# Predicting the Genotoxicity of Secondary and Aromatic Amines Using Data Subsetting To Generate a Model Ensemble

Brian E. Mattioni, Gregory W. Kauffman, and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, 152 Davey Laboratory, University Park, Pennsylvania 16802

Laura L. Custer, Stephen K. Durham, and Greg M. Pearl

Bristol-Myers Squibb, Princeton, New Jersey 08453

Received January 23, 2003

Binary quantitative structure−activity relationship (QSAR) models are developed to classify a data set of 334 aromatic and secondary amine compounds as genotoxic or nongenotoxic based on information calculated solely from chemical structure. Genotoxic endpoints for each compound were determined using the SOS Chromotest in both the presence and absence of an S9 rat liver homogenate. Compounds were considered genotoxic if assay results indicated a positive genotoxicity hit for either the S9 inactivated or S9 activated assay. Each compound in the data set was encoded through the calculation of numerical descriptors that describe various aspects of chemical structure (e.g. topological, geometric, electronic, polar surface area). Furthermore, five additional descriptors that focused on the secondary and aromatic nitrogen atoms in each molecule were calculated specifically for this study. Descriptor subsets were examined using a genetic algorithm search engine interfaced with a *k*-Nearest Neighbor fitness evaluator to find the most information-rich subsets, which ultimately served as the final predictive models. Models were chosen for their ability to minimize the total number of misclassifications, with special attention given to those models that possessed fewer occurrences of positive toxicity hits being misclassified as nontoxic (false negatives). In addition, a subsetting procedure was used to form an ensemble of models using different combinations of compounds in the training and prediction sets. This was done to ensure that consistent results could be obtained regardless of training set composition. The procedure also allowed for each compound to be externally validated three times by different training set data with the resultant predictions being used in a "majority rules" voting scheme to produce a consensus prediction for each member of the data set. The individual models produced an average training set classification rate of 71.6% and an average prediction set classification rate of 67.7%. However, the model ensemble was able to correctly classify the genotoxicity of 72.2% of all prediction set compounds.

## INTRODUCTION

Modern drug discovery efforts have been focusing upon utilization of high throughput screening coupled with the concurrent assessment of a compound's advantages and liabilities at early stages in the development process. The use of predictive *in silico* models is at the forefront of several strategies for identifying liabilities early in the drug discovery process.[1−9] Particularly useful is predicting toxicological liabilities, such as carcinogenicity, mutagenicity, hepatotoxicity, and teratogenicity, because in vivo and in vitro toxicity testing requires a substantial investment of both time and money. The current paradigm for predictive *in silico* toxicity assessment is to utilize toxic biophores, or associated chemical structure,[10−13] to predict general toxic liabilities, such as carcinogenicity. Depending upon the methods used for generating these toxic biophores, the subsequent alerts are generally either too sensitive or too specific to have significant impact upon the drug discovery process.

The use of physicochemical parameters for predicting the toxic effects of chemicals dates back to reports by Meyer and Overton.[14,15] They showed a correlation between the olive oil/water partition coefficient of a chemical and the minimal concentration required to retard the mobility of tadpoles. This indicated the importance of lipophilicity in certain toxicological events. As a result, the use of the 1-octanol/water partition coefficient (lipophilicity) in many QSAR and pattern-recognition studies of toxicity has become standard.[16] Other parameters which have proven useful for studies of toxicity include Hammett's sigma,[17] Taft's steric index,[18] reduction potential, hydrogen-bonding, the difference between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies,[19] and shape measures.[20,21] Many graph-invariant and polar surface area descriptors have also found extensive use in correlative and pattern-recognition studies of toxicity.[22,23]

To develop a predictive toxicity model, one has to start with a clean data set where all compounds share at least one toxic mechanism. It should be noted that several compounds with potentially competing mechanisms were intentionally

* Corresponding author phone: (814)865-3739; fax: (814)865-3314; e-mail: pcj@psu.edu.

10.1021/ci034013i CCC: $25.00 
Published on Web 04/30/2003

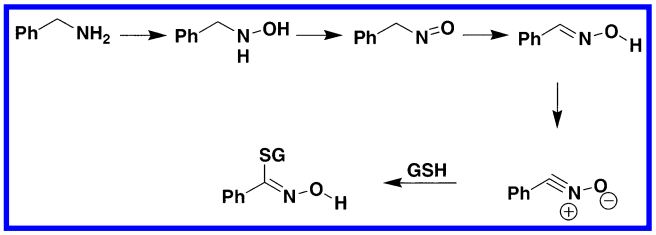


**Figure 1.** Proposed genotoxic amine biotransformation pathway.

included to challenge the computational methodology. To start, we have chosen to model the genotoxicity liability of aromatic and secondary amines. The amine compounds were selected due to their presence in pharmaceuticals and their potential, via CYP450 metabolism, to form reactive metabolites such as nitro and nitroso derivatives,[19,24−28] see Figure 1. The mechanism of action for amines is hypothesized to involve N-hydroxylation, typically mediated by cytochrome P450 1A2, and then subsequently undergoing O-esterification biotransformation.[27] The resulting metabolic product may then produce a reactive nitrenium ion, which is capable of binding to cellular nucleophiles such as DNA. In addition to this biotransformation pathway, Newcomb and Toy have recently published a detailed Cyp P450 hydroxylation mechanism (Figure 2), which should be included when investigating the genotoxicity of amines due to hydroxylation.[28]

To that end, we have designed a library of amines (aromatic and secondary aliphatic) and processed these compounds in the SOS Chromotest. The data set was designed that adequately represents the chemical space of amines. The neighborhood around the hypothesized toxic biophore should be represented, while excluding compounds with known biophores associated with the toxicity under investigation. These data set limitations have typically made it impossible to develop local predictive toxicity models because there is not sufficient data available in the published literature. Assuming sufficient data existed, serious questions remain regarding combination of data produced under divergent protocols in multiple laboratories and methods used to assign compounds to either a training or predictive data set. To alleviate these modeling issues, an amine data set was created from commercially available compounds. These compounds were tested in the SOS Chromotest to assess the potential genotoxicity.

The development of predictive models for amine toxicity remains a challenging task.[19,29−37] Much of the published literature in this area has focused on quantitative predictions of toxicity endpoints, while very few have focused on *classifying* compounds as toxic or nontoxic. In addition, much of this work has used small, congeneric data sets to build models, thereby simplifying the task via reduced structural diversity.

Benigni and co-workers reported classification models for mutagenicity assessment using a set of 126 aromatic and heteroaromatic amines. A classification rate of 76.2% was

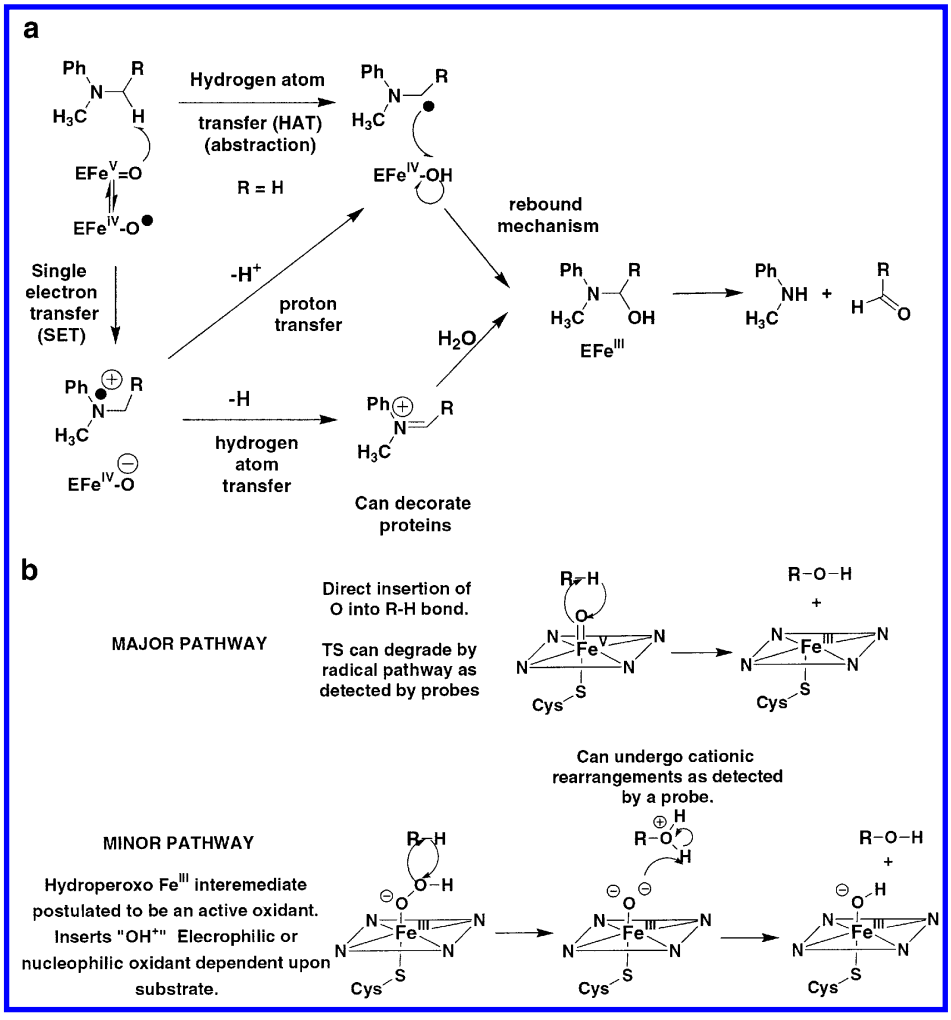


**Figure 2.** (a) Alternate hypothesized CYP peroxidases metabolic pathway and (b) major pathway.

achieved for both the TA98 and TA100 bacterial strains.[32] In a subsequent paper, they showed that partitioning the data into chemically similar subclasses containing 15−61 amines could improve modeling for some subclasses. Classification rates in the latter study ranged from 62.8% correct for single ring compounds to 100% correct for a small series of diphenylmethanes.[19] Benigni et al. and Franke et al. have also demonstrated the ability to distinguish between carcinogenic and noncarcinogenic amines using rodent carcinogenicity data in two recent publications.[30,31] In both reports, several data sets containing 49−66 amines were used to construct classification models with accuracy rates varying from 73.0% to 84.8%.

The current project is a joint venture between researchers at Bristol-Myers Squibb (BMS) and the Jurs group at Penn State University. Genotoxicity endpoints were determined for a structurally diverse set of 334 aromatic and secondary amine compounds at BMS and were submitted to the Jurs group for model development. The broad goal of this work was to identify a model or set of models that delivered accurate and economical predictions of genotoxicity for compounds based solely on features of their molecular structure. Thousands of models were screened using a genetic algorithm (GA) search engine interfaced with a *k*-Nearest Neighbor (*k*NN) fitness evaluator. Models were chosen for their ability to minimize the total number of misclassifications, with special attention given to those models that possessed fewer occurrences of positive toxicity hits being misclassified as nontoxic (false negatives). Previous studies using methods developed by the Jurs Group−ADAPT software package−have proven successful for modeling chemical toxicity data. These include the prediction of the following: carcinogenic activity,[38−41] *Tetrahymena* acute toxicity,[22] fathead minnow acute toxicity,[23] acute mammalian toxicity,[42,43] clastogenicity,[44] and genotoxic activity.[45,46] It was believed that the success of ADAPT for these QSAR applications could be extended to the difficult task of predicting the associated genetic toxicity of amine compounds.

