Quantitative Structure−Activity Relationship Modeling of Juvenile Hormone Mimetic Compounds for *Culex Pipiens* Larvae, with a Discussion of Descriptor-Thinning Methods

Subhash C. Basak,*,† Ramanathan Natarajan,† Denise Mills,† Douglas M. Hawkins,‡ and Jessica J. Kraker‡

Natural Resources Research Institute, Center for Water and Environment, University of Minnesota−Duluth,
5013 Miller Trunk Hwy, Duluth, Minnesota 55811, and School of Statistics, University of Minnesota,
224 Church Street SE, Minneapolis, Minnesota 55455

Received May 25, 2005

Quantitative structure−activity relationship (QSAR) modelers often encounter the problem of multicollinearity owing to the availability of large numbers of computable molecular descriptors. Sparsity of the variables while using descriptors such as atom pairs increases the complexity. Three different predictor-thinning methods, namely, a modified Gram−Schmidt algorithm, a marginal soft thresholding algorithm, and LASSO (least absolute shrinkage and selection operator), were utilized to reduce the number of descriptors prior to developing linear models. Juvenile hormone (JH) activity of 304 compounds on *Culex pipiens* larvae was taken as the model data set, and predictor trimming of a large number of diverse descriptors comprising 268 global molecular descriptors (topostructural, topochemical, and geometrical), 13 quantum chemical descriptors, and 915 atom pairs (substructural counts) was applied prior to linear regression by the ridge regression method. The data set ($N = 304$) was split into five calibration data sets of random samples of sizes 60/110/160/210/260, and the remaining 244/194/144/94/44 compounds were used for validations. LASSO was not found to be a very effective method in handling a large set of descriptors because the number of predictors retained could not exceed the number of observations. The results indicated that the modified Gram−Schmidt algorithm could be used to trim the number of predictors in the global molecular descriptor set where collinearity of the descriptors was the major concern. On the contrary, the soft thresholding approach was found to be an effective tool in subset selection from a diverse set of descriptors having both sparsity and multicollinearity, as in the case of the combined set of atom pairs and global molecular descriptors. The final model developed after variable selection was dominated more by atom pairs, which indicated the important structural moieties that affect JH activity of the compounds. The success of the method reiterates the fact that QSAR or quantitative structure−property relationship (QSPR) models can be developed for a diverse set of compounds using properly parametrized and diverse sets of descriptors, of course, with the selection of the appropriate statistical tools.

1. INTRODUCTION

Quantitative structure−activity relationship (QSAR) modeling has become an indispensable tool in drug design, and its implementation considerably reduces the time and cost required in lead optimization. The QSAR approach plays an equally important role in developing structure−toxicity relationships (QSTRs), which facilitate the prediction of toxicity of chemicals from structural inputs alone, without any need for the determination of physicochemical properties.[1,2] In general, QSAR modeling can be used to predict toxicity, biological activity, environmental hazard, and physicochemical properties such as partition coefficients and so forth for existing compounds as well as those that are not yet synthesized, that is, those that are from virtual libraries.

To build a QSAR model, the candidate chemical structures must be transformed into numerical descriptors that encode their structural features. One important method of numerical characterization of molecular structures was achieved by applying the principles of graph theory to molecular structure, and this has resulted in numerous topological descriptors based on molecular graphs. The number of available topological descriptors[3,4] is well over 300, and the number is ever increasing. To encode information about the three-dimensional geometry of molecules, geometrical or 3-D descriptors[3] were added to the list. Quantum chemical computations at various levels of complexity yield a set of parameters that is being used in model building.[5] Substructures such as atom pairs[6] can also be used as molecular descriptors because they have been found to perform well in QSAR models. The number of atom pairs available for a set of compounds may be in the hundreds. Hence, the QSAR modeler is overwhelmed with a large number, say, over 1000, of structural descriptors to model a few hundred compounds and even less in many instances. When there is a large pool of descriptor variables available wherein many are intercorrelated, the QSAR, QSTR, or quantitative structure−property relationship (QSPR) developed may suffer from overfitting.

* Corresponding author e-mail: sbasak@nrri.umn.edu; phone: +1-218-720-4230; fax: +1-218-720-4328.

† University of Minnesota−Duluth.

‡ University of Minnesota−Minneapolis.

10.1021/ci050215y CCC: $33.50
Published on Web 10/14/2005

The ultimate purpose of developing a regression model is not only to predict properties but also to provide an interpretation of the structural features responsible for the activity or property of interest.[7] This interpretation is very important in drug design in order to convey to the bench chemist which structural features are crucial for lead optimization. Many investigators try to interpret the linear models from the coefficients of the predictors; such attempts are very dangerous. When orthogonal descriptors are not used, the descriptor coefficients in a regression equation present a composite picture, and thus, the coefficients need not represent the individual trends. Several methods such as ridge regression (RR),[8,9] partial least-squares projections (PLS),[10] principal component regression (PCR),[11] the pair-wise correlation method,[12−14] stepwise linear regression,[15] and best subset selection[16] have been developed to alleviate this problem. There are claims and counterclaims disputing the superiority of one method or another.[17,18] When the number of descriptors available is on the order of 1000, as in the current study, conventional stepwise regression would look for a subset of statistically significant predictors. Typically, no more than a dozen of the predictors would be selected, and there is a considerable body of evidence to indicate that subsetting methods such as stepwise regression do not work well (see Rencher and Punn[19]), particularly for such a large number of predictors as we have. There are large elements of chance going into the handful of predictors in the final regression, and this leads to a substantial overstatement of their apparent statistical significance. However, it seems plausible that some predictor thinning can be done safely, provided this is not carried to the extreme of stepwise regression. In the present study, to both lighten computation and improve interpretability, thinning of the descriptors using a Gram−Schmidt orthogonalization method[20] and a marginal soft threshold method was applied in a stepwise selection of 50, 100, 150, 200, or 250 of the predictors. The performance of a moderate pruning of the predictors and its effect on the fit of the models are discussed.

2. MATERIALS AND METHODS

2.1. Bioactivity Data. To test the applicability of introducing a new step, namely, descriptor thinning, before doing linear modeling, a highly reliable data set of adequate size was a primary prerequisite. Bioactivity data were preferred to physicochemical properties because the mechanism of bioactivity is more complex and has implications in drug design; moreover, it is more relevant to the authors' area of research. Activity of juvenile hormone (JH) mimetic compounds on *Culex pipiens* mosquito larvae compiled from the papers by Niwa et al.[21−24] and Hayashi et al.[25,26] appeared to satisfy the criteria as there were 304 compounds in the data set with the same type of bioassay. The structures of the compounds and their juvenile hormone activity reported in terms of pI_{50} (M), that is, the logarithm of the reciprocal of the concentration at which 50% inhibition of metamorphosis was observed, are given in Table 1. There are 314 compounds listed in the table, but 6 of them had missing data (marked as N/A) and 4 had censored values (pI_{50} < 4.5). These 10 compounds were not included in the modeling. Although some structure−activity modeling was reported by Hayashi et al.,[26,27] this was done about 15 years ago when the Hansch[28] approach was the primary method of QSAR, using descriptors such as *Sterimol* parameters and log *P* as the measure of hydrophobicity. This approach had its own limitations due to the use of condensed parameters; also, the diverse set of data could not be modeled well. In several instances, indicator variables were used to explain the data variation due to a particular group on the scaffold. In contrast, the present investigation has the luxury of considering the entire data set comprised of dissimilar compounds owing to the availability of the large and diverse descriptor pool.

2.2. Calculation of Molecular Descriptors. A large number of structural descriptors were calculated using molecular structures represented by SMILES code as the only input. A large set of topological descriptors including Wiener number,[29] molecular connectivity indices calculated using Randic[30] as well as Kier and Hall[31] formulations, frequencies of paths of varying length, Bonchev and Trinajstic[32] information indices based on distance matrices of molecular graphs, neighborhood complexity indices defined by Basak et al.[33] for hydrogen-filled graphs, and Balaban's *J* indices[34] were calculated using *POLLY version 2.3*.[35] The triplet indices developed by Balaban and co-workers[36] are otherwise known as real-number local vertex invariants (LOVIs) because these indices are the solutions of linear equations obtained from a matrix (adjacency matrix **A**, distance matrix **D**, or unity matrix **1**), a main diagonal column vector (vertex degree *V* and atomic number *Z*), or a free term column vector. The LOVIs were calculated using in-house software *Triplet*. Additional topological descriptors, along with a large set of electrotopological or E-state indices,[37] hydrogen bonding indices, and κ shape indices[38] were calculated using *Molconn-Z version 3.5*.[39] The complete list of calculated indices can be found with their abbreviations and a brief description in some of our earlier publications.[40] Quantum chemical descriptors such as solvent accessible volume, molecular surface area, energy of the highest occupied molecular orbital (E_{HOMO}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), and the HOMO−LUMO gap ($E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}}$) were calculated using Chem3D Ultra.[41] In addition to these descriptors, atom pairs were calculated using the program *APProbe*.[42]

2.3. Initial Data Transformation and Reduction. To handle the large differences in the magnitudes of the various descriptors, the descriptors were transformed as $\log_e(N + x)$, where N is the numerical value of the descriptor and $x = 1$ when $N > -1$, which is true for most of the descriptors. However, some of the Molconn-Z parameters have values of ≤ -1. For those descriptors, the value of x is the smallest whole number which results in a positive sum for $(N + x)$. The CORR procedure of the SAS statistical package[43] was used to identify pairs of perfectly correlated descriptors ($R = 1$), and only one descriptor of each such pair was retained. In addition, any descriptors possessing a constant value for all compounds within the data set were omitted. The final descriptor set contained 920 atom pairs (AP) and 281 global indices (inclusive of all topological and geometrical and quantum chemical parameters).

