QSAR and Classification of Murine and Human Soluble Epoxide Hydrolase Inhibition by Urea-Like Compounds

Nathan R. McElroy and Peter C. Jurs*

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

Christophe Morisseau and Bruce D. Hammock

Department of Entomology and U. C. D. Cancer Center, University of California, Davis, California 95616

Received June 21, 2002

A data set of 348 urea-like compounds that inhibit the soluble epoxide hydrolase enzyme in mice and humans is examined. Compounds having IC_{50} values ranging from 0.06 to >500 μM (murine) and 0.10 to >500 μM (human) are categorized as active or inactive for classification, while quantitation is performed on smaller compound subsets ranging from 0.07 to 431 μM (murine) and 0.11 to 490 μM (human). Each compound is represented by calculated structural descriptors that encode topological, geometrical, electronic, and polar surface features. Multiple linear regression (MLR) and computational neural networks (CNNs) are employed for quantitative models. Three classification algorithms, *k*-nearest neighbor (*k*NN), linear discriminant analysis (LDA), and radial basis function neural networks (RBFNN), are used to categorize compounds as active or inactive based on selected data split points. Quantitative modeling of human enzyme inhibition results in a nonlinear, five-descriptor model with root-mean-square errors (log units of IC_{50} [μM]) of 0.616 ($r^2 = 0.66$), 0.674 ($r^2 = 0.61$), and 0.914 ($r^2 = 0.33$) for training, cross-validation, and prediction sets, respectively. The best classification results for human and murine enzyme inhibition are found using *k*NN. Human classification rates using a seven-descriptor model for training and prediction sets are 89.1% and 91.4%, respectively. Murine classification rates using a five-descriptor model for training and prediction sets are 91.5% and 88.6%, respectively.

Introduction

Soluble epoxide hydrolases (sEH) are ubiquitous enzymes that catalyze the hydrolysis of epoxides to their corresponding 1,2-diols by the addition of water.[1,2] Found in many plant and animal species, these enzymes play an important role in the conversion of lipophilic compounds to more hydrophilic and reactive metabolites in biological systems. In some cases, the hydrolysis of epoxides to diols serves as a detoxification step by creating more polar, excretable products,[3] or compounds necessary for biological processes.[1] In other cases, the resulting diols or their metabolites can be more harmful to a system than the parent epoxides. Epoxy fatty acids, such as leukotoxin and isoleukotoxin, are hydrolyzed to more toxic diols[4] and have been associated with symptoms of multiple organ failure in burn victims[5] and acute respiratory distress syndrome.[6] Epoxyeicosatrienoic acids (EETs) are metabolites of arachidonic acids that undergo hydrolysis by sEH to form dihydroxyeicosatrienoic acids (DHETs).[7−9] EETs help regulate blood pressure, but an increase in DHETs, particularly 14,-15-DHET, a product of sEH hydrolysis, has been associated with induced hypertension in pregnant women.[8] Research has shown that inhibition of sEH can reduce blood pressure in mice[7] and rats.[8] The search for good inhibitors of sEH may lead to promising pharmaceutical targets for the treatment of hypertension and other physiological symptoms in the human body.

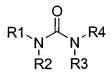
The crystal structure of the murine sEH enzyme and its interaction with alkylurea inhibitors has been investigated.[3,10] The active site resides in the corner of an L-shaped tunnel surrounded by hydrophobic channels that are open to solvents.[10] This allows an inhibitor to approach from either side. An alkylurea carbonyl oxygen can accept a hydrogen bond from a hydroxyl group on an active site tyrosine. Likewise, hydrogen bond interactions can occur between an alkylurea amine group and an aspartate carboxyl group at the active site. Both interactions mimic those encountered in the ring-opening of an epoxide at the active site,[10] and previous studies have investigated the roles of aspartates and tyrosines in this mechanism.[3,11,12]

Several types of compounds have been investigated as possible inhibitors of epoxide hydrolases such as chalcone oxides,[13,14] heavy metals,[15] and alkylurea compounds,[16,17] but the focus of this study is the use of alkylurea compounds. The backbone structure for the majority of compounds used in this study (288 out of 348) and some representative compounds are shown in Table 1. The inhibition of sEH is primarily dependent upon hydrogen-bond acceptance of the inhibitor and hydrogen-bond donation of the active site. The degree to which this can occur when considering the structure is dependent upon the substituent groups attached to the inhibitor nitrogens (R_1−R_4). Because of the shape and size of the active site tunnel as well as its hydrophobicity, varying substituent groups will affect how the urea moiety of the molecule sits within the tunnel, thereby affecting inhibition.

* To whom correspondence should be addressed. Phone: (814) 865−3739. Fax: (814) 865−3314. E-mail: pcj@psu.edu.

