Inter*net* Electronic Journal of
Molecular Design

December 2004, Volume 3, Number 12, Pages 802–821

Editor: Ovidiu Ivanciuc

Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against *Pimephales promelas* and *Tetrahymena pyriformis*

Ovidiu Ivanciuc

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–0857

Received: June 30, 2003; Revised: October 1, 2003; Accepted: November 7, 2003; Published: December 31, 2004

Citation of the article:
O. Ivanciuc, Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against *Pimephales promelas* and *Tetrahymena pyriformis*, *Internet Electron. J. Mol. Des.* **2004**, *3*, 802–821, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against *Pimephales promelas* and *Tetrahymena pyriformis*

Ovidiu Ivanciuc*

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–0857

Received: June 30, 2003; Revised: October 1, 2003; Accepted: November 7, 2003; Published: December 31, 2004

Internet Electron. J. Mol. Des. **2004**, *3* (12), 802–821

Abstract

Motivation. The prediction of the mechanism of action (MOA) using structural descriptors has major applications in selecting the appropriate quantitative structure–activity relationships (QSAR) model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimes.

Method. The SVM (support vector machines) algorithm was recently proposed as an efficient and flexible classification method for various bioinformatics and cheminformatics applications. In this study we have investigated the application of SVM for the classification of 337 organic compounds from eight MOA classes (nonpolar narcosis, polar narcosis, ester narcosis, amine narcosis, weak acid respiratory uncoupling, electrophilicity, proelectrophilicity, and nucleophilicity). The MOA classification was based on three indices, namely: log K_{ow}, the octanol-water partition coefficient; log $1/IGC_{50}$, the 50% inhibitory growth concentration against *Tetrahymena pyriformis*; log $1/LC_{50}$, the 50% lethal concentration against *Pimephales promelas*. The prediction power of each SVM model was evaluated with a leave–5%–out cross–validation procedure.

Results. In order to find classification models with good predictive power, we have investigated a large number of SVM models obtained with the dot, polynomial, radial basis function, neural, and anova kernels. The MOA classification performances of SVM models depend strongly on the kernel type and various parameters that control the kernel shape. The discrimination between nonpolar narcotic compounds and the other chemicals can be obtained with radial and anova SVM models, with a prediction accuracy of 0.80. The separation of less reactive compounds (polar, ester, and amine narcotics) from more reactive compounds (electrophiles, proelectrophiles, and nucleophiles) is obtained with a slightly higher error (prediction accuracy 0.71, obtained with radial SVM models).

Conclusions. SVM models that use as input parameters hydrophobicity and experimental toxicity against *Pimephales promelas* and *Tetrahymena pyriformis* represent an effective MOA classification method for a large diversity of organic compounds. This approach can be used to predict the aquatic toxicity mechanism and to select the appropriate QSAR model for new chemical compounds.

Keywords. Support vector machines; structure–toxicity relationships; quantitative structure–activity relationships; QSAR; aquatic toxicity; mechanism of action; *Tetrahymena pyriformis*; *Pimephales promelas*.

* Correspondence author; E–mail: ovidiu_ivanciuc@yahoo.com.

1 INTRODUCTION

Because numerous organic chemicals can be environmental pollutants, considerable efforts were directed towards the study of the relationships between a compound's structure and its toxicity [1]. Significant progress has been made to classify chemical compounds according to their mechanism of toxicity and to screen them for their environmental risk assessment. Using theoretical descriptors computed from the molecular structure and various classification algorithms, the prediction of the mechanism of action (MOA) has major applications in identifying chemicals with similar toxicity mechanism, in selecting the appropriate quantitative structure–activity relationships (QSAR) model, and in extrapolating toxic effects across species and exposure regimes when limited experimental data are available [2–18].

In a recent investigation, Ren, Frymier, and Schultz [16] assembled a set of 337 organic compounds from eight MOA classes (nonpolar narcosis, polar narcosis, ester narcosis, amine narcosis, weak acid respiratory uncoupling, electrophilicity, proelectrophilicity, and nucleophilicity). Ren *et al*. used three indices for MOA classification, namely the octanol-water partition coefficient, the 50% inhibitory growth concentration against *Tetrahymena pyriformis*, and the 50% lethal concentration against *Pimephales promelas*. Using discriminant analysis and logistic regression, they were able to discriminate nonpolar narcotic compounds from other MOA compounds. However, the separation of less reactive compounds (polar, ester, and amine narcotics) from more reactive compounds (electrophiles, proelectrophiles, and nucleophiles) was less successful.

Support vector machines (SVM) represent a new class of machine learning algorithms that found numerous applications in various classification and regression models, particularly in MOA prediction [17,18]. In this study we have investigated the application of SVM for the recognition of the aquatic toxicity mechanism for the compounds previously explored by Ren *et al*. [16].

2 MATERIALS AND METHODS

2.1 Chemical Data

In this study we have investigated the application of SVM for the classification of 337 organic compounds from eight MOA classes [16] (126 nonpolar narcosis, 79 polar narcosis, 23 ester narcosis, 13 amine narcosis, 13 weak acid respiratory uncoupling, 69 electrophilicity, 8 proelectrophilicity, and 6 nucleophilicity). The MOA classification was based on three indices taken from [16], namely: $\log K_{ow}$, the octanol-water partition coefficient; $\log 1/IGC_{50}$, the 50% inhibitory growth concentration against *Tetrahymena pyriformis*; $\log 1/LC_{50}$, the 50% lethal concentration against *Pimephales promelas*. The name of the compounds, $\log K_{ow}$, $\log 1/IGC_{50}$, $\log 1/LC_{50}$, and MOA are presented in Table 1.

Table 1. Structure of the chemical compounds, the octanol-water partition coefficient log K_{ow}, the 50% inhibitory growth concentration against *Tetrahymena pyriformis* log $1/IGC_{50}$, the 50% lethal concentration against *Pimephales promelas* log $1/LC_{50}$, and mechanism of toxic action (1, nonpolar narcosis; 2, polar narcosis; 3, ester narcosis; 4, amine narcosis; 5, weak acid respiratory uncoupling; 6, electrophilicity; 7, proelectrophilicity; 8, nucleophilicity) [16].

No	Compound	CAS No	log K_{ow}	log $1/IGC_{50}$	log $1/LC_{50}$	MOA
1	ethylcarbamate	51–79–6	–0.15	–1.65	–1.77	1
2	benzamide	55–21–0	0.64	–0.91	–0.74	1
3	ethanol	64–17–5	–0.31	–2.31	–2.49	1
4	methanol	67–56–1	–0.74	–2.72	–2.96	1
5	2–propanol	67–63–0	0.05	–1.88	–2.21	1
6	acetone	67–64–1	–0.24	–2.15	–2.10	1
7	dimethylsulfoxide	67–68–5	–1.35	–2.49	–2.64	1
8	1–propanol	71–23–8	0.25	–1.75	–1.87	1
9	1–butanol	71–36–3	0.84	–1.43	–1.37	1
10	1–pentanol	71–41–0	1.51	–1.07	–0.73	1
11	acetonitrile	75–05–8	–0.34	–2.28	–1.60	1
12	2–methyl–2–propanol	75–65–0	0.35	–1.79	–1.94	1
13	2–methyl–1–propanol	78–83–1	0.76	–1.37	–1.29	1
14	2–butanone	78–93–3	0.29	–1.75	–1.65	1
15	naphthalene	91–20–3	3.35	–0.12	1.32	1
16	quinoline	91–22–5	2.03	0.09	0.22	1
17	*N*,*N*–diethylaniline	91–66–7	3.31	0.67	0.96	1
18	biphenyl	92–52–4	3.98	1.05	1.90	1
19	benzothiazole	95–16–9	2.02	–0.03	0.33	1
20	1,2–dichlorobenzene	95–50–1	3.38	0.53	1.19	1
21	3,4–dichlorotoluene	95–75–0	3.95	1.22	1.74	1
22	3–pentanone	96–22–0	0.82	–1.46	–1.25	1
23	isopropylbenzene	98–82–8	3.66	0.69	1.28	1
24	acetophenone	98–86–2	1.63	–0.46	–0.13	1
25	α,α–dimethylbenzenepropanol	103–05–9	2.42	–0.07	0.39	1
26	*N*,*N*–dimethylbenzylamine	103–83–3	1.98	0.12	0.55	1
27	*n*–butylbenzene	104–51–8	4.26	1.25	1.83	1
28	2–ethyl–1–hexanol	104–76–7	2.81	0.17	0.66	1
29	5–ethyl–2–methylpyridine	104–90–5	2.17	–0.18	0.17	1
30	propionitrile	107–12–0	0.16	–1.97	–1.44	1
31	2–pentanone	107–87–9	1.31	–1.44	–1.16	1
32	4–methyl–2–pentanone	108–10–1	1.31	–1.21	–0.73	1
33	bromobenzene	108–86–1	2.99	0.08	0.94	1
34	4–methylpyridine	108–89–4	1.22	–0.88	–0.64	1
35	cyclohexanol	108–93–0	1.23	–0.77	–0.85	1
36	3–methylpyridine	108–99–6	1.2	–0.99	–0.19	1
37	2–methylpyridine	109–06–8	1.11	–1.01	–0.98	1
38	pyrrole	109–97–7	0.75	–1.09	–0.50	1
39	5–methyl–2–hexanone	110–12–3	1.88	–0.65	–0.14	1
40	2–heptanone	110–43–0	1.98	–0.49	–0.06	1
41	6–methyl–5–hepten–2–one	110–93–0	2.21	–0.45	0.83	1
42	2–octanone	111–13–7	2.37	–0.15	0.55	1
43	1–hexanol	111–27–3	2.03	–0.38	0.02	1
44	1–heptanol	111–70–6	2.62	0.10	0.53	1
45	1–octanol	111–87–5	3.07	0.58	0.98	1
46	2–undecanone	112–12–9	4.09	1.50	3.06	1
47	decylalcohol	112–30–1	4.57	1.34	1.82	1
48	1–tridecanol	112–70–9	5.58	2.37	2.59	1
49	benzophenone	119–61–9	3.18	0.87	1.08	1
50	1,2,4–trichlorobenzene	120–82–1	4.02	1.08	1.78	1
51	*N*,*N*–dimethylaniline	121–69–7	2.31	0.28	0.36	1
52	dibenzofuran	132–64–9	4.12	1.42	1.96	1
53	1–nonanol	143–08–8	4.02	0.86	1.40	1

Table 1. (Continued)