2.4. Linear Regression Models. Calculated molecular descriptors were grouped into three sets of predictors, namely, (i) the atom pair descriptors (AP), (ii) the global molecular descriptors (DES), which included all molecular descriptors except the atom pairs, and (iii) the combined set of descriptors (BTH), consisting of both AP and DES. The

Table 1. Structures and Juvenile Hormone (JH) Activity of JH Mimetics on *Culex pipiens* Larvae

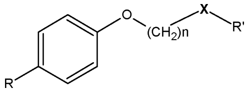
	R	*n*	X	R′	p*I* 50 (M)
			I a 		
1	$-CH_2CH(CH_2CH_3)_2$	3	$-CH_2-$	$-CH_2CH_2CH_3$	5.18
2	$-CH_2CH(CH_2CH_3)_2$	1	$-CH=N-O-$	$-CH_2CH_2CH_3$	8.40
3	$-CH_2CH(CH_2CH_3)_2$	2	$-CH=N-O-$	$-CH_2CH_3$	9.53
4	$-CH_2CH(CH_2CH_3)_2$	3	$-CH=N-O-$	$-CH_3$	6.66
5	$-CH_2CH(CH_2CH_3)_2$	2	$-CH=N-O-$	$-CH(CH_3)_2$	10.76
6	$-CH_2CH(CH_2CH_3)_2$	2	$-O-$	$-CH_2CH_2CH_2CH_3$	6.78
7	$-CH_2CH(CH_2CH_3)_2$	3	$-O-$	$-CH_2CH_2CH_3$	8.17
8	$-CH_2CH(CH_2CH_3)_2$	4	$-O-$	$-CH_2CH_3$	6.68
9	$-CH_2CH(CH_2CH_3)_2$	5	$-O-$	$-CH_3$	5.71
10	$-CH_2CH(CH_2CH_3)_2$	3	$-O-$	$-CH_2CH(CH_3)_2$	8.81
11	$-CH_2CH(CH_2CH_3)_2$	2	$-O-N=CH-$	$-CH_2CH_3$	9.88
12	$-CH_2CH(CH_2CH_3)_2$	2	$-O-N=CH-$	$-CH(CH_3)_2$	9.71
13	$-CH_2CH(CH_2CH_3)_2$	3	$-O-N=C-$	$(CH_3)_2$	8.65
14	$-CH_2CH(CH_2CH_3)_2$	2	$-HN-O-$	$-CH_2CH_2CH_3$	7.33
15	$-CH_2CH(CH_2CH_3)_2$	3	$-HN-O-$	$-CH_2CH_3$	8.43
16	$-CH_2CH(CH_2CH_3)_2$	4	$-HN-O-$	$-CH_3$	6.30
17	$-CH_2CH(CH_2CH_3)_2$	3	$-HN-O-$	$-CH(CH_3)_2$	8.97
18	$-CH_2CH(CH_2CH_3)_2$	2	$-O-NH-$	$-CH_2CH(CH_3)_2$	8.66
19	$-CH_2CH(CH_2CH_3)_2$	1	$-CH(CH_3)OC(=O)-NH-$	$-CH(CH_3)_2$	7.95
20	$-CH_2CH(CH_2CH_3)_2$	1	$-CH_2OC(=O)-N(CH_3)-$	$-CH_2CH_3$	8.59
21	$-CH_2CH(CH_2CH_3)_2$	3	$-C(CH_3)OC(=O)-N(CH_3)-$	$-CH(CH_3)_2$	8.25
22	$-CH_2CH(CH_2CH_3)_2$	2	$-CH_2OC(=O)-NH-$	$-CH_3$	6.50
23	$-CH_2CH(CH_2CH_3)_2$	2	$-CH_2OC(=O)-N(CH_3)-$	$-CH_3$	7.77
24	$-CH_2CH(CH_2CH_3)_2$	2	$-CH(CH_3)OC(=O)-NH-$	$-CH_3$	7.11
25	$-CH_2CH(CH_2CH_3)_2$	2	$-NHC(=O)O-$	$-CH_2CH_3$	7.54
26	$-CH_2CH(CH_2CH_3)_2$	2	$-NHC(=O)O-$	$-CH(CH_3)_2$	8.12
27	$-CH_2CH(CH_2CH_3)_2$	2	$-N(CH_3)C(=O)O-$	$-CH(CH_3)_2$	7.86
28	$-CH_2CH(CH_2CH_3)_2$	3	$-NHC(=O)O-$	$-CH_3$	6.31
29	$-CH_2CH(CH_2CH_3)_2$	3	$-N(CH_3)C(=O)O-$	$-CH_3$	6.79
30	$-CH_2CH(CH_2CH_3)_2$	2	$-NHC(=O)NH-$	$-CH(CH_3)_2$	6.97
31	$-CH_2CH(CH_2CH_3)_2$	3	$-NHC(=O)NH-$	$-CH_3$	4.83
32	$-CH_2C(CH_3)_3$	1	$-CH=N-O-$	$-CH_2CH_2CH_3$	6.54
33	$-CH_2C(CH_3)_3$	2	$-CH=N-O-$	$-CH_2CH_3$	8.57
34	$-CH_2C(CH_3)_3$	2	$-CH=N-O-$	$-CH(CH_3)_2$	10.03
35	$-CH_2C(CH_3)_3$	2	$-O-$	$-CH_2CH_2CH_2CH_3$	6.36
36	$-CH_2C(CH_3)_3$	3	$-O-$	$-CH_2CH_2CH_3$	8.01
37	$-CH_2C(CH_3)_3$	4	$-O-$	$-CH_2CH_3$	6.66
38	$-CH_2C(CH_3)_3$	5	$-O-$	$-CH_3$	5.29
39	$-CH_2C(CH_3)_3$	3	$-O-$	$-CH_2CH(CH_3)_2$	9.43
40	$-CH_2C(CH_3)_3$	2	$-O-N=CH$	$-CH_2CH_3$	9.50
41	$-CH_2CH(CH_3)_2$	1	$-CH=N-O-$	$-CH_2CH_2CH_3$	5.39
42	$-CH_2CH(CH_3)_2$	2	$-CH=N-O-$	$-CH_2CH_3$	8.29
43	$-CH_2CH(CH_3)_2$	2	$-CH=N-O-$	$-CH(CH_3)_2$	10.85
44	$-CH_2CH(CH_3)_2$	2	$-O-$	$-CH_2CH_2CH_2CH_3$	6.44
45	$-CH_2CH(CH_3)_2$	3	$-O-$	$-CH_2CH_2CH_3$	7.86
46	$-CH_2CH(CH_3)_2$	4	$-O-$	$-CH_2CH_3$	6.48
47	$-CH_2CH(CH_3)_2$	5	$-O-$	$-CH_3$	5.35
48	$-CH_2CH(CH_3)_2$	2	$-HN-O-$	$-CH_2CH_2CH_3$	6.57
49	$-CH_2CH(CH_3)_2$	3	$-HN-O-$	$-CH_2CH_3$	7.81
50	$-O-CH(CH_2CH_3)_2$	1	$-CH=N-O-$	$-CH_2CH_2CH_3$	7.83
51	$-O-CH(CH_2CH_3)_2$	2	$-CH=N-O-$	$-CH_2CH_3$	8.19
52	$-O-CH(CH_2CH_3)_2$	3	$-CH=N-O-$	$-CH_3$	6.23
53	$-O-CH(CH_2CH_3)_2$	2	$-CH=N-O-$	$-CH(CH_3)_2$	9.30
54	$-O-CH(CH_2CH_3)_2$	2	$-O-$	$-CH_2CH_2CH_2CH_3$	5.49
55	$-O-CH(CH_2CH_3)_2$	3	$-O-$	$-CH_2CH_2CH_3$	6.80
56	$-O-CH(CH_2CH_3)_2$	4	$-O-$	$-CH_2CH_3$	5.98
57	$-O-CH(CH_2CH_3)_2$	5	$-O-$	$-CH_3$	5.04
58	$-CH_2CH(CH_2CH_3)_2$	1	$-CH=CHC(=O)O-$	$-CH_2CH_3$	5.74
59	$-CH_2CH(CH_2CH_3)_2$	1	$-CH=C(CH_3)HC(=O)O-$	$-CH_2CH_3$	5.84
60	$-CH_2CH(CH_2CH_3)_2$	2	$-C(CH_3)OC(=O)$	$-CH(CH_3)_2$	5.32
61	$-CH_2CH(CH_2CH_3)_2$	2	$-O-C(=O)CH=CH$	$(CH_3)_2$	6.82
62	$-CH_2CH(CH_2CH_3)_2$	1	$-C(=O)NH-$	$-CH_2CH_2CH_2CH_3$	5.09
63	$-CH_2CH(CH_2CH_3)_2$	2	$-C(=O)NH-$	$-CH_2CH_2CH_2CH_3$	4.81
64	$-CH_2CH(CH_2CH_3)_2$	3	$-C(=O)NH-$	$-CH_2CH_2CH_2CH_3$	5.87
65	$-CH_2CH(CH_2CH_3)_2$	4	$-C(=O)NH-$	$-CH_3$	5.15
66	$-CH_2CH(CH_2CH_3)_2$	1	$C(CH3)=CHC(=O)NH-$	$-CH(CH_3)_2$	6.99
67	$-CH_2CH(CH_2CH_3)_2$	2	$-OC=(O)NH-$	$-CH_2CH_3$	7.57
68	$-CH_2CH(CH_2CH_3)_2$	2	$-OC=(O)NH-$	$-CH(CH_3)_2$	8.62