10.1021/jm020269o CCC: $25.00
Published on Web 02/20/2003

Table 1. Representative Compounds of the Prominent Structural Backbone



Compound	R_1	R_2	R_3	R_4
20	$(CH_2)_5CH_3$	H	H	CH_3
87	C_6H_{11}	H	H	$C(CH_3)_2CH_2C(CH_3)_3$
91	C_6H_{11}	H	H	$CH_2C(O)OC(CH_3)_3$
160	C_6H_{11}	H	H	C_6H_4-4-Cl
172	C_6H_{11}	H	H	C_6H_4-4-OH
189	C_6H_5	H	CH_3	C_6H_5
220	C_6H_3-3,4-Cl_2	H	H	C_6H_3-3,4-Cl
225	C_6H_3-3,4-Cl_2	$(CH_2)_5COOH$	CH_3	CH_3
328	$(CH_2)_{11}CH_3$	H	H	$CH_2C(CH_3)_2C(CH_3)_3$
329	$(CH_2)_{11}CH_3$	H	H	C_5H_9

The goal of this study is to create robust QSAR models and binary classification models that predict and categorize inhibition values of alkylurea compounds toward murine and human sEH. The use of quantitative and classification models can augment and narrow the search for future drug compounds, and the methodologies discussed below have been successfully applied in modeling several physical and biological properties such as aqueous solubility,[18,19] glass transition temperatures,[20] multidrug-resistance reversal activity values,[21] and enzyme inhibition.[22−24]

Experimental Section

Apparatus. Melting points were determined with a Thomas-Hoover apparatus (A. H. Thomas Co., Philadelphia, PA) and are uncorrected. Infrared (IR) spectra were recorded on a Mattson Galaxy Series FTIR 3000 spectrometer (Madison, WI). Mass spectra were measured by LC-MS: Waters 2790 liquid chromatograph (Milford, MA) equipped with a 30 × 2.1 mm 3 μm C18 Xterra column (Waters) and a Micromass Quattro Ultima (Manchester, UK) mass spectrometer. ^{1}H and ^{13}C NMR were acquired on a QE−300 (General Electric).

Synthesis. Of the 348 compounds used in this study, 55 were obtained from Aldrich Chemical Co., Inc. (Milwaukee, WI), 21 were obtained from Chem Service (West Chester, PA), and 4 from Lancaster (Windham, NH) (see Table 3 for details). A total of 98 compounds used in this study have been described in previous publications.[16,17,25−28] Other compounds were synthesized by the condensation of the appropriate isocyanate and amine following described methodology.[16] Reaction products were purified by recrystallization. In addition to sharp melting points and single spot on silica gel thin-layer chromatography (TLC), the products were characterized using ^{1}H NMR (General Electric QE−300), infrared, and electrospray mass spectrometry. As examples, synthesis of compounds **160**, **220**, and **329** are described below.

***N*-Cyclohexyl-*N*′-4-chlorophenylurea (160).** To a stirred warm solution of 0.638 g (5.0 mmol) of 4-chloroaniline in 40 mL of hexane was added 0.71 g (5.7 mmol) of cyclohexyl isocyanate dissolved in 5 mL of hexane. After stirring at room-temperature overnight, a white solid was obtained, which was recrystallized twice from hexane. The resulting white crystal (1.06 g; yield: 84%) had a melting point of 223.0−224.0 °C. IR (KBr) 3342 (s, NH), 3282 (m, NH), 1629 (s, C═O), and 1568 (s, amide II) cm^{-1}. ^{1}H NMR (DMSO-d_6/TMS): δ 8.41 (s, 1H, N′H), 7.40 (dt, J = 8.9 Hz, 2.0 Hz, 2H, C-2′,6′), 7.24 (dt, J = 8.9 Hz, 2.0 Hz, 2H, C-3′,5′), 5.32 (d, J = 8.1 Hz, 1H, NH), 3.4 (m, 1H, cyclohexyl), 1.8 (m, 2H, cyclohexyl), 1.7 (m, 2H, cyclohexyl), 1.5 (m, 1H, cyclohexyl), 1.3 (m, 2H, cyclohexyl), 1.2 (m, 3H, cyclohexyl) ppm; ^{13}C NMR (DMSO-d_6): δ 154.3 (C═O), 139.6 (C-1′), 128.4 (C-3′,5′), 124.4 (C-4′), 119.1 (C-2′,6′), 47. (C-1), 33.0 (C-2,6), 25.3 (C-4), 24.4 (C-3,5) ppm. LC-MS *m*/*z* (relative intensity): 505.1 (19, $[2M + H]^+$), 253.0 (100, $[M + H]^+$).