## EXPERIMENTAL PROCEDURES

**Data Collection.** The compounds used for model development were selected based upon (1) availability from Aldrich Chemical Co. and (2) passed an expanded Lipinski drug likeness filter.[47] The purest form of the compounds was selected for running in the assay, and most of the compounds obtained have >98% purity. The data set is comprised of compounds selected based upon chemical diversity and supplemented by manual addition of chemically interesting groups. The chemical diversity was based upon clustering of the electrotopological indices, molecular weight, and various atom and ring counts. Compound selection was then limited to three compounds from a specific class. The compounds chosen from the clusters were limited to five example compounds per cluster where priority was given to compounds that did not contain genotoxic alerts, as predicted by DEREK 3.6.[10] The data set was also enriched by adding simple chemical substitutions; an example of these additions would be to include examining the effect of methyl versus *tert*-butyl substituted amines or effects of halogen substitution.

The SOS Chromotest was used to determine the genotoxic liability for all of the compounds in the amine data set. The assay measures induction of a *lacZ* reporter gene in response to DNA damage.[48] The SOS pathway plays a leading role in the way *E. coli* respond to genotoxic damage.[49] Because this pathway responds to a broad spectrum of genotoxic substances, SOS induction can be used as an early monitor for DNA damage. *E. coli* were modified with a *lacZ* reporter gene under transcriptional control an SOS repair gene. Normally, SOS repair genes are repressed, but in response to DNA damage, these genes are induced resulting in production of $\beta$-galactosidase, the gene product of *lacZ*. Fold increases in gene induction are determined by measuring $\beta$-galactosidase activity using a *o*-nitrophenyl-$\beta$-d-galactopyranoside (ONGP). The assay has been used extensively with many different chemical classes. A review of published data between 1982 and 1992 demonstrated that for 1776 compounds, the SOS Chromotest had 90% concordance with the Ames mutagenicity test.[50]

**Data Processing.** The average atom-pair similarity[51] over all 334 compounds was approximately 0.213 (on a 0 to 1 scale), indicating a high degree of structural diversity in this data set. Three-dimensional structures of the compounds were obtained using the CONCORD software package (Tripos Inc., St. Louis, MO) with no subsequent geometry optimization. IMAX values for the inactivated assay ranged from 0.81 to 11.66, and for the S9 activated assay they ranged from 0.85 to 10.57. Table 1 lists the compound number (specific to this study), the compound name, and an indicator of toxic (+) or nontoxic (−) for both the inactivated and S9 activated assays. For modeling purposes, if a compound was determined to be toxic by either the activated or inactivated assays, then the overall toxicity profile for that compound was "toxic". Using this convention, the data set contained 197 nontoxic (59.0%) and 137 (41.0%) toxic compounds.

Models were developed with descriptors calculated from only the parent compound structures. If the observed toxicity for a compound is the result of only a metabolite, then the potential exists for descriptors derived from the parent structure to not capture the important structural features necessary for an accurate prediction. Therein lies one of the challenges and assumptions of this study: Can an accurate model or set of models be constructed to predict genotoxicity of parent compounds ***or*** their metabolized byproducts? If robust models could be developed using this approach, then the impact on genotoxicity assessment of compounds would be substantial with respect to reduced time and expense for individual compound screening.

**Software**. This work was performed using the Automated Data Analysis and Pattern-Recognition Toolkit (ADAPT)[52,53] software package. Descriptor calculation, genetic algorithm,[54−57] and *k*NN[58] routines were implemented in FORTRAN and run on a Digital ALPHA Station 500 with the OSF/1 UNIX operating system. All incidental data handling was done using a 1.6 GHz Pentium 4 desktop PC equipped with 256 MB of RAM.

**Descriptor Generation and Objective Feature Selection.** Descriptors that encode topological, geometric, and electronic aspects of molecular structure were calculated for each compound. It should be noted that salts **211**, **212**, and **216**

**Table 1.** Compound Numbers, Names, ± S9 Activated Toxicity Assay Results, Predictions for Each Compound, and Majority Prediction over a Total of Three Individual Predictions[a]

| | | toxicity[b] | | predictions | | | | |
|---|---|---|---|---|---|---|---|---|
| no. | name | −S9 | +S9 | 1 | 2 | 3 | maj. | missed |
| 1 | aminopterin | − | − | − | − | − | − | |
| 2 | dipyridamole | − | − | − | − | − | − | |
| 3 | perphenazine | − | − | − | − | + | − | |
| 4 | sulfaquinoxaline | − | − | − | + | − | − | |
| 5 | nomega-methyltryptamine | + | − | − | − | − | − | * |
| 6 | tubercidin | − | − | + | − | + | + | * |
| 7 | 4-aminophenyl sulfone | − | + | − | − | − | − | * |
| 8 | 1,1′-dianthrimide | − | − | − | − | − | − | |
| 9 | disperse orange 11 | − | − | − | − | − | − | |
| 10 | 1-(methylamino)anthraquinone | − | − | − | − | − | − | |
| 11 | 1-aminoanthraquinone | − | − | + | − | − | − | |
| 12 | 8-anilino-1-naphthalenesulfonic acid | − | + | − | − | − | − | * |
| 13 | ethyl 2-aminobenzoate | − | − | − | − | − | − | |
| 14 | 3-chloro-2-methylaniline | − | − | − | + | − | − | |
| 15 | 2,6-dimethylaniline | − | − | + | + | + | + | * |
| 16 | 2-chloro-6-methylaniline | − | − | − | − | − | − | |
| 17 | 2,4,6-trimethylaniline | − | − | + | − | − | − | |
| 18 | 2-amino-4-chlorobenzoic acid | − | − | − | − | − | − | |
| 19 | *o*-anisidine | − | − | + | + | + | + | * |
| 20 | *N*-phenyl-1-naphthylamine | + | + | + | − | + | + | |
| 21 | 2-aminobiphenyl | − | + | + | + | − | + | |
| 22 | 9-aminoacridine | − | − | + | + | − | + | * |
| 23 | 4,4′-bis(diethylamino)benzophenone | − | − | − | − | − | − | |
| 24 | *N*-phenylanthranilic acid | − | − | − | + | − | − | |
| 25 | 7-diethylamino-4-methylcoumarin | − | − | − | − | − | − | |
| 26 | ethoxyquin | − | − | − | − | + | − | |
| 27 | 2-aminonaphthalene | + | + | + | + | + | + | |
| 28 | *N*,*N*-diethylaniline | − | − | + | − | − | − | |
| 29 | 3-(diethylamino)phenol | − | + | − | − | − | − | * |
| 30 | 2,4-diamino-6-phenyl-1,3,5-triazine | − | − | + | + | + | + | * |
| 31 | 2-(4-aminophenyl)−6-methylbenzothiazole | − | − | + | + | − | + | * |
| 32 | 2-chlorophenothiazine | − | − | − | − | − | − | |
| 33 | 1-phenylpiperazine | − | − | − | − | − | − | |
| 34 | *N*-benzyl-*N*-ethylaniline | + | − | − | + | + | + | |
| 35 | phenothiazine | − | − | − | − | − | − | |
| 36 | benzidine | − | − | + | + | + | + | * |
| 37 | *N*-(4-hydroxyphenyl)−2-naphthylamine | − | − | − | − | − | − | |
| 38 | ethyl 4-(butylamino)benzoate | − | − | − | − | − | − | |
| 39 | 2-amino-6-ethoxybenzothiazole | + | + | + | + | + | + | |
| 40 | *N*-ethyl-*o*-toluidine | + | − | + | + | + | + | |
| 41 | *o*-phenetidine | − | + | + | + | + | + | |
| 42 | 5-chloro-*o*-anisidine | − | − | + | + | + | + | * |
| 43 | 2-amino-6-chlorobenzothiazol | − | + | + | + | + | + | |
| 44 | 2-chloroaniline | + | + | + | + | + | + | |
| 45 | 2-aminophenol | + | + | + | + | + | + | |
| 46 | 3,4-dimethylaniline | − | − | − | + | + | + | * |
| 47 | 4-chloro-2-methylaniline | − | − | − | − | − | − | |
| 48 | 2,5-dimethylaniline | − | − | − | + | − | − | |
| 49 | 5-chloro-2-methylaniline | − | − | − | − | − | − | |
| 50 | 3-dimethylaminophenol | − | − | − | − | + | − | |
| 51 | *N*,*N*,*N*′,*N*′-tetramethyl-1,4-phenylenediamine | − | + | + | − | + | + | |
| 52 | *N*-phenyl-1,4-phenylenediamine | − | − | − | − | − | − | |
| 53 | 4,4′-methylenedianiline | + | + | + | + | + | + | |
| 54 | *N*-ethyl-*m*-toluidine | + | − | + | + | + | + | |
| 55 | 4-methoxy-2-methylaniline | + | − | − | + | + | + | |
| 56 | 2,5-dimethoxyaniline | − | − | + | + | + | + | * |
| 57 | *N*-phenylglycine | − | − | + | + | − | + | * |
| 58 | *N*-phenylbenzylamine | − | + | − | + | + | + | |
| 59 | *p*-anisidine | + | + | − | + | + | + | |
| 60 | 4-(methylthio)aniline | + | − | − | + | + | + | |
| 61 | 4-bromoaniline | − | − | + | + | − | + | * |
| 62 | 4-chloroaniline | − | − | + | + | − | + | * |
| 63 | 3-chloroaniline | − | − | + | + | + | + | * |
| 64 | piperazine | − | − | − | − | − | − | |
| 65 | diisopropanolamine | + | − | − | − | − | − | * |
| 66 | 2-(2-aminoethylamino)ethanol | − | − | + | + | − | + | * |
| 67 | diethanolamine | − | − | − | + | − | − | |
| 68 | bis(2-methoxyethyl)amine | − | − | − | − | − | − | |
| 69 | 2-aminoanthraquinone | − | + | − | − | − | − | * |
| 70 | *N*-ethyl-1-naphthylamine | − | − | + | + | − | + | * |
| 71 | *N*-methylanthranilic acid | − | − | − | − | − | − | |