Table 1 (Continued)

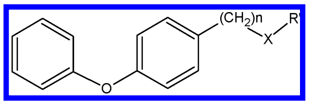
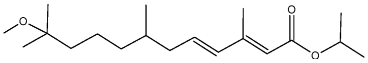
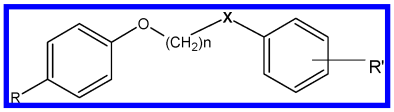
	R	*n*	X	R′	p*I* 50 (M)
69	-CH$_2$C(CH$_3$)$_3$	1	-CH=CHC(=O)O-	-CH$_2$CH$_3$	5.82
70	-CH$_2$C(CH$_3$)$_3$	2	-OC=(O)NH-	-CH$_2$CH$_3$	6.87
71	-CH$_2$C(CH$_3$)$_3$	2	-OC=(O)NH-	-CH(CH$_3$)$_2$	8.23
72	-CH$_2$C(CH$_3$)$_3$	2	-NHC(=O)O-	-CH$_2$CH$_3$	7.70
73	-CH$_2$C(CH$_3$)$_3$	2	-NHC(=O)O-	-CH(CH$_3$)$_2$	7.78
			Ib  		
74		1	-C(=O)O-	-CH$_2$CH$_2$CH$_2$CH$_3$	<4.5
75		2	-C(=O)O-	-CH$_2$CH$_2$CH$_3$	4.91
76		3	-C(=O)O-	-CH$_2$CH$_3$	<4.5
77		3	-OC(=O)-	-CH$_2$CH$_2$CH$_3$	<4.5
78		3	-OC(=O)-	-CH$_2$CH$_3$	5.02
79		4	-OC(=O)-	-CH$_3$	<4.5
80		2	-NHC(=O)O-	-CH$_2$CH$_3$	8.82
81	 				9.50
			II  		
82	-CH$_2$CH(CH$_2$CH$_3$)$_2$	4	O	H	N/A
83	-CH$_2$CH(CH$_2$CH$_3$)$_2$	3	CH$_2$	H	4.24
84	-CH$_2$CH(CH$_2$CH$_3$)$_2$	3	-O-	H	7.56
85	-CH$_2$CH(CH$_2$CH$_3$)$_2$	3	-O-	3-CH$_3$	6.48
86	-CH$_2$CH(CH$_2$CH$_3$)$_2$	3	-O-	4-CH$_3$	6.74
87	-CH$_2$CH(CH$_2$CH$_3$)$_2$	3	-O-	2-CH$_3$	5.11
88	-CH$_2$CH(CH$_2$CH$_3$)$_2$	3	-O-	F$_5$	N/A
89	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	H	7.15
90	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	2-CH$_3$	N/A
91	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	3-CH$_3$	7.17
92	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	4-CH$_3$	5.96
93	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	3-CH$_2$CH$_3$	7.58
94	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	4-CH$_2$CH$_3$	5.70
95	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	3-CH(CH$_3$)$_2$	7.78
96	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	4-CH(CH$_3$)$_2$	N/A
97	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	3-C(CH$_3$)$_3$	N/A
98	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-O-	3-CH(CH$_3$)$_2$	7.88
99	-CH$_2$CH(CH$_2$CH$_3$)$_2$	2	-CH$_2$-	H	6.70
100	-CH$_2$CH(CH$_2$CH$_3$)$_2$	1	-CH$_2$-	H	5.41
101	-CH$_2$CH(CH$_2$CH$_3$)$_2$	1	-CH$_2$-	3-OCH$_3$	4.96
102	-CH$_2$CH(CH$_2$CH$_3$)$_2$	1	-CH$_2$-	4-OCH$_3$	6.80
103	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	H	6.63
104	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH$_3$	8.03
105	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH$_2$CH$_3$	8.65
106	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH$_2$CH$_2$CH$_3$	8.07
107	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH$_2$CH$_2$CH$_2$CH$_3$	6.24
108	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	2-OCH$_3$	N/A
109	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	3-OCH$_3$	6.90
110	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-OCH$_3$	8.52
111	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-OCH$_2$CH$_3$	8.74
112	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-OCH$_2$CH$_3$	6.80
113	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH(CH$_3$)$_2$	8.80
114	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-C(CH$_3$)$_3$	8.18
115	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH(CH$_3$)(CH$_2$CH$_3$)	8.24
116	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$ -	4-C(CH$_3$)$_2$(CH$_2$CH$_3$)	7.35
117	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH$_2$CH(CH$_3$)$_2$	8.33
118	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-CH$_2$C(CH$_3$)$_3$	6.45
119	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-O-CH(CH3)2	9.04
120	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-O-C$_5$H$_9$	6.72
121	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	naphthyl	6.90
122	-C(CH$_3$)$_3$	0	-CH$_2$-	4-OCH$_2$CH$_3$	7.62
123	-C(CH$_3$)$_3$	0	-CH$_2$-	4-OCH(CH$_3$)$_2$	7.58
124	-O-CH(CH$_3$)$_2$	0	-CH$_2$-	4-OCH(CH$_3$)$_2$	6.96
125	-O-CH(CH$_3$)(C$_2$H$_5$)	0	-CH$_2$-	4-OCH(CH$_3$)$_2$	7.12
126	-O-CH(CH$_2$CH$_3$)$_2$	0	-CH$_2$-	4-OCH(CH$_3$)$_2$	7.61
127	-CH$_2$CH(CH$_2$CH$_3$)$_2$	0		4-OCH$_2$CH$_3$	5.07

Table 1 (Continued)

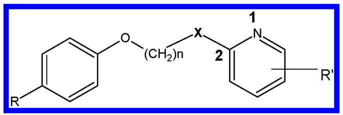
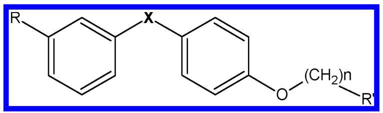
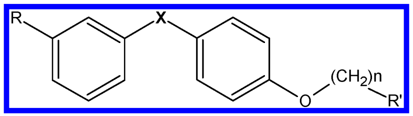
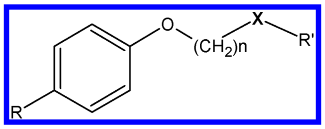
	R	*n*	X	R′	p*I* 50 (M)
128	$-CH_2CH(CH_2CH_3)_2$	0		$3-OCH_2CH_3$	5.19
129	$-CH_2CH(CH_2CH_3)_2$	0		$4-OCH_2CH_2CH_3$	5.13
130	$-CH_2CH(CH_2CH_3)_2$	0		$3-OCH_2CH_2CH_3$	5.26
			III^a  		
131	$-CH_2CH(CH_2CH_3)_2$	3	-O-	H	8.44
132	$-CH_2CH(CH_2CH_3)_2$	3	-O-	$4-CH_3$	7.63
133	$-CH_2CH(CH_2CH_3)_2$	3	-O-	$5-CH_3$	7.88
134	$-CH_2C(CH_3)_3$	3	-O-	H	8.22
135	$-CH_2CH(CH_2CH_3)_2$	2	-O-	H	10.04
136	$-CH_2CH(CH_2CH_3)_2$	1	$-CH(CH_3)-O-$	H	9.64
137	$-CH_2CH(CH_2CH_3)_2$	2	-O-	$4-CH_3$	7.80
138	$-CH_2CH(CH_2CH_3)_2$	2	-O-	$5-CH_3$	9.10
139	$-CH_2CH(CH_2CH_3)_2$	2	-O-	H (3*)	6.73
140	$-CH_2CH(CH_2CH_3)_2$	2	-O-	H (4*)	5.04
141	$-CH_2CH(CH_2CH_3)_2$	2	-O-	H (1,3-diazine*)	8.69
142	$-CH_2C(CH_3)_3$	2	-O-	H	9.34
143	$-CH_2C(CH_3)_3$	1	$-CH(CH_3)-O-$	H	8.72
144	$-CH_2C(CH_3)_3$	1	$-CH(CH_3)-O-$	$5-CF_3$	7.45
145	$-CH_2CH(CH_2CH_3)_2$	1	$-CH(CH_3)-O-$	H	7.92
146	$-CH_2CH(CH_2CH_3)_2$	0	$-CH_2-$	H	4.93
147	$-CH_2CH(CH_2CH_3)_2$	0	$-CH_2-$	H(4*)	5.70
148	$-CH_2CH(CH_2CH_3)_2$	0	$-CH_2-$	$2-OCH_3$(5*)	7.94
			IV  		
149	H	2	-O-	$-OCH_2CH_2CH_2CH_3$	8.08
150	H	3	-O-	$-OCH_2CH_2CH_3$	9.47
151	H	4	-O-	$-OCH_2CH_3$	7.93
152	H	5	-O-	$-OCH_3$	6.42
153	H	3	-O-	$-OCH_2CH{=}CH_2$	9.07
154	H	3	-O-	$-OCH_2CH{\equiv}CH$	9.01
155	H	3	-O-	$-OCH_2CH(CH_3)_2$	9.73
156	H	3	-O-	$-OCH(CH_3)CH_2CH_3$	8.90
157	H	2	-O-	$-CH(CH_3)OCH_2CH_2CH_3$	9.60
158	H	2	-O-	$-CH(CH_3)OCH_2CH(CH_3)_2$	9.68
159	H	1	-O-	$-CH(CH_3)CH_2OCH_2CH_2CH_3$	8.30
160	H	1	-O-	$-CH(CH_3)CH_2OCH_2CH(CH_3)_2$	8.42
161	H	4	-O-	$-OCH_2CH_2CH_3$	8.21
162	H	4	-O-	$-OCH_2CH(CH_3)_2$	8.08
163	CH_3	3	-O-	$-OCH_2CH_2CH_3$	9.72
164	CH_3	3	-O-	$-OCH_2CH(CH_3)_2$	10.49
165	CH_3	3	-O-	$-OCH_2CH(CH_3)_2$	10.28
166	H	2	$-CH_2-$	$-OCH_2C(CH_3)_3$	8.07
167	H	3	$-CH_2-$	$-OCH_2CH_2CH_3$	9.18
168	H	3	$-CH_2-$	$-OCH_2CH_2CH_3$	9.37
169	H	2	$-CH_2-$	$-OCH_2CH(CH_3)_2$	7.19
170	H	2	$-CH_2-$	$-OCH(OCH_2)_2CH_2$	6.66
171	H	3	$-CH_2-$	$-OC_6H_5$	7.88
172	H	4	$-CH_2-$	$-OC_6H_5$	6.09
173	H	6	$-CH_2-$	$-CH_3$	7.17
			V  		
174	H	2	-O-	$-NHOCH_2CH_2CH_3$	8.28
175	H	2	-O-	$-NHOCH_2CH(CH_3)_2$	8.51
176	$3-CH_3$	2	-O-	$-NHOCH_2CH_2\ CH_3$	8.76
177	H	2	$-CH_2-$	$-NHOCH_2CH_2CH_3$	8.24
178	H	2	$-CH_2-$	$-NHOCH_2CH(CH_3)_2$	8.66
179	H	3	-O-	$-NHOCH_2CH_3$	9.53
180	H	3	-O-	$-NHOCH(CH_3)_2$	9.95
181	$3-CH_3$	3	-O-	$-NHOCH_2CH_3$	9.85

