that both the training set and the test set contain a representative amount of compounds showing different toxicity; in addition, all the compounds pass through the test phase. About 24 molecules were thus placed in each test set. The partitions obtained so far will be indicated as A, B, C, and D. This gave

us the opportunity to carry out a severe test to quantify the robustness and consistency of the method. The procedure of DFG selection has been repeated for each partition.

To facilitate the understanding of the DFG we will describe four examples: DFG 7, 9, 11, and 22. DFG 7 is defined by the oxygen atom and the aromatic carbon that is connected to it; the oxygen is also connected to an aliphatic carbon, the aromatic carbon is also connected to two other aromatic carbons. DFG 9 is defined by the oxygen atom and the aliphatic carbon connected to it; the oxygen is also connected to an aromatic carbon, the aliphatic carbon is unsubstituted. DFG 11 is defined by two aromatic carbons and the bond is part of only one ring; one aromatic C is connected to another aromatic carbon that has only one other aromatic bond, the other C is connected to two aromatic carbons, one mono–substituted, the second di–substituted. DFG 22 is defined by two aromatic carbons and the bond is part of only one ring; one aromatic C is connected to another aromatic carbon that has only one other aromatic bond, the other C is connected to two aromatic carbons, both of them are di–substituted.

3 RESULTS AND DISCUSSION

We performed several analyses using our descriptor set and biological data. Initially, the analysis was done including all compounds in one dataset; in fact, this approach has been used in all the previous studies concerning the aromatic amine genotoxicity. We used different statistical approaches, including: multilinear regression, with or without variable selection, and artificial neural network (multi layer perceptron), as nonlinear analysis. The results are shown in Table 1. In the last line correlation coefficients are reported; they show a good prediction (from 0.77 to 0.91) that remains sufficiently good even in the LOO calculation (0.67, or 0.70 eliminating compound **92** that represents a special case, v.i.). The LOO calculation was performed for the MLR analysis using all the variables; we did not make similar calculations for the other models because our validation is based on the more significant leave–many out partitioning described in section 2. This result should be compared to the result reported by Maran *et al*. [27] (0.83, and 0.81 cv), that is the best correlation found in the literature. In their paper Maran *et al*. used quantum chemical descriptors plus the variable accounting for the number of rings present in each structure; this last variable explains 68% of the variance. This result clearly demonstrates that the molecular skeleton is important in this dataset of compounds that, in principle, react following a common mechanism involving the amino group. The use of quantum descriptors modulates the activity through the modeling of both transport and electronic features. However, Maran *et al*., through an accurate use of several other descriptors that can be assigned to the transport effect, demonstrated that the use of the ring variable is sufficient to account for this effect. As a consequence, the role of the quantum chemical variables should be related to the modulation of the electronic interaction of the compound with the biological target.