**Table 1** (Continued)

| | | toxicity[b] | | predictions | | | | |
|---|---|---|---|---|---|---|---|---|
| no. | name | −S9 | +S9 | 1 | 2 | 3 | maj. | missed |
| 72 | *o*-tolidine | − | − | − | + | + | + | * |
| 73 | 3-(ethylamino)-*p*-cresol | − | − | + | − | − | − | |
| 74 | 2-methoxy-5-methylaniline | − | + | + | + | − | + | |
| 75 | *N*,*N*-dimethylaniline, redistilled | − | − | + | − | − | − | |
| 76 | diphenylamine | − | − | − | + | − | − | |
| 77 | *N*-(4-hydroxyphenyl)glycine | − | + | − | + | − | − | * |
| 78 | 2-anilinoethanol | − | − | + | − | − | − | |
| 79 | sulfisoxazole | − | − | − | + | − | − | |
| 80 | 1,4-bis[(4-methylphenyl)amino]-9,10-anthracenedione | − | − | − | − | − | − | |
| 81 | 1,4-diaminoanthraquinone | + | + | − | − | + | − | * |
| 82 | 1,5-diaminoanthraquinone | − | − | − | + | − | − | |
| 83 | 4,4′-benzylidenebis(*N*,*N*)-dimethylaniline | − | − | + | − | + | + | * |
| 84 | 2,6-diaminoanthraquinone | − | − | + | + | + | + | * |
| 85 | 3-amino-9-ethylcarbazole | + | + | + | + | + | + | |
| 86 | 1-naphthylamine | + | + | + | + | + | + | |
| 87 | phenoxazine | + | + | − | + | − | − | * |
| 88 | 2-aminobenzothiazole | − | − | + | + | + | + | * |
| 89 | 4-phenoxyaniline | − | − | − | − | − | − | |
| 90 | 4,4′-thiodianiline | − | + | + | − | + | + | |
| 91 | 2,6-diaminopyridine | − | − | − | − | − | − | |
| 92 | 2,6-dimethylmorpholine | − | − | + | − | − | − | |
| 93 | bis(hexamethylene)triamine | − | − | + | − | + | + | * |
| 94 | 2-aminofluorene | + | + | + | + | + | + | |
| 95 | *p*-phenetidine | − | − | + | + | + | + | * |
| 96 | 2-amino-6-fluorobenzothiazole | + | + | + | + | + | + | |
| 97 | 2-fluoroaniline | − | + | + | + | + | + | |
| 98 | 2,4-difluoroaniline | − | + | + | + | + | + | |
| 99 | 2-amino-5-fluorobenzoic acid | − | + | + | − | − | − | * |
| 100 | 1,8-diaminonaphthalene | + | + | + | − | + | + | |
| 101 | iminodibenzyl | − | − | + | − | − | − | |
| 102 | acridine orange base | − | − | + | − | + | + | * |
| 103 | indoline | + | + | + | + | + | + | |
| 104 | 2-aminopyridine | − | − | − | − | + | − | |
| 105 | 2,7-diaminofluorene | − | + | + | + | + | + | |
| 106 | L-abrine | − | − | − | + | − | − | |
| 107 | 4-(dimethylamino)benzophenone | − | − | − | − | − | − | |
| 108 | *m*-anisidine | − | + | + | + | + | + | |
| 109 | *N*-methyldiphenylamine | − | − | + | − | + | + | * |
| 110 | 2,6-diethylaniline | + | − | − | − | − | − | * |
| 111 | *N*-allylaniline | − | − | + | + | + | + | * |
| 112 | 3-bromoaniline | − | + | + | − | + | + | |
| 113 | 3-aminophenol | − | − | + | + | + | + | * |
| 114 | 3-aminophenyl sulfone | − | − | + | + | + | + | * |
| 115 | triphenylamine | − | − | + | − | − | − | |
| 116 | 1-aminoanthracene | + | + | + | + | − | + | |
| 117 | 4,4′-diaminobenzophenone | − | − | − | − | − | − | |
| 118 | *N*-(2-carboxyphenyl)glycine | − | − | − | − | − | − | |
| 119 | 2-aminoanthracene | − | + | + | + | + | + | |
| 120 | *N*,*N*-dibutylaniline | − | − | − | − | − | − | |
| 121 | *N*-methyl-*N*-phenylbenzylamine | + | − | + | − | + | + | |
| 122 | 2-bromoaniline | + | + | + | + | + | + | |
| 123 | 2-chloro-4-methylaniline | − | + | + | + | + | + | |
| 124 | 2-chloro-1,4-phenylenediamine | + | + | + | + | + | + | |
| 125 | *m*-phenetidine | + | + | + | + | − | + | |
| 126 | 1,2,3,4-tetrahydroquinoline | + | − | + | + | + | + | |
| 127 | *N*-isopropylaniline | − | − | + | − | − | − | |
| 128 | 4-chloro-*N*-methylaniline | − | − | − | + | + | + | * |
| 129 | 2-aminobenzimidazole | + | − | + | − | + | + | |
| 130 | 2,4-diamino-6-(hydroxymethyl)pteridine | − | − | − | − | − | − | |
| 131 | 9-aminophenanthrene | + | + | + | + | + | + | |
| 132 | 1-(2-fluorophenyl)piperazine | − | − | − | − | − | − | |
| 133 | 5-chloro-2-(methylamino)benzophenone | − | − | − | − | − | − | |
| 134 | 8-azaadenine | − | − | − | − | − | − | |
| 135 | 5-aminoisoquinoline | + | + | − | + | + | + | |
| 136 | 2-amino-4-phenylphenol | − | − | + | + | + | + | * |
| 137 | 2-amino-4-*tert*-butylphenol | − | + | + | − | + | + | |
| 138 | 10-methylphenothiazine | − | − | − | − | − | − | |
| 139 | 2-amino-4-methylbenzothiazole | + | + | + | + | − | + | |
| 140 | 1-aminoisoquinoline | − | − | + | + | + | + | * |
| 141 | 1-aminopyrene | − | − | − | + | − | − | |
| 142 | 2-amino-1-methylbenzimidazole | − | + | + | + | + | + | |
| 143 | *N*-phenylethylenediamine | + | − | − | − | − | − | * |

**Table 1** (Continued)

| | | toxicity[b] | | predictions | | | | |
|---|---|---|---|---|---|---|---|---|
| no. | name | −S9 | +S9 | 1 | 2 | 3 | maj. | missed |
| 144 | 5-amino-2-methoxyphenol | − | + | + | + | − | + | |
| 145 | 2-amino-6-methoxybenzothiazole | − | + | + | + | + | + | |
| 146 | 1,2-diaminoanthraquinone | − | − | − | + | − | − | |
| 147 | *N*-cyclohexylaniline | − | − | − | − | − | − | |
| 148 | dipentylamine | − | − | + | − | − | − | |
| 149 | 1,5-diaminonaphthalene | − | + | − | − | + | − | * |
| 150 | 1-(4-fluorophenyl)piperazine | − | − | − | − | − | − | |
| 151 | 1-[(2-hydroxyethyl)amino]-4-(methylamino)-9,10-anthracenedione | − | − | − | − | − | − | |
| 152 | 4-amino-3-methylbenzoic acid | − | − | − | + | − | − | |
| 153 | 2-amino-6-methylbenzothiazole | + | − | + | + | + | + | |
| 154 | 2,3-diaminotoluene | + | + | − | + | + | + | |
| 155 | 3-aminobenzophenone | + | − | + | + | − | + | |
| 156 | 5-amino-*o*-cresol | − | + | + | + | + | + | |
| 157 | 6-amino-*m*-cresol | + | + | + | + | + | + | |
| 158 | 4-amino-*m*-cresol | − | − | + | + | + | + | * |
| 159 | 2,3,5,6-tetramethyl-1,4-phenylenediamine | − | − | + | − | + | + | * |
| 160 | 5-*tert*-butyl-*o*-anisidine | + | + | + | + | + | + | |
| 161 | 4-aminobenzyl cyanide | + | − | + | + | + | + | |
| 162 | 9-amino-6-chloro-2-methoxyacridine | − | − | − | − | − | − | |
| 163 | 2,4-diamino-6,7-diisopropylpteridine | − | + | + | + | − | + | |
| 164 | pentaethylenehexamine | + | + | + | − | − | − | * |
| 165 | 4,4′-methylene-bis(2,6-dimethylaniline) | + | − | + | − | − | − | * |
| 166 | *N*,*N*-diisopropylaniline | − | − | − | − | − | − | |
| 167 | 2-(phenylsulfonyl)aniline | − | − | − | − | − | − | |
| 168 | *N*,*N*′-bis(2-hydroxyethyl)ethylenediamine | − | − | − | − | − | − | |
| 169 | 2,3-difluoroaniline | + | + | + | + | + | + | |
| 170 | 2-(methylamino)pyridine | − | − | − | − | + | − | |
| 171 | 4-(butylamino)benzoic acid | − | − | − | − | − | − | |
| 172 | 5-chloro-2-methylamino-*N*-phenethylbenzamide | − | − | − | − | − | − | |
| 173 | 5-aminoindole | + | + | + | + | + | + | |
| 174 | 4-aminoindole | + | + | + | + | + | + | |
| 175 | 2-amino-4-methoxybenzothiazole | + | − | + | + | + | + | |
| 176 | harmane-1,2,3,4-tetrahydro-3-carboxylic acid | − | − | − | − | − | − | |
| 177 | clozapine | − | − | − | − | + | − | |
| 178 | *N*-methyl-*p*-anisidine | + | − | + | − | + | + | |
| 179 | perhydroisoquinoline | − | − | + | − | − | − | |
| 180 | 2,5-dimethyl-1,4-phenylenediamine | + | + | − | + | + | + | |
| 181 | α-(methylaminomethyl)benzyl alcohol | − | − | − | + | + | + | * |
| 182 | 2-anilinopyridine | + | + | + | − | − | − | * |
| 183 | 2-acetylphenothiazine | − | − | − | + | − | − | |
| 184 | 2-amino-7-bromofluorene | − | + | + | + | + | + | |
| 185 | 2-benzylaminopyridine | − | − | − | − | − | − | |
| 186 | 6-aminoindazole | − | − | − | + | + | + | * |
| 187 | solvent blue 59 | − | − | − | − | − | − | |
| 188 | *N*-methyldodecylamine | − | − | + | − | − | − | |
| 189 | 6-amino-2-mercaptobenzothiazole | − | − | + | − | + | + | * |
| 190 | 4,4′-vinylidene-bis(*N*,*N*-dimethylaniline) | − | + | − | − | − | − | * |
| 191 | 5-amino-2-methylindole | + | − | + | + | + | + | |
| 192 | 2-[4-(dimethylamino)phenyl]−6-methylbenzothiazole | + | − | + | − | + | + | |
| 193 | 2-aminoterephthalic acid | − | − | + | − | − | − | |
| 194 | 4-(methylamino)benzoic acid | − | − | − | − | − | − | |
| 195 | oracet blue b | − | − | − | − | − | − | |
| 196 | 5-anilino-1,2,3,4-thiatriazole | − | − | − | + | − | − | |
| 197 | pindolol | + | + | − | + | − | − | * |
| 198 | tolfenamic acid | − | − | − | − | − | − | |
| 199 | *N*-methyl-*N*-phenylbenzotriazolemethanamine | + | + | + | + | + | + | |
| 200 | 1-(3-methoxyphenyl)piperazine | − | − | − | + | − | − | |
| 201 | 2-(propylamino)ethanol | − | − | − | − | − | − | |
| 202 | 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole | − | − | − | + | − | − | |
| 203 | 2-amino-6-(methylsulfonyl)benzothiazole | − | − | + | − | + | + | * |
| 204 | 6-methoxy-1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole-1-carboxylic acid | − | − | − | − | − | − | |
| 205 | 4-(diethylamino)benzophenone | − | − | − | − | − | − | |
| 206 | sangivamycin | − | − | − | − | + | − | |
| 207 | methyl 3-amino-4-methylbenzoate | − | + | − | − | − | − | * |
| 208 | 3,3′-methylenedianiline | + | − | + | + | + | + | |
| 209 | 2-amino-4-chlorobenzothiazole | + | − | + | + | + | + | |
| 210 | 6-methoxy-1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole | − | − | + | + | − | + | * |
| 211 | bis-benzimide (trihydrochloride salt) | + | + | + | + | − | + | |
| 212 | bis-benzimide Hoechst no. 33342 (trihydrochloride salt) | − | − | + | − | − | − | |
| 213 | 4-amino-2-chloro-6,7-dimethoxyquinazoline | + | + | − | + | − | − | * |