Table 1 (Continued)

	R	*n*	X	R′	p*I* 50 (M)
182	3-CH_3	3	-O-	-$NHOCH(CH_3)_2$	10.00
183	3-CH_3	3	-O-	-$NHOCH(CH_3)_2$	8.92
184	H	3	-CH_2-	-$NHOCH(CH_3)_2$	9.49
185	3-CH_3	3	-O-	-$NHOCH_2CH_3$	9.60
186	3-CH_3	3	-O-	-$NHOCH(CH3)_2$	9.95
187	H	2	-O-	-$NHCH_2CH_2CH_3$	9.55
188	3-CH_3	2	-O-	-$NHCH_2CH_2CH_3$	9.71
189	3-CH_3	2	-O-	-$NHCH(CH_3)_2$	10.00
190	H	3	-O-	-$NHCH_2CH_3$	8.34
191	H	3	-O-	-$NHOCH(CH_3)_2$	8.66
192	H	2	-O-	-$NHCH_2CH_2CH_2CH_3$	6.88
193	H	3	-O-	-$NHCH_2CH_2CH_3$	7.68
194	H	4	-O-	-$NHCH_2CH_3$	7.36
195	H	3	-O-	-$NHCH_2CH(CH_3)_2$	8.37
196	H	2	-O-	-$N(CH_3)CH_2CH_2CH_2CH_3$	7.23
197	H	3	-O-	-$N(CH_3)CH_2CH_2CH_3$	7.93
198	H	4	-O-	-$N(CH_3)CH_2CH_3$	7.55
199	H	1	-O-	-$CH{=}NOCH_2CH(CH_3)_2$	8.72
200	3-CH_3	2	-CH_2-	-$CH{=}NOCH_2CH_3$	9.86
201	3-CH_3	2	CH_2-	-$CH{=}NOCH(CH_3)_2$	10.06
202	H	3	-O-	-$ON{=}CHCH_3$	8.55
203	H	3	-O-	-$ONCH(CH_3)_2$	8.76
204	H	2	-O-	-$ON{=}CHCH_2CH_3$	9.59
205	3-CH_3	2	-O-	-$ON{=}CHCH_2CH_3$	10.31
206	3-CH_3	2	-O-	-$ON{=}CHCH(CH_3)_2$	10.35

VI



207	-OCH_3	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.13
208	-OCH_2CH_3	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.27
209	-$OCH_2CH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.82
210	-$OCH_2CH_2CH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.39
211	-$OCH_2(CH_2)_3CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.21
212	-$OCH_2(CH_2)_4CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.46
213	-$OCH_2(CH_2)_5CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.29
214	-$OCH_2(CH_2)_6CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.29
215	-$OCH(CH_3)_2$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	7.81
216	-$CH_2CH(CH_3)CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.25
217	-$OCH_2CH(CH_3)CH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.79
218	-$OCH(CH_3)CH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	7.57
219	-$OCH(CH_2CH_3)CH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	7.83
220	-$OCH_2CH{=}CH_2$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.94
221	-$OCH_2CH{\equiv}CH$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.20
222	-$OCH_2CH{=}C(CH_3)_2$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.62
223	-$OCH_2CH_2CH_2OH$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.79
224	-$OCH_2CH_2OCH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.90
225	-$OCH_2CH_2OCH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.48
226	-OCH_2-C_6H_{11}	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.66
227	-OCH_2-C_6H_5	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	7.27
228	-$O(CH_2)_2CH(CH_3)(OCH_3)CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	5.77
229	-O-(2-epoxy-3-methyl)-butyl	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	7.04
230	-$OCH_2CH_2CH_3$	2	-$CH{=}N$-O-	-CH_2CH_3	7.91
231	-$OCH(CH_3)_2$	2	-$CH{=}N$-O-	-CH_2CH_3	8.79
232	-$OCH(CH_3)_2$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	8.15
233	-$OCH(CH_3)CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	9.18
234	-$OCH(CH_2CH_3)CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	9.50
235	-$OCH_2CH{=}C(CH_3)_2$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	7.38
236	-$O(CH_2)_2CH(CH_3)(OCH_3)CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	7.14
237	-$CH_2CH_2CH_2CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.44
238	-$CH_2(CH_2)_3CH_3$	1	-$CH{=}N$-O-	-$CH_2CH_2CH_3$	6.02
239	-$CH_2CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	7.28
240	-$CH_2CH_2CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	7.34
241	-$CH_2CH(CH_3)CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	10.85
242	-$CH(CH_3)CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	6.73
243	-$CH_2CH_2(C_2H_5)CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	10.76
244	-$CH_2CH(CH_3)CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	8.46
245	-$CH_2C(CH_3)_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	10.03
246	-$C(CH_3)_2CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	7.21
247	-$CH_2CH(n$-$C_3H_7)CH_2CH_2CH_3$	2	-$CH{=}N$-O-	-$CH(CH_3)_2$	9.78

Table 1 (Continued)

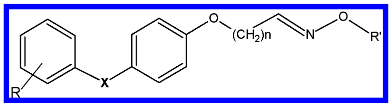
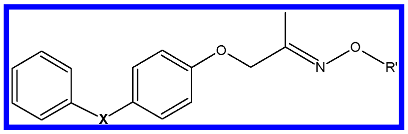
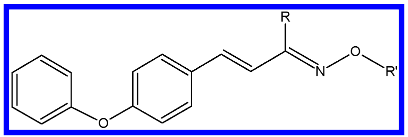
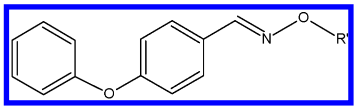
	R	*n*	X	R′	p*I* 50 (M)
248	$-CH_2-C_3H_5$	2	-CH=N-O-	$-CH(CH_3)_2$	8.71
249	$-CH_2$-1-methyl-C_3H_4	2	-CH=N-O-	$-CH(CH_3)_2$	9.82
250	$-CH_2-C_6H_{11}$	2	-CH=N-O-	$-CH(CH_3)_2$	9.37
			VIIa		
					