***N*,*N*′-Bis(3,4-dichlorophenyl)urea (220).** To a stirred solution of 0.53 g (3.3 mmol) of 3,4-dichloroaniline in 15 mL of chloroform was added 0.56 g (3.0 mmol) of 3,4-dichlorophenyl isocyanate dissolved in 5 mL of chloroform. After stirring at room temperature for 1 h, a white solid was obtained, which was recrystallized twice from hexane. The resulting white crystal (0.96 g; yield: 91%) had a decomposition point of 270.0−271.0 °C. IR (KBr) 3293 (m, NH), 3274 (m, NH), 1624 (s, C═O), and 1567 (s, amide II) cm^{-1}. ^{1}H NMR ($CDCl_3$/TMS): δ 9.54 (s, 2H, NH), 8.26 (s, 2H, C-2,2′), 7.93 (d, J = 9.0 Hz, 2H, C-5,5′), 7.74 (d, J = 8.9 Hz, 2H, C-6,6′) ppm. LC-MS *m*/*z* (relative intensity): 701.5 (25, $[2M + H]^+$), 351.3 (100, $[M + H]^+$).

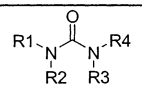
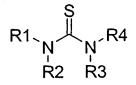
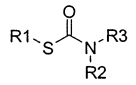
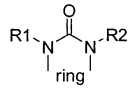
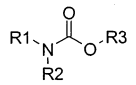
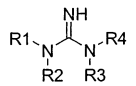
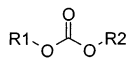
***N*-Cyclopentyl-*N*′-dodecylurea (329).** To a stirred cold solution of 0.26 g (3.0 mmol) of cyclopentylamine in 30 mL of hexane was added 0.42 g (2.0 mmol) of dodecyl isocyanate dissolved in 5 mL of hexane. After stirring at room temperature for 1 h, a white solid was obtained, which was recrystallized twice from hexane. The resulting white crystal (0.55 g; yield: 93%) had a melting point of 95.0−96.0 °C. IR (KBr) 3346 (m, NH), 3329 (m, NH), 1671 (s, C═O), and 1570 (s, amide II) cm^{-1}. ^{1}H NMR (DMSO-d_6/TMS): δ 5.44 (t, J = 7.4 Hz, 1H, N′H), 4.87 (d, J = 8.1 Hz, 1H, NH), 3.5 (m, 1H, cyclopentyl), 3.09 (q, J = 6.9 Hz, 2H, CH_2, C-1′), 1.9 (m, 2H, cyclopentyl), 1.7 (m, 2H, cyclopentyl), 1.6 (m, 1H, cyclopentyl), 1.4 (m, 2H, CH_2, C-2′), 1.2 (bm, 19H), 1.1 (m, 2H, C-11′), 0.85 (t, J = 6.5 Hz, 3 H, CH_3) ppm; ^{13}C NMR (DMSO-d_6): δ 157.8 (C═O), 48.6 (C-1), 40.3 (C-1′), 34.2 (C-2,5), 31.9 (C-2′), 30.5 (C-3′), 29.6 (C-4′,5′,6′,7′), 29.5 (C-8′), 29.3 (C-9′), 27.1 (C-10′), 25.2 (C-3,4), 22.6 (C-11′), 14.1 (C-12′) ppm. LC-MS *m*/*z* (relative intensity): 593.5 (21, $[2M + H]^+$), 297.3 (100, $[M + H]^+$).

Enzyme Preparation. Recombinant mouse sEH (MsEH) and human sEH (HsEH) were produced in a baculovirus expression system[29,30] and purified by affinity chromatography.[31] The preparations were at least 97% pure as judged by sodium dodecyl sulfate−polyacrylamide gel electrophoresis and scanning densitometry. No detectable esterase or glutathione transferase activities were observed. Esterases, as well as glutathione transferases, interfere with the high throughput screening assay used to obtain rank order of the compounds.[32] Protein concentration was quantified using the Pierce BCA assay (Pierce, Rockford, IL) using bovine serum albumin (BSA) as the calibrating standard.

IC_{50} Assay Conditions. IC_{50}s were determined as previously described[16] using racemic 4-nitrophenyl-*trans*-2,3-epoxy-3-phenylpropyl carbonate as substrate.[32] Enzymes (0.10 μM MsEH or 0.20 μM HsEH) were incubated with inhibitors for 5 min in pH 7.4 sodium phosphate buffer at 30 °C prior to substrate introduction ([S] = 40 μM). Activity was assessed by measuring the appearance of the 4-nitrophenolate anion at 405 nm at 30 °C during 1 min (Spectramax 200; Molecular Devices, Inc., Sunnyvale, CA). Assays were performed in triplicate. By definition, IC_{50} are concentrations of inhibitor that reduce enzyme activity by 50%. IC_{50} were determined by regression of at least five datum points with a minimum of two points in the linear region of the curve on either side of the IC_{50}. The curve was generated from at least three separate runs, each in triplicate.