**Table 1** (Continued)

| | | toxicity[b] | | predictions | | | | |
|---|---|---|---|---|---|---|---|---|
| no. | name | −S9 | +S9 | 1 | 2 | 3 | maj. | missed |
| 214 | 6-ethyl-*o*-toluidine | + | − | − | + | + | + | |
| 215 | 2,2′-oxydianiline | − | − | + | − | + | + | * |
| 216 | minaprine (dihydrochloride salt) | + | − | − | − | − | − | * |
| 217 | 2-amino-4-methylbenzonitrile | + | − | + | + | + | + | |
| 218 | 3-(2′-benzimidazolyl)-7-*N*,*N*-diethylaminocoumarin | − | − | + | − | + | + | * |
| 219 | 2-amino-5,6-dimethylbenzimidazole | − | − | − | + | + | + | * |
| 220 | 2-amino-5,6-dimethylbenzothiazole | − | + | + | + | + | + | |
| 221 | 2-(2-aminophenyl)indole | + | + | + | − | + | + | |
| 222 | 4-pentylaniline | + | + | − | − | + | − | * |
| 223 | 1-(2-pyridyl)piperazine | − | − | − | − | − | − | |
| 224 | 1-(2-methoxyphenyl)piperazine | + | − | − | − | − | − | * |
| 225 | 6-norlysergic acid diethylamide | − | − | − | + | − | − | |
| 226 | 3-(2′-benzothiazolyl)-7-diethylaminocoumarin | − | − | − | − | − | − | |
| 227 | 5-phenyl-*o*-anisidine | − | − | + | + | + | + | * |
| 228 | (S)-(−)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole-3-carboxylic acid | − | − | − | − | − | − | |
| 229 | 2-amino-4-(ethylsulfonyl)phenol | − | − | − | − | − | − | |
| 230 | 5-(2-aminoethylamino)-1-naphthalenesulfonic acid | + | + | − | + | − | − | * |
| 231 | 8-(2-aminoethylamino)-1-naphthalenesulfonic acid | − | − | − | − | − | − | |
| 232 | 5-methoxy-2-methylaniline | − | − | + | + | + | + | * |
| 233 | 4′-piperazinoacetophenone | − | − | − | − | + | − | |
| 234 | 3,8-diamino-6-phenylphenanthridine | − | + | + | + | + | + | |
| 235 | 9,10-diaminophenanthrene | − | − | − | + | − | − | |
| 236 | *N*-ethyl-*N*-isopropylaniline | − | − | + | − | − | − | |
| 237 | 3,3′,5,5′-tetramethylbenzidine | − | − | − | − | + | − | |
| 238 | 1,2,3,4-tetrahydro-2,2,4,7-tetramethylquinoline | + | − | − | + | − | − | * |
| 239 | *N*-phenylbenzotriazolemethanamine | + | − | + | + | + | + | |
| 240 | 2-chloro-4,6-dimethylaniline | + | + | − | − | − | − | * |
| 241 | *N*-(2-amino-4-chlorophenyl)anthranilic acid | − | − | − | − | − | − | |
| 242 | astemizole | − | − | − | + | − | − | |
| 243 | 9-(methylaminomethyl)anthracene | − | − | + | + | − | + | * |
| 244 | enoxacin | + | + | − | − | − | − | * |
| 245 | S-2-benzothiazolyl 2-amino-α-(methoxyimino)-4-thiazolethiolacetate | + | − | + | + | + | + | |
| 246 | 1-(5-isoquinolinesulfonyl)piperazine | − | − | − | − | − | − | |
| 247 | 1-(5-isoquinolinesulfonyl)-3-methylpiperazine | − | + | − | − | − | − | * |
| 248 | 1-(5-isoquinolinesulfonyl)-2-methylpiperazine | − | − | − | − | − | − | |
| 249 | 2-chloro-N6-cyclopentyl-1-deazaadenosine | − | − | − | + | − | − | |
| 250 | 6-Amino-2-naphthoic acid | − | − | − | − | + | − | |
| 251 | 1,3-dihydro-1,3,3-trimethylspiro[2H-indole-2,3′-[3H]phenanthr[9,10-b][1,4]oxazine] | − | − | − | − | + | − | |
| 252 | 5-chloro-1,3-dihydro-1,3,3-trimethylspiro[2H-indole-2,3′-[3H]phenanthr[9,10-b][1,4]oxazine] | − | − | − | − | + | − | |
| 253 | α-methyl-*N*-phenyl-1H-benzotriazole-1-methanamine | + | + | + | + | + | + | |
| 254 | *trans*-2-[4-(dimethylamino)styryl]benzothiazole | − | − | − | + | + | + | * |
| 255 | N,α-diphenylbenzotriazolemethanamine | + | + | + | + | + | + | |
| 256 | 6-methyl-2-thiouracil | + | − | − | − | − | − | * |
| 257 | folic acid | − | − | − | − | − | − | |
| 258 | 3-amino-1,2,4-triazole | − | − | − | − | − | − | |
| 259 | cytarabine | − | − | − | − | − | − | |
| 260 | 1-naphthylamine-8-sulfonic acid | − | − | − | − | − | − | |
| 261 | 8-chlorotheophylline | + | − | − | − | − | − | * |
| 262 | 2,3-xylidine | + | + | + | + | + | + | |
| 263 | 2-trifluoromethylaniline | − | − | + | − | − | − | |
| 264 | *o*-aminobenzenesulfonic acid | + | − | − | − | − | − | * |
| 265 | *o*-toluidine | − | − | − | + | − | − | |
| 266 | 3-chloro-*p*-toluidine | + | + | − | + | + | + | |
| 267 | 3-amino-A,A,A-trifluorotoluene | + | + | + | + | − | + | |
| 268 | 2-amino-1-phenol-4-sulfonic acid | + | + | + | − | − | − | * |
| 269 | *N*-methylaniline | − | − | − | − | − | − | |
| 270 | *N*-phenyl-*m*-anisidine | − | − | − | − | − | − | |
| 271 | thiocarbanilide | + | − | − | − | − | − | * |
| 272 | formanilide | − | − | − | − | − | − | |
| 273 | allylthiourea | − | − | + | + | − | + | * |
| 274 | *N*-methylethanolamine | + | − | − | + | − | − | * |
| 275 | tetraethylenepentamine | + | + | − | − | − | − | * |
| 276 | 1-naphthylamine-7-sulfonic acid | − | − | + | + | − | + | * |
| 277 | phenyl 4-aminosalicylate | + | − | + | + | − | + | |
| 278 | melphalan | + | − | − | + | + | + | |
| 279 | cyclen | + | + | + | + | − | + | |
| 280 | 1,7-dioxa-4,10-diazacyclododecane | + | + | − | + | + | + | |
| 281 | 5-fluorouridine | − | − | − | − | − | − | |

**Table 1** (Continued)

| | | toxicity[b] | | predictions | | | | |
|---|---|---|---|---|---|---|---|---|
| no. | name | −S9 | +S9 | 1 | 2 | 3 | maj. | missed |
| 282 | 3,5-bis(trifluoromethyl)aniline | − | − | + | + | − | + | * |
| 283 | carbutamide | + | − | − | − | − | − | * |
| 284 | brucine | + | − | − | − | + | − | * |
| 285 | 4-(trifluoromethyl)aniline | + | + | + | + | − | + | |
| 286 | 2-oxazolidone | + | + | + | + | + | + | |
| 287 | 5,6-dihydrouracil | + | − | − | − | − | − | * |
| 288 | 8-aminoquinoline | + | + | + | + | + | + | |
| 289 | 3-aminoquinoline | + | + | + | + | + | + | |
| 290 | sulfameter | − | − | − | − | − | − | |
| 291 | O6-methyl-2′-deoxyguanosine | − | − | − | − | + | − | |
| 292 | 5-iodoridine | − | − | − | − | − | − | |
| 293 | 3-methylxanthine | − | − | − | − | − | − | |
| 294 | dioctylamine | + | − | − | − | + | − | * |
| 295 | didecylamine | + | − | + | + | − | + | |
| 296 | 5-(ethylisopropyl)amiloride | − | − | − | − | − | − | |
| 297 | 5-dimethylamiloride | − | − | − | − | − | − | |
| 298 | 5-(*N*,*N*-hexamethylene)amiloride | − | − | + | − | − | − | |
| 299 | guanazole | − | − | − | − | − | − | |
| 300 | 3-(methylthio)aniline | + | − | + | + | + | + | |
| 301 | neopterin D-erythro-form | − | − | − | − | − | − | |
| 302 | hycanthone | − | − | − | − | − | − | |
| 303 | *N*-methylhomoveratrylamine | − | − | − | − | − | − | |
| 304 | 1,4,7-triazacyclononane | − | − | + | + | − | + | * |
| 305 | methyl 2-amino-5-chlorobenzoate | − | − | − | − | − | − | |
| 306 | diaveridine | + | − | + | + | + | + | |
| 307 | 4-(diethylamino)benzoic acid | − | − | − | − | − | − | |
| 308 | althiazide | − | − | − | − | − | − | |
| 309 | nornicoti | − | − | − | + | − | − | |
| 310 | 3-amino-2-naphthoic acid | − | − | − | − | + | − | |
| 311 | 2′,3′-dideoxyuridine | − | − | − | − | − | − | |
| 312 | *N*-desmethylclozapine | + | − | − | + | + | + | |
| 313 | *N*-methylglucamine | − | − | + | − | − | − | |
| 314 | 1-aminofluorene | + | + | + | + | + | + | |
| 315 | proglumide | − | − | − | − | − | − | |
| 316 | 3-deazaadenosine | − | + | − | − | + | − | * |
| 317 | 2,4,6-triphenylaniline | − | − | − | − | − | − | |
| 318 | 3,3-dimethyl-1-*p*-tolyltriazene | + | + | + | + | + | + | |
| 319 | isoproterenol | − | − | − | − | + | − | |
| 320 | *N*,*N*′-bis(3-aminopropyl)ethylenediamine | − | − | − | + | + | + | * |
| 321 | 3,5-dimethylanthranilic acid | − | − | − | − | − | − | |
| 322 | 1,4,8,12-tetraazacyclopentadecane | + | + | + | − | + | + | |
| 323 | benomyl | − | − | − | − | − | − | |
| 324 | (S)-(−)-1,1′-binaphthyl-2,2′-diamine | − | − | + | + | − | + | * |
| 325 | 1-methyl-L-tryptophan | − | − | − | − | − | − | |
| 326 | 2-aminoacridone | − | − | + | − | + | + | * |
| 327 | 2-amino-4-chlorobenzonitrile | + | + | − | + | + | + | |
| 328 | albuterol | − | − | − | − | + | − | |
| 329 | pipemidic acid | + | + | − | − | − | − | * |
| 330 | 3-amino-*o*-cresol | − | + | + | − | − | − | * |
| 331 | *N*-(*tert*-butoxycarbonyl)-L-leucine methyl ester | − | − | − | − | − | − | |
| 332 | ganciclovir | − | − | − | − | − | − | |
| 333 | 5-iodo-2′,3′-dideoxyuridine | − | − | − | − | − | − | |
| 334 | N(4)-octadecyl-1-arabinofuranosylcytosine | − | − | − | − | − | − | |

[a] All predictions display results when each compound appeared in *only* the prediction sets. [b] Nontoxic = (−), toxic = (+).

were treated as only the parent compound for descriptor calculation purposes. Topological descriptors include atom, bond, and fragment counts as well as complex path and adjacency information derived from a variety of graph-invariant techniques.[59−68] These descriptors serve to capture the overall size, connectivity, and degree of branching of the molecules without an explicit knowledge of the three-dimensional coordinates. Geometric descriptors provide information about the bulk size and shape of the molecules. Examples include the solvent-accessible surface area,[69] molecular volume,[69] and the gravitational indices.[70] Atomic charge information for each structure was calculated by performing a single SCF iteration using AM1[71] parameters in MOPAC 6.0.[72,73] The heat of formation, dipole moment, and the LUMO−HOMO energy gap were also extracted from the MOPAC output and incorporated as descriptors for each molecule.