No	Compound	CAS No	log K_{ow}	log 1/IGC$_{50}$	log 1/LC$_{50}$	MOA
54	5–nonanone	502–56–7	2.97	0.07	0.66	1
55	2–tolunitrile	529–19–1	2.21	–0.24	0.42	1
56	*N*–amylbenzene	538–68–1	4.9	1.79	1.94	1
57	3–methyl–2–butanone	563–80–4	0.56	–1.17	–1.00	1
58	2–hexanone	591–78–6	1.38	–1.34	–0.63	1
59	2–tridecanone	593–08–8	5.08	2.12	2.74	1
60	2–decanone	693–54–9	3.77	0.58	1.44	1
61	2–nonanone	821–55–6	3.16	0.66	0.97	1
62	*cis*–3–hexen–1–ol	928–96–1	1.4	–0.81	–0.58	1
63	4–phenylpyridine	939–23–1	2.58	0.66	0.98	1
64	octylcyanide	2243–27–8	3.12	0.62	1.40	1
65	undecylcyanide	2437–25–4	4.9	1.90	2.63	1
66	5–chloro–2–pyridinol	4214–79–3	1.76	–0.75	–0.94	1
67	1–hexen–3–ol	4798–44–1	1.34	–0.81	0.52	1
68	1,2–bis–(4–pyridyl)–ethane	4916–57–8	1.59	–0.03	0.09	1
69	2–dimethylaminopyridine	5683–33–0	1.43	–0.55	–0.02	1
70	2–dodecanone	6175–49–1	4.55	1.67	2.19	1
71	4–bromophenyl–3–pyridylketone	14548–45–9	2.96	0.82	1.11	1
72	4–benzoylpyridine	14548–46–0	1.98	–0.09	0.25	1
73	6–chloro–2–picoline	18368–63–3	1.89	–0.48	–0.26	1
74	acrylamide	79–06–1	–0.78	–0.79	–0.19	1
75	nitrobenzene	98–95–3	1.85	0.14	0.02	1
76	3–nitrotoluene	99–08–1	2.45	0.42	0.73	1
77	dibutylfumarate	105–75–9	3.96	1.49	2.52	1
78	2–methyl–3–butyn–2–ol	115–19–5	0.28	–1.49	–1.59	1
79	acetoneoxime	127–06–0	0.12	–1.25	–0.88	1
80	2–phenyl–3–butyn–2–ol	127–66–2	1.88	–0.18	0.11	1
81	4–pentyn–2–ol	2117–11–5	0.12	–1.63	0.38	1
82	3–bromothiophene	872–31–1	2.62	–0.04	1.42	1
83	1,4–dicyanobutane	111–69–3	–0.32	–1.54	–1.25	1
84	3–methyl–1–pentyn–3–ol	77–75–8	1.07	–1.32	–1.09	1
85	cyclohexanoneoxime	100–64–1	1.19	–0.80	–0.26	1
86	*N*,*N*–diethylacetamide	685–91–6	1.75	–1.54	–1.11	1
87	anthranilamide (2–aminobenzamide)	88–68–6	0.35	–0.87	–0.46	1
88	2–acetyl–1–methylpyrrole	932–16–1	0.84	–0.69	–0.11	1
89	2',3',4'–trichloroacetophenone	13608–87–2	3.21	1.34	2.05	1
90	2',4'–dichloroacetophenone	2234–16–4	2.73	0.62	1.21	1
91	4–nitrophenylphenylether	620–88–2	3.83	1.58	1.91	1
92	triphenylphosphineoxide	791–28–6	2.83	0.77	0.71	1
93	triphenylphosphate	115–86–6	4.59	1.81	2.57	1
94	carbon tetrachloride	56–23–5	2.83	–0.02	0.57	1
95	chloroacetonitrile	107–14–2	0.45	0.85	1.75	1
96	malononitrile	109–77–3	–0.6	0.22	2.07	1
97	1,2,4–trichlorobenzene	120–82–1	4.02	1.08	1.78	1
98	*trans*–3–hexen–1–ol	928–97–2	1.4	–0.78	–0.43	1
99	*n*–propylsulfide	111–47–7	2.96	0.00	0.74	1
100	toluene	108–88–3	2.73	–0.50	0.41	1
101	4–xylene	106–42–3	3.15	0.12	1.08	1
102	butylsulfide	544–40–1	4.02	1.04	1.61	1
103	1–benzoylacetone	93–91–4	2.52	0.57	2.17	1
104	acetaldoxime	107–29–9	–0.13	–0.89	–0.11	1
105	2–butanoneoxime	96–29–7	0.65	–1.07	–0.99	1
106	γ–decanolactone	706–14–9	2.72	0.49	0.98	1
107	1–bromohexane	111–25–1	3.8	0.94	1.68	1
108	5–methyl–2–hexanone	110–12–3	1.88	–0.65	–0.14	1

Table 1. (Continued)

No	Compound	CAS No	log K_{ow}	log 1/IGC_{50}	log 1/LC_{50}	MOA
109	1–bromoheptane	629–04–9	4.36	1.49	2.09	1
110	2,3–benzofuran	271–89–6	2.67	–0.11	0.93	1
111	1–bromooctane	111–83–1	4.89	1.87	2.36	1
112	2–methylimidazole	693–98–1	–0.4	–0.91	–0.54	1
113	2,4,5–tribromoimidazole	2034–22–2	1.96	1.43	1.58	1
114	1–methylpiperazine	109–01–3	–0.1	–0.96	–1.36	1
115	2,4,5–trimethyloxazole	20662–84–4	0.58	–1.01	–0.61	1
116	cyclohexanone	108–94–1	0.81	–1.23	–0.87	1
117	3,3–dimethyl–2–butanone	75–97–8	1.2	–1.44	0.06	1
118	2–methyl–2,4–pentanediol	107–41–5	–0.68	–1.96	–1.96	1
119	1–(2–aminioethyl)piperazine	140–31–8	–0.6	–0.79	–1.23	1
120	methyl–4–chlorobenzoate	1126–46–1	2.76	0.42	1.19	1
121	1,6–dicyanohexane	629–40–3	0.59	–0.77	–0.59	1
122	3–methylindole	83–34–1	2.6	0.37	1.17	1
123	flavone	525–82–6	3.56	1.41	1.80	1
124	2–ethylpyridine	100–71–0	1.69	–0.87	–0.59	1
125	2,4–dimethyl–3–pentanol	600–36–2	1.93	–0.71	–0.15	1
126	3–(3–pyridyl)–1–propanol	2859–67–8	0.12	–0.84	–0.04	1
127	4–chloro–3–methylphenol	59–50–7	3.1	0.80	1.29	2
128	aniline	62–53–3	0.9	0.11	–0.09	2
129	2–hydroxybenzamide	65–45–2	1.28	–0.24	0.13	2
130	dicumarol	66–76–2	1.9	1.70	1.82	2
131	4–*tert*–pentylphenol	80–46–6	3.83	1.23	1.80	2
132	2,4,6–trichlorophenol	88–06–2	3.69	1.41	1.64	2
133	2–hydroxybenzaldehyde	90–02–8	1.81	0.42	1.73	2
134	1–naphthol	90–15–3	2.84	0.75	1.49	2
135	2–phenylphenol	90–43–7	3.09	1.09	1.44	2
136	3,5–dibromosalicylaldehyde	90–59–5	3.42	1.65	2.52	2
137	salicylaldoxime	94–67–7	1.1	0.25	1.63	2
138	2,4–dihydroxybenzaldehyde	95–01–2	1.33	0.73	1.02	2
139	2–methylphenol	95–48–7	1.98	–0.29	0.89	2
140	2–chloroaniline	95–51–2	1.88	–0.25	1.35	2
141	2–chlorophenol	95–57–8	2.15	0.18	0.97	2
142	4–*tert*–butylphenol	98–54–4	3.31	0.91	1.47	2
143	benzylamine	100–46–9	1.09	–0.24	0.02	2
144	*N*–methylaniline	100–61–8	1.66	0.06	0.03	2
145	2–cyanopyridine	100–70–9	0.45	–0.83	–0.84	2
146	4–acetamidophenol	103–90–2	0.46	–0.82	–0.73	2
147	4–butylaniline	104–13–2	3.18	1.07	1.17	2
148	nonylphenol	104–40–5	5.76	2.47	3.20	2
149	2,4–dimethylphenol	105–67–9	2.35	0.14	0.87	2
150	4–methylphenol	106–44–5	1.97	–0.16	0.82	2
151	4–chlorophenol	106–48–9	2.39	0.54	1.32	2
152	4–methylaniline	106–49–0	1.39	–0.05	–0.17	2
153	resorcinol	108–46–3	0.8	–0.65	0.34	2
154	phenol	108–95–2	1.5	–0.32	0.46	2
155	pyridine	110–86–1	0.65	–1.32	–0.10	2
156	2,4–dichlorophenol	120–83–2	3.17	1.04	1.32	2
157	3–ethoxy–4–hydroxybenaldehyde	121–32–4	1.58	0.02	0.28	2
158	3–methoxy–4–hydroxybenaldehyde	121–33–5	1.21	–0.03	0.09	2
159	4–ethylphenol	123–07–9	2.50	0.21	1.07	2
160	2,6–di(*tert*)butyl–4–methylphenol	128–37–0	4.17	1.79	2.78	2
161	phenyl–4–aminosalicylate	133–11–9	3.15	1.41	1.62	2
162	3–methoxysalicylaldehyde	148–53–8	1.37	0.38	1.80	2
163	3–methoxyphenol	150–19–6	1.58	–0.33	0.23	2
164	4–methoxyphenol	150–76–5	1.34	–0.14	0.05	2

Table 1. (Continued)