Data Sets. This study employed four data sets, comprised of subsets of the 348 urea-like compounds, each having an IC_{50} inhibition value (μM) for human sEH and/or murine sEH. Common structural scaffolds and their frequencies in the data are shown in Table 2. The observed IC_{50} values (in log units of μM) for human and murine inhibition are shown for each compound in Table 3. Structural information and observed IC_{50} (μM) values with associated errors are included in the Supporting Information. Of these compounds, 288 (83%) have the base structure shown in Table 1 with varying substituent

Table 2. Generic Scaffolds of the 348 Compounds

Scaffold	Frequency
	288
	12
	10
	8
	7
	5
	3
	2
miscellaneous	13

groups (R_1-R_4). The remaining compounds have a similar structural basis with substitutions (i.e., replacing the oxygen with sulfur, replacing one of the nitrogens with sulfur, etc.) as illustrated in Table 2. Oxygen atoms are found in 326 compounds, nitrogen in 345, sulfur in 29, chlorine in 50, and ring structures in 266 compounds. A total of four data sets were created to accommodate quantitative and classification modeling for human and murine data points.

Data set 1 contains all 207 compounds for which a quantitative IC_{50} value was available for human sEH. Compounds for which the lower limit of detection was reported were excluded, as were compounds with IC_{50} values reported as inequalities. Most of the compounds with IC_{50} results that are reported as inequalities are non-urea-like compounds, which holds true for data set 2 excluded compounds. The 207 compounds have a molecular weight range of 144 to 404 amu (mean = 250 amu). Inhibition values range from 0.11 to 490 μM (−0.96 to 2.69 log units), with a mean inhibition value of 45.46 μM (0.69 log units). Relative experimental errors range from 0.8% to 33.3% of observed values (mean = 6.8%). The compounds were split into a training set (TSET), cross-validation set (CVSET), and external prediction set (PSET) with 167, 19, and 21 members, respectively.

Data set 2 contains all 186 compounds for which a quantitative IC_{50} value was available for murine sEH. Compounds for which the lower limit of detection was reported were excluded, as were compounds with IC_{50} values reported as inequalities. The 186 compounds have a molecular weight range of 101 to 404 amu (mean = 246 amu). Inhibition values range from 0.07 to 431 μM (−1.16 to 2.63 log units), with a mean inhibition value of 50.05 μM (0.66 log units). Relative experimental errors range from 0.5% to 29.6% of observed values (mean = 7.6%). The 186 compounds were distributed among a TSET, CVSET, and PSET of 150, 17, and 19 members, respectively. For both data sets 1 and 2, compounds were placed pseudo-randomly into the three subsets with the stipulation that PSET values adequately represented the entire range of IC_{50} values without being extreme values on either end of that range.

Data set 3 contains all 339 compounds for which an IC_{50} value was available for human sEH. The 339 compounds have a molecular weight range of 60 to 527 amu (mean = 246 amu). A histogram of inhibition values was created to determine a split point between active and inactive compounds based on compound distribution (see Figure S1 in the Supporting Information). A cutoff of 100 μM (2.00 log units) was chosen as the best compromise between (1) having adequate numbers of compounds in each class, and (2) finding a split point that would support the highest quality models because many of the compounds were poorly soluble above this cutoff concentration. A total of 222 compounds with IC_{50} values of <100 μM are considered active, and 117 compounds with $IC_{50} \geq 100$ μM are considered inactive. The compounds were split into a TSET, CVSET, and PSET of 269, 35, and 35 compounds, respectively.

Data set 4 contains all 339 compounds for which an IC_{50} value was available for murine sEH. The 339 compounds have a molecular weight range of 60 to 527 amu (mean = 246 amu). A histogram was created for these data, and the best split point was determined to be 195 μM (2.29 log units) based on the two criteria above (see Figure S2 in the Supporting Information). A total of 235 compounds with IC_{50} <195 μM are considered active, and 104 compounds with $IC_{50} \geq 195$ μM are considered inactive. These compounds were placed into a TSET, CVSET, and PSET of 269, 35, and 35 members, respectively. For both data sets 3 and 4, compounds were placed randomly into the three subsets with the stipulation that the active-inactive global ratios were maintained in the subsets.

All quantitative modeling and classification routines were performed on a DEC 3000 AXP Model 500 workstation running the UNIX operating system. The Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software package[33,34] was used for descriptor generation and linear model building. In-house simulated annealing,[35] genetic algorithm,[36] CNN,[37] and classification routines were used to develop linear and nonlinear quantitative and classification models. The following steps describe the methodology used for building quantitative and classification models.

Structure Entry and Modeling. All compounds were sketched on a Pentium-III PC using HyperChem (Hypercube, Inc. Waterloo, ON, Canada), which stored two-dimensional connectivity information for use in ADAPT. All nontopological descriptors required accurate three-dimensional structure information, therefore the two-dimensional structures were passed to the semiempirical molecular orbital package MOPAC[38] for optimization. The PM3 Hamiltonian[39] was used to find low-energy three-dimensional geometries while the AM1 Hamiltonian[40] was used for charge information. Previous work has supported this approach of using different Hamiltonians to extract appropriate structural information.[41]

Descriptor Generation. To construct predictive or classification models, it was first necessary to encode meaningful information about the structural environment of each compound. A total of 250 descriptors were calculated for each compound using ADAPT: 150 topological, 30 geometric, 10 electronic, and 60 hybrid descriptors. Topological descriptors required only a simple 2-D sketch of the molecule and encoded information about atom types, bond types, and connectivity. Examples included path counts,[42,43] molecular connectivity,[44−46] distance edge descriptors,[47] and fragment counts. Geometric descriptors provided information about molecular size and shape; therefore, they were calculated using three-dimensional coordinates of the structures. Some examples include molecular surface area and volume[48] and moments of inertia.[49] Electronic descriptors provided information about the electronic environment of the compound such as partial atomic charges, highest occupied and lowest unoccupied molecular orbital energies, and electronegativity. Hybrid or polar surface descriptors revealed information about partially charged surface areas[50] and hydrogen bonding characteristics of the molecule.[51] Specific details on descriptors that were chosen in models are given below.