251	JH I				6.29
252	JH III				6.15
253	H	1	$-CH_2-$	$-CH_2CH_3$	7.09
254	H	1	$-CH_2-$	$-CH_2CH_2CH_3$	7.97
255	H	1	$-CH_2-$	$-CH(CH_3)_2$	7.72
256	H	1	$-CH_2-$	$-CH_2CH=CH_2$	8.14
257	H	1	$-CH_2-$	$-CH_2CH\equiv CH$	7.46
258	H	1	$-CH_2-$	$-CH_2CH_2H_2CH_3$	7.41
259	H	1	$-CH_2-$	$-CH_2CH(CH_3)CH_3$	8.48
260	H	1	$-CH_2-$	$-CH(CH_3)CH_2CH_3$	7.42
261	H	1	$-CH_2-$	$-CH_2CH_2CH_2CH_2CH_3$	6.72
262	H	1	$-CH_2-$	$-CH_2CH(CH_3)CH_2CH_3$	8.10
263	H	1	$-CH_2-$	$-C_5H_9$	8.10
264	H	1	$-CH_2-$	$-C_6H_{11}$	8.12
265	H	1	$-CH_2-$	$-CH_2C_6H_5$	6.75
266	H	2	$-CH_2-$	$-CH_2CH_3$	8.76
267	H	2	$-CH_2-$	$-CH_2CH_2CH_3$	8.37
268	H	2	$-CH_2-$	$-CH_2(CH_3)_2$	9.68
269	H	2	$-CH_2-$	$-C_5H_9$	8.42
270	H	2	-O-	H	6.58
271	H	2	-O-	$-CH_2CH_3$	9.09
272	H	2	-O-	$-CH_2(CH_3)_2$	9.76
273	3-CH_3	2	-O-	$-CH_2CH_3$	9.46
274	3-CH_3	2	-O-	$-CH_2(CH_3)_2$	10.00
275	H	1	-O-	$-CH_2CH_2CH_3$	7.54
276	2-CH_3	1	-O-	$-CH_2CH_2CH_3$	7.36
277	2-OCH_3	1	-O-	$-CH_2CH_2CH_3$	6.81
278	2-F	1	-O-	$-CH_2CH_2CH_3$	7.35
279	2-Cl	1	-O-	$-CH_2CH_2CH_3$	6.89
280	2-CF_3	1	-O-	$-CH_2CH_2CH_3$	6.18
281	3-CH_3	1	-O-	$-CH_2CH_2CH_3$	8.71
282	3-CH_2CH_3	1	-O-	$-CH_2CH_2CH_3$	7.85
283	3-OCH_3	1	-O-	$-CH_2CH_2CH_3$	7.41
284	3-F	1	-O-	$-CH_2CH_2CH_3$	8.27
285	3-Cl	1	-O-	$-CH_2CH_2CH_3$	8.12
286	3-CF_3	1	-O-	$-CH_2CH_2CH_3$	7.47
287	3-NO_2	1	-O-	$-CH_2CH_2CH_3$	7.75
288	4-CH_3	1	-O-	$-CH_2CH_2CH_3$	6.84
289	4-CH_2CH_3	1	-O-	$-CH_2CH_2CH_3$	6.42
290	4-$CH_2CH_2CH_3$	1	-O-	$-CH_2CH_2CH_3$	6.36
291	4-OCH_3	1	-O-	$-CH_2CH_2CH_3$	7.34
292	4-OCH_2CH_3	1	-O-	$-CH_2CH_2CH_3$	6.86
293	4-F	1	-O-	$-CH_2CH_2CH_3$	7.58
294	4-Cl	1	-O-	$-CH_2CH_2CH_3$	7.26
295	4-NO_2	1	-O-	$-CH_2CH_2CH_3$	7.08
296	2,3-diCH_3	1	-O-	$-CH_2CH_2CH_3$	7.90
297	2,5-diCH_3	1	-O-	$-CH_2CH_2CH_3$	6.61
298	3,5-diCH_3	1	-O-	$-CH_2CH_2CH_3$	7.98
299	2,3,5-triCH_3	1	-O-	$-CH_2CH_2CH_3$	6.22
300	2,3,6-triCH_3	1	-O-	$-CH_2CH_2CH_3$	5.68
			VIIb		
					
301			-O-	$-CH_2CH_3$	7.30
302			-O-	$-CH_2CH_2CH_3$	7.53
303			$-CH_2-$	$-CH_2CH_3$	6.99
304			$-CH_2-$	$-CH_2CH_2CH_3$	7.02
305			$-CH_2-$	$-CH_2CH=CH_2$	7.14

Table 1 (Continued)

	R	*n*	X	R′	p*I* 50 (M)
			VIIc		
					
306	$-CH_3$			$-CH_2CH_3$	6.58
307	$-CH_3$			$-CH_2CH_2CH_3$	6.28
308	$-CH_3$			$-CH_2CH{=}CH_2$	6.33
309	$-CH_3$			$-CH_2CH \equiv CH$	6.72
310	H			$-CH_2CH_3$	6.76
311	H			$-CH_2CH_2CH_3$	6.59
312	H			$-CH_2CH{=}CH_2$	6.38
			VIId		
					
313				$-CH_2CH_2CH_2CH_2CH_3$	5.46
314				$-CH_2CH_2CH(CH_3)_2$	6.26

[a] If the position of attachment on the pyridine ring is different, it is denoted in the R′ column in parentheses (*x**).

Table 2. Results of Linear Regression Analysis Using All the Descriptors in Each Predictor Set

predictors type used	linear regression model used (q^2)		
	RR	PCR	PLS
AP	0.600	0.094	0.515
DES	0.337	−0.028	0.375
BTH	0.621	−0.021	0.549

statistical software LinMods[44] was used to fit RR, PCR, and PLS models to each class of descriptors. A 10-fold cross validation was performed, and the q^2 values are reported for the linear models.

3. RESULTS AND DISCUSSION

3.1. Initial Data Analysis. The preliminary analysis consisted of modeling the activity using RR, PCR, and PLS, using all descriptors, in LinMods. The application of these methods to chemometric data has been described in depth in previous papers; one notable example is that of Frank and Friedman.[17] The leave-one-out cross-validated squared multiple correlation q^2 of these regressions are given in Table 2. The use of all available descriptors in the AP and BTH predictor sets for prediction illustrated the difficulty of handling regression computations with this large number of predictors; even though the regression code is specifically adapted for data sets with huge numbers of predictors, the computation was intolerably long. Another practical difficulty is that any interpretation of so many predictors (in terms of explaining the chemical mechanism) would have been quite complicated. To eliminate these practical difficulties in using a large number of diverse sets of descriptors, predictor thinning was attempted. Three predictor thinning methods, namely, modified Gram−Schmidt orthogonalization, marginal soft thresholding, and LASSO[45] (least absolute shrinkage and selection operator), were considered for subset selection and are explained in the next section.

3.2. Predictor Thinning. During the past few years, increasing interest has been shown in the problem of modeling collinear, high-dimensional data. One method specifically designed for selecting subsets of predictors from such data is LASSO.[45] It is similar in spirit to ridge regression[8] in the sense that it fits a regression model that is constrained by certain restrictions applied to the coefficient vector. Unlike ridge regression, though, LASSO can shrink some parameters to zero, allowing predictor selection to take place. However, the number of predictors retained in the model cannot exceed the number of observations, which is problematic for many data sets in practice.

A simple approach to predictor thinning avoids this restriction on the number of retained predictors. The marginal soft threshold method simply selects the predictors which are most highly correlated with the response, ignoring any correlations with other predictors. A regression method (such as ridge regression, principal component regression, or partial least squares) applicable to collinear data can then be implemented.

An intuitively appealing revision of this method would be to account for the collinearity among the predictors *during* the selection of the predictors. This idea leads to implementation of an algorithm known as modified Gram−Schmidt orthogonalization.[20] This algorithm is typically used to numerically stabilize the calculation of regression coefficients for ordinary least-squares regression, but we used it for a different purpose here—that of variable selection. The basic idea of this method is that each predictor is added to the model on the basis of its correlation with the response *after* effects of previously included predictors are removed from the remaining predictors. In its original form, this method cannot select more predictors than the number of observations; this difficulty can be overcome by the inclusion of a very small ridge parameter, though this is not considered in this study, for the sake of simplicity. The process begins by selecting the predictor which is most highly correlated with the response, as with the marginal soft threshold method. The other predictors are then revised to be orthogonal to this selected predictor; that is, the remaining predictors are