No	Compound	CAS No	log K_{ow}	log 1/IGC$_{50}$	log 1/LC$_{50}$	MOA
165	α,α,α,4–tetrafluoro–*o*–toluidine	393–39–5	2.51	–0.02	0.78	2
166	2,4,6–trimethylphenol	527–60–6	2.73	0.42	1.02	2
167	4–ethylaniline	589–16–2	1.96	0.03	0.22	2
168	4–ethoxy–2–nitroaniline	616–86–4	2.39	0.76	0.85	2
169	3–acetamidophenol	621–42–1	0.73	–0.16	–0.87	2
170	4–propylphenol	645–56–7	3.20	0.64	1.09	2
171	3–cyano–4,6–dimethyl–2–hydroxypyridine	769–28–8	1.77	–0.70	–0.02	2
172	4–phenoxyphenol	831–82–3	3.35	1.36	1.58	2
173	2–amino–5–bromopyridine	1072–97–5	1.29	0.49	–0.01	2
174	4–acetylpyridine	1122–54–9	0.48	–0.87	–0.14	2
175	methyl–4–cyanobenzoate	1129–35–7	1.54	–0.05	0.54	2
176	3–benzyloxyaniline	1484–26–0	2.77	0.88	1.34	2
177	2–allylphenol	1745–81–9	2.55	0.33	0.95	2
178	methyl–2,4–dihydroxybenzoate	2150–47–2	1.94	0.61	0.57	2
179	α,α,α–tetrafluoro–*m*–toluidine	2157–47–3	2.51	0.77	0.77	2
180	2,3,6–trimethylphenol	2416–94–6	2.67	0.28	1.22	2
181	5–bromovanillin	2973–76–4	1.92	0.62	0.59	2
182	*N*–ethylbenzylamine	14321–27–8	1.82	0.22	0.37	2
183	4–octylaniline	16245–79–7	5.12	2.43	3.23	2
184	6–chloro–2–pyridinol	16879–02–0	0.93	–0.29	–0.22	2
185	2,6–diisopropylaniline	24544–04–5	3.18	0.78	1.10	2
186	4–hexyloxyaniline	39905–57–2	3.65	1.38	1.80	2
187	ethyl–4–aminobenzoate	94–09–7	1.86	0.70	0.67	2
188	carbazole	86–74–8	3.72	0.91	2.33	2
189	2–bromo–3–pyridinol	6602–32–0	1.46	0.15	–0.43	2
190	2–chloro–3–pyridinol	6636–78–8	1.38	–0.01	–0.68	2
191	2–amino–4–chloro–6–methylpyrimidine	5600–21–5	0.83	–0.47	–0.01	2
192	5–bromosalicylaldehyde	1761–61–1	2.80	1.11	2.19	2
193	5–chlorosalicylaldehyde	635–93–8	2.65	1.01	2.31	2
194	2,4–diaminotoluene	95–80–7	0.14	–0.66	–1.07	2
195	3'–aminoacetophenone	99–03–6	0.86	–0.82	–0.45	2
196	1,4–bis(3–aminopropyl)piperazine	7209–38–3	–0.62	–0.66	–1.19	2
197	2,2'–methylenebis(3,4,6–trichlorophenol)	70–30–4	7.54	3.04	4.29	2
198	2,3,4–trichloroaniline	634–67–3	3.33	1.35	1.74	2
199	2–chloro–4–methylaniline	615–65–6	2.41	0.18	0.60	2
200	3–cyano–4,6–dimethyl–2–hydroxypyridine	769–28–8	1.77	–0.70	–0.03	2
201	phenyl–4–aminosalicylate	133–11–9	3.15	1.41	1.74	2
202	3–benzyloxyaniline	1484–26–0	2.77	0.88	1.34	2
203	2,4,6–triiodophenol	609–23–4	4.50	2.67	2.59	2
204	4,6–dimethyl–2–hydroxybenzaldehyde	708–76–9	1.86	0.62	1.83	2
205	2,2'–methylenebis(4–chlorophenol)	97–23–4	4.89	3.09	2.94	2
206	2,4–dinitrophenol	51–28–5	1.54	1.06	1.24	3
207	hexylaldehyde	66–25–1	1.78	–0.19	0.66	3
208	methylacetate	79–20–9	0.18	–2.00	–0.73	3
209	dibutyladipate	105–99–7	3.55	0.79	1.85	3
210	diethylsebacate	110–40–7	3.90	1.35	1.98	3
211	2–(ethylamino)ethanol	110–73–6	–0.46	–0.76	–1.22	3
212	diethyladipate	141–28–6	1.75	–0.24	1.08	3
213	diethylbenzylmalonate	607–81–8	2.76	0.71	1.66	3
214	diethylphthalate	84–66–2	2.47	0.23	0.84	3
215	dibutylphthalate	84–74–2	4.72	1.60	2.40	3
216	ethylbenzoate	93–89–0	2.64	–0.01	1.14	3
217	propylacetate	109–60–4	1.24	–1.24	0.23	3
218	ethylhexanoate	123–66–0	2.83	0.06	1.21	3
219	butylacetate	123–86–4	1.82	–0.49	0.81	3

Table 1. (Continued)

No	Compound	CAS No	log K_{ow}	log 1/IGC$_{50}$	log 1/LC$_{50}$	MOA
220	dimethylphthalate	131–11–3	1.56	–0.44	0.21	3
221	ethylacetate	141–78–6	0.73	–1.30	–0.42	3
222	hexylacetate	142–92–7	2.83	–0.04	1.52	3
223	*tert*–butylacetate	540–88–5	1.76	–1.49	–0.45	3
224	diethylmalonate	105–53–3	0.96	–1.03	1.02	3
225	isovaleraldehyde	590–86–3	1.23	–0.33	1.42	3
226	dimethylnitroterephthalate	5292–45–5	1.66	0.43	1.56	3
227	dibutylsuccinate	141–03–7	3.60	0.51	1.71	3
228	dibutylisophthalate	84–69–5	4.11	1.44	2.49	3
229	propylamine	107–10–8	0.48	–0.71	–0.72	4
230	butylamine	109–73–9	0.86	–0.57	–0.56	4
231	amylamine	110–58–7	1.49	–0.48	–0.31	4
232	hexylamine	111–26–2	2.06	–0.22	0.25	4
233	heptylamine	111–68–2	2.57	0.21	0.72	4
234	octylamine	111–86–4	2.90	0.61	1.40	4
235	nonylamine	112–20–9	3.57	1.70	1.82	4
236	decylamine	2016–57–1	4.10	2.06	2.18	4
237	undecylamine	7307–55–3	4.63	2.33	2.91	4
238	(+/–)–*sec*–butylamine	33966–50–6	0.74	–0.67	–0.58	4
239	1,2–diaminopropane	78–90–0	–0.91	–0.56	–1.13	4
240	1,3–diaminopropane	109–76–2	–1.43	–0.70	–1.21	4
241	(+,–)–1,2–dimethylpropylamine	598–74–3	1.10	–0.71	–0.51	4
242	pentachlorophenol	87–86–5	5.18	2.07	3.04	5
243	2,4–dinitroaniline	97–02–9	1.72	0.72	1.07	5
244	2,4,6–tribromophenol	118–79–6	4.08	2.03	1.70	5
245	2,5–dinitrophenol	329–71–5	1.86	0.95	1.74	5
246	4,6–dinitro–2–methylphenol	534–52–1	2.12	1.73	2.11	5
247	2,6–dinitrophenol	573–56–8	1.33	0.54	0.67	5
248	pentabromophenol	608–71–9	4.85	2.66	3.72	5
249	2,3,4,6–tetrachlorophenol	935–95–5	4.45	2.21	2.35	5
250	4–phenylazophenol	1689–82–3	3.18	1.66	2.23	5
251	pentachloropyridine	2176–62–7	3.53	1.71	2.73	5
252	2,3,5,6–tetrachloroaniline	3481–20–7	4.10	1.89	2.93	5
253	4–*tert*–butyl–2,6–dinitrophenol	4097–49–8	3.36	1.80	2.65	5
254	2,3,4,5–tetrachlorophenol	4901–51–3	4.21	2.72	2.75	5
255	allylisothiocyanate	57–06–7	1.94	2.06	3.06	6
256	2–methyl–1,4–naphthquinone	58–27–5	2.20	1.54	3.19	6
257	4–phenoxybenzaldehyde	67–36–7	3.96	1.26	1.63	6
258	methylmethacrylate	80–62–6	1.38	–1.22	–0.41	6
259	1–chloro–2–nitrobenzene	88–73–3	2.52	0.63	0.73	6
260	allylmethacrylate	96–05–9	1.68	–0.68	2.11	6
261	2–methylbutanal	96–17–3	1.14	–0.39	0.94	6
262	4–dimethlaminobenzaldehyde	100–10–7	1.81	0.23	0.51	6
263	1,4–dinitrobenzene	100–25–4	1.47	1.30	2.37	6
264	benzaldehyde	100–52–7	1.48	–0.01	0.92	6
265	4–chlorobenzaldehyde	104–88–1	2.13	0.40	1.81	6
266	isobutylacrylate	106–63–8	2.22	0.29	1.79	6
267	2–propenal	107–02–8	–0.01	1.41	3.45	6
268	3–butenenitrile	109–75–1	0.40	–1.48	–0.43	6
269	valeraldehyde	110–62–3	1.36	–0.13	0.81	6
270	4–amino–2–nitrophenol	119–34–6	0.96	0.88	0.63	6
271	2,4–nitrotoluene	121–14–2	1.98	0.64	0.88	6
272	1–chloro–3–nitrobenzene	121–73–3	2.47	0.73	0.92	6
273	2–chloro–4–nitroaniline	121–87–9	2.05	0.75	0.93	6
274	4–isopropylbenzaldehyde	122–03–2	2.97	0.67	1.35	6
275	2–methylvaleraldehyde	123–15–9	1.67	–0.47	0.73	6

Table 1. (Continued)