In addition to the ADAPT descriptors mentioned above, DRAGON[52] software was used to calculate approximately 230 descriptors from four categories: BCUTs (Burden CAS, University of Texas);[53] molecular walk counts;[54] constitutional descriptors;[51] and Weighted Holistic Invariant Molecular (WHIM) descriptors.[55,56]

Table 3. Observed and Predicted log IC_{50} (μM) Values and Predicted Class Labels for Soluble Epoxide Hydrolase Inhibitors

no.	set[a]	synthesis[b]	human log IC_{50} (μM) obsd[c]	human log IC_{50} (μM) calcd[d]	human cls. calcd[e]	murine log IC_{50} (μM) obsd[c]	murine cls. calcd[e]
1	3,4	Aldrich	N/A	x	−	N/A	−
2	3,4	Aldrich	N/A	x	−	N/A	−
3	3,4	Aldrich	N/A	x	−	N/A	−
4	3,4	Aldrich	N/A	x	−	N/A	−
5	3,4	Aldrich	N/A	x	−	N/A	−
6	3,4	Aldrich	N/A	x	−	N/A	−
7	3,4	ref 28	N/A	x	−	N/A	−
8	3,4	Aldrich	N/A	x	−	N/A	−
9	3,4	Aldrich	N/A	x	−	N/A	−
10	1,2,3,4	new	1.78	1.67	+	1.50	−*
11	3,4	new	N/A	x	+*	N/A	−
12	3,4	new	N/A	x	+*	N/A	−
13	1,2,3,4	new	0.04	0.31	+	0.41	+
14	1,2,3,4	new	0.00	0.18	+	−0.10	+
15	3,4	new	N/A	x	−	N/A	−
16	2,3,4	new	N/A	x	−	2.52	−
17	3,4	new	N/A	x	−	N/A	−
18	1,2,3,4	new	−0.96	−0.29	+	−0.74	+
19	1,2,3,4	ref 17	1.03	1.36	+	0.85	+
20	2,3,4	new	N/A	x	+*	2.35	+*
21	1,2,3,4	new	1.62	1.98	+	1.13	+
22	1,2,3,4	new	0.63	1.16	+	0.23	+
23	1,2,3,4	new	0.49	0.81	+	0.05	+
24	1,2,3,4	new	0.16	0.67	+	−0.07	+
25	1,2,3,4	new	−0.24	−0.36	+	0.00	+
26	1,2,3,4	new	−0.60	0.03	+	−0.51	+
27	1,2,3,4	ref 17	0.74	1.18	+	0.02	+
28	1,2,3,4	ref 17	2.65	1.44	+*	2.13	+
29	1,2,3,4	ref 17	1.52	1.23	+	1.11	+
30	1,2,3,4	ref 17	−0.17	0.88	+	−0.14	+
31	1,2,3,4	ref 17	0.78	1.27	+	0.08	+
32	1,2,3,4	ref 17	−0.14	0.33	+	−1.05	+
33	1,2,3,4	ref 17	0.77	0.87	+	0.03	+
34	1,3,6	ref 17	−0.62	0.11	+	N/A	+x
35	1,2,3,4	ref 17	0.23	0.29	+	−0.28	+
36	1,2,3,4	ref 17	1.53	1.19	+	0.98	+
37	1,3,4	ref 17	−0.92	−0.03	+	−1.22	+
38	1,2,3,4	ref 17	−0.24	0.06	+	0.36	+
39	1,2,3,4	ref 17	−0.74	−0.34	+	−0.92	+
40	1,2,3,4	ref 28	0.23	0.65	+	−0.42	+
41	3,4	ref 17	N/A	x	−	N/A	+*
42	2,3,4	ref 17	N/A	x	+*	1.89	+
43	1,2,3,4	ref 17	1.68	0.63	+	0.76	+
44	1,2,3,4	ref 28	0.59	0.36	+	0.20	+
45	3,4	new	−1.00	x	+	−1.22	+
46	1,2,3,4	new	−0.21	−0.18	+	−0.68	+
47	1,2,3,4	new	−0.52	−0.55	+	−0.92	+
48	5,6	ref 28	N/A	x	+x	N/A	+x
49	1,2,3,4	new	−0.13	0.00	−*	−0.70	+
50	3,4	ref 17	N/A	x	−	N/A	−
51	3,4	ref 28	N/A	x	−	N/A	+*
52	3,4	new	N/A	x	−	N/A	−
53	2,3,4	new	N/A	x	−	2.30	−
54	1,3,4	ref 28	−0.59	0.04	+	−1.30	+
55	1,2,3,4	ref 28	1.18	−0.17	+	−0.30	+
56	3,4	new	N/A	x	−	N/A	+*
57	1,2,3,4	ref 28	−0.19	−0.10	−*	−0.85	+
58	2,3,4	new	N/A	x	+*	1.54	+
59	1,2,3,4	new	1.36	0.02	+	−0.10	+
60	3,4	new	N/A	x	−	N/A	−
61	2,3,4	new	N/A	x	−	1.85	−*
62	2,3,4	new	N/A	x	−	2.00	−*
63	3,4	new	N/A	x	−	N/A	−
64	3,4	ref 17	N/A	x	−	N/A	−
65	3,4	new	N/A	x	−	N/A	−
66	3,4	new	N/A	x	−	N/A	−
67	3,4	ref 17	N/A	x	−	N/A	−