Table 1. Analyses of the complete dataset

Comp.	log R_{Exp} [a]	MLR log R_{Pred} [b]	ANN log R_{Pred} [c]	Δ [d]	MLR common variables log R_{Pred} [e]	MLR forward insertion log R_{Pred} [f]	ANN forward insertion log R_{Pred} [g]	MLR backward removal log R_{Pred} [h]	ANN backward removal log R_{Pred} [j]
1	2.62	2.20	1.91	0.71	1.12	1.47	1.65	1.44	1.51
2	–2.05	–2.39	–2.24	0.19	–1.97	–1.94	–1.95	–1.97	–2.08
3	–2.00	–2.62	–2.11	0.11	–2.75	–2.75	–2.47	–2.74	–2.33
4	–2.30	–2.27	–2.44	0.14	–1.97	–1.96	–2.06	–1.97	–2.08
5	–0.60	–1.23	–0.63	0.03	–0.88	–0.89	–0.95	–0.86	–1.10
6	1.13	0.89	1.18	–0.05	0.50	1.02	1.37	1.03	1.15
7	2.62	2.21	2.09	0.53	2.91	2.91	2.83	2.90	2.77
8	2.88	3.35	2.99	–0.11	3.86	3.97	3.30	3.96	3.19
9	–1.14	–2.62	–2.11	0.97	–2.75	–2.75	–2.47	–2.74	–2.33
10	0.75	1.01	1.09	–0.34	0.99	0.99	0.91	0.98	0.97
11	–0.67	–0.47	–0.12	–0.55	0.14	0.14	–0.02	0.14	0.04
12	3.16	2.76	2.04	1.12	1.89	1.88	2.03	1.89	1.99
13	–0.55	–0.10	0.49	–1.04	–0.28	–0.33	–0.08	–0.31	0.03
14	–1.32	–1.69	–1.55	0.23	–1.97	–1.93	–1.71	–1.97	–2.08
15	0.89	1.08	1.40	–0.51	0.46	0.80	1.11	0.80	1.02
16	0.81	1.19	0.61	0.20	–0.28	–0.29	0.20	–0.31	0.03
17	–2.22	–2.15	–1.93	–0.29	–1.97	–1.94	–1.95	–1.97	–2.08
18	0.48	1.68	0.60	–0.12	1.12	1.47	1.65	1.44	1.51
19	3.31	2.78	2.94	0.37	2.79	2.71	3.11	2.72	3.16
20	1.93	1.08	1.40	0.53	0.46	0.80	1.11	0.80	1.02
21	–0.62	–0.29	–0.56	–0.06	–0.24	–0.10	–0.25	–0.08	–0.30
22	–0.14	–0.82	–0.06	–0.08	–0.94	–1.00	–0.70	–0.95	–0.79
23	–1.96	–2.39	–2.24	0.28	–1.97	–1.94	–1.95	–1.97	–2.08
24	0.60	–0.55	–0.34	0.94	0.46	–0.44	–0.33	–0.44	–0.10
25	–2.52	–2.68	–2.38	–0.14	–1.97	–1.94	–1.95	–1.97	–2.08
26	1.52	–0.60	0.53	0.99	–0.20	0.13	0.10	0.15	0.33
27	0.41	0.93	0.57	–0.16	1.12	1.47	1.65	1.44	1.51
28	2.38	1.57	2.21	0.17	1.43	1.43	2.21	1.44	2.39
29	–2.40	–2.15	–1.93	–0.47	–1.97	–1.94	–1.95	–1.97	–2.08
30	–0.92	–0.29	–0.56	–0.36	–0.24	–0.10	–0.25	–0.08	–0.30
31	–2.10	–2.85	–2.44	0.34	–1.97	–1.94	–1.95	–1.97	–2.08
32	0.55	1.17	0.77	–0.22	1.35	1.35	1.08	1.35	1.12
33	0.31	–0.36	–0.40	0.71	–0.66	–0.72	–0.99	–0.68	–0.87
34	–2.00	–1.82	–2.10	0.1	–1.97	–1.94	–1.95	–1.97	–2.08
35	–3.00	–2.83	–2.39	–0.61	–1.97	–1.94	–1.95	–1.97	–2.08
36	–2.70	–2.22	–2.59	–0.11	–1.97	–1.94	–1.95	–1.97	–2.08
37	–1.60	–1.32	–1.05	–0.55	–0.66	–0.72	–0.99	–0.68	–0.87
38	0.01	0.71	0.20	–0.19	–0.28	–0.29	0.20	–0.31	0.03
39	3.23	3.54	3.21	0.02	3.81	3.74	3.43	3.72	3.47
40	–0.89	–0.29	–0.56	–0.33	–0.24	–0.10	–0.25	–0.08	–0.30
41	3.35	3.35	2.99	0.36	3.86	3.97	3.30	3.96	3.19
42	–2.15	–1.45	–1.24	–0.91	–0.66	–0.72	–0.99	–0.68	–0.87
43	–2.52	–2.38	–2.27	–0.25	–1.97	–1.96	–2.06	–1.97	–2.08
44	2.46	2.33	2.59	–0.13	2.45	2.45	2.83	2.45	2.98
45	–3.32	–2.65	–2.44	–0.88	–1.97	–1.96	–2.06	–1.97	–2.08
46	2.98	1.57	2.21	0.77	1.43	1.43	2.21	1.44	2.39
47	–1.30	–0.22	–0.49	–0.81	–0.28	–0.33	–0.08	–0.31	0.03
48	3.50	3.52	2.61	0.89	2.91	2.91	2.83	2.90	2.77
49	–0.69	–1.16	–0.70	0.01	–1.97	–1.93	–1.71	–1.97	–2.08
50	1.18	1.41	1.43	–0.25	1.12	0.85	1.12	0.82	0.99
51	1.12	1.77	2.22	–1.1	2.01	2.02	1.83	1.99	1.89
52	–2.67	–1.86	–2.26	–0.41	–1.73	–1.72	–2.43	–1.73	–2.09
53	–3.00	–2.39	–2.24	–0.76	–1.97	–1.94	–1.95	–1.97	–2.08

Table 1. (Continued)

Comp.	log R_{Exp} [a]	MLR log R_{Pred} [b]	ANN log R_{Pred} [c]	Δ [d]	MLR common variables log R_{Pred} [e]	MLR forward insertion log R_{Pred} [f]	ANN forward insertion log R_{Pred} [g]	MLR backward removal log R_{Pred} [h]	ANN backward removal log R_{Pred} [j]
54	–1.30	–0.29	–0.56	–0.74	–0.24	–0.10	–0.25	–0.08	–0.30
55	–0.92	–0.41	–0.25	–0.67	–0.24	–0.10	–0.25	–0.08	–0.30
56	0.20	0.25	0.27	–0.07	–0.03	–0.04	–0.10	–0.02	0.00
57	–1.03	–0.36	–0.40	–0.63	–0.66	–0.72	–0.99	–0.68	–0.87
58	–0.54	–0.78	–1.11	0.57	–1.97	–1.93	–1.71	–1.97	–2.08
59	–2.70	–2.61	–2.19	–0.51	–1.97	–1.94	–1.95	–1.97	–2.08
60	–1.14	–0.09	–1.07	–0.07	–0.66	–0.72	–0.99	–0.68	–0.87
61	–1.49	–1.00	–0.58	–0.91	–0.90	–0.77	–0.98	–0.72	–0.99
62	0.04	0.25	0.27	–0.23	–0.03	–0.04	–0.10	–0.02	0.00
63	0.43	1.00	0.55	–0.12	–0.20	0.14	0.39	0.15	0.33
64	3.80	3.54	3.21	0.59	3.81	3.74	3.43	3.72	3.47
65	–3.00	–2.46	–2.40	–0.60	–2.62	–2.61	–2.36	–2.61	–2.58
66	–0.80	–0.80	–2.02	1.22	–1.97	–1.96	–2.06	–1.97	–2.08
67	–1.17	–0.60	–1.22	0.05	–0.88	–0.87	–0.99	–0.86	–1.10
68	0.69	–0.10	0.49	0.20	–0.28	–0.33	–0.08	–0.31	0.03
69	–2.70	–2.04	–1.87	–0.83	–1.97	–1.96	–2.06	–1.97	–2.08
70	–3.00	–2.61	–2.53	–0.47	–1.97	–1.94	–1.95	–1.97	–2.08
71	0.15	0.23	–0.13	0.28	–0.28	–0.29	0.20	–0.31	0.03
72	–1.24	–1.24	–0.95	–0.29	–1.97	–1.96	–2.06	–1.97	–2.08
73	0.38	–0.69	–0.72	1.10	–1.31	–1.39	–1.31	–1.32	–1.34
74	–0.99	–1.05	–1.20	0.21	–0.66	–0.69	–0.86	–0.68	–0.87
75	3.00	1.79	1.91	1.09	1.12	1.47	1.65	1.44	1.51
76	–0.39	–0.22	–0.49	0.10	–0.28	–0.33	–0.08	–0.31	0.03
77	–1.77	–1.28	–1.31	–0.46	–1.24	–1.25	–1.46	–1.23	–1.55
78	1.02	–0.10	0.49	0.53	–0.28	–0.33	–0.08	–0.31	0.03
79	1.04	–0.10	0.49	0.55	–0.28	–0.33	–0.08	–0.31	0.03
80	–0.01	0.41	0.11	–0.12	0.33	0.32	0.06	0.34	0.15
81	0.23	–0.59	–0.74	0.97	–0.66	–0.69	–0.86	–0.68	–0.87
82	–2.22	–1.74	–2.28	0.06	–1.97	–1.94	–1.95	–1.97	–2.08
83	–3.14	–1.86	–2.26	–0.88	–1.73	–1.72	–2.43	–1.73	–2.09
84	–0.48	–0.55	–0.34	–0.14	0.46	–0.44	–0.33	–0.44	–0.10
85	–0.49	–1.86	–1.11	0.62	–1.97	–1.94	–1.95	–1.97	–2.08
86	3.77	2.33	2.59	1.18	2.45	2.45	2.83	2.45	2.98
87	0.20	–0.22	–0.49	0.69	–0.28	–0.33	–0.08	–0.31	0.03
88	1.18	1.45	1.37	–0.19	1.89	1.88	2.03	1.89	1.99
89	–1.04	–0.62	–0.85	–0.19	–0.20	–1.09	–0.92	–1.08	–0.79
90	0.87	0.69	0.76	0.11	0.87	0.85	1.03	0.88	0.99
91	–1.42	–0.74	–0.48	–0.94	0.50	–0.21	–0.51	–0.20	–0.41
92	1.83	3.72	2.73	–0.90	3.74	3.74	3.00	3.74	3.06
93	1.43	2.76	2.04	–0.61	1.89	1.88	2.03	1.89	1.99
94	–1.77	–1.85	–2.12	0.35	–0.66	–0.69	–0.86	–0.68	–0.87
95	3.97	1.77	2.22	1.75	2.01	2.02	1.83	1.99	1.89
r^2 [k]		**0.85**	**0.91**		**0.77**	**0.78**	**0.84**	**0.78**	**0.84**
LOO r^2 [l]	**0.67/ 0.70**								