Table 3. 10-fold q^2 Values for Ridge Regression Fits for Various Subsets of the BTH Data Set

method		number of predictors used				
		250	200	150	100	50
260 Compounds						
(1) Gram−Schmidt, followed by RR	true 10-fold q^2	0.304	0.317	0.319	0.302	0.166
	hold-out R^2	0.642	0.621	0.601	0.567	0.455
(2) marginal soft threshold, followed by RR	true 10-fold q^2	0.525	0.509	0.491	0.495	0.446
	hold-out R^2	0.659	0.646	0.686	0.651	0.535
(3) *LASSO*	true 10-fold q^2	0.492	0.528	0.478	0.388	0.329
	hold-out R^2	0.497	0.658	0.691	0.681	0.620
210 Compounds						
(1) Gram−Schmidt, followed by RR	true 10-fold q^2		0.254	0.276	0.230	0.147
	hold-out R^2		0.398	0.377	0.346	0.213
(2) marginal soft threshold, followed by RR	true 10-fold q^2	0.434	0.424	0.417	0.426	0.406
	hold-out R^2	0.531	0.512	0.512	0.495	0.456
(3) *LASSO*	true 10-fold q^2		0.433	0.478	0.472	0.352
	hold-out R^2		0.169	0.367	0.523	0.406
160 Compounds						
(1) Gram−Schmidt, followed by RR	true 10-fold q^2			0.158	0.139	0.070
	hold-out R^2			0.072	0.056	0.004
(2) marginal soft threshold, followed by RR	10-fold q^2	0.480	0.472	0.469	0.453	0.447
	hold-out R^2	0.489	0.461	0.457	0.433	0.344
(3) *LASSO*	10-fold q^2			0.292	0.423	0.427
	hold-out R^2			0.419	0.406	0.433
110 Compounds						
(1) Gram−Schmidt, followed by RR	true 10-fold q^2				0.088	0.013
	hold-out R^2				0.153	0.078
(2) marginal soft threshold, followed by RR	true 10-fold q^2	0.349	0.349	0.335	0.326	0.275
	hold-out R^2	0.468	0.465	0.461	0.460	0.405
(3) *LASSO*	true 10-fold q^2				0.151	0.218
	hold-out R^2				0.461	0.436
60 compounds						
(1) Gram−Schmidt, followed by RR	true 10-fold q^2					0.011
	hold-out R^2					0.022
(2) marginal soft threshold, followed by RR	true 10-fold q^2	0.338	0.309	0.255	0.242	0.196
	hold-out R^2	0.396	0.413	0.399	0.363	0.285
(3) *LASSO*	true 10-fold q^2					0.235
	hold-out R^2					0.320

reformed to include the residual effects of those predictors, after accounting for the effect of the selected predictor. The one of these remaining predictors that is the best linear estimator of the response is then selected, and this process continues until a desired number of predictors is selected. This set of predictors is then used to fit a regression model (as previously mentioned).

Among the different regression models available to fit the data, we preferred to use ridge regression because it was found to outperform either principal component regression or partial least squares in the initial data analysis using all the available descriptors. The same trend was observed after variable thinning, so only ridge regression models were considered for further regression modeling.

A brief comment regarding the methods discussed here in conjunction with usual ridge regression is deemed necessary. In general, ridge regression is a very powerful modification of usual (ordinary least-squares) multivariate regression. Ridge regression has the important advantages of having easily derived statistical properties (for example, formulaic standard error estimates[8] and influence measures[46]) and of being applicable even in the face of highly collinear predictors; in fact, it is often used in cases where the number of predictors (p) exceeds the number of observations (n). In contrast, the estimated coefficients for an ordinary least-squares regression are highly variable (meaning that only slight changes in the observed predictor values might mean dramatically different regression estimates) under collinear circumstances and not even computable when p exceeds n.

No variable trimming is necessitated before the fitting of a ridge regression model; the equation for calculating the coefficients of the regression inherently incorporates the correlations among the variables. All of the predictors are viewed as working together to explain the response. On the other hand, for obtaining a more tractable model and a more interpretable set of predictors, it would be useful to have a smaller (reduced) set of predictors on which to fit the regression model. This must be balanced against the danger of too extreme a reduction of included predictors, since predictor selection will add bias to the estimated coefficients.[19] Thus, for these data, we consider rather broad thinning to between 50 and 250 predictors. While some bias will be added here, the benefit in interpretation may outweigh this effect.

Since the data set compiled for the present study is relatively large, there is an opportunity to compare these predictor selection methods. The large number of available compounds allows for the utilization of hold-out samples of varying sizes from the BTH data set, along with different choices for the number of predictors included. The data are broken up in five ways: a random sample of 60/110/160/210/260 compounds is selected as calibration data, and the remaining 244/194/144/94/44 compounds are used for validation. For each of these five divisions of data, the three

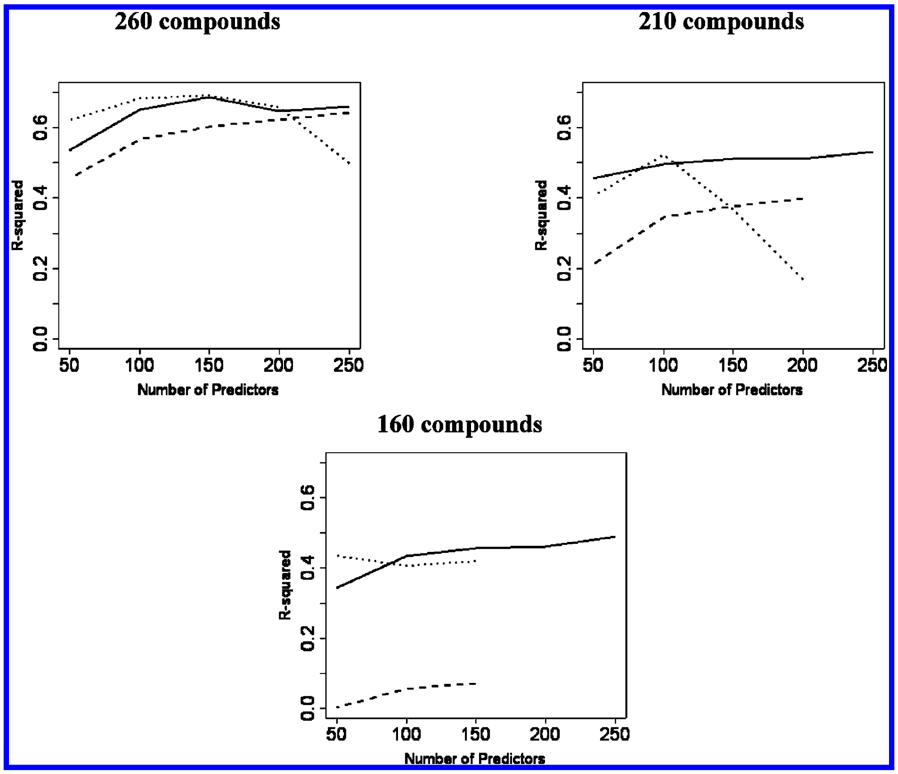


Figure 1. R^2 from the validation sets (solid lines) for the marginal soft threshold method, Gram-Schmidt-based method (dashed lines), and LASSO method (dotted lines) are shown for the subsets of 260, 210, and 160 compounds.

previously described methods of predictor thinning are applied. Subsets of 50, 100, 150, 200, and 250 predictors are chosen by each method (though this is not always feasible, depending on the size of the calibration data), and a ridge regression model is fit to the calibration data. The 10-fold cross-validated q^2 values (see, for example, Shao[47]) for each of these model fits are presented in Table 3, along with the observed R^2 values for prediction of the validation data. The freeware package *R*[48] was used to write a program for calculating 10-fold q^2 values for ridge regression, and the package *PowerMV*[49] was used for calculating 10-fold q^2 values for LASSO. An important comment on this calculation is that care must be taken to incorporate the "leave-out" step for cross validation early enough in the model-fitting process. Otherwise, the supposed ("naïve") q^2 can overestimate[50] the true R^2. The proper application of the cross validation is acknowledged by calling the q^2 value "true".

Almost immediately, there is a concerning observation: the Gram−Schmidt method tends to perform dramatically worse than the soft threshold or LASSO methods, particularly when the number of compounds is low. This behavior, illustrated in Figure 1, is due to a tendency to "overfit" the data[50] because of the *type* of predictors involved. Though the Gram−Schmidt orthogonalization seems intuitively appealing, it is not appropriate here; note that the correlations among the DES set are generally high, while very few of the predictors in the AP set have high correlation, see Figure 2. The extreme sparsity of the atom pairs (see Figure 3) combined with low multicollinearities among them resulted in a model which tried to fit itself to individual observations (even when cross validation was implemented). This can lead to very problematic extrapolations for future predictions.

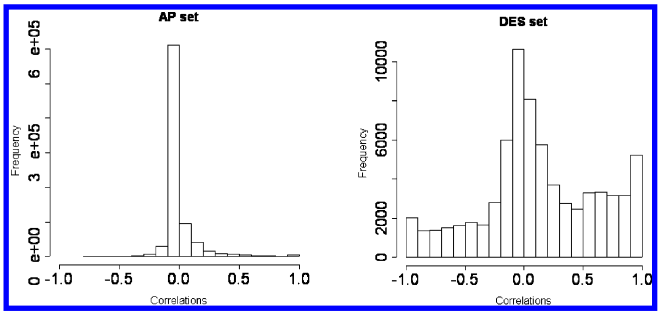


Figure 2. Frequency of correlations for the AP and DES sets. (The DES set contains many more correlated predictors than the AP set).

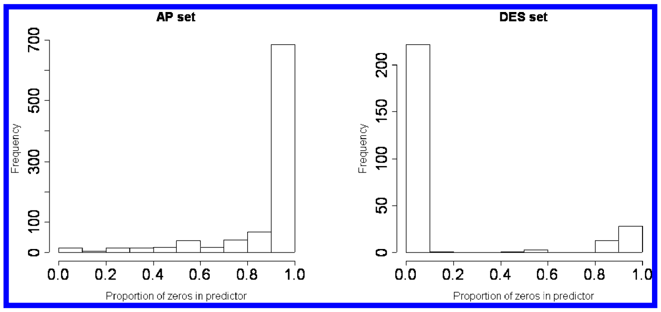


Figure 3. Histograms to compare the sparsity (proportions of zero) of the atom pairs (AP) and global molecular descriptors (DES). [Most predictors in the AP set have a very high proportion of zero observations (i.e., are sparse), while most predictors in the DES set have a very low proportion of zero observations.]