Table 3. (Continued)

no.	set[a]	synthesis[b]	human log IC_{50} (μM) obsd[c]	human log IC_{50} (μM) calcd[d]	human cls. calcd[e]	murine log IC_{50} (μM) obsd[c]	murine cls. calcd[e]
68	3,4	new	N/A	x	−	N/A	−
69	3,4	new	N/A	x	−	N/A	−
70	3,4	ref 17	N/A	x	−	N/A	−
71	3,4	new	N/A	x	−	N/A	−
72	3,4	new	N/A	x	−	N/A	−
73	3,4	ref 28	N/A	x	+*	N/A	+*
74	1,2,3,4	new	1.44	1.65	+	1.29	+
75	1,2,3,4	ref 25	1.62	1.53	+	1.71	+
76	1,2,3,4	new	0.92	1.99	+	0.49	−*
77	1,2,3,4	new	1.60	1.11	+	1.23	+
78	1,2,3,4	ref 16	0.59	1.11	+	0.20	+
79	1,2,3,4	new	0.70	0.78	+	−0.22	+
80	1,2,3,4	new	0.34	0.98	+	−0.92	+
81	1,2,3,4	new	1.23	0.54	+	0.62	+
82	1,2,3,4	new	1.51	1.36	+	1.06	+
83	1,2,3,4	new	1.91	2.11	+	1.72	+
84	1,2,3,4	new	1.98	1.41	+	1.87	+
85	1,2,3,4	new	0.71	0.75	+	0.11	+
86	1,2,3,4	new	2.18	1.21	+*	1.60	+
87	1,2,3,4	new	0.49	−0.26	+	−1.15	+
88	1,2,3,4	new	2.45	1.35	+*	2.13	+
89	1,2,3,4	new	1.33	0.98	+	0.52	+
90	1,2,3,4	new	1.85	1.53	+	1.52	+
91	1,2,3,4	new	1.28	0.18	+	0.64	+
92	1,2,3,4	new	2.55	1.04	−	2.09	+
93	1,3,4	new	−0.85	0.65	+	−1.30	+
94	1,3,4	new	−0.04	0.48	+	−1.22	+
95	2,3,4	ref 27	−1.15	x	+	−0.96	+
96	1,3,4	ref 27	0.40	0.57	+	−1.30	+
97	1,2,3,4	ref 27	2.40	0.91	+*	1.95	+
98	1,2,3,4	ref 27	1.08	1.07	+	0.45	+
99	3,4	new	−1.00	x	+	−1.30	+
100	3,4	new	−1.00	x	+	−1.30	+
101	3,4	ref 27	−1.00	x	+	−1.30	+
102	3,4	ref 27	−1.00	x	+	−1.30	+
103	3,4	new	−1.00	x	+	−1.30	+
104	1,2,3,4	new	−0.62	−0.38	+	−0.96	+
105	1,2,3,4	new	1.33	1.46	+	1.37	+
106	1,2,3,4	ref 27	2.14	1.83	+*	1.38	+
107	1,2,3,4	ref 27	0.81	−0.05	+	0.43	+
108	1,2,3,4	ref 27	0.23	0.64	+	−1.00	+
109	1,2,3,4	new	1.30	1.04	+	0.79	+
110	1,2,3,4	new	0.95	0.57	+	0.71	+
111	1,2,3,4	Aldrich	−0.80	−0.03	+	−1.05	+
112	3,4	new	−1.00	x	+	−1.30	+
113	3,4	new	−1.00	x	+	−1.30	+
114	1,3,4	new	−0.80	−0.24	+	−1.22	+
115	3,4	new	−1.00	x	+	−1.30	+
116	3,4	new	−1.00	x	+	−1.30	+
117	1,2,3,4	new	−0.26	0.88	+	−0.38	+
118	3,4	new	−1.00	x	+	−1.30	+
119	1,2,3,4	new	0.92	0.63	+	−0.05	+
120	1,2,3,4	new	1.25	0.78	+	−0.15	+
121	3,4	new	−1.00	x	+	−1.30	+
122	3,4	ref 27	−1.00	x	+	−1.30	+
123	3,4	ref 27	−1.05	x	+	−1.30	+
124	3,4	ref 27	−1.00	x	+	−1.30	+
125	3,4	ref 27	−1.00	x	+	−1.30	+
126	3,4	ref 16	−1.00	x	+	−1.30	+
127	3,4	ref 16	−1.00	x	+	−1.30	+
128	1,3,4	ref 27	−0.74	−0.83	+	−1.30	+
129	1,2,3,4	new	1.20	1.12	+	0.61	+
130	1,2,3,4	ref 25	0.14	0.06	+	−0.12	+
131	1,3,4	new	−0.74	−0.82	+	−1.30	+
132	1,3,4	ref 28	0.28	0.03	+	−1.30	+
133	3,4	ref 28	−1.00	x	+	−1.30	+
134	1,3,4	ref 28	0.57	0.07	+	−1.30	+