No	Compound	CAS No	log K_{ow}	log 1/IGC$_{50}$	log 1/LC$_{50}$	MOA
276	2,4–pentanedione	123–54–6	0.24	–0.27	–0.24	6
277	butyraldehyde	123–72–8	0.88	–0.37	0.65	6
278	ethylacrylate	140–88–5	1.32	0.52	1.60	6
279	2–aminoethanol	141–43–5	–1.31	–1.01	–1.53	6
280	1–fluoro–4–nitrobenzene	350–46–9	1.89	0.10	0.70	6
281	2–fluorobenzaldehyde	446–52–6	1.76	0.08	1.96	6
282	3–pyridincarboxaldehyde	500–22–1	0.57	–0.15	0.82	6
283	2–methylbenzaldehyde	529–20–4	2.26	–0.01	0.36	6
284	2–nitrobenzaldehyde	552–89–6	1.74	0.17	0.96	6
285	4–nitrobenzaldehyde	555–16–8	1.56	0.20	1.18	6
286	2,4–dimethoxybenzaldehyde	613–45–6	1.79	–0.06	0.92	6
287	methyl–4–nitrobenzoate	619–50–1	1.94	0.39	0.88	6
288	4–nitrobenzamide	619–80–7	0.82	0.18	0.10	6
289	pentafluorobenzaldehyde	653–37–2	2.39	0.82	2.25	6
290	pentafluoroaniline	771–60–8	1.87	0.26	0.69	6
291	2–hydroxyethylacrylate	818–61–1	–0.21	0.69	1.38	6
292	2–hydroxyethylmethacrylate	868–77–9	0.47	–1.08	–0.24	6
293	2,4–dichlorobenzaldehyde	874–42–0	3.08	1.04	1.99	6
294	2–hydroxypropylacrylate	999–61–1	0.35	0.65	1.59	6
295	tetrachlorocatechol	1198–55–6	4.29	1.70	2.29	6
296	2–ethoxyethylmethacrylate	2370–63–0	1.40	–0.78	0.76	6
297	benzylmethacrylate	2495–37–6	2.82	0.65	1.58	6
298	*N*–hexylacrylate	2499–95–8	3.44	0.71	2.16	6
299	cyclohexylacrylate	3066–71–5	2.83	0.76	2.02	6
300	2,4,5–trimethoxybenzaldehyde	4460–86–0	1.19	–0.10	0.60	6
301	isopropylmethacrylate	4655–34–9	2.25	–0.88	0.53	6
302	2–amino–5–chlorobenzonitrile	5922–60–1	1.79	0.44	0.73	6
303	1,3,5–trichloro–2,4–dinitrobenzene	6284–83–9	2.65	2.19	3.09	6
304	2–chloro–5–nitrobenzaldehyde	6361–21–3	2.25	0.53	1.69	6
305	4–ethoxybenzaldehyde	10031–82–0	2.31	0.07	0.73	6
306	3–hydroxy–2–nitropyridine	15128–82–2	0.92	0.87	–0.08	6
307	1,3–dichloro–4,6–dinitrobenzene	28689–08–9	2.49	2.60	3.72	6
308	5–hydroxy–2–nitrobenzaldehyde	42454–06–8	1.75	0.44	0.60	6
309	methyl–2,5–dichlorobenzoate	2905–69–3	3.16	0.81	1.17	6
310	2,3–dibromopropanol	96–13–9	0.63	–0.49	0.49	6
311	3,5–dibromo–4–hydroxybenzonitrile	1689–84–5	2.88	1.16	1.38	6
312	2,4–dichlorobenzamide	2447–79–2	1.37	–0.36	0.30	6
313	diethylchloromalonate	14064–10–4	1.64	0.63	2.31	6
314	*N*–vinylcarbazole	1484–13–5	4.95	2.24	4.78	6
315	1,3–dichloro–4,6–dinitrobenzene	3698–83–7	2.65	2.57	3.80	6
316	2,2,2–trichloroethanol	115–20–8	1.42	–0.46	–0.30	6
317	2–chloroethanol	107–07–3	0.03	–1.42	0.34	6
318	1–chloro–2–propanol	127–00–4	0.14	–1.49	–0.41	6
319	3–chloro–1–propanol	627–30–5	0.15	–1.40	–0.93	6
320	pentafluoroaniline	771–60–8	1.87	0.20	0.69	6
321	2,2–dichloroacetamide	683–72–7	0.19	–0.98	–0.27	6
322	2–chloro–6–methylbenzonitrile	6575–09–3	2.80	0.46	1.00	6
323	4–dimethylaminocinnamaldehyde	6203–18–5	1.62	0.52	1.47	6
324	2–propyn–1–ol	107–19–7	–0.38	–0.74	1.56	7
325	2–butyn–1,4–diol	110–65–6	–1.83	–1.19	0.21	7
326	catechol	120–80–9	0.88	0.75	1.08	7
327	2–butyn–1–ol	764–01–2	0.37	–1.43	0.84	7
328	3–butyn–1–ol	927–74–2	–0.18	–1.84	0.29	7
329	2–decyn–1–ol	4117–14–0	3.54	0.99	2.16	7
330	3–butyn–2–ol	65337–13–5	0.14	–0.40	0.78	7
331	1,5–hexadien–3–ol	924–41–4	0.69	0.25	0.41	7

Table 1. (Continued)

No	Compound	CAS No	log K_{ow}	log 1/IGC$_{50}$	log 1/LC$_{50}$	MOA
332	1–amino–2–propanol	78–96–6	–0.96	–0.93	–1.53	8
333	hexanoic acid	142–62–1	1.92	–0.20	–0.44	8
334	triethanolamine	102–71–6	–1.00	–1.75	–1.90	8
335	1–amino–2–propanol	78–96–6	–0.96	–0.93	–1.53	8
336	diethanolamine	111–42–2	–1.43	–1.03	–1.65	8
337	*N*,*N*–diethylethanolamine	100–37–8	0.48	–1.50	–1.18	8

2.2 MOA Classification with Support Vector Machines

Support vector machines were developed by Vapnik [19–21] as an effective algorithm for determining an optimal hyperplane to separate two classes of patterns [22–32]. In the first step, using various kernels that perform a nonlinear mapping, the input space is transformed into a higher dimensional feature space. Then, a maximal margin hyperplane (MMH) is computed in the feature space by maximizing the distance to the hyperplane of the closest patterns from the two classes. The patterns that determine the separating hyperplane are called support vectors.

In the first experiment we have used SVM models to discriminate between nonpolar narcotic compounds (chemicals that have baseline toxicity) and the other compounds that have excess toxicity (representing the following MOAs: polar narcosis, ester narcosis, amine narcosis, weak acid respiratory uncoupling, electrophilicity, proelectrophilicity, and nucleophilicity). From the total set of 337 compounds, 126 represent the SVM class +1 (nonpolar narcotic) and 211 represent the SVM class –1 (all other MOA classes).

The chemicals that exhibit excess toxicity belong to seven MOA classes and their toxicity has a wide range of variation. For these compounds it is useful to further separate less reactive and more reactive compounds. In the second experiment we have developed SVM models that discriminate between less reactive compounds (SVM class +1, formed by polar narcotics, ester narcotics, amine narcotics) and more reactive compounds (SVM class –1, formed by weak acid respiratory uncouplers, electrophiles, proelectrophiles, and nucleophiles). From the total of 211 compounds with excess toxicity 115 are less reactive and 96 are more reactive compounds.

All SVM models from the present paper for the classification of polar and nonpolar pollutants were obtained with mySVM [33], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem Links [34] at http://www.biochempress.com. Before computing the SVM model, the input vectors were scaled to zero mean and unit variance. The prediction power of each SVM model was evaluated with a leave–5%–out (L5%O) cross–validation procedure, and the capacity parameter C took the values 10, 100, and 1000. We present below the kernels and their parameters used in this study.

The dot kernel. The inner product of x and y defines the dot kernel:

$$K(x, y) = x \cdot y \tag{1}$$

The polynomial kernel. The polynomial of degree *d* (values 2, 3, 4, and 5) in the variables *x* and *y* defines the polynomial kernel:

$$K(x, y) = (x \cdot y + 1)^d \tag{2}$$

The radial kernel. The following exponential function in the variables *x* and *y* defines the radial basis function kernel, with the shape controlled by the parameter γ (values 0.5, 1.0, and 2.0):

$$K(x, y) = \exp(-\gamma \| x - y \|^2) \tag{3}$$

The neural kernel. The hyperbolic tangent function in the variables *x* and *y* defines the neural kernel, with the shape controlled by the parameters *a* (values 0.5, 1.0, and 2.0) and *b* (values 0, 1, and 2):

$$K(x, y) = \tanh(ax \cdot y + b) \tag{4}$$

The anova kernel. The sum of exponential functions in *x* and *y* defines the anova kernel, with the shape controlled by the parameters γ (values 0.5, 1.0, and 2.0) and *d* (values 1, 2, and 3):

$$K(x, y) = \left(\sum_i \exp(-\gamma(x_i - y_i)) \right)^d \tag{5}$$

3 RESULTS AND DISCUSSION

The performances of SVM classifiers in structure–activity studies depend on the combination of several parameters, such as kernel type and various parameters that control the kernel shape. Because there are no clear guidelines on selecting the most effective kernel for a certain classification problem, we have tested five kernel types (dot, polynomial, radial basis function, neural, and anova), for a total of 78 SVM model for each MOA experiment (see Table 2 for kernel type, parameters, and *C* values).

Using data from [16], the first MOA experiment uses SVM to discriminate between nonpolar narcotic compounds and all other compounds that have excess toxicity. The SVM models were obtained with three descriptors, namely log K_{ow}, log 1/IGC_{50}, and log 1/LC_{50}. The calibration and prediction results presented in Table 2 were obtained for the group of 337 compounds, separated into 126 nonpolar narcotics (SVM class +1) and 211 compounds with excess toxicity (SVM class –1). The calibration results reported in Table 2 are: TP_c, true positive in calibration, the number of +1 patterns (nonpolar compounds) computed in class +1; FN_c, false negative in calibration, the number of +1 patterns computed in class –1; TN_c, true negative in calibration, the number of –1 patterns (excess toxicity compounds) computed in class –1; FP_c, false positive in calibration, the number of –1 patterns computed in class +1; SV_c, number of support vectors in calibration; BSV_c, number of bounded support vectors in calibration; AC_c, calibration accuracy.