However, such concerns do not apply to the DES data set; the predictors are highly correlated (see Figure 2) but are not sparse (see Figure 3). In fact, Table 4 suggests that Gram−Schmidt selection has more consistent predictive behavior for the DES data set than for the BTH set. We finally decided, though, not to implement any variable trimming here since there are already only 281 predictors in this DES data set; thus, trimming to 200 or 150 predictors did not seem especially beneficial.

Table 4. Comparison of Gram–Schmidt and Marginal Soft Thresholding Methods for Global Molecular Descriptors Set (DES)[a]

subset method		number of predictors 250	200	150	100	50
(1) Gram–Schmidt	true 10-fold q^2	0.420	0.414	0.401	0.373	0.349
	hold-out R^2	0.438	0.400	0.384	0.362	0.239
(2) Marginal soft threshold	true 10-fold q^2	0.387	0.392	0.404	0.387	0.367
	hold-out R^2	0.453	0.450	0.458	0.458	0.405

[a] Hold-out R^2 values are calculated using a random holdout sample of size 100.

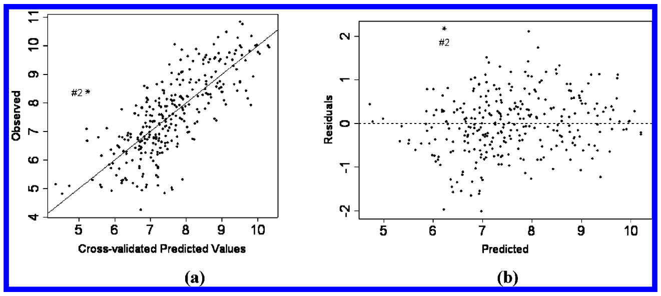


Figure 4. Observed vs cross-validated fit (a) and the residual plot (b).

Thus, we determined that there were two promising avenues for predictive models: (1) using the full DES set of variables or (2) using the soft threshold method on the combined (BTH) set of variables; we could also use the LASSO method, but since the soft threshold method results in models with very similar predictive ability to LASSO, we opted for the simpler soft threshold method. Using option 1, we obtain a 10-fold q^2 of 0.407. Option 2 is used to select a set of 250 predictors; the ridge regression model has 10-fold q^2 of 0.558; this higher value for the predictive ability of this model, along with the fact that it uses predictors from the combined data set, supports it as the preferred model.

3.3. Modeling the JH Activity. After the evaluation of the applicability of descriptor trimming, we decided to fit the JH activity of all 304 compounds using 250 predictors selected from the combined set of predictors by the marginal soft threshold method, and ridge regression was the choice of linear model. However, initial analysis discovered six of these variables whose coefficient estimates in the regression model were highly variable. In addition, these six variables together accomplished essentially the same purpose (differentiating sequence #207 and #208 from the remaining observations), so they were dropped from the model, leaving 244 predictor variables.

The true predictive ability of a regression model is usually visualized by a plot of the observed versus the cross-validated fitted values. The plot for the current study is given in Figure 4a. These values, along with usual residual analysis (see residual plot in Figure 4b), suggest that there is a data point (#2 in sequence) that is overestimated by the model (this point is starred on the plots). Figure 5a and b show the corresponding plots after removal of the observation. As further analysis was unable to detect the reason for the high prediction of the other point, this suggests that either a predictor may have been missed during the selection or there was not a predictor available that is capable of explaining the difference of this observation (it may, in fact, simply be an unusual observation). However, since a few coefficients changed fairly dramatically after excluding this observation, we removed it for the final analysis. Thus, the final fitted model is a ridge regression model fit on a subset of 244 of the original predictors, excluding sequence observation #2, with 10-fold $q^2 = 0.595$.

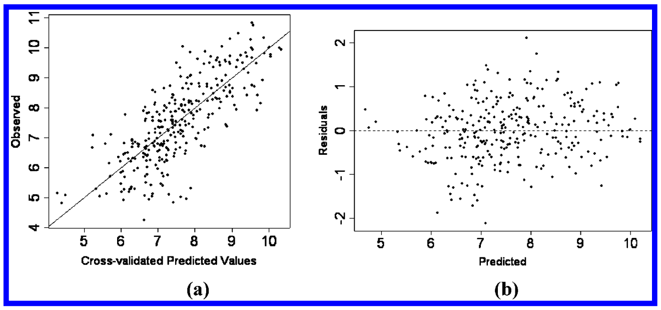


Figure 5. Observed vs cross-validated fit (a) and the residual plot (b) after removal of observation #2.

Table 5. Top 15 Variables with the Highest t-values

variable	*t* value	brief description of the variable
SssO	5.79	electrotopological state of -O-
C0X2−8-C1X2	−5.65	$-CH_2-$(6 intervening vertices)$=CH_2$
$AS1_5$	−4.95	triplet index from adjacency matrix, distance sum, and vertex degree; operation $y = 5$
$AS1_1$	−4.83	triplet index from adjacency matrix, distance sum, and vertex degree; operation $y = 1$
C0X1−2-O0X2	−4.63	H_3C-O-
C0X2−7-C1X2	−4.59	$-CH_2-$(5 intervening vertices)$=CH_2$
C0X1−7-C0X2	4.58	CH_3-(5 intervening vertices)$-CH_2-$
AZV_2	−4.47	triplet index from adjacency matrix, atomic number, and vertex degree; operation $y = 2$
C1X2−8-O0X2	4.42	$CH_2=$(6 intervening vertices)-O-
C1X3−8-N1X2	−4.36	-CH=(6 intervening vertices)=N-
N1X2−3-O0X2	4.13	-N=CH-O-
C1X3−2-C1X3	−4.01	>C=C<
C0X3−13-C0X3	3.95	>CH-(11 intervening vertices)-CH<
C1X3−5-N1X2	−3.90	>C=(3 intervening vertices)=N-
C0X2−6-O0X2	3.74	$-CH_2-$(4 intervening vertices)-O-
P_2	−3.74	path count of length 2

The 16 predictors with largest *t* values are shown in Table 5 along with brief descriptions of each of them. A majority of them are atom pairs (AP set) and a few are triplet indices. The electrotopological state of oxygen (−O−) was found to be the most important factor that affects JH activity. The E-state index of an atom is a measure of the intrinsic state which is perturbed by every other atom in the molecule and, hence, takes into account the valence electronegativity of the atom and its local chemical environment. All the compounds in the data set invariably have the −O− group in the ether, ester, oxime, or carbamate moiety. Picking up the E-state index of oxygen (−O−) reflects the chemical nature (ether, ester, oxime, or carbamate) and the graph distances of each molecule as the most important variable, and this adds support to the robustness of the model and validates the variable selection protocol used in the study. In addition to this, four of the atom pairs that contain oxygen (see Table 5) are indicated as important moieties that affect the bioactivity (JH activities) of the compounds. Though most of the compounds in the data set have two types of oxygen atoms, namely, −O− and >C=O, it is interesting to note that the carbonyl oxygen (>C=O) had not been picked up, whereas the fragments with −O− alone were picked up with high *t* values. The earlier models by Hayashi et al.[26,27] also indicated the alkoxy oxygen as the site of electrostatic interaction from their studies on the electrostatic potential of these compounds. The same authors suggested a molecular dimension of approximately 21 Å for the JH activity of these compounds. The best linear model developed by us in this study seems to have selected this feature in the form of the atom pair C0X3-13-C0X3 (>C−(11 intervening vertexes)−C<). This atom pair appears to correspond to the above-mentioned molecular dimension because all four juvenile hormones (JH 0, JH−I, JH−II, and JH−III) have a C15 carbon backbone which includes an epoxy ring. In addition to these, the moieties (substructures) that affect the JH activity of the compounds could be identified by other atom pairs (please see Table 5). The presence of triplet indices derived from both adjacency and distance matrices in the top descriptors also indicates the importance of shape of the ligand. Another interesting observation is that none of the semiempirical quantum chemical parameters, such as E_{HOMO}, E_{LUMO}, or $E_{\mathrm{HOMO}} - E_{\mathrm{LUMO}}$, were selected in the final model.

In the earlier attempts of SAR modeling of JH activity, individual models were built for each class of structurally related compounds using a small set of descriptors. In contrast to this, we modeled diverse sets of structures together using a diverse set of computable descriptors, deleting only 2 out of 304 observations as outliers, and still, all the important features responsible for JH activity were picked up in the final model, and more importantly, this is achieved after descriptor thinning, and this bears testimony to our philosophy of QSAR, “A diverse set of compounds can be modeled with a diverse set of descriptors”. This phenomenon was observed in several of our earlier studies on partition coffiecients,[40,51,52] vapor pressure,[53,54] mutagenicity,[55] and the boiling point[56] of organic compounds. In the QSAR modeling of blood/air[51,52] and tissue/air partition coefficients[40] of structurally and physicochemically diverse (hydrophobic and hydrophilic) compounds, we showed that (a) a collection of diverse descriptors is capable of giving good quality QSARs for progressively heterogeneous sets of halogenated compounds and (b) a set of diverse descriptors yielded good quality models for hydrophobic and hydrophilic compounds as well as the combined set of hydrophobic and hydrophilic chemicals, where physicochemically based algorithms failed in the case of the hydrophilic subset. The same trend was observed[55] for a large, diverse set of 508 mutagens/nonmutagens, where the diverse set of calculated descriptors gave models of similar quality for the entire data set as well as different structural classes derived from the heterogeneous set.