[a] Experimental values expressed as the logarithm of the number of revertants per nanomole. [b] MLR predicted values using all 28 variables. [c] ANN (16–0.9–30) predicted values using all 28 variables. [d] Residuals of the ANN prediction using all 28 variables. [e] MLR predicted values using 7 variables common to all other MLR predictions. [f] MLR predicted values using 8 variables selected by insertion. [g] ANN (8–0.9–30) predicted values using 8 variables selected by insertion. [h] MLR predicted values using 9 variables selected by removal. [j] ANN (8–0.9–30) predicted values using 9 variables selected by removal. [k] Regression coefficient. [l] Regression coefficient of Leave–one–out analysis; all cases and excluding compound **92**.

When using fragments we implicitly consider that the chemical structure contains all the information needed to describe the interaction of the compound with the environment. Therefore, the use of structural descriptors concerns both the reactivity and the transport. The use of fragments consequently needs a dataset where the structural characteristics are well represented, because the presence of the entire variability of the chemical structures is fundamental for the model application. The aromatic amine dataset is certainly a good example, because it is sufficiently homogeneous.

However, when modeling biological activity even small structural differences can highly influence the biological response. Another important point concerns the lack of interaction between fragments, *i.e.*, each fragment is analyzed as a single entity. It is thus impossible to evaluate long distance effects. The principal outcome of all these considerations is that an analysis that uses fragments can give a good result only if the differences in activity are well represented by the structural description.

In the present study the result is appreciably good. In fact all the analyses made on the whole dataset give good structure–activity correlations (see Table 1). The use of all the 28 fragments gives good predictions by both MLR and ANN analyses. In addition, decreasing the number of fragments to 8 or 9 (through the use of the forward insertion, or the backward elimination options) still yields good predictions, mostly by ANN. Finally, using the 7 variables present in the majority of the analyses it is possible to get a nice result, in line with all the MLR analyses.

Table 2. Analyses of the dataset partitions

Results of set partitions	MLR R^2 [b]	ANN R^2 [c]	MLR R^2 Common variables [d]	MLR R^2 [e]	ANN R^2 [f]	MLR R^2 [g]	ANN R^2 [h]
Training A	0.87	0.91		0.82	0.86	0.82	0.83
Test A	0.61	0.74	0.73	0.63	0.73	0.65	0.72
Training B	0.89	0.87		0.77	0.82	0.79	0.84
Test B	0.55	0.72	0.70	0.65	0.66	0.50	0.57
Training C	0.89	0.91		0.82	0.86	0.82	0.86
Test C	0.55	0.55	0.55	0.45	0.64	0.55	0.53
Test 3 w/o **92** [a]	0.62	0.56	0.66	0.51	0.67	0.58	0.54
Training D	0.85	0.89		0.79	0.84	0.79	0.84
Test D	0.79	0.66	0.74	0.75	0.75	0.75	0.76

[a] Regression coefficients excluding compound **92**. [b] MLR regression coefficients using all 28 variables. [c] ANN (16–0.9–30) regression coefficients using all 28 variables. [d] MLR regression coefficients using 7 variables common to all other MLR predictions. [e] MLR regression coefficients using 8 variables selected by insertion. [f] ANN (8–0.9–30) regression coefficients using 8 variables selected by insertion. [g] MLR regression coefficients using 9 variables selected by removal. [h] ANN (8–0.9–30) regression coefficients using 9 variables selected by removal.