Table 3. (Continued)

no.	set[a]	synthesis[b]	human log IC_{50} (μM) obsd[c]	human log IC_{50} (μM) calcd[d]	human cls. calcd[e]	murine log IC_{50} (μM) obsd[c]	murine cls. calcd[e]
135	1,2,3,4	new	1.86	0.65	+	1.37	+
136	3,4	new	−1.00	x	+	−1.30	+
137	1,3,4	new	−0.28	−0.02	+	−1.22	+
138	1,3,4	ref 25	−0.80	−0.04	+	−1.22	+
139	1,3,4	ref 16	0.54	0.19	+	1.40	+
140	1,3,4	ref 28	−0.24	−0.28	+	−1.30	+
141	1,2,3,4	new	−0.82	0.41	+	−1.15	+
142	1,2,3,4	new	0.40	0.46	+	−0.14	+
143	3,4	new	−1.00	x	+	−1.30	+
144	1,2,3,4	new	−0.32	−0.79	+	−0.07	+
145	1,2,3,4	new	−0.85	−0.87	+	−1.10	+
146	1,3,4	new	−0.89	−0.82	+	−1.22	+
147	1,2,3,4	new	0.40	−0.80	+	1.04	+
148	3,4	ref 27	−1.00	x	+	−1.30	+
149	1,2,3,4	ref 27	−0.77	−0.45	+	−0.89	+
150	1,2,3,4	ref 27	−0.64	−0.17	+	−0.15	+
151	1,3,4	new	0.00	−0.79	+	−1.30	+
152	3,4	ref 27	N/A	x	+	−1.30	+
153	3,4	ref 27	−1.00	x	+	−1.22	+
154	1,3,4	ref 27	−0.57	0.06	+	−1.22	+
155	2,3,4	ref 27	−1.15	x	+	−1.05	+
156	3,4	ref 27	−1.00	x	+	−1.22	+
157	1,3,4	ref 27	−0.64	−0.74	+	−1.22	+
158	1,3,4	ref 27	−0.80	−0.77	+	−1.30	+
159	1,2,3,4	ref 27	−0.37	0.74	+	−0.11	+
160	1,2,3,4	new	−0.40	−0.41	+	−0.72	+
161	1,3,4	new	−0.72	−0.37	+	−1.30	+
162	1,2,3,4	new	−0.70	−0.46	+	−1.15	+
163	1,2,3,4	ref 28	−0.72	−0.43	+	−1.15	+
164	3,4	new	−1.00	x	+	−1.22	+
165	1,2,3,4	ref 27	−0.96	−0.38	+	−0.77	+
166	1,2,3,4	ref 27	−0.92	−0.38	+	−1.15	+
167	1,2,3,4	ref 27	0.34	−0.13	+	−0.77	+
168	1,3,4	new	0.18	−0.10	+	−1.30	+
169	1,3,4	ref 27	0.18	0.17	+	−1.30	+
170	1,2,3,4	ref 27	−0.44	0.27	+	−0.77	+
171	1,2,3,4	new	0.63	0.39	+	0.36	+
172	1,2,3,4	ref 27	1.05	−0.06	+	−0.10	+
173	1,2,3,4	ref 27	1.99	0.68	+	1.64	+
174	1,2,3,4	ref 27	0.32	−0.10	+	0.04	+
175	1,2,3,4	Chem Service	−0.44	0.17	+	−0.80	+
176	1,2,3,4	new	1.59	1.34	+	1.16	+
177	3,4	ref 27	−1.00	x	+	−1.30	+
178	3,4	ref 27	−1.00	x	+	−1.30	+
179	3,4	ref 27	−1.00	x	+	−1.30	+
180	3,4	new	−1.00	x	+	−1.30	+
181	1,2,3,4	new	1.07	0.98	+	0.85	+
182	1,2,3,4	Lancaster	2.34	2.14	+*	2.35	+*
183	1,2,3,4	new	0.85	1.35	−*	0.90	+
184	1,2,3,4	new	2.62	1.19	−	2.26	+
185	1,2,3,4	new	0.98	0.94	+	0.13	+
186	1,2,3,4	new	1.83	1.63	+	1.81	+
187	1,2,3,4	new	1.19	0.98	+	0.75	+
188	1,2,3,4	Aldrich	0.08	0.82	+	0.36	+
189	1,2,3,4	ref 16	2.11	0.86	−	2.29	+*
190	3,4	Aldrich	N/A	x	−	N/A	−
191	3,4	Aldrich	N/A	x	−	N/A	−
192	1,2,3,4	new	0.48	0.56	+	0.90	+
193	1,2,3,4	new	−0.26	0.25	+	−0.68	+
194	3,4	new	N/A	x	−	N/A	−
195	1,2,3,4	new	2.30	2.09	−	2.31	+*
196	1,2,3,4	new	0.58	1.00	+	0.37	−*
197	1,2,3,4	new	0.88	0.67	+	−0.23	+
198	1,2,3,4	new	0.34	0.69	+	−0.60	+
199	1,2,3,4	new	1.48	1.64	+	1.06	+
200	1,2,3,4	new	1.18	1.70	+	1.23	−*
201	1,2,3,4	new	0.53	1.15	+	−0.23	−*