An additional set of descriptors that combine solvent-accessible surface area and partial atomic charge information are called hybrid descriptors. Examples include the charged partial surface area (CPSA) descriptors,[74] hydrogen-bonding descriptors,[57,75,76] and a new set of atom-specific CPSA descriptors. The atom-specific CPSA descriptors limit the original concept of CPSA properties to nitrogen, oxygen, sulfur, and halogen atoms *individually*.[77] Five descriptors for each of the four atom types are calculated as

$$\text{DESC-1} = \sum_{i=1}^{N} \text{SA}_i \quad (1)$$

$$\text{DESC-2} = \left(\sum_{i=1}^{N} \text{SA}_i\right)(Q_\text{T}) \quad (2)$$

$$\text{DESC-3} = \sum_{i=1}^{N} (\text{SA}_i)(Q_i) \quad (3)$$

$$\text{DESC-4} = \frac{\sum_{i=1}^{N} \text{SA}_i}{\text{SA}_{total}} \quad (4)$$

$$\text{DESC-5} = \frac{\sum_{i=1}^{N} (\text{SA}_i)(Q_i)}{\text{SA}_{total}} \quad (5)$$

where for a given atom type, $N$ is the number of atoms, $\text{SA}_i$ is the solvent-accessible surface area of the $i$th atom, $Q_i$ is the partial atomic charge on the $i$th atom, $\text{SA}_{total}$ is the total molecular surface area of the molecule, and $Q_\text{T}$ is the sum of partial atomic charges over the entire molecule. For clarity, "DESC" is a dummy label to illustrate the definition of specific atom types (e.g. NITR, OXYG, SULF, or HALO). This assessment of molecular surface area properties focuses only on the exposure of polar functionality, thus excluding contributions from hydrophobic groups.

Five additional descriptors that focused on the secondary and aromatic nitrogen atoms in each molecule were calculated specifically for this study.[77] The partial charges on the most negatively charged and least negatively charged secondary and aromatic nitrogen atoms were extracted from MOPAC output. Also, the solvent-accessible surface areas of the most accessible and the least accessible secondary and aromatic nitrogen atoms were calculated using the descriptor routine Surface Area and VOLume (SAVOL).[69] Finally, the partial charge on the most solvent-accessible nitrogen was also determined. It was hypothesized that the information content in these new descriptors would be useful for correlating the interactions of the nitrogen atoms—and ultimately the toxicity—to the molecular structure of the parent compound.

With the addition of AlogPS values[78] as a molecular hydrophobic measure, a total of 111 topological, 11 geometric, 10 electronic/quantum chemical, 51 CPSA, 23 hydrogen-bonding, and 5 nitrogen-specific descriptors were calculated—a total of 211 descriptors for each compound in the data set. The entire descriptor pool was first evaluated using objective feature selection to maximize the information content of the descriptors considered during the model-building phase. This approach requires no knowledge of the experimental data; rather, statistical properties of the descriptors themselves are used as criteria for inclusion in the reduced pool. An identical test was used to remove descriptors that contained the same value for greater than 84% of the compounds. This eliminated descriptors that lacked sufficient information to be useful for discriminating a vast majority of the data. Pairwise correlations were calculated for the remaining descriptors to identify those that provided redundant information. If two descriptors had a correlation coefficient of 0.84 or greater, one of the two descriptors was randomly removed. The values of 84% for the identical test and 0.84 for the pairwise correlation test were determined empirically to arrive at a reduced descriptor pool size that satisfied the following three criteria: (1) have a descriptor-to-training sample ratio of less than 0.6,[79] (2) contain less than 75 descriptors (an empirically determined threshold for our implementation of the GA), and (3) maintain diversity among the descriptor types present in the pool. A final reduced pool of 69 descriptors remained after the objective feature selection methods. Of these, 34 were topological, 7 were electronic/quantum chemical, 23 were CPSA and hydrogen-bonding, 4 were the secondary and aromatic nitrogen atom-specific descriptors, and the final one was AlogPS. This pool of descriptors was held constant throughout the entire model building process. Small subsets of descriptors were evaluated for each of the training sets, and the most predictive subsets were prioritized using a *k*NN classifier.

**Computational Methodology.** To ensure robust model development, an approach similar to model bagging[80] was employed. To establish diversity, nine models were built using a subsetting procedure, which entailed retraining with subsets of the original compound set. The data set was pseudorandomly split into three groups; the only constraint being that the same proportion of toxic and nontoxic compounds be present in each of the groups and that the proportion be approximately equal to that of the entire data set. A leave-one-group-out training procedure was used where one group served as the prediction set (~112 members), while the remaining compounds served as the training set (~222 members). *k*NN models were developed for all three combinations of training sets and prediction sets until every compound was used as a prediction set member once. The pseudorandom splitting procedure was repeated two more times to ultimately deliver nine models that were generated from nine unique representations of the training data. More importantly, this method of set formation allows each compound to be externally validated three times from different sets of training data, rather than from a single training set as is commonly practiced. This approach was used to ensure that consistent results could be obtained regardless of training set composition and to guarantee that results were not a consequence of a single training and prediction set—a serious concern when modeling structurally diverse data sets.[81]

The *k*NN method classifies samples based on a distance metric in an *n*-dimensional feature space. The ADAPT implementation of *k*NN uses an Euclidean distance metric given by eq 6

$$\mathbf{d}_{ij} = \sqrt{\sum_{k=1}^{n} [\mathbf{D}(i,k) - \mathbf{D}(j,k)]^2} \quad (6)$$

where $\mathbf{d}_{ij}$ is the distance from the $i$th to the $j$th compounds, $n$ is the number of descriptors, and $\mathbf{D}$ is the matrix of compounds × descriptors. The descriptors were first transformed on the range {0,1} to eliminate the bias of descriptor magnitudes. A leave-one-out (LOO) training method was

**Table 2.** Training and Prediction Set Classification Rates for the Nine Individual Models

| model | training set % (~222 members) | prediction set % (~112 members) | descriptors |
|---|---|---|---|
| 1 | 72.1 | 67.0 | SYMM, PND-3, DELH, ELEC, NITR-5, SAAA-2 |
| 2 | 71.6 | 66.1 | LOGP, MOLC-9, NDB, NRA, MDEC-24, ESUM, NITR-1, NITR-5, HALO-3 |
| 3 | 71.0 | 66.4 | SAHI, MDEO-22, MDEN-13, DELH, SAAA-2 |
| 4 | 76.1 | 67.9 | LOGP, SAHI, SYMM, EAVE, CARB, RNCS, RDTA |
| 5 | 70.3 | 67.9 | N5CH, 3SP2, MDEC-44, CARB, NITR-5 |
| 6 | 69.2 | 66.4 | LOGP, PND-1, PND-3, NITR-5, WNSA-2, CHAA-2 |
| 7 | 70.7 | 69.6 | QSHI, NSB, 2SP2, SYMM, MDEN-13, HARD, CARB |
| 8 | 68.9 | 68.8 | QLO, SAHI, MOLC-9, DPSA-3, SCAA-3 |
| 9 | 74.6 | 69.1 | SAHI, V5C, SYMM, PND-3, PND-6, ELOW, QNEG, HOMO, FPSA-3, SCAA-3 |
| average | 71.6 | 67.7 | |
| SD | ±2.4 | ±1.3 | |

used, which entails calculating the distance between sample *i* and its *k* nearest neighbors and assigning the majority class of the nearest neighbors to sample *i*. For two-class problems, setting *k* equal to an odd number of neighbors prevents a tie from occurring.

A GA/*k*NN search was used to identify several high-quality models for each training set/prediction set combination. The best subsets were those that produced the fewest number of misclassifications for the training data. The model chosen as best was then validated by calculating the distance between each prediction set member and its nearest neighbors in the training set and assigning a class prediction based on majority rule.

## RESULTS AND DISCUSSION

Subsets of 3−10 descriptors from the reduced pool of 69 were evaluated by the GA search. The number of nearest neighbors (*k*) was systematically varied from 1, 3, and 5 to find the value that gave optimal results. The best models found in this study contained as few as five descriptors and as many as 10 descriptors, with a value of *k* equal to three consistently giving the best results. Training and prediction set classification accuracies for the nine models are shown in Table 2. The average training set classification rate was 71.6 ± 2.4%, and the average prediction classification rate was 67.7 ± 1.3%. The low standard deviations for both the training and prediction sets demonstrate that consistent predictions were obtained for all models regardless of set membership. This may also indicate that an equivalent amount of chemical diversity was represented in each of the training sets—an anticipated result of averaging over many models.

A frequency plot summarizing the various descriptor classes represented in the nine models is shown in Figure 3. Topological descriptors appeared three times as frequently as any other descriptor class. These descriptors are attractive for modeling purposes because of quick computation time and the breadth of information they encode. Several of these descriptors provide connectivity information with regard to only nitrogen/heteroatoms or offer a measure of molecular flexibility or rigidity. Inclusion of the electronic/quantum chemical descriptors suggests that the electronic environment of molecules can be important when discriminating between toxic and nontoxic compounds. Not surprisingly, the appearance of the atom-specific CPSA (NITR-1, NITR-5, HALO-3) and nitrogen-specific descriptors (SAHI, QSHI, QLO), which were developed specifically for this study, verifies the idea that encoding specific accessibility and

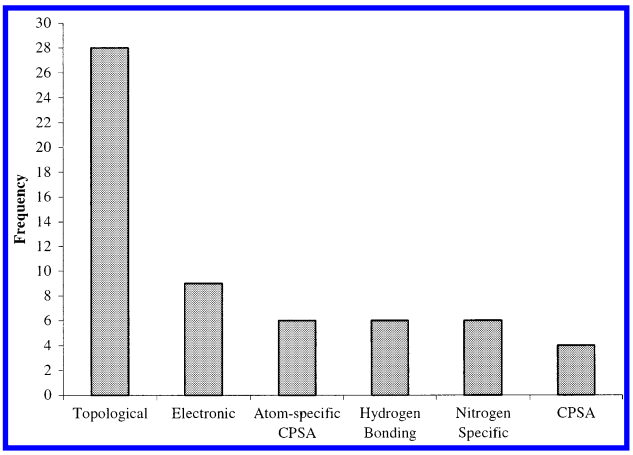


**Figure 3.** Frequency plot for descriptor classes appearing in the models for this study.

electronic properties around the chemical entities known to elicit toxic profiles can provide valuable information for correlative modeling. Finally, the traditional CPSA and hydrogen-bonding descriptors add proof that polar surface area properties lend important information about compound toxicity. This validates previous work that has shown the importance of these descriptors in addressing toxicological data sets.[22,23]

An extensive summary of the 40 unique descriptors present in at least one of the nine models is given in Table 3. Several of the descriptors appeared in multiple models implying that the information provided by these descriptors was particularly important when predicting the genetic toxicity of the compounds. SYMM, NITR-5, and SAHI appeared in four models, while PND-3, CARB, and LOGP appeared three times. Last, MOLC-9, MDEN-13, DELH, SCAA-3, and SAAA-2 are present in two of the nine models. The remaining descriptors listed in Table 3 appeared in only a single model.