The validation of the model has been performed following two approaches: the first is a leave–one–out validation; the second, much more significant, is performed dividing the dataset into two

parts, one for the model training, and the other for the model testing. This last approach was repeated four times using four different partitions (see Table 2). In all validations, the predictions of the training sets are very good, fully comparable to the whole set result. This confirms that the model is reliable and that the use of fragments can be applied to the present problem. In contrast, the results of the test sets are of lower quality and, more important, of different value (from 50% to 80% predictivity). In particular, partition C gives the worst test prediction and partition D gives by far the prediction of the best quality. It is also remarkable that the use of the subset of common variables gives a result for each test set similar to those obtained using the specific variables of the current partition.

Looking at some of the bad predicted compounds we can note that their structure contains some special features that can explain the outcome. For example, compound **92** is the only example of a benzophenanthrene compound. It is correctly predicted only in one analysis (28 variables, ANN) and it is particularly bad predicted when it is part of the test set (in partition C). A similar behavior can be observed for compound **95**. In this case, as mentioned by Maran *et al*., the reason of the high toxicity of this compound in comparison with those of the same class (**10**, **32**, **51**, **56**, **62**, and **80**) is not easily explicable.

Looking at the fragments selected in the regression procedure, it is possible to note that in all the partitions and in the whole set the selected fragments are often constant. In particular, fragments 1, 11, 13, 14, 17, 22, and 27 are almost always selected. Those fragments can be then considered as the most significant in the description of the structure variability. Interestingly, all these fragments concern special aromatic patterns, with or without heteroatoms; moreover, no fragment containing non–aromatic heteroatoms is selected. This outcome is in agreement with our expectations; in fact, this dataset is characterized by the differences in the carbon backbone that is mainly made by the aromatic substructure.

A feature of this dataset is its wide variability of toxicity values. This variability is not due to different modes of action; in fact, the mechanism of the genotoxicity of aromatic amines is connected to the presence of the amino group that, either after activation or directly, is responsible of the action. Therefore, the value variability should be related to the modulation of the remaining parts of the structure. Because all the compounds belong to the aromatic group, they have many substructures in common; consequently, some of their DFGs cannot be used to single them out. Obviously, the number of times (the occurrence number) a DFG is present in a molecule can be sufficiently informative (*e.g*., there are six unsubstituted aromatic C–C bonds in benzene, but only two in para–substituted benzene).

Nevertheless, other more specific characteristics are more useful. The variable selection that is automatically performed by Statistica shows the selection of one DFG that is important to signal the number of presence of common features (*e.g.*, DFG no. 1) and of several other DFGs that permit the assembly of different compound classes. For example, DFG no. 13 is characteristic of quinolines and phenazines; or DFG no. 17 is characteristic of biphenyls and fluorenes; those last are then separated by DFG no. 27. The structure descriptions that are produced by the model are sufficiently significant to separate the compounds by toxicity.

To further make clear the presentation we will comment one result. Consider compounds no. **7** and **17** (Figure 2), that have very different toxicities (2.62 and –2.22, respectively). Their complete descriptions (excluding single fragments) are shown in Table 3.

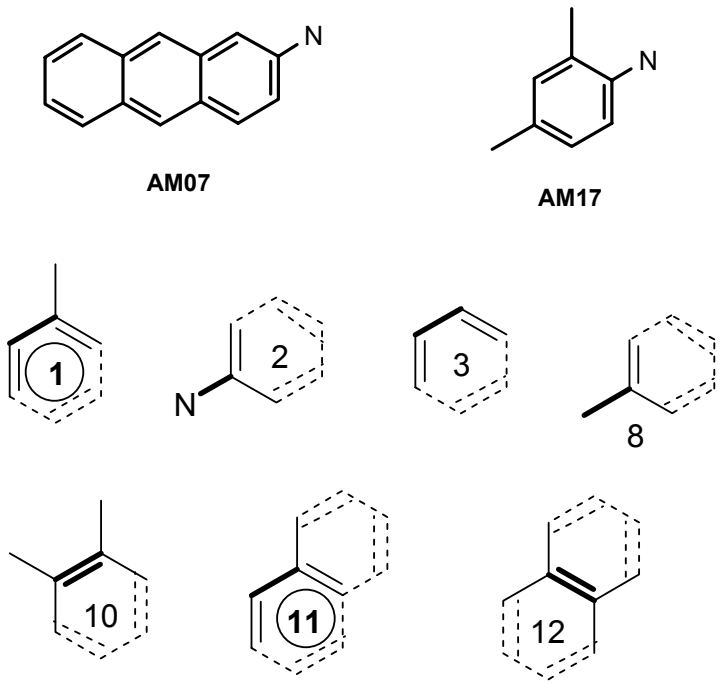

Figure 2. Compounds **7** and **17** and the corresponding DFGs.

Table 3. DFG Description of Compounds **7** and **17**

Occurrence for **7**	2	1	4	0	0	8	2
Occurrence for **17**	4	1	1	2	1	0	0
DFG [a]	**1**	2	3	8	10	**11**	12

[a] Identification label of DFG. Bold numbers indicate selected DFGs.

It is clear that DFG no. 2 is useless. Then, we can see two DFGs (1 and 3) that are common to both compounds; in this case the selection of any of them does not change the final result, thus DGF

1 is selected. Finally, there are two DFGs typical of compounds **7** (11 and 12) and two typical of compound **17** (8 and 10). To separate the two compounds it is sufficient to use one DFG; when decreasing the number of variables (forward insertion or backward elimination) the program selects DFG no. 11. The calculated toxicity values are: using all DFGs, 2.21 and –2.15, respectively; using only DFGs nos. 1 and 11, 2.91 and –1.97, respectively. The results are comparable.