Table 3. (Continued)

no.	set[a]	synthesis[b]	human log IC_{50} (μM) obsd[c]	human log IC_{50} (μM) calcd[d]	human cls. calcd[e]	murine log IC_{50} (μM) obsd[c]	murine cls. calcd[e]
202	1,2,3,4	new	−0.32	0.18	+	−0.70	+
203	3,4	new	N/A	x	−	N/A	−
204	2,3,4	new	−1.00	x	+	−1.05	+
205	1,2,3,4	new	1.46	1.19	−*	1.56	+
206	1,2,3,4	Lancaster	1.40	1.45	+	2.15	+
207	2,3,4	new	N/A	x	−	2.04	+
208	1,2,3,4	new	0.59	1.75	−*	1.27	+
209	1,2,3,4	new	1.30	1.76	−*	0.13	+
210	1,2,3,4	new	1.76	1.23	+	2.27	−*
211	1,2,3,4	new	1.45	1.19	+	1.48	+
212	1,2,3,4	new	1.12	1.18	+	1.43	+
213	1,2,3,4	new	0.46	0.40	+	0.41	+
214	1,2,3,4	new	0.58	0.38	+	−0.41	+
215	1,3,4	new	1.70	1.63	+	N/A	−
216	1,2,3,4	new	0.85	0.65	+	1.11	+
217	1,2,3,4	new	−0.16	0.19	+	0.11	+
218	1,2,3,4	new	1.34	0.26	+	1.72	+
219	1,2,3,4	new	−0.92	−0.17	+	−0.44	+
220	1,2,3,4	new	−0.59	−0.49	+	−0.15	+
221	1,2,3,4	ref 26	1.67	1.74	+	2.03	−*
222	1,2,3,4	new	0.10	−0.32	+	0.16	+
223	3,4	new	N/A	x	−	N/A	+*
224	1,2,3,4	new	0.68	0.79	+	0.77	+
225	3,4	new	N/A	x	+*	N/A	−
226	1,2,3,4	new	−0.22	0.05	+	1.26	+
227	1,2,3,4	new	−0.70	−0.40	+	0.41	+
228	1,2,3,4	new	2.34	1.25	−	2.11	+
229	1,2,3,4	new	1.15	1.33	−*	2.08	+
230	1,2,3,4	Aldrich	1.23	1.42	+	1.94	+
231	3,4	new	N/A	x	−	N/A	−
232	1,2,3,4	new	1.99	0.96	−*	2.31	−
233	1,2,3,4	new	−0.55	0.55	+	−1.00	−*
234	3,4	new	N/A	x	−	N/A	+*
235	1,2,3,4	new	2.43	0.87	+*	2.28	+
236	1,2,3,4	new	1.70	0.59	+	1.57	+
237	3,4	new	N/A	x	−	N/A	+*
238	1,2,3,4	new	1.57	1.17	−*	1.11	+
239	1,2,3,4	new	0.05	0.07	+	−0.15	+
240	3,4	new	N/A	x	+*	N/A	+*
241	5,6	new	N/A	x	+x	N/A	+x
242	3,4	Chem Service	N/A	x	+*	N/A	−
243	1,2,3,4	Chem Service	1.60	1.72	+	2.49	−
244	1,2,3,4	Chem Service	1.17	1.26	+	2.49	−
245	1,2,3,4	new	2.55	2.12	−	2.27	+
246	1,2,3,4	new	1.33	1.59	+	1.39	+
247	3,4	ref 25	−1.00	x	+	−1.30	+
248	3,4	Lancaster	N/