The SYMM descriptor assesses the topological symmetry of the compounds by examining the number of unique atoms relative to the total atom count for each molecule. Unique atoms in this sense are those that possess unique molecular environments based on topological (atom type, connectivity) considerations. Molecules with SYMM values on the higher end of the {0,1} scale imply low symmetry, while SYMM values closer to zero indicate high symmetry. Based on average values for this descriptor, it was noted that toxic compounds were generally less symmetric than the nontoxic compounds in this data set. NITR-5 is a nitrogen-specific CPSA descriptor (see eq 5 above). This descriptor captures

**Table 3.** Symbols and Brief Explanations of the Descriptors Used in the Nine Models

| descriptor symbol | chemical meaning |
|---|---|
| | Topological |
| LOGP | calculated logarithm of *n*-octanol/water partition coefficient[78] |
| NSB | count of single bonds |
| NDB | count of double bonds |
| NRA | count of ring atoms |
| 2SP2 | count of $sp^2$-hybridized carbon atoms attached to two carbon atoms |
| 3SP2 | count of $sp^2$-hybridized carbon atoms attached to three carbon atoms |
| N5CH | count of fifth-order chains $\chi$ index[61,65] |
| V5C | valence-corrected fifth-order cluster $\chi$ index[61,65] |
| MOLC-9 | average distance sum connectivity topological index J[59] |
| MDEC-24 | molecular distance edge index between secondary/quaternary carbon atoms[66] |
| MDEC-44 | molecular distance edge index between quaternary carbon atoms[66] |
| MDEO-22 | molecular distance edge index between secondary oxygen atoms[66] |
| MDEN-13 | molecular distance edge index between primary/tertiary nitrogen atoms[66] |
| ESUM | sum of E-state values over all heteroatoms[67] |
| EAVE | average E-state value over all heteroatoms[67] |
| PND-1 | superpendentic index over all pendant atoms[60] |
| PND-3 | superpendentic index over all pendant nitrogen atoms[60] |
| PND-6 | superpendentic index over all pendant halogen atoms[60] |
| SYMM | topological symmetry index |
| ELOW | through-space distance between atoms with minimum and maximum E-state values (topological/geometry hybrid)[67] |
| | Electronic |
| DELH | heat of formation |
| ELEC | electronegativity, $^1/_2$(HOMO+LUMO) |
| HARD | absolute hardness, $^1/_2$(HOMO−LUMO) |
| CARB | average charge on carbonyl carbon atoms |
| HOMO | energy of the highest occupied molecular orbital |
| QNEG | partial charge on the most negative atom |
| | Atom-Specific CPSA[77] |
| NITR-1 | sum of surface areas over all nitrogen atoms |
| NITR-5 | sum of atomic charge weighted nitrogen surface area divided by total molecular surface area |
| HALO-3 | sum of atomic charge weighted halogen surface area |
| | CPSA[74] |
| RNCS | relative negatively charged surface area |
| WNSA-2 | surface weighted negatively charged partial surface area |
| DPSA-3 | difference in negatively and positively charge partial surface area |
| FPSA-3 | fractional positively charged partial surface area |
| | Hydrogen Bond[57,75,76] (Acceptor Atom = N, O, F, S) |
| SAAA-2 | ratio of acceptor atom surface area to total number of acceptor atoms |
| RDTA | ratio of number of donor hydrogen atoms to acceptor atoms |
| CHAA-2 | ratio of acceptor atom partial charge to total number of acceptor atoms |
| SCAA-3 | ratio of partial charge weighted acceptor atom surface area to total molecular surface area |
| | Secondary and Aromatic Nitrogen-Specific[77] |
| SAHI | surface area of the most accessible nitrogen atom |
| QSHI | partial charge of the most accessible nitrogen atom |
| QLO | partial charge on the most negatively charged nitrogen atom |

electronic polar surface area contributions of all nitrogen atoms relative to the solvent accessible surface area of the entire molecule. The general trend seen was that toxic compounds had lower (more negative) values for this descriptor when compared to the nontoxic class. This result indicates that toxic compounds have more negatively charged nitrogen atoms as well as more accessible nitrogen atoms, thus when combined in eq 5, the magnitude of the descriptor increases in the negative direction. This ultimately demonstrates that toxic compounds possess nitrogen atoms with high solvent accessibility but which are also highly reactive when taking charge considerations into account, which is a chemically intuitive result. SAHI is simply the solvent-accessible surface area of the most-accessible secondary or aromatic nitrogen atom. Interestingly, compounds with more accessible nitrogen atoms were generally more toxic implying that surface accessibility of nitrogen atoms contributes greatly to the resultant toxicity—another relatively intuitive result.

The superpendentic index calculated between pendant nitrogen atoms is represented by PND-3. This descriptor encodes the mean topological distance between terminal substituents in a molecule—in this case, only terminal, or primary nitrogen atoms. Higher values for the descriptor indicate a large number of terminal vertices or long path lengths between the terminal atoms in the molecule. Toxic compounds possessed lower values for this descriptor than nontoxic compounds. An interesting observation is that 151 of the compounds—106 of which are nontoxic compounds—do not possess a terminal nitrogen atom. This suggests that a primary amine attached to an aromatic system can be a major contributor to the toxicity of a compound. The average charge on carbonyl carbon atoms is supplied by CARB. The general trend seen with this descriptor was that nontoxic compounds had higher values compared to the toxic compounds. More specifically, only 63% of the nontoxic compounds (83 of 131 nontoxic compounds) in the training sets had a value of zero, while 88% of the toxic compounds (80 of 91 toxic compounds) had a value of zero for CARB. This suggests that CARB may serve largely as an indicator variable for a carbonyl substructure, rather than providing useful electronic information. To test this theory, nonzero values for the descriptor were replaced with a value of 1.0, thereby transforming the CARB descriptor into an indicator variable for the presence of a carbonyl group. The three models containing CARB were retrained using the new carbonyl indicator variable, and results were compared to those using the original values. Nearly identical results were seen for the training set classification rates, while no change was seen with the classification rates for the prediction set, thereby adding substantial proof that CARB is being used to indicate the presence of a carbonyl substructure. Alternatively, more carbonyls may also be serving to delocalize electron density from nearby nitrogen atoms—an electronic tendency observed for nontoxic compounds in the NITR-5 descriptor. The inclusion of LOGP is not surprising due to the fact that hydrophobicity has been a useful parameter in many QSAR studies of toxicity. Descriptor averages for the toxic and nontoxic classes showed conflicting trends, however. In two of the three models that contained the hydrophobicity descriptor, toxic compounds had marginally higher LOGP values, while one of the models showed that the nontoxic class possessed more hydrophobic compounds than the toxic class. Chemical intuition would argue in favor of the former case where toxicity is associated with more hydrophobic and thus more tissue soluble compounds. This example illustrates that even the most important parameters in pattern-recognition studies are susceptible to training set

**Table 4.** Confusion Matrix for Compounds When They Appeared in Three Prediction Sets Using the "Majority Rules" Predictions

| | predicted | | |
|---|---|---|---|
| actual | nontoxic | toxic | class % |
| nontoxic | 146 | 51 | 74.1 |
| toxic | 42 | 95 | 69.3 |
| | overall = 72.2% | | |

selection and that producing an ensemble of models will have a profound impact on averaging out discrepancies such as the one just described.

Many descriptors appeared less frequently in the models. One graph-invariant descriptor, MOLC-9, is encoding the relative size and degree of branching within each molecule, where higher values of MOLC-9 indicate a higher degree of branching. The class averages for this descriptor revealed that toxic compounds had higher values for MOLC-9 compared to the nontoxic class. Molecular distance edge (MDE) information is supplied by MDEN-13, which calculates mean topological distances between all primary and tertiary nitrogen atoms in a molecule. One requirement for this descriptor is that a compound must have more than one nitrogen atom to have a nonzero value. The trend observed was that nontoxic compounds yielded higher values than the toxic compounds. The molecular heat of formation appeared as a descriptor in two models. Toxic compounds tended to have higher values than nontoxic compounds for this descriptor. Finally, the descriptors SCAA-3 and SAAA-2 are encoding polar surface area specific to hydrogen-bond acceptor atoms (N, O, S, F). Nontoxic compounds possessed a lower average SCAA-2 descriptor value than toxic compounds, while average values for SAAA-2 were nearly identical for both toxic and nontoxic compounds. Unfortunately, this translates to minimal interpretability of the SAAA-2 descriptor for discrimination between the two compound classes.

A main advantage of the subsetting approach used in this work is that each compound is used for external validation of three unique training sets. The three individual predictions were averaged, and a global prediction was ultimately determined as the majority class. Predictions for each of the three models and the majority model are included in Table 1. Misclassified compounds are labeled with an asterisk in the "missed" column. A confusion matrix for the model ensemble is shown in Table 4. In this table, true class membership is shown along the columns and predicted class membership along the rows. A total classification rate of 72.2% (241 of 334 compounds) was achieved using this approach, thereby demonstrating the ability of a model ensemble to outperform any one of its individual constituent models. Furthermore, the use of more models could allow the user to assign a level of confidence to the average prediction for a given compound based upon the number of individual constituent models that correctly predicted its true class membership.

Table 5 shows a comparison between the results generated with the model ensemble presented in this work and three commercially available expert systems (DEREK,[10] MultiCASE,[82] TOPKAT[83]), which use structural alerts to predict the toxicological profile of a compound. It should be noted that DEREK, MultiCASE, and TOPKAT are parametrized

**Table 5.** Comparison between the Model Presented in This Study and Toxicity Predictions Obtained Using Three Commercially Available Expert Systems

| | DEREK (v 5.0) (%) | MultiCASE (A2H) (%) | TOPKAT (v 5.0) (%) | model ensemble (%) |
|---|---|---|---|---|
| % concordance[a] | 41 | 59 | 60 | 72 |
| false positives[b] | 59 | 25 | 22 | 15 |
| false negatives[c] | 0 | 16 | 18 | 13 |
| sensitivity[d] | 100 | 61 | 54 | 69 |
| specificity[e] | 0 | 57 | 63 | 74 |

[a] (Total number of correct predictions/number of compounds) * 100. [b] (Total number of false positive predictions/number of compounds) * 100. [c] (Total number of false negative predictions/number of compounds) * 100. [d] (Total number of correct positive predictions/number of positive compounds) * 100. [e] (Total number of correct negative predictions/number of negative compounds) * 100.

to predict Ames mutagenicity, but there is ~90% concordance between the SOS Chromotest and the Ames assay.[50] The classification rates (% concordance) obtained from the three software packages were less than the classification rates produced by the ensemble *k*NN model presented in the current study. In addition, with the exception of DEREK (which labeled every compound as toxic), the ensemble was more sensitive in classifying the toxic compounds correctly while also demonstrating better specificity in regards to the nontoxic class. These results reveal one of the inherent limitations in using only structural alerts for secondary and aromatic amine toxicity predictions.