We can finally note that in our best model the greatest error is < 1.8 unit, a result that is better than that of Maran *et al*. [27]. Also the standard error is lower [0.58 against 0.66, using MLR (forward insertion model) with 8 variables].

4 CONCLUSIONS

The application of our fragmentation procedure has permitted the realization of several models of the genotoxicity of a set of aromatic amines. The results demonstrate that the use of a small subset of fragments (7–8) is sufficient to describe the variability of toxicity for this dataset. The compounds in the dataset show highly different genotoxicities; however, they follow a single mechanism of action. The consequence is that the toxicity level is due to differences in structure features that, without changing the mechanism, affect the potency. This characteristic is particularly welcome by model based on fragments.

Our results confirm this hypothesis showing good predictions. More interesting is to note that the number of fragments needed to have a good prediction can be reduced without losing accuracy, a detail that is in agreement with the redundancy of the structural description in this application. Finally, it is also noteworthy the similar results obtained using both linear and non–linear models. Also this outcome confirms the homogeneous behavior of the dataset.

In this perspective, we can even think that the application of the fragment model can be used to cluster different modes of action if the weight of the fragments can be used to this end. This, also considering our previous application [26], supports our efforts in modeling compound toxicity using fragment–only description. In the future, the approach can be refined introducing a method to consider the fragment relative position in the structure; however, this task is not easy because the variability could be so high to prevent a statistically sound application.

Supplementary Material

The molecular structures of compounds **1-95** used in the QSAR models are given in Figure A1.

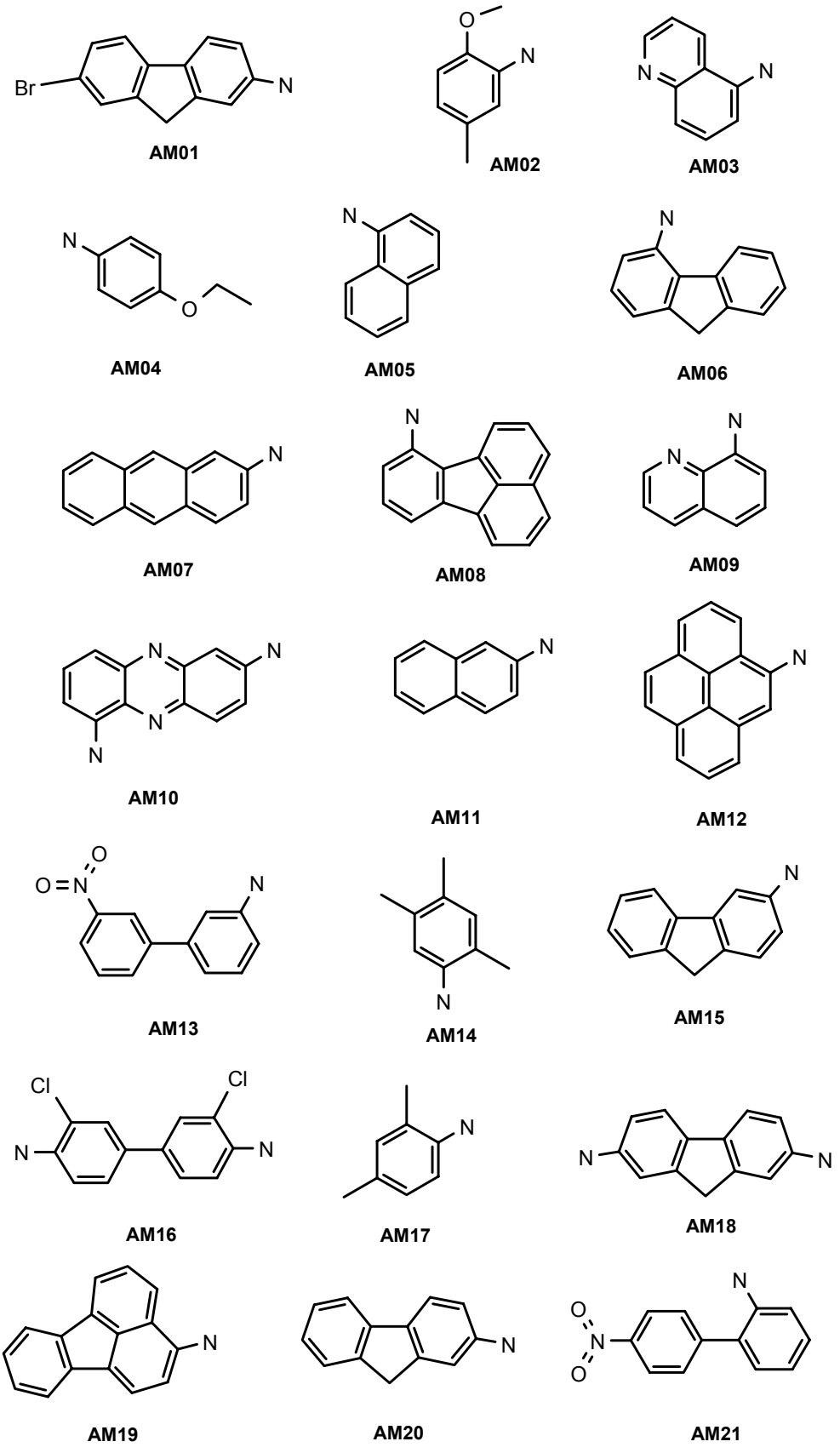

Figure A1. Aromatic amines dataset.