4. CONCLUSIONS

While the use of the Gram−Schmidt algorithm is intuitively appealing as a method to account for multicollinearity in the predictor selection stage of the model fit, it can result in overfitting of the model to the data. In particular, very sparse data appear to receive no benefit from this procedure. The Gram−Schmidt algorithm can be used for selection from the global molecular descriptors (DES), though these predictors do not provide as close of a fit as those chosen from the full data set. Thus, we find that soft thresholding is the preferred method for subset selection; we obtain a reasonably strong predictive model for a very diverse set of data, with a more interpretable subset of predictors than the full ridge regression model. The linear ridge regression model developed for the JH activity of 304 compounds was interpreted on the basis of the top 16 parameters selected from the *t* values. This indicated that the presence of specific moieties containing nitrogen and oxygen atoms is important in addition to the shape and size of the ligands. Almost all of the features that were suggested as important for JH activity by earlier studies were picked up in the top parameters. We achieved this by modeling a diverse set of data with a diverse set of computable molecular descriptors.

ACKNOWLEDGMENT

This is contribution number 388 from the Center for Water and the Environment of the Natural Resources Research Institute, University of Minnesota−Duluth. The research reported in this paper is supported by the United States Department of Agriculture, Agricultural Research Service, Specific Cooperative Agreement 58-1275-4-313.

REFERENCES AND NOTES

(1) Karcher, W.; Devillers, J. *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Kluwer: The Netherlands, 1997.

(2) Gute, B. D.; Basak, S. C. Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131.

(3) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1999.

(4) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: Weinheim, Germany, 2000; Methods and Principles in Medicinal Chemistry, Vol. 11.

(5) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(6) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 64−73.

(7) Stanton, D. T. On the Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1423−1433.

(8) Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55−67.

(9) Hoerl, A. E.; Kennard, R. W. Ridge regression: Applications to nonorthogonal problems. *Technometrics* **1970**, *12*, 69−82.

(10) Wold, S. Discussion: PLS in chemical practice. *Technometrics* **1993**, *35*, 136−139.

(11) Massy, W. F. Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.* **1965**, *60*, 234−246.

(12) Héberger, K.; Rajkó, R. Generalization of pair-correlation method (PCM) for nonparametric variable selection. *J. Chemom*. **2002**, *16*, 436−443.

(13) Héberger, K.; Andrade, J. M. Procrustes rotation and pairwise correlation: a parametric and a nonparametric method for variable selection. *Croat. Chem. Acta* **2004**, *77*, 117−125.

(14) Héberger, K.; Rajkó, R. Variable selection using pair-correlation method. Environmental applications. *SAR QSAR Environ. Res*. **2002**, *13*, 541−554.

(15) Draper, N. R.; Smith, H. In *Applied Regression Analysis*, 2nd ed.; John Wiley & Sons Inc.: New York, 1981; pp 294−379.

(16) Miller, A. J. *Subset Selection in Regression*; Chapman and Hall: London, 1990; pp 43−82.

(17) Frank, I. E.; Friedman, J. H. A Statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109−135.

(18) Wold, S. Discussion: PLS in chemical practice. *Technometrics* **1993**, *35*, 136−139.

(19) Rencher, A.; Punn, F. Inflation of R^2 in best subset regression. *Technometrics* **1980**, *22*, 49−53.

(20) Thisted, R. A. *Elements of Statistical Computing*; Chapman and Hall: New York, 1988.

(21) Niwa, A.; Iwamura, H.; Nakagawa, Y.; Fujita, T. Development of (phenoxyphenoxy)- and (benzylphenoxy)alkanaldoxime *O*-ethers as

potent insect juvenile hormone mimics and their quantitative structure−activity relationsips. *J. Agric. Food Chem*. **1989**, *37*, 378−384.

(22) Niwa, A.; Iwamura, H.; Nakagawa, Y.; Fujita, T. Development of (phenoxyphenoxy)- and (benxylphenoxy)propyl ethers as potent insect juvenile hormone mimetics. *J. Agric. Food Chem*. **1989**, *37*, 462−467.

(23) Niwa, A.; Iwamura, H.; Nakagawa, Y.; Fujita, T. Development of (4-alkoxyphenoxy)- and (4-alkoxyphenoxy)alkanaldoxime *O*-ethers as potent insect juvenile hormone mimics and their structure−activity relationships. *J. Agric. Food Chem*. **1989**, *37*, 467−472.

(24) Niwa, A.; Iwamura, H.; Nakagawa, Y.; Fujita, T. Development of N,O-disubstituted hydroxylamines and *N*,*N*-disubstituted amines as insect juvenile hormone mimetics and the role of the nitrogenous function for activity. *J. Agric. Food Chem*. **1990**, *38*, 514−520.

(25) Hayashi, T.; Iwamura, H.; Fujita, T. Development of 4-alkylphenyl aralkyl ethers and related compounds as potent insect juvenile hormone mimetics and structural aspects of their activity. *J. Agric. Food Chem.* **1990**, *38*, 1965−971.

(26) Hayashi, T.; Iwamura, H.; Fujita, T. Insect juvenile hormone mimetic activity of (4-substituted)phenoxyalkyl compounds with various nitrogenous and oxygenous functions and its relationship to their electrostatic and stereochemical properties. *J. Agric. Food Chem*. **1991**, *39*, 2029−2038.

(27) Hayashi, T.; Iwamura, H.; Fujita, T.; Takakusa, N.; Yamada, T. Structural Requirements for Activity of Juvenile Hormone Mimetic Compounds against Various Insects. *J. Agric. Food Chem*. **1991**, *39*, 2039−2045.

(28) Hansch, C. In *Correlation Analysis in Chemistry*; Chapman, N. B., Shorter, J., Eds.; Plenum Press: New York, 1978; pp 397−438.

(29) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(30) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(31) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, U. K., 1986.

(32) Bonchev, D. and Trinajstic, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.

(33) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the Second International Conference on Mathematical Modelling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri−Rolla: Rolla, MO, 1979; pp 851−856.

(34) Balaban, A. T. Highly discriminating distance-based topological indices. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(35) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY*, version 2.3; University of Minnesota: Minneapolis, Minnesota, 1988.

(36) Filip, P. A.; Balaban, T. S.; Balaban, A. T. A new approach for devising local graph invariants: Derived topological indices with low degeneracy and good correlational ability. *J. Math. Chem*. **1987**, *1*, 61−83.

(37) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.

(38) Kier, L. B.; Hall, L. H. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1999; pp 455−489.

(39) *Molconn-Z*, version 3.5; Hall Associates Consulting: Quincy, MA, 2000.

(40) Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. Prediction of tissue: air partition coefficients: A comparison of structure-based and property-based methods. *SAR QSAR Environ. Res.* **2002**, *13*, 649−665.

(41) *Chem3D Ultra*, version 8; CambridgeSoft Corporation: Cambridge, MA, 200X.

(42) *APProbe*; University of Minnesota: Minneapolis, Minnesota, 1993.

(43) In *SAS/STAT User Guide*, release 6.03: SAS Institute, Inc.: Cary, NC, 1988.

(44) Hawkins, D. M. *LinMods Program*; School of Statistics, University of Minnesota: Minneapolis, MN.

(45) Tibshirani, R. Regression shrinkage and selection via the *LASSO*. *J. R. Stat. Soc. B* **1996**, *58*, 267−288.

(46) Walker, E.; Birch, J. B. Influence Measures in Ridge Regression. *Technometrics* **1988**, *30*, 221−227.

(47) Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486-494.

(48) Gentleman, R.; Ihaka, R. *Comprehensive R Archive Network*, version 2.0.1; http://www.r-project.org.

(49) Liu, J.; Feng, J.; Young, S. *PowerMV*, version 0.61; National Institute of Statistical Sciences. http://www.niss.org/PowerMV (Feb. 3, 2005).

(50) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

(51) Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. Prediction of human blood:air partition coefficient: A comparison of structure-based and property-based methods. *Risk Anal.* **2003**, *23*, 1173−1184.

(52) Basak, S. C.; Mills, D.; El-Masri, H. A.; Mumtaz, M. M.; Hawkins, D. M. Predicting blood:air partition coefficients using theoretical molecular descriptors. *Environ. Toxicol. Pharmacol.* **2004**, *16*, 45−55.

(53) Basak, S. C.; Mills, D. Quantitative structure−property relationships (QSPRs) for the estimation of vapor pressure: A hierarchical approach using mathematical structural descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 692−701.

(54) Basak, S. C.; Mills, D. Development of quantitative structure−activity relationship models for vapor pressure estimation using computed molecular descriptors. *ARKIVOC* **2005**, submitted for publication.

(55) Basak, S. C.; Mills, D.; Gute, B. D.; Hawkins, D. M. Predicting mutagenicity of congeneric and diverse sets of chemicals using computed molecular descriptors: A hierarchical approach. In *Quantitative Structure−Activity Relationship (QSAR) Models of Mutagens and Carcinogens*; Benigni, R., Ed.; CRC Press: Boca Raton, FL, 2003; pp 207−234.

(56) Basak, S. C.; Mills, D. Use of mathematical structural invariants in the development of QSPR models. *MATCH* **2001**, *44*, 15−30.

CI050215Y