Table 2. Results for SVM classification of nonpolar narcotic compounds (SVM calss +1) from other compounds (SVM calss –1) using as descriptors log K_{ow}, log 1/IGC$_{50}$ and log 1/LC$_{50}$. [a]

Exp	*C*	*K*			TP$_c$	FN$_c$	TN$_c$	FP$_c$	SV$_c$	BSV$_c$	AC$_c$	TP$_p$	FN$_p$	TN$_p$	FP$_p$	SV$_p$	BSV$_p$	AC$_p$
1	10	D			78	48	186	25	195	191	0.78	79	47	186	25	185.8	181.0	0.79
2	100				78	48	186	25	194	190	0.78	80	46	186	25	185.2	179.7	0.79
3	1000				65	61	151	60	72	19	0.64	59	67	122	89	85.3	20.2	0.54
			d															
4	10	P	2		81	45	185	26	176	166	0.79	82	44	184	27	165.8	155.4	0.79
5	100		2		80	46	176	35	173	162	0.76	76	50	183	28	165.6	151.3	0.77
6	1000		2		33	93	151	60	283	2	0.55	27	99	154	57	261.0	1.1	0.54
7	10		3		80	46	187	24	163	141	0.79	78	48	182	29	157.2	133.6	0.77
8	100		3		81	45	174	37	238	114	0.76	76	50	172	39	225.2	106.3	0.74
9	1000		3		61	65	100	111	301	0	0.48	38	88	163	48	285.4	0.3	0.60
10	10		4		79	47	189	22	173	117	0.80	78	48	178	33	158.8	113.8	0.76
11	100		4		68	58	162	49	215	2	0.68	59	67	136	75	212.8	1.5	0.58
12	1000		4		41	85	191	20	205	0	0.69	67	59	123	88	202.2	0.0	0.56
13	10		5		32	94	188	23	226	1	0.65	42	84	173	38	204.8	1.8	0.64
14	100		5		80	46	175	36	224	0	0.76	48	78	160	51	203.4	0.0	0.62
15	1000		5		80	46	175	36	224	0	0.76	48	78	160	51	203.4	0.0	0.62
			γ															
16	10	R	0.5		95	31	185	26	176	140	0.83	85	41	177	34	167.8	134.4	0.78
17	100		0.5		96	30	192	19	164	113	0.85	87	39	175	36	154.6	105.3	0.78
18	1000		0.5		104	22	190	21	151	87	0.87	83	43	168	43	144.1	83.8	0.74
19	10		1.0		97	29	190	21	172	122	0.85	89	37	180	31	165.1	114.8	0.80
20	100		1.0		102	24	193	18	158	89	0.88	89	37	173	38	150.2	81.7	0.78
21	1000		1.0		109	17	202	9	142	56	0.92	86	40	172	39	133.8	52.0	0.77
22	10		2.0		98	28	197	14	170	98	0.88	88	38	178	33	166.2	91.2	0.79
23	100		2.0		108	18	201	10	153	60	0.92	86	40	171	40	145.3	54.2	0.76
24	1000		2.0		119	7	206	5	131	23	0.96	90	36	169	42	126.2	21.4	0.77
			a	*b*														
25	10	N	0.5	0.0	75	51	98	113	158	158	0.51	49	77	130	81	152.1	149.7	0.53
26	100		0.5	0.0	75	51	98	113	158	158	0.51	49	77	130	81	151.7	149.3	0.53
27	1000		0.5	0.0	75	51	98	113	158	158	0.51	49	77	130	81	151.6	149.2	0.53
28	10		1.0	0.0	44	82	130	81	165	163	0.52	52	74	122	89	155.3	153.6	0.52
29	100		1.0	0.0	44	82	130	81	165	163	0.52	49	77	123	88	154.8	152.9	0.51
30	1000		1.0	0.0	44	82	130	81	165	163	0.52	47	79	124	87	154.8	152.8	0.51
31	10		2.0	0.0	46	80	129	82	166	163	0.52	53	73	125	86	157.2	155.4	0.53
32	100		2.0	0.0	45	81	130	81	165	163	0.52	53	73	125	86	156.7	154.9	0.53
33	1000		2.0	0.0	45	81	130	81	164	162	0.52	50	76	126	85	156.7	154.8	0.52
34	10		0.5	1.0	67	59	81	130	180	180	0.44	47	79	104	107	171.1	169.8	0.45
35	100		0.5	1.0	67	59	81	130	180	180	0.44	50	76	103	108	171.0	169.8	0.45
36	1000		0.5	1.0	67	59	81	130	180	180	0.44	53	73	102	109	170.9	169.8	0.46
37	10		1.0	1.0	57	69	98	113	178	178	0.46	39	87	122	89	169.9	168.1	0.48
38	100		1.0	1.0	38	88	123	88	178	176	0.48	40	86	120	91	169.7	167.9	0.47
39	1000		1.0	1.0	31	95	147	64	61	52	0.53	40	86	120	91	169.7	167.9	0.47
40	10		2.0	1.0	44	82	130	81	165	163	0.52	52	74	125	86	155.9	154.3	0.53
41	100		2.0	1.0	45	81	130	81	164	162	0.52	55	71	119	92	155.5	154.1	0.52
42	1000		2.0	1.0	44	82	130	81	165	163	0.52	49	77	124	87	155.7	154.0	0.51
43	10		0.5	2.0	80	46	64	147	180	180	0.43	68	58	81	130	170.0	169.4	0.44
44	100		0.5	2.0	38	88	122	89	179	177	0.47	68	58	80	131	166.6	166.2	0.44
45	1000		0.5	2.0	38	88	122	89	179	177	0.47	68	58	79	132	168.9	168.4	0.44
46	10		1.0	2.0	62	64	89	122	176	176	0.45	57	69	95	116	168.2	167.7	0.45
47	100		1.0	2.0	62	64	89	122	176	176	0.45	54	72	99	112	168.1	167.4	0.45
48	1000		1.0	2.0	62	64	89	122	176	176	0.45	55	71	99	112	167.9	167.4	0.46
49	10		2.0	2.0	39	87	124	87	176	174	0.48	48	78	109	102	166.7	165.6	0.47
50	100		2.0	2.0	39	87	124	87	176	174	0.48	47	79	112	99	166.6	165.3	0.47
51	1000		2.0	2.0	38	88	123	88	176	174	0.48	47	79	112	99	166.6	165.3	0.47

Table 2. (Continued)

Exp	*C*	*K*	γ	*d*	TP_c	FN_c	TN_c	FP_c	SV_c	BSV_c	AC_c	TP_p	FN_p	TN_p	FP_p	SV_p	BSV_p	AC_p
52	10	A	0.5	1	85	41	183	28	198	178	0.80	85	41	177	34	189.2	168.6	0.78
53	100		0.5	1	92	34	176	35	184	164	0.80	87	39	172	39	175.2	152.2	0.77
54	1000		0.5	1	96	30	177	34	189	152	0.81	84	42	172	39	172.8	143.9	0.76
55	10		1.0	1	92	34	182	29	198	174	0.81	81	45	168	43	189.2	165.2	0.74
56	100		1.0	1	97	29	180	31	183	153	0.82	79	47	173	38	173.3	143.1	0.75
57	1000		1.0	1	91	35	186	25	186	140	0.82	80	46	170	41	171.1	131.3	0.74
58	10		2.0	1	93	33	178	33	197	165	0.80	81	45	170	41	189.2	156.4	0.74
59	100		2.0	1	93	33	186	25	181	139	0.83	79	47	173	38	172.7	132.2	0.75
60	1000		2.0	1	92	34	192	19	195	124	0.84	78	48	170	41	177.9	118.7	0.74
61	10		0.5	2	95	31	190	21	169	127	0.85	87	39	182	29	159.4	119.2	0.80
62	100		0.5	2	99	27	194	17	154	99	0.87	87	39	174	37	148.2	95.1	0.77
63	1000		0.5	2	73	53	169	42	212	26	0.72	68	58	150	61	199.3	24.0	0.65
64	10		1.0	2	98	28	194	17	166	105	0.87	86	40	176	35	154.7	97.4	0.78
65	100		1.0	2	105	21	194	17	158	77	0.89	80	46	173	38	146.4	72.5	0.75
66	1000		1.0	2	94	32	180	31	185	20	0.81	77	49	158	53	153.2	22.9	0.70
67	10		2.0	2	102	24	196	15	164	78	0.88	89	37	175	36	156.3	72.0	0.78
68	100		2.0	2	114	12	201	10	145	48	0.93	80	46	167	44	137.0	39.8	0.73
69	1000		2.0	2	122	4	205	6	135	20	0.97	85	41	153	58	125.3	17.2	0.71
70	10		0.5	3	99	27	193	18	155	100	0.87	86	40	170	41	149.4	92.7	0.76
71	100		0.5	3	106	20	197	14	138	72	0.90	84	42	172	39	135.2	67.5	0.76
72	1000		0.5	3	74	52	146	65	215	3	0.65	58	68	149	62	191.1	2.9	0.61
73	10		1.0	3	106	20	199	12	151	70	0.91	84	42	173	38	143.5	65.3	0.76
74	100		1.0	3	114	12	204	7	130	38	0.94	84	42	167	44	123.2	35.1	0.74
75	1000		1.0	3	101	25	182	29	144	2	0.84	81	45	152	59	138.2	2.6	0.69
76	10		2.0	3	117	9	203	8	142	32	0.95	85	41	164	47	139.0	31.2	0.74
77	100		2.0	3	125	1	208	3	133	13	0.99	83	43	158	53	126.7	12.2	0.72
78	1000		2.0	3	126	0	209	2	112	1	0.99	86	40	164	47	109.7	0.9	0.74

[a] The table reports the experiment number Exp, capacity parameter *C*, kernel type *K* (dot D; polynomial P; radial basis function R; neural N; anova A) and corresponding parameters, calibration results (TP_c, true positive in calibration; FN_c, false negative in calibration; TN_c, true negative in calibration; FP_c, false positive in calibration; SV_c, number of support vectors in calibration; BSV_c, number of bounded support vectors in calibration; AC_c, calibration accuracy) and L20%O prediction results (TP_p, true positive in prediction; FN_p, false negative in prediction; TN_p, true negative in prediction; FP_p, false positive in prediction; SV_p, average number of support vectors in prediction; BSV_p, average number of bounded support vectors in prediction; AC_p, prediction accuracy).

Similarly with other multivariate statistical models, support vector machines can be calibrated to perfectly discriminate two groups of patterns, but only a cross–validation test can demonstrate the prediction power of an SVM model. In our prediction experiments we have randomly generated 20 cross–validation sets, each set containing +1 and –1 compounds in a ratio as close as possible to the one from the total set of 337 compounds. For each SVM model we present in Table 2 the following leave–5%–out cross–validation statistics: TP_p, true positive in prediction; FN_p, false negative in prediction; TN_p, true negative in prediction; FP_p, false positive in prediction; SV_p, average number of support vectors in prediction; BSV_p, average number of bounded support vectors in prediction; AC_p, prediction accuracy.

Using the leave–one–out (LOO) cross–validation technique, Ren [16] used discriminant analysis and logistic regression to separate the same groups of nonpolar narcotic compounds and all other compounds that have excess toxicity. Although the LOO cross–validation technique is an easier statistical test than the L5%O test used here, we will compare our results with those from [16].