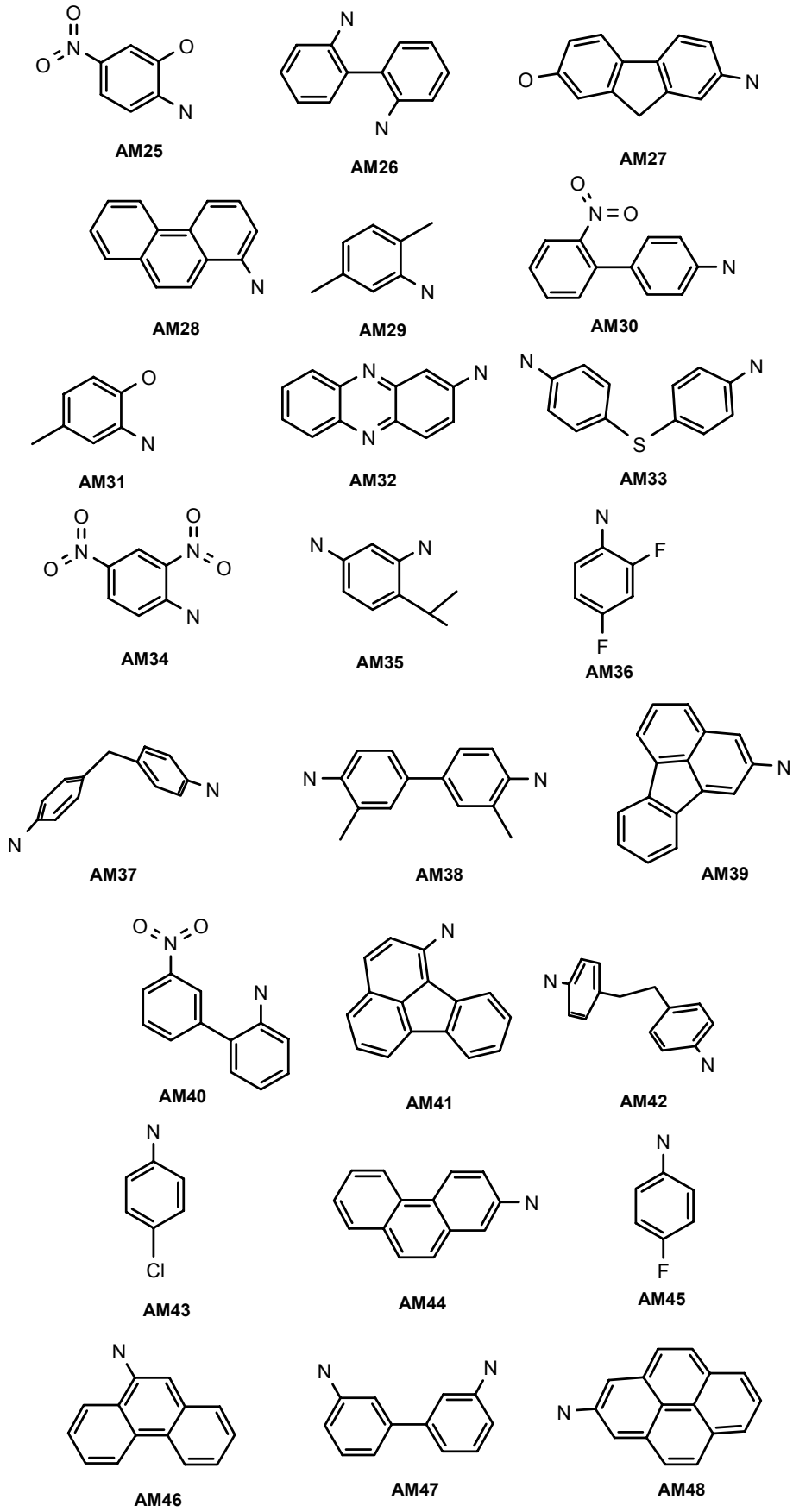

Figure A1. (Continued).