To further test the validity and strength of the model ensemble, an additional set of amine compounds was ordered and processed with the SOS Chromotest several months after the original model development. This second "blind" prediction set consisted of the 18 compounds shown in Table 6. By averaging predictions for the nine individual models, the ensemble was able to correctly classify 17 of the 18 additional compounds (94.4%). Individual model predictions along with the resultant consensus prediction are also shown in Table 6. Classification rates for the individual models ranged from ~61% to ~89% demonstrating that a consensus approach can average out discrepancies from individual models to provide accurate predictions for query compounds.

A more thorough examination of the misclassifications revealed some interesting results that may rationalize poor predictions for some compounds. To reiterate, the average atom-pair similarity index for the data set was approximately 0.2. Sixteen of the 93 misclassified compounds (**65**, **66**, **93**, **159**, **164**, **256**, **261**, **273**, **274**, **275**, **282**, **284**, **287**, **294**, **304**, and **320**) possessed similarity values below the average indicating that these compounds were structurally dissimilar to the training data used to develop the model. This, in essence, is considered model extrapolation. One assumption of a robust QSAR model is that query compounds should encompass the same structure/descriptor space as the compounds used to develop the model. Extreme cases of structural dissimilarity are likely underrepresented in each of the individual models leading to poor predictions on average. That is to say, the best models chosen by the GA/*k*NN routine were comprised of descriptors that provided sufficient predictions for a *majority* of the compounds while structural features of the outliers were perhaps overlooked. It should be noted that many of the compounds in question

**Table 6.** Compound Number, Assay Result, and Predictions for the 18 Additional Amine Compounds Serving as the Second "Blind" Prediction Set

| | toxicity | | predictions | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| no.[a] | −S9 | +S9 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | consensus |
| 1 | − | − | − | − | + | − | − | − | − | − | + | − |
| 2 | − | − | − | − | − | − | − | − | − | − | − | − |
| 3 | − | − | − | − | − | − | − | − | − | − | − | − |
| 4 | + | − | − | − | + | + | + | + | − | + | + | + |
| 5 | − | − | + | − | + | − | − | − | + | − | + | − |
| 6 | + | + | − | − | − | − | + | − | − | − | + | −[b] |
| 7 | + | + | + | + | + | + | + | + | + | + | + | + |
| 8 | − | − | − | − | − | − | + | − | + | − | + | − |
| 9 | − | − | − | − | + | − | − | + | − | − | − | − |
| 10 | − | − | + | − | + | − | + | − | − | + | − | − |
| 11 | + | + | + | − | + | + | + | + | + | + | + | + |
| 12 | − | − | − | − | − | + | − | − | − | − | + | − |
| 13 | − | − | + | + | + | − | − | − | − | − | − | − |
| 14 | − | + | + | + | + | + | + | + | + | + | + | + |
| 15 | − | − | + | − | − | − | − | + | − | − | − | − |
| 16 | − | − | − | − | − | + | + | − | − | − | − | − |
| 17 | − | − | + | − | + | − | − | − | − | − | − | − |
| 18 | + | + | + | + | + | + | − | + | − | + | + | + |

[a] Compound names: (1) 1-naphthylamine-5-sulfonic acid, (2) 5-(2,5-dihydroxybenzylamino)-2-hydroxybenzoic acid, (3) 2′-chloro-5′-methyl-4′-nitrobenzanilide, (4) 3,5-xylidine, (5) 1-naphthylamine-4-sulfonic acid, (6) ethylenediamine-*N*,*N*′-diacetic acid, (7) 2,4-dichloroaniline, (8) 3-amino-4-methylbenzoic acid, (9) heptamethyleneimine, (10) diphenyl-*p*-phenylenediamine, (11) 2-aminodiphenylene oxide, (12) 1-amino-2,4-dibromoanthraquinone, (13) 3-amino-9-fluorenone, (14) 2-amino-*m*-cresol, (15) 2,4,6-tri-*tert*-butylaniline, (16) diethylenetriamine, (17) dihexylamine, (18) 1-amino-4-bromonaphthalene. [b] Misclassification.

**Table 7.** Misclassifications by Toxicity Grouping

| experimental | | no. of | total | misclassified by all |
|---|---|---|---|---|
| −S9 | +S9 | samples | misclassified | three predictions |
| − | − | 197 | 51 (25.9%) | 18 (9.1%) |
| − | + | 35 | 12 (34.3%) | 7 (20.0%) |
| + | − | 44 | 17 (38.6%) | 12 (27.2%) |
| + | + | 58 | 13 (22.4%) | 4 (6.9%) |

by this argument were among the small percentage of acyclic and nonaromatic structures in the data set.

Table 7 summarizes the distribution of misclassifications by experimentally determined toxicity classes. These results are further broken down into the number of compounds missed by the *majority* of the three predictions and the number of compounds missed by *all* three predictions. To summarize, only 9.1% of the compounds showing clear nontoxicity (−/−) were missed by all three models, while a total of 16.8% of compounds showing some form of toxicity were missed by all three models. Table 7 illustrates, however, that the model performs better when the −S9 and +S9 activated assays are in agreement on nontoxicity or toxicity. The fuzziness inherent in mixed toxicity profiles (−/+ or +/−) indeed poses a problem for superior predictions, albeit a reasonable majority of compounds are still being correctly classified. One interesting conclusion that should be noted is that compounds that demonstrate toxicity only in the metabolized assay (−/+) were predicted slightly better than compounds that showed toxicity only in the inactivated assay (+/−). This deviates from our initial belief that descriptors encoding information about only the parent molecules would likely result in greater confusion for the (−/+) samples since the associated toxicity is only indirectly related to the parent structure.

**Table 8.** Prediction Set Classification Rate Averages and Standard Deviations for the Random Models Developed to Test for Chance Correlations

| model | prediction set %[a] |
|---|---|
| 1 | 52.1 ± 3.3 |
| 2 | 50.0 ± 6.0 |
| 3 | 49.6 ± 5.1 |
| 4 | 52.5 ± 4.6 |
| 5 | 53.8 ± 5.8 |
| 6 | 51.6 ± 4.4 |
| 7 | 53.0 ± 5.3 |
| 8 | 49.9 ± 5.3 |
| 9 | 52.5 ± 3.2 |
| average[b] | 51.7 |
| SD[b] | ±4.8 |

[a] Results were obtained by averaging prediction set rates for a total of 10 random scramblings for each individual model. [b] Total average and standard deviation was obtained by averaging results from 90 random models, 10 for each of the nine runs.

Some generalizations can also be made about the limitations of this model based on substructural units that were commonly missed. Twenty compounds in the data set contained an O═S═O substructure; 10 of these were misclassified. All seven of the truly toxic compounds containing the sulfone moiety were missed, while three of the truly nontoxic compounds were misclassified as toxic. Another group of compounds that the model had difficulty classifying were simple aniline derivatives, including halogenated, tolyl, xylyl, and cresol variations. The common thread among these compounds was a phenyl ring, with as few as two substituents, as many as four substituents, and in various ortho, meta, and para substitution patterns. Of the 19 aniline derivatives that were misclassified, only four of the compounds were truly toxic. All of these were mislabeled as nontoxic compounds, while 15 of the nontoxic derivatives were classified as toxic. It would be reasonable to conclude that a compromise was reached during the model selection process that aimed to ensure more complex molecules be correctly classified at the cost of missing some of the simpler molecules. A local model of simple aniline derivatives would likely incorporate more graph theoretic descriptors that are better equipped at distinguishing between structural isomers, thus producing a more accurate model for these compounds.

The model also had difficulty providing accurate predictions for the acyclic molecules. Only 16 molecules in the data set contained no ring structure and half of these compounds were misclassified. It is safe to say that this structural subclass was grossly underrepresented, and the presence of more samples would likely boost the prediction accuracy for these compounds.

**Testing for Chance Correlations.** Random class assignments for this data set would statistically propagate to approximately 52%, while prediction set errors for all single models presented were 68% on average. Due to this limited margin, additional experiments were conducted to ensure that chance correlations were not responsible for the observed classification rates. The true class labels for all individual models were randomly scrambled 10 times such that each molecule was assigned the experimental toxicity profile of another compound. The GA/*k*NN algorithm was then used

to identify the best models containing 5−10 descriptors. If the real models indeed convey a true quantitative structure−activity relationship, then the models built with random class labels should yield very poor validation results. The prediction set average and standard deviation for the random models is shown in Table 8. With respect to the validation results of the nine models listed in Table 2, the average prediction set classification accuracies using scrambled data were 52.1%, 50.0%, 49.6%, 52.5%, 53.8%, 51.6%, 53.0%, 49.9%, and 52.5%. This demonstrates that randomizing the experimental values adversely affects the predictive ability of models for screening the genotoxicity endpoints of the compounds. While not entirely conclusive, these results indicate that chance effects likely played a minimal role, if any, in the model formation process.

## CONCLUSION

The best *k*NN models found in this study produced average classification rates of 71.6 ± 2.4% and 67.7 ± 1.3%, respectively, for the nine different training and prediction sets. The low standard deviations demonstrate that consistent results can be obtained regardless of the compounds used for model development. The model ensemble was able to correctly predict the genotoxicity for the entire data set with a 72.2% accuracy rate proving that using a committee of models can produce superior results when compared to the individual models used to develop the ensemble. It was also shown that the ensemble model was able to provide more accurate predictions than three commercially available expert systems: DEREK, MultiCASE, and TOPKAT. An additional experiment involving a second "blind" prediction set also demonstrated that the ensemble could accurately (∼94%) predict the genotoxicity of 18 query compounds tested several months after the original model development. The chemical meanings of the descriptors that appeared frequently in the models were discussed with possible explanations as to their link between chemical structure and genetic toxicity. The results also showed that the models had better success in labeling the genotoxicity of compounds when assay results were in agreement for the S9 inactivated/activated assays.

## ACKNOWLEDGMENT

The authors would like to acknowledge Sondra Livingston-Carr, Oneal Puri, and Jennifer Price.

## REFERENCES AND NOTES

(1) Johnson, D. E.; Wolfgang, G. H. I. Predicting Human Safety: Screening and Computational Approaches. *Drug Discov. Today* **2000**, *5*, 445−454.
(2) Fink, S. I.; Leo, A.; Yamakawa, M.; Hansch, C.; Quinn, F. R. The Quantitative Structure-Selectivity Relationship of Anthracycline Antitumor Activity and Cardiac Toxicity. *Farmaco* **1980**, *35*, 965−979.
(3) Quinn, F. R.; Neiman, Z.; Beisler, J. A. Toxicity Quantitative Structure−Activity Relationships of Colchicines. *J. Med. Chem.* **1981**, *24*, 636−639.
(4) Cronin, M. T. D. Computational Methods for the Prediction of Drug Toxicity. *Curr. Opin. Drug Discov. Devel.* **2000**, *3*, 292−297.
(5) Matthews, E. J.; Benz, R. D.; Contrera, J. F. Use of Toxicological Information in Drug Design. *J. Mol. Graph. Model.* **2000**, *18*, 605−615.
(6) Pearl, G.; Livingston-Carr, S.; Durham, S. Integration of Computational Analysis as a Sentinel Tool in Toxicological Assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247−255.
(7) Durham, S.; Pearl, G. Computational Methods to Predict Drug Safety Liabilities. *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 110−115.
(8) Richard, A. M. Structure-Based Methods for Predicting Mutagenicity and Carcinogenicity: Are We There Yet? *Mutat. Res. − Fundam. Mol. Mech. Mut.* **1998**, *400*, 493−507.
(9) Greene, N. Computer Systems for the Prediction of Toxicity: An Update. *Adv. Drug Deliver. Rev.* **2002**, *54*, 417−431.
(10) Sanderson, D. M.; Earnshaw, C. G. Computer-Prediction of Possible Toxic Action from Chemical-Structure − The DEREK System. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.
(11) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.
(12) Woo, Y. T.; Lai, D. Y.; Argus, M. F.; Arcos, J. C. Development of Structure Activity Relationship Rules for Predicting Carcinogenic Potential of Chemicals. *Toxicol. Lett.* **1995**, *79*, 219−228.
(13) Bacha, P. A.; Gruver, H. S.; Hartog, B. K. D.; Tamura, S. Y.; Nutt, R. F. Rule Extraction from a Mutagenicity Data Set Using Adaptively Grown Phylogenetic-like Trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1104−1111.
(14) Meyer, H. Lipoidtheorief der Narkosen. *Arch. Exp. Pathol. Pharmakol.* **1899**, *42*, 109.
(15) Overton, E. *Studien uber die Narkosen*; Gustav Fischer: Jena, Germany, 1901.
(16) Rekker, R. F. *The Hydrophobic Fragmental Constant*; Elsevier: New York, 1977.
(17) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96−103.
(18) Taft, R. W. Separation of Polar, Steric, and Resonance Effects in Reactivity. In *Steric Effects in Organic Chemistry*; Newman, M. S., Ed.; John Wiley & Sons: New York, 1956.
(19) Benigni, R.; Passerini, L.; Gallo, G.; Giorgi, F.; Cotta-Ramusino, M. QSAR Models for Discriminating Between Mutagenic and Nonmutagenic Aromatic and Heteroaromatic Amines. *Environ. Mol. Mutagen.* **1998**, *32*, 75−83.
(20) Chignell, C. F. Overview of Molecular Parameters that Relate to Biological Activity in Toxicology. In *Structure−Activity Correlation as a Predictive Tool in Toxicology: Fundamentals, Methods, and Applications*; Golberg, L., Ed. Hemisphere: Washington, DC, 1983.
(21) Lewis, D. F. W. Computer-Assisted Methods in the Evaluation of Chemical Toxicity. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1992.
(22) Serra, J. R.; Jurs, P. C.; Kaiser, K. L. E. Linear Regression and Computational Neural Network Prediction of *Tetrahymena* Acute Toxicity for Aromatic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **2001**, *14*, 1535−1545.
(23) Eldred, D. V.; Weikel, C. L.; Jurs, P. C.; Kaiser, K. L. E. Prediction of Fathead Minnow Acute Toxicity of Organic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **1999**, *12*, 670−678.
(24) Ashby, J.; Tennant, R. W. Chemical-Structure, Salmonella Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis Among 222 Chemicals Tested in Rodents by the United-States NCI/NTP. *Mutat. Res.* **1988**, *204*, 17−115.
(25) Ashby, J.; Paton, D.; Lefevre, P. A.; Styles, J. A.; Rose, F. L. Evaluation of Two Suggested Methods of Deactivating Organic Carcinogens by Molecular Modification. *Carcinogenesis* **1982**, *3*, 1277−1282.
(26) Kugler-Steigmeier, M. E.; Friederich, U.; Graf, U.; Lutz, W. K.; Maier, P.; Schlatter, C. Genotoxicity of Aniline Derivatives in Various Short-Term Tests. *Mutat. Res.* **1989**, *211*, 279−289.
(27) Colvin, M. E.; Hatch, F. T.; Felton, J. S. Chemical and Biological Factors Affecting Mutagen Potency. *Mutat. Res. − Fundam. Mol. Mech. Mut.* **1998**, *400*, 479−492.
(28) Newcomb, M.; Toy, P. H. Hypersensitive Radical Probes and the Mechanisms of Cytochrome P450-catalyzed Hydroxylation Reactions. *Acc. Chem. Res.* **2000**, *33*, 449−455.
(29) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity if Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37−52.
(30) Benigni, R.; Passerini, L. Carcinogenicity of the Aromatic Amines: from structure−activity relationships to mechanisms of action and risk assessment. *Mutat. Res. − Rev. Mutat. Res.* **2002**, *511*, 191−206.
(31) Franke, R.; Gruska, A.; Giuliani, A.; Benigni, R. Prediction of Rodent Carcinogenicity of Aromatic Amines: A quantitative structure−activity relationships model. *Carcinogenesis* **2001**, *22*, 1561−1571.
(32) Benigni, R.; Andreoli, C.; Giuliani, A. QSAR Models for Both Mutagenic Potency and Activity: Application to Nitroarenes and Aromatic Amines. *Environ. Mol. Mutagen.* **1994**, *24*, 208−219.
(33) Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Struc-

ture: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 671−678.
(34) Hatch, F. T.; Knize, M. G.; Colvin, M. E. Extended Quantitative Structure−Activity Relationships for 80 Aromatic and Heterocyclic Amines: Structural, Electronic, and Hydropathic Factors Affecting Mutagenic Potency. *Environ. Mol. Mutagen.* **2001**, *38*, 268−291.
(35) Cash, G. G. Prediction of the Genotoxicity of Aromatic and Heteroaromatic Amines Using Electrotopological State Indices. *Mutat. Res. − Gen. Tox. Environ. Mut.* **2001**, *491*, 31−37.
(36) Karelson, M.; Sild, S.; Maran, U. Nonlinear QSAR treatment of genotoxicity. *Mol. Simulat.* **2000**, *24*, 229−242.
(37) Maran, U.; Karelson, M.; Katritzky, A. R. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3−10.
(38) Yuan, M.; Jurs, P. C. Computer Assisted Structure−Activity Studies of Chemical Carcinogens. Polycyclic aromatic hydrocarbons. *Toxicol. Appl. Pharmacol.* **1980**, *52*, 294−312.
(39) Yuta, K.; Jurs, P. C. Computer Assisted Structure−Activity Studies of Chemical Carcinogens. Aromatic amines. *J. Med. Chem.* **1981**, *24*, 241−251.
(40) Chou, J. T.; Jurs, P. C. Computer Assisted Structure−Activity Studies of Chemical Carcinogens. An *N*-nitroso compound data set. *J. Med. Chem.* **1979**, *22*, 792−797.
(41) Jurs, P. C.; Noor Hasan, M.; Henry, D. R.; Stouch, T. R.; Whalen-Pederson, E. K. Computer-Assisted Studies of Molecular Structure and Carcinogenic Activity. *Fundam. Appl. Toxicol.* **1983**, *3*, 343−349.
(42) Eldred, D. V.; Jurs, P. C. Prediction of Acute Mammalian Toxicity of Organophosphorous Pesticide Compounds from Molecular Structure. *SAR QSAR Environ. Res.* **1999**, *10*, 75−99.
(43) Johnson, S. R.; Jurs, P. C. Prediction of Acute Mammalian Toxicity from Molecular Structure for a Diverse Set of Substituted Anilines Using Regression Analysis and Computational Neural Networks. In *Computer-Assisted Lead Finding and Optimization*; Waterbeemd, B., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997.
(44) Serra, J. R.; Thompson, E. D.; Jurs, P. C. Development of Binary Classification of Structural Chromosome Aberrations for a Diverse Set of Organic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **2003**, *16*, 153−163.
(45) Stouch, T. R.; Jurs, P. C. Computer-Assisted Studies of Molecular Structure and Genotoxic Activity by Pattern Recognition Techniques. *Environ. Health Perspec.* **1985**, *61*, 329−343.
(46) Mosier, P. D.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Thiophene Derivatives from Molecular Structure. *Chem. Res. Toxicol.* **2003**, accepted for publication.
(47) Mitchell, T.; Showell, G. A. Design Strategies for Building Drug-like Chemical Libraries. *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 314−318.
(48) Hofnung, M.; Quillardet, P. The SOS Chromotest, a Colorimetric Assay Based on the Primary Cellular Responses to Genotoxic Agents. *Ann. N. Y. Acad. Sci.* **1988**, *534*, 817−825.
(49) Sutton, M. D.; Smith, B. T.; Godoy, V. G.; Walker, G. C. The SOS Response: Recent Insights into *umuDC*-Dependent Mutagenesis and DNA Damage Tolerance. *Annu. Rev. Genetics* **2000**, *34*, 479−497.
(50) Quillardet, P.; Hofnung, M. The SOS Chromotest − A Review. *Mutat. Res.* **1993**, *297*, 235−279.
(51) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(52) Jurs, P. C.; Chou, J. T.; Yuan, M. Studies of Chemical Structure-Biological Activity Relations Using Pattern Recognition. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.
(53) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. A Computer System for Structure−Activity Studies Using Chemical Structure Information Handling and Pattern Recognition Techniques. In *Chemometrics: Theory and Application*; Kowalski, B. R., Ed.; American Chemical Society: Washington, DC, 1977.
(54) Lucasius, C. B.; Kateman, G. Understanding and Using Genetic Algorithms Part 1. Concepts, properties, and context. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 1−33.
(55) Hibbert, D. B. Genetic Algorithms in Chemistry. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 277−293.
(56) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(57) Wessel, M. D. Computer-Assisted Development of Quantitative Structure−Property Relationships and Design of Feature Selection Routines. Ph.D. Thesis, Department of Chemistry, The Pennsylvania State University, University Park, PA, 1997.
(58) Cover, T. M.; Hart, P. E. Nearest Neighbor Pattern Classification. *IEEE T Inform. Theory* **1967**, *IT-13*, 21−27.
(59) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(60) Gupta, S.; Singh, M.; Madan, A. K. Superpendentic Index: A Novel Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 272−277.
(61) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(62) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
(63) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.
(64) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.
(65) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd., John Wiley & Sons: Letchworth, Hertfordshire, England, 1986.
(66) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, *λ*. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(67) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(68) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(69) Pearlman, R. S. Molecular Surface Areas and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(70) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.
(71) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(72) *MOPAC*, v. 6.0; Stewart, J. P. P. Quantum Chemistry Program Exchange, Program 455; Indiana University: Bloomington, IN.
(73) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(74) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(75) Pimentel, G. I.; McClellan, A. L. *The Hydrogen Bond*; Freeman: San Francisco, 1960.
(76) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; Van Nostrand Reinhold: New York, 1971.
(77) Kauffman, G. W. The Development of Predictive Models for Physical and Biological Properties from Molecular Structure and The Analysis of Data from a Conducting Polymer Chemiresistive Sensor Array. Ph.D. Thesis, Department of Chemistry, The Pennsylvania State University, University Park, PA, 2002.
(78) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. P. Prediction of *n*-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-state Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407−1421.
(79) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative-Structure Property Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(80) Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123−140.
(81) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(82) Klopman, G. Predicting Toxicity through a Computer Automated Structure Evaluation Program. *Environ. Health Perspec.* **1984**, *61*, 269.
(83) Enslein, K.; Gombar, V. K.; Blake, B. W. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat. Res.* **1994**, *305*, 47−61.

CI034013I