The LOO results for the discriminant analysis are: TP_p = 90, FN_p = 36, TN_p = 168, FP_p = 43, AC_p = 0.77. Slightly better results were obtained by Ren *et al*. [16] with logistic regression: TP_p = 78, FN_p = 48, TN_p = 186, FP_p = 25, AC_p = 0.78. The results from Table 2 show that SVM models obtained with the dot kernel give better results (Exp. 1 and 2 from Table 2, both with AC_p = 0.79). The overfitting of SVM models is evident for the polynomial kernel. When the degree of the polynomial kernel increases from 2 to 5, AC_p decreases from 0.76 to 0.62. These cross–validation results show that SVM models can be overfitted, and the only practical method to identify the optimum model is by comparing prediction statistics. Radial kernels constantly give good prediction results, with AC_p between 0.76 and 0.80. The SVM model with the maximum prediction (Exp. 19 from Table 2, TP_p = 89, FN_p = 37, TN_p = 180, FP_p = 31, AC_p = 0.80) was selected for further analysis.

The 37 false negative predictions in the L5%O cross–validation test are: **26**, *N*,*N*–dimethylbenzylamine; **41**, 6–methyl–5–hepten–2–one; **46**, 2–undecanone; **49**, benzophenone; **63**, 4–phenylpyridine; **64**, octylcyanide; **67**, 1–hexen–3–ol; **68**, 1,2–bis–(4–pyridyl)–ethane; **69**, 2–dimethylaminopyridine; **71**, 4–bromophenyl–3–pyridylketone; **75**, nitrobenzene; **76**, 3–nitrotoluene; **77**, dibutylfumarate; **81**, 4–pentyn–2–ol; **82**, 3–bromothiophene; **85**, cyclohexanoneoxime; **87**, anthranilamide (2–aminobenzamide); **88**, 2–acetyl–1–methylpyrrole; **89**, 2',3',4'–trichloroaceto–phenone; **90**, 2',4'–dichloroacetophenone; **91**, 4–nitrophenylphenylether; **92**, triphenylphosphine–oxide; **95**, chloroacetonitrile; **96**, malononitrile; **103**, 1–benzoylacetone; **104**, acetaldoxime; **106**, γ–decanolactone; **110**, 2,3–benzofuran; **113**, 2,4,5–tribromoimidazole; **114**, 1–methylpiperazine; **117**, 3,3–dimethyl–2–butanone; **119**, 1–(2–aminioethyl)piperazine; **120**, methyl–4–chlorobenzoate; **121**, 1,6–dicyanohexane; **122**, 3–methylindole; **123**, flavone; **126**, 3–(3–pyridyl)–1–propanol. These compounds are structurally diverse, with two significant classes of outliers, namely ketones (**41**, **46**, **49**, and **117**) and halogenoketones (**71**, **89**, and **90**). These results suggest that the recognition of nonpolar narcotic compounds can be improved only by using supplementary structural descriptors.

The 31 false positive predictions are (the MOA is indicated in parenthesis; see Table 1 caption for notation): **131**, 4–*tert*–pentylphenol (2); **145**, 2–cyanopyridine (2); **167**, 4–ethylaniline (2); **170**, 4–propylphenol (2); **171**, 3–cyano–4,6–dimethyl–2–hydroxypyridine (2); **185**, 2,6–diisopropylaniline (2); **199**, 2–chloro–4–methylaniline (2); **200**, 3–cyano–4,6–dimethyl–2–hydroxypyridine (2); **210**, diethylsebacate (3); **215**, dibutylphthalate (3); **223**, *tert*–butylacetate (3); **227**, dibutylsuccinate (3); **228**, dibutylisophthalate (3); **231**, amylamine (4); **232**, hexylamine (4); **233**, heptylamine (4); **241**, (+,–)–1,2–dimethylpropylamine (4); **242**, pentachlorophenol (5); **257**, 4–phenoxybenzaldehyde (6); **258**, methylmethacrylate (6); **283**, 2–methylbenzaldehyde (6); **294**, 2–hydroxypropylacrylate (6); **295**, tetrachlorocatechol (6); **301**, isopropylmethacrylate (6); **316**, 2,2,2–trichloroethanol (6); **319**, 3–chloro–1–propanol (6); **321**, 2,2–dichloroacetamide (6); **325**, 2–butyn–1,4–diol (7); **333**, hexanoic acid (8); **334**, triethanolamine (8); **337**, *N*,*N*–diethylethanolamine (8).

Table 3. Results for SVM classification of less reactive compounds (SVM class +1, formed by polar narcotics, ester narcotics, amine narcotics) and more reactive compounds (SVM class –1, formed by weak acid respiratory uncouplers, electrophiles, proelectrophiles, and nucleophiles). using as descriptors log K_{ow}, log 1/IGC_{50} and log 1/LC_{50}. [a]

Exp	*C*	*K*			TP_c	FN_c	TN_c	FP_c	SV_c	BSV_c	AC_c	TP_p	FN_p	TN_p	FP_p	SV_p	BSV_p	AC_p
1	10	D			97	18	50	46	151	144	0.70	97	18	46	50	144.2	139.1	0.68
2	100				97	18	51	45	152	145	0.70	97	18	47	49	144.3	137.7	0.68
3	1000				65	50	45	51	69	28	0.52	79	36	39	57	91.2	30.6	0.56
			d															
4	10	P	2		101	14	38	58	154	136	0.66	97	18	38	58	144.0	131.8	0.64
5	100		2		101	14	41	55	156	136	0.67	96	19	40	56	145.3	130.5	0.64
6	1000		2		35	80	59	37	159	2	0.45	51	64	48	48	158.4	1.6	0.47
7	10		3		95	20	52	44	148	124	0.70	93	22	43	53	143.2	117.4	0.64
8	100		3		58	57	61	35	188	31	0.56	86	29	42	54	166.6	91.3	0.61
9	1000		3		22	93	82	14	145	0	0.49	50	65	53	43	149.2	0.3	0.49
10	10		4		103	12	56	40	135	101	0.75	89	26	49	47	129.8	91.0	0.65
11	100		4		97	18	32	64	164	0	0.61	85	30	39	57	156.2	1.4	0.59
12	1000		4		97	18	32	64	164	0	0.61	72	43	46	50	155.7	0.0	0.56
13	10		5		28	87	86	10	138	5	0.54	76	39	47	49	144.5	15.4	0.58
14	100		5		35	80	77	19	156	0	0.53	63	52	55	41	145.7	0.0	0.56
15	1000		5		35	80	77	19	156	0	0.53	63	52	55	41	145.7	0.0	0.56
			γ															
16	10	R	0.5		100	15	65	31	146	118	0.78	83	32	51	45	141.0	111.0	0.64
17	100		0.5		107	8	73	23	132	85	0.85	88	27	54	42	127.2	79.2	0.67
18	1000		0.5		109	6	77	19	118	55	0.88	91	24	58	38	112.6	51.5	0.71
19	10		1.0		105	10	73	23	141	98	0.84	84	31	55	41	134.9	90.0	0.66
20	100		1.0		108	7	76	20	125	56	0.87	92	23	57	39	118.2	51.4	0.71
21	1000		1.0		111	4	84	12	115	31	0.92	83	32	57	39	107.2	27.7	0.66
22	10		2.0		107	8	77	19	141	63	0.87	94	21	55	41	133.2	58.2	0.71
23	100		2.0		114	1	85	11	126	29	0.94	78	37	55	41	119.3	24.6	0.63
24	1000		2.0		115	0	92	4	115	10	0.98	76	39	56	40	107.7	8.4	0.63
			a	*b*														
25	10	N	0.5	0.0	64	51	46	50	105	102	0.52	63	52	45	51	99.9	97.6	0.51
26	100		0.5	0.0	64	51	45	51	104	102	0.52	62	53	44	52	99.0	96.3	0.50
27	1000		0.5	0.0	59	56	48	48	102	102	0.51	63	52	44	52	99.0	96.2	0.51
28	10		1.0	0.0	65	50	44	52	108	105	0.52	55	60	56	40	96.3	94.3	0.53
29	100		1.0	0.0	62	53	43	53	108	106	0.50	60	55	54	42	95.7	93.7	0.54
30	1000		1.0	0.0	63	52	44	52	107	104	0.51	62	53	52	44	95.5	93.3	0.54
31	10		2.0	0.0	62	53	42	54	109	107	0.49	57	58	52	44	98.9	97.0	0.52
32	100		2.0	0.0	62	53	43	53	109	107	0.50	64	51	44	52	99.5	97.8	0.51
33	1000		2.0	0.0	62	53	43	53	108	106	0.50	61	54	45	51	97.8	96.0	0.50
34	10		0.5	1.0	66	49	45	51	104	102	0.53	60	55	48	48	97.4	96.2	0.51
35	100		0.5	1.0	65	50	45	51	103	101	0.52	59	56	47	49	96.5	95.2	0.50
36	1000		0.5	1.0	65	50	45	51	103	101	0.52	59	56	47	49	96.5	95.2	0.50
37	10		1.0	1.0	63	52	55	41	96	96	0.56	66	49	52	44	94.8	93.5	0.56
38	100		1.0	1.0	63	52	55	41	96	96	0.56	64	51	54	42	94.3	93.0	0.56
39	1000		1.0	1.0	63	52	44	52	106	104	0.51	67	48	55	41	96.8	95.5	0.58
40	10		2.0	1.0	71	44	51	45	91	89	0.58	64	51	59	37	90.5	88.5	0.58
41	100		2.0	1.0	71	44	51	45	91	89	0.58	52	63	51	45	92.1	90.3	0.49
42	1000		2.0	1.0	71	44	51	45	91	89	0.58	51	64	52	44	93.4	91.7	0.49
43	10		0.5	2.0	76	39	38	58	106	106	0.54	72	43	41	55	100.2	99.8	0.54
44	100		0.5	2.0	76	39	38	58	106	106	0.54	73	42	41	55	99.5	99.3	0.54
45	1000		0.5	2.0	76	39	38	58	106	106	0.54	71	44	38	58	99.7	99.5	0.52
46	10		1.0	2.0	53	62	52	44	102	102	0.50	59	56	48	48	97.7	96.6	0.51
47	100		1.0	2.0	53	62	52	44	102	102	0.50	60	55	49	47	97.2	96.0	0.52
48	1000		1.0	2.0	53	62	52	44	102	102	0.50	60	55	49	47	97.0	96.0	0.52
49	10		2.0	2.0	63	52	44	52	106	104	0.51	58	57	51	45	101.2	99.9	0.52
50	100		2.0	2.0	63	52	44	52	106	104	0.51	60	55	48	48	100.2	98.5	0.51
51	1000		2.0	2.0	63	52	44	52	106	104	0.51	65	50	49	47	100.3	98.5	0.54

Table 3. (Continued)

Exp	*C*	*K*	γ	*d*	TP_c	FN_c	TN_c	FP_c	SV_c	BSV_c	AC_c	TP_p	FN_p	TN_p	FP_p	SV_p	BSV_p	AC_p
52	10	A	0.5	1	90	25	65	31	154	131	0.73	85	30	50	46	144.4	123.7	0.64
53	100		0.5	1	91	24	66	30	140	119	0.74	83	32	56	40	134.7	111.5	0.66
54	1000		0.5	1	90	25	69	27	142	116	0.75	79	36	55	41	132.2	106.3	0.64
55	10		1.0	1	90	25	70	26	146	120	0.76	80	35	56	40	138.1	114.7	0.64
56	100		1.0	1	93	22	74	22	141	112	0.79	76	39	57	39	132.3	103.5	0.63
57	1000		1.0	1	93	22	73	23	133	100	0.79	78	37	58	38	126.8	92.8	0.64
58	10		2.0	1	88	27	76	20	143	108	0.78	75	40	56	40	134.9	102.8	0.62
59	100		2.0	1	93	22	75	21	136	96	0.80	75	40	59	37	128.2	88.4	0.64
60	1000		2.0	1	92	23	77	19	133	86	0.80	80	35	58	38	123.8	79.5	0.65
61	10		0.5	2	99	16	71	25	137	99	0.81	83	32	54	42	130.9	93.0	0.65
62	100		0.5	2	103	12	76	20	124	74	0.85	82	33	57	39	118.3	68.2	0.66
63	1000		0.5	2	109	6	77	19	112	48	0.88	85	30	57	39	105.8	43.9	0.67
64	10		1.0	2	105	10	75	21	126	79	0.85	85	30	55	41	123.0	70.5	0.66
65	100		1.0	2	109	6	79	17	115	47	0.89	80	35	54	42	108.5	41.6	0.64
66	1000		1.0	2	110	5	82	14	106	29	0.91	77	38	51	45	100.8	25.8	0.61
67	10		2.0	2	108	7	80	16	125	51	0.89	84	31	53	43	119.8	46.1	0.65
68	100		2.0	2	113	2	85	11	110	24	0.94	76	39	54	42	107.0	22.1	0.62
69	1000		2.0	2	115	0	94	2	98	6	0.99	72	43	54	42	94.3	4.6	0.60
70	10		0.5	3	107	8	75	21	128	69	0.86	85	30	55	41	120.9	63.5	0.66
71	100		0.5	3	111	4	83	13	111	44	0.92	81	34	56	40	106.0	37.9	0.65
72	1000		0.5	3	83	32	68	28	134	8	0.72	76	39	57	39	113.8	8.8	0.63
73	10		1.0	3	111	4	82	14	126	43	0.91	80	35	51	45	116.9	38.2	0.62
74	100		1.0	3	114	1	88	8	110	19	0.96	76	39	54	42	102.9	17.0	0.62
75	1000		1.0	3	114	1	95	1	94	5	0.99	72	43	57	39	88.8	3.6	0.61
76	10		2.0	3	114	1	91	5	116	16	0.97	76	39	53	43	109.2	14.7	0.61
77	100		2.0	3	114	1	95	1	104	3	0.99	75	40	54	42	98.3	2.5	0.61
78	1000		2.0	3	115	0	96	0	91	1	1.00	74	41	53	43	92.1	0.5	0.60

[a] See Table 2 for notations

The false positive predictions for excess toxicity compounds are distributed in all seven MOAs, but two classes are particularly difficult to discriminate, namely amine narcotics (MOA 4, with 4 compounds out of 13), and nucleophiles (MOA 8, with 3 compounds out of 6). A possible explanation for the high prediction error for these two classes of compounds is the relative small number of chemicals for MOAs 4 and 8. With too few examples for training, the classification function cannot "learn" the features that characterize these MOAs. However, other two MOAs with a small number of chemicals (weak acid respiratory uncoupling compounds, MOA 5, which has 13 compounds and one false positive prediction; proelectrophilic compounds, MOA 7, which has 8 compounds and one false positive prediction) have a low number of prediction errors, indicating that the classification errors for MOAs 4 and 8 are due to the input descriptors that are not able to characterize the structural features of these two classes. The L5%O prediction statistics for the SVM model is slightly higher than the LOO statistics for discriminant analysis and logistic regression [16]. This difference is significant because the L5%O is more difficult than the LOO cross–validation.

The neural kernel for SVM discrimination of nonpolar narcotic compounds and all other compounds that have excess toxicity gives poor predictions, with AC_p between 0.44 to 0.53. The

prediction statistics are much better for the anova kernel, with AC_p between 0.61 to 0.80. The AC_p values for the anova kernel show that the parameters γ and *d* have a great influence on the quality of the SVM model. The dependence between the kernel parameters (γ and *d*) and AC_p is non–linear, indicating that a good model can be identified only by sampling a significant fraction of the space defined by the γ and *d* parameters. The best AC_p obtained with the anova kernel (Exp. 61 in Table 2, $TP_p = 87$, $FN_p = 39$, $TN_p = 182$, $FP_p = 29$, $AC_p = 0.80$) is identical with that obtained in Exp. 19 with a radial kernel.

The second group of experiments investigates the SVM discrimination between less reactive compounds (115 compounds; SVM class +1, formed by polar narcotics, ester narcotics, amine narcotics) and more reactive compounds (96 compounds; SVM class –1, formed by weak acid respiratory uncouplers, electrophiles, proelectrophiles, and nucleophiles). For the same classification problem, Ren *et al*. [16] obtained the following LOO predictions: discriminant analysis, $TP_p = 94$, $FN_p = 21$, $TN_p = 39$, $FP_p = 57$, $AC_p = 0.63$; logistic regression, $TP_p = 96$, $FN_p = 19$, $TN_p = 52$, $FP_p = 44$, $AC_p = 0.70$. These results show that the separation between less reactive and more reactive compounds is more difficult than discriminating between nonpolar narcotic compounds and other compounds that have excess toxicity. In Table 3 we present the statistics for the SVM models obtained for the discrimination between less reactive and more reactive compounds.

The L5%O predictions obtained with dot kernels (Exp. 1 and 2 from Table 3, $AC_p = 0.68$) are better than those obtained with discriminant analysis and slightly worse than those obtained with logistic regression. The use of polynomial kernels does not increase the prediction power of the SVM models (AC_p between 0.49 and 0.65). The best predictions are obtained with the radial kernel (Exp. 18, 20, and 22, all with $AC_p = 0.71$); there is a slight difference for the number of FN and FP in these three experiments. The neural kernel gives bad predictions, with AC_p between 0.49 and 0.58. The anova kernel can perfectly separate the two classes in calibration (fitting), in Exp. 78 from Table 3, with $TP_c = 115$, $FN_c = 0$, $TN_c = 96$, $FP_c = 0$, $AC_c = 1$, but the predictions for this experiment are of low quality: $TP_p = 74$, $FN_p = 41$, $TN_p = 53$, $FP_p = 43$, $AC_p = 0.60$. Overall, the bad predictions obtained with the anova kernel (AC_p between 0.60 and 0.67) are surprising, because in previous studies we found that the anova kernel gives best predictions [18,35–38].

From the three models that give good predictions (Exp. 18, 20, and 22) we selected for further analysis Exp. 18 from Table 3 because it has slightly better calibration statistics ($TP_c = 109$, $FN_c = 6$, $TN_c = 77$, $FP_c = 19$, $AC_c = 0.88$). The 24 false negative predictions in the L5%O cross–validation test are (the MOA is indicated in parenthesis; see Table 1 caption for notation): **130**, dicumarol (2); **133**, 2–hydroxybenzaldehyde (2); **136**, 3,5–dibromosalicylaldehyde (2); **138**, 2,4–dihydroxybenzal–

dehyde (2); **148**, nonylphenol (2); **151**, 4–chlorophenol (2); **153**, resorcinol (2); **160**, 2,6–di(*tert*)–butyl–4–methylphenol (2); **174**, 4–acetylpyridine (2); **175**, methyl–4–cyanobenzoate (2); **182**, *N*–ethylbenzylamine (2); **188**, carbazole (2); **192**, 5–bromosalicylaldehyde (2); **193**, 5–chlorosalicyl–aldehyde (2); **197**, 2,2'–methylenebis–(3,4,6–trichlorophenol) (2); **204**, 4,6–dimethyl–2–hydroxy–benzaldehyde (2); **208**, methylacetate (3); **215**, dibutylphthalate (3); **221**, ethylacetate (3); **224**, diethylmalonate (3); **226**, dimethylnitro–terephthalate (3); **236**, decylamine (4); **237**, undecylamine (4); **240**, 1,3–diaminopropane (4).

The 38 false positive predictions are (the MOA is indicated in parenthesis; see Table 1 caption for notation): **242**, pentachlorophenol (5); **245**, 2,5–dinitrophenol (5); **248**, pentabromophenol (5); **249**, 2,3,4,6–tetrachlorophenol (5); **250**, 4–phenylazophenol (5); **254**, 2,3,4,5–tetrachlorophenol (5); **256**, 2–methyl–1,4–naphthquinone (6); **257**, 4–phenoxybenzaldehyde (6); **258**, methylmethacrylate (6); **259**, 1–chloro–2–nitrobenzene (6); **261**, 2–methylbutanal (6); **270**, 4–amino–2–nitrophenol (6); **272**, 1–chloro–3–nitrobenzene (6); **273**, 2–chloro–4–nitroaniline (6); **274**, 4–isopropylbenzaldehyde (6); **275**, 2–methylvaleraldehyde (6); **276**, 2,4–pentanedione (6); **278**, ethylacrylate (6); **281**, 2–fluorobenzaldehyde (6); **283**, 2–methylbenzaldehyde (6); **286**, 2,4–dimethoxybenzaldehyde (6); **292**, 2–hydroxyethylmethacrylate (6); **293**,2,4–dichlorobenzaldehyde (6); **295**, tetrachlorocatechol (6); **296**, 2–ethoxyethylmethacrylate (6); **297**, benzylmethacrylate (6); **301**, isopropylmethacrylate (6); **305**, 4–ethoxybenzaldehyde (6); **306**, 3–hydroxy–2–nitropyridine (6); **309**, methyl–2,5–dichlorobenzoate (6); **311**, 3,5–dibromo–4–hydroxybenzonitrile (6); **312**, 2,4–dichlorobenzamide (6); **316**, 2,2,2–trichloroethanol (6); **322**, 2–chloro–6–methylbenzonitrile (6); **324**, 2–propyn–1–ol (7); **327**, 2–butyn–1–ol (7); **329**, 2–decyn–1–ol (7); **337**, *N*,*N*–diethylethanolamine (8).

The 24 false negative predictions for the 115 less reactive compounds (SVM class +1) are distributed in all three MOAs: MOA 2, polar narcotics, with 16 outliers out of 79 compounds; MOA 3, ester narcotics, with 5 outliers out of 23 compounds; MOA 4, amine narcotics, with 3 outliers out of 13 compounds. The 38 false positive predictions for the 96 more reactive compounds (SVM class –1) are distributed in all four MOAs: MOA 5, weak acid respiratory uncouplers, with 6 outliers out of 13 compounds; MOA 6, electrophiles, with 28 outliers out of 69 compounds; MOA 7, proelectrophiles, with 3 outliers out of 8 compounds; MOA 8, proelectrophiles, with 1 outlier out of 6 compounds. The prediction accuracy for class +1 ($Ac^{+}_{p} = TP_p/(TP_p + FN_p) = 91/115 = 0.79$ is much larger than prediction accuracy for class –1 ($Ac^{-}_{p} = TN_p/(TN_p + FP_p) = 58/96 = 0.60$, indicating that the three descriptors used in the SVM model ($\log K_{ow}$, $\log 1/IGC_{50}$, and $\log 1/LC_{50}$) are not the best combination for predicting more reactive compounds.

4 CONCLUSIONS

The prediction of the mechanism of action (MOA) using structural descriptors has major applications in selecting the appropriate quantitative structure–activity relationships (QSAR) model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimes. The SVM (support vector machines) algorithm was recently proposed as an efficient and flexible classification method for various bioinformatics and cheminformatics applications. In this study we have investigated the application of SVM for the classification of 337 organic compounds from eight MOA classes (nonpolar narcosis, polar narcosis, ester narcosis, amine narcosis, weak acid respiratory uncoupling, electrophilicity, proelectrophilicity, and nucleophilicity). The MOA classification was based on three indices, namely: $\log K_{ow}$, the octanol-water partition coefficient; $\log 1/IGC_{50}$, the 50% inhibitory growth concentration against *Tetrahymena pyriformis*; $\log 1/LC_{50}$, the 50% lethal concentration against *Pimephales promelas*. The prediction power of each SVM model was evaluated with a leave–5%–out cross–validation procedure.

In order to find classification models with good predictive power, we have investigated a large number of SVM models obtained with the dot, polynomial, radial basis function, neural, and anova kernels. The MOA classification performances of SVM models depend strongly on the kernel type and various parameters that control the kernel shape. The discrimination between nonpolar narcotic compounds and the other chemicals can be obtained with radial and anova SVM models, with a prediction accuracy of 0.80. The separation of less reactive compounds (polar, ester, and amine narcotics) from more reactive compounds (electrophiles, proelectrophiles, and nucleophiles) is obtained with a slightly higher error (prediction accuracy 0.71, obtained with radial SVM models). SVM models that use as input parameters hydrophobicity and experimental toxicity against *Pimephales promelas* and *Tetrahymena pyriformis* represent an effective MOA classification method for a large diversity of organic compounds. This approach can be used to predict the aquatic toxicity mechanism and to select the appropriate QSAR model for new chemical compounds.

Supplementary Material

The mySVM model files for Exp. 19 in Table 2 and Exp. 18 in Table 3 are available as supplementary material.

Note Added in Proof

Table 1 from Ren, Frymier, and Schultz [16] has 12 duplicated entries: compounds **50** and **97** (CAS RN 120–82–1); **161** and **201** (CAS RN 133–11–9); **171** and **200** (CAS RN 769–28–8); **176** and **202** (CAS RN 1484–26–0); **290** and **320** (CAS RN 771–60–8); **332** and **335** (CAS RN 78–96–6). These compounds are duplicated also in Table 1 from this paper, but the accuracy values do not change by deleting the duplicated compounds.

5 REFERENCES

[1] A. R. Katritzky, D. B. Tatham, and U. Maran, Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure–Toxicity Relationships, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1162–1176.

[2] H. J. M. Verhaar and C. J. Van Leeuwen, and J. L. M. Hermens, Classifying Environmental Pollutants. 1: Structure–Activity Relationships for Prediction of Aquatic Toxicity, *Chemosphere* **1992**, *25*, 471–491.

[3] S. P. Bradbury, Predicting Modes of Toxic Action From Chemical Structure: An Overview, *SAR QSAR Environ. Res*. **1994**, *2*, 89–104.

[4] O. G. Mekenyan and G. D. Veith, The Electronic Factor in QSAR: MO–Parameters, Competing Interactions, Reactivity and Toxicity, *SAR QSAR Environ. Res*. **1994**, *2*, 129–143.

[5] S. P. Bradbury, Quantitative Structure–Activity Relationships and Ecological Risk Assessment: An Overview of Predictive Aquatic Toxicology Research, *Toxicol. Lett*. **1995**, *79*, 229–237.

[6] S. Karabunarliev, O. G. Mekenyan, W. Karcher, C. L. Russom, and S. P. Bradbury, Quantum–Chemical Descriptors for Estimating the Acute Toxicity of Electrophiles to the Fathead Minnow (*Pimephales promelas*): An Analysis Based on Molecular Mechanisms, *Quant. Struct.–Act. Relat*. **1996**, *15*, 302–310.

[7] C. L. Russom, S. P. Bradbury, S. J. Broderium, D. E. Hammermeister, and R. A. Drummond, Predicting Modes of Toxic Action From Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*), *Environ. Toxicol. Chem*. **1997**, *16*, 948–967.

[8] A. P. Bearden and T. W. Schultz, Structure–Activity Relationships for *Pimephales* and *Tetrahymena*: A Mechanism of Action Approach, *Environ. Toxicol. Chem*. **1997**, *16*, 1311–1317.

[9] A. B. A. Boxall, C. D. Watts, J. C. Dearden, G. M. Bresnen, and R. Scoffin, Classification of Environmental Pollutants Into General Mode of Toxic Action Classes Based on Molecular Descriptors, in: *Quantitative Structure–Activity Relationships in Environmental Sciences VII*, Eds. F. C. Fredenslund and G. Schüürmann, SETAC Press, Pensacola, Florida, USA, 1997, pp. 315–327.

[10] A. P. Bearden and T. W. Schultz, Comparison of *Tetrahymena* and *Pimephales* Toxicity Based on Mechanism of Action, *SAR QSAR Environ. Res*. **1998**, *9*, 127–153.

[11] S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, and S. P. Bradbury, A Comparative Study of Molecular Similarity, Statistical, and Neural Methods for Predicting Toxic Modes of Action, *Environ. Toxicol. Chem*. **1998**, *17*, 1056–1064.

[12] T. W. Schultz, Structure–Toxicity Relationships for Benzenes Evaluated with *Tetrahymena pyriformis*, *Chem. Res. Toxicol*. **1999**, *12*, 1262–1267.

[13] S. Ren and T. W. Schultz, Identifying the Mechanism of Aquatic Toxicity of Selected Compounds by Hydrophobicity and Electrophilicity Descriptors, *Toxicol. Lett*. **2002**, *129*, 151–160.

[14] E. Urrestarazu Ramos, W. H. J. Vaes, H. J. M. Verhaar, and J. L. M. Hermens, Quantitative Structure–Activity Relationships for the Aquatic Toxicity of Polar and Nonpolar Narcotic Pollutants, *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 845–852.

[15] S. Ren, Classifying Class I and Class II Compounds by Hydrophobicity and Hydrogen Bonding Descriptors, *Environ. Toxicol*. **2002**, *17*, 415–423.

[16] S. Ren, P. D. Frymier, and T. W. Schultz, An exploratory study of the use of multivariate techniques to determine mechanisms of toxic action, *Ecotox. Environ. Safety* **2003**, *55*, 86-97.

[17] O. Ivanciuc, Support Vector Machine Identification of the Aquatic Toxicity Mechanism of Organic Compounds, *Internet Electron. J. Mol. Des*. **2002**, *1*, 157–172, http://www.biochempress.com.

[18] O. Ivanciuc, Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2003**, *2*, 195–208, http://www.biochempress.com.

[19] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979.

[20] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[21] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.

[22] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowledge Discov.* **1998**, *2*, 121–167.
[23] B. Schölkopf, K. –K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers, *IEEE Trans. Signal Process.* **1997**, *45*, 2758–2765.
[24] V. N. Vapnik, An Overview of Statistical Learning Theory, *IEEE Trans. Neural Networks* **1999**, *10*, 988–999.
[25] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods*: *Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
[26] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
[27] K.–R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, An Introduction to Kernel–Based Learning Algorithms, *IEEE Trans. Neural Networks* **2001**, *12*, 181–201.
[28] C.–C. Chang and C.–J. Lin, Training ν–Support Vector Classifiers: Theory and Algorithms, *Neural Comput.* **2001**, *12*, 2119–2147.
[29] I. Steinwart, On the Influence of the Kernel on the Consistency of Support Vector Machines, *J. Machine Learning Res.* **2001**, *2*, 67–93, http://www.jmlr.org.
[30] A. Ben–Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, Support Vector Clustering, *J. Machine Learning Res.* **2001**, *2*, 125–137, http://www.jmlr.org.
[31] R. Collobert and S. Bengio, SVMTorch: Support Vector Machines for Large–Scale Regression Problems, *J. Machine Learning Res.* **2001**, *1*, 143–160, http://www.jmlr.org.
[32] O. L. Mangasarian and D. R. Musicant, Lagrangian Support Vector Machines, *J. Machine Learning Res.* **2001**, *1*, 161–177, http://www.jmlr.org.
[33] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.
[34] BioChem Links, http://www.biochempress.com.
[35] O. Ivanciuc, Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons, *Internet Electron. J. Mol. Des.* **2002**, *1*, 203–218, http://www.biochempress.com.
[36] O. Ivanciuc, Structure–Odor Relationships for Pyrazines with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2002**, *1*, 269–284, http://www.biochempress.com.
[37] O. Ivanciuc, Support Vector Machines for Cancer Diagnosis from the Blood Concentration of Zn, Ba, Mg, Ca, Cu, and Se, *Internet Electron. J. Mol. Des.* **2002**, *1*, 418–427, http://www.biochempress.com.
[38] O. Ivanciuc, Support Vector Machines Classification of Black and Green Teas Based on Their Metal Content, *Internet Electron. J. Mol. Des.* **2003**, *2*, 348–357, http://www.biochempress.com.