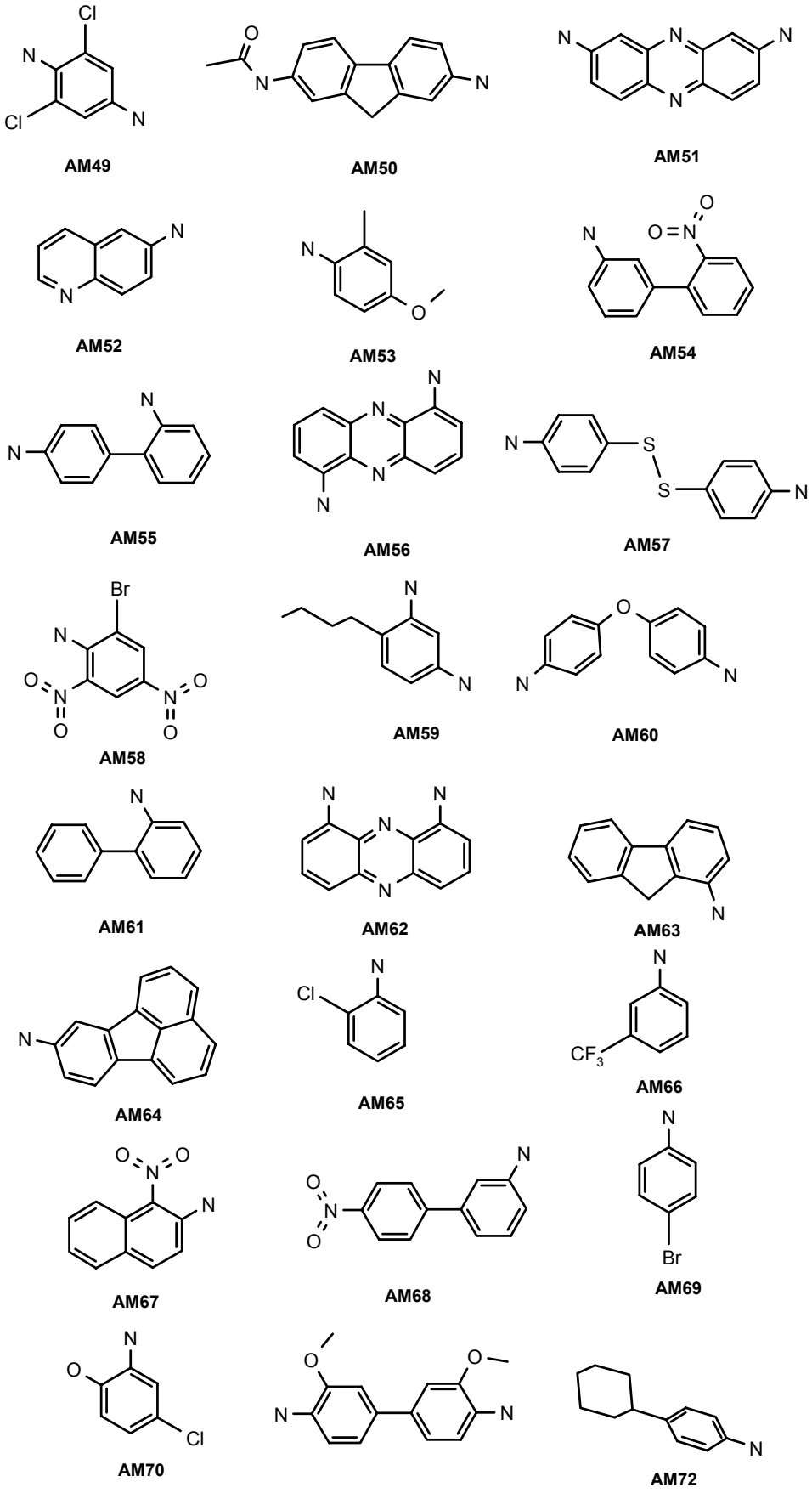

Figure A1. (Continued).

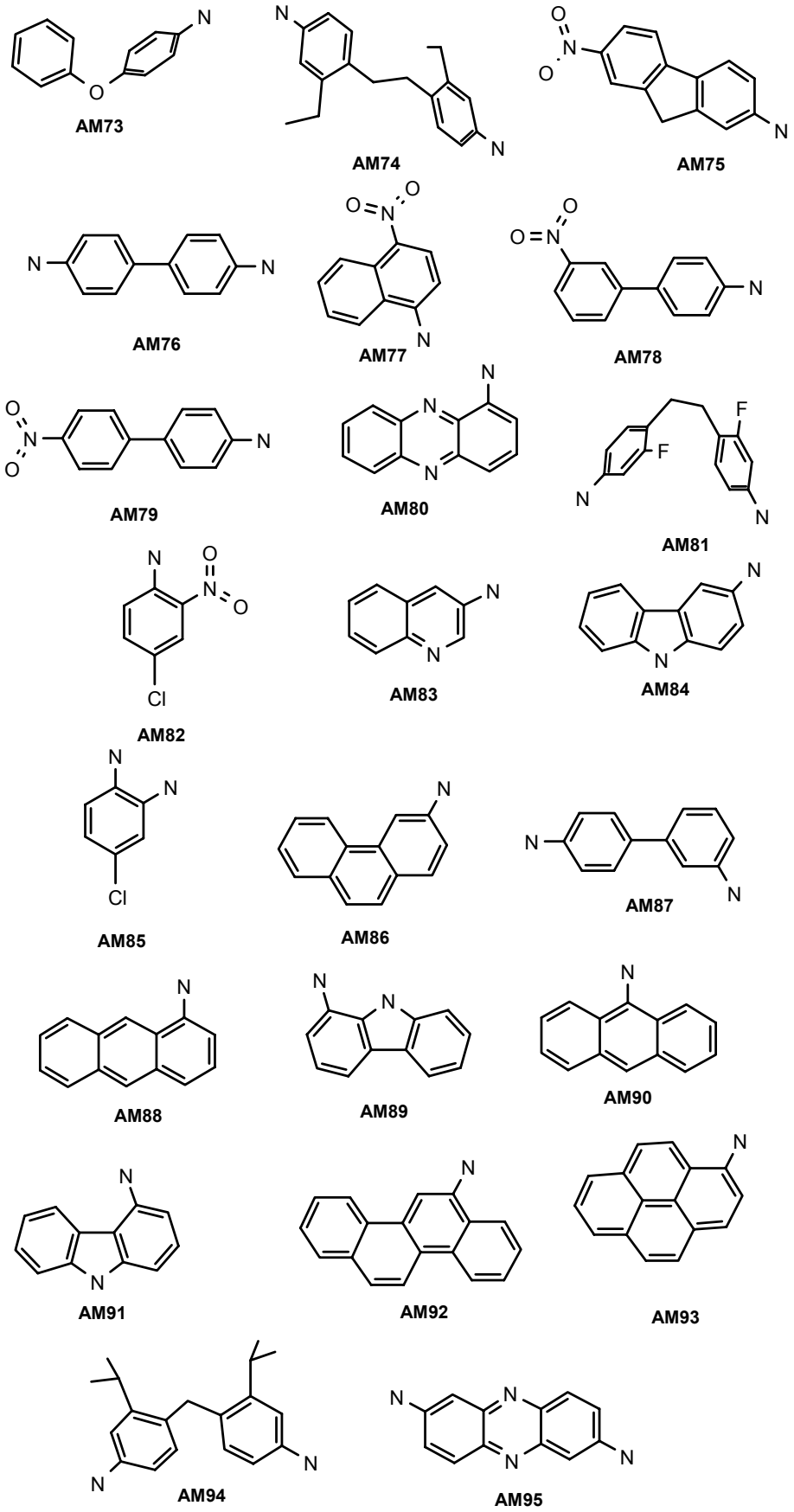

Figure A1. (Continued).

5 REFERENCES

[1] G. C. Gini and A. R. Katritzky, *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*, AAAI Press, Menlo Park CA, USA, 1999.
[2] A. M. Richard, Structure–based methods for predicting mutagenicity and carcinogenicity: are we there yet? *Mutat. Res*. **1998**, *400*, 493–507.
[3] E. Benfenati and G. Gini, Computational predictive programs (expert systems) in toxicology, *Toxicology* **1997**, *119*, 213–225
[4] T. Sugimura, Overview of carcinogenic heterocyclic amines, *Mutat. Res*. **1997**, *376*, 211–219.
[5] K. I Skog, M. A. Johansson and M. I. Jagerstad, Carcinogenic heterocyclic amines in model systems and cooked food: a review on formation, occurrence and intake, *Food Chem. Toxicol*. **1998**, *36*, 879–896
[6] R. Benigni, A. Giuliani, R. Franke and A. Gruska, Quantitative structure–activity relationships of mutagenic and carcinogenic aromatic amines. *Chem. Rev*. **2000**, *100*, 3696–3714.
[7] M. E. Colvin, F. T. Hatch and J. S. Felton, Chemical and biological factors affecting mutagen potency. *Mutat. Res*. **1998**, *400*, 479–492.
[8] K. Chung, L. Kirkovsky, A. Kirkovsky and W. P. Purcell, Review of mutagenicity of monocyclic aromatic amines: quantitative structure–activity relationships. *Mutat. Res*. **1997**, *387*, 1–16.
[9] F. T. Hatch, M. G. Knize and M. E. Colvin, Extended Quantitative Structure–Activity Relationships for 80 Aromatic and Heterocyclic Amines: Structural, Electronic, and Hydropathic Factors Affecting Mutagenic Potency, *Env. Mol. Mutagenesis* **2001**, *38*, 268–291.
[10] G. G. Cash, Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices, *Mut. Res*. **2001**, *491*, 31–37.
[11] S. C. Basak, D.R. Mills, Prediction of mutagenicity utilizing a hierarchical QSAR approach, *Sar Qsar Environ. Res*. **2001**, *12*, 481–496.
[12] S. C. Basak, D. R. Mills, A. T. Balaban, B. D. Gute, Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach, *J. Chem. Inf. Comp. Sci*. **2001**, *41*, 671–678.
[13] M. Karelson, S. Sild, U. Maran, Non–Linear QSAR Treatment of Genotoxicity, *Mol. Simulat*. **2000**, *24*, 229–242.
[14] U. Maran, S. Sild, QSAR modeling of genotoxicity on non–congeneric sets of organic compounds, *Artif. Intell. Rev*. **2003**, *20*, 13–38.
[15] U. Maran, S. Sild, QSAR Modeling of Mutagenicity on Non–congeneric Sets of Organic Compounds, In Artificial Intelligence Methods and Tools for Systems Biology, Dubitzky, W.; Azuaje, F. (Eds.), Kluwer Academic Publishers, Boston/Dordrecht/London Copyright 2004, pp 19–36.
[16] S. C. Basak, D. Mills, B. D. Gute, D. M Hawkins, Predicting Mutagenicity of Congeneric and Diverse Sets of Chemicals Using Computed Molecular Descriptors: A Hierarchical Approach. In: Benigni R (ed) Quantitative Structure–Activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton, FL, 2003, pp 207–234.
[17] G. G. Cash, B. Anderson, K. Mayo, S. Bogaczyk, J. Tunkel, Predicting genotoxicity of aromatic and eteroaromatic amines using electrotopological state indices, *Mutat. Res*. **2005**, *585*, 170–83.
[18] B. E. Mattioni, G. W. Kauffman, P. C. Jurs, L. L. Custer, S. K. Durham and G. M. Pearl, Predicting the Genotoxicity of Secondary and Aromatic Amines Using Data Subsetting To Generate a Model Ensemble, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 949–963.
[19] G. Klopman, M. R. Frierson and H. S. Rosenkranz, Computer analysis of toxicological data bases: mutagenicity of aromatic amines in Salmonella tester strains, *Environ. Mutagen*. **1985**, *7*, 625–653
[20] D. F. V. Lewis, C. Ioannides, R. Walker and D. V. Parke, Quantitative structure–activity relationships and COMPACT analysis of a series of food mutagens, *Food Addit. Contam*. **1995**, *12*, 715–723.
[21] Y. P. Zhang, G. Klopman and H. S. Rosenkranz, Structural bases of the mutagenicity of heterocyclic amines formed during the cooking process, *Environ. Mol. Mutagen*. **1993**, *21*, 100–115.
[22] H.–U. Aeschbacher and R. J. Turesky, Mammalian cell mutagenicity and metabolism of heterocyclic aromatic amines, *Mutat. Res*. **1991**, *259*, 235–250.
[23] G. Kalopissis, Structure–activity relationships of aromatic amines in the Ames *Salmonella typhimurium* assay, *Mutat. Res*. **1991**, *246*, 45–66.
[24] G. Kalopissis, Structure–activity relationships of aromatic diamines in the Ames *Salmonella typhimurium* assay. Part II, *Mutat. Res*. **1992**, *269*, 9–26.
[25] *Statistica release 6.1*; *StatSoft*, Vigonza (PD), Italy.
[26] M. Casalegno, E. Benfenati and G. Sello, An Automated Group Contribution Method in Predicting Aquatic Toxicity: The Diatomic Fragment Approach, *Chem. Res. Toxicol*. **2005**, *18*, 740–746.
[27] U. Maran, M. Karelson and A. R. Katritzky, A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines, *Quant. Struct.–Act. Relat*. **1999**, *18*, 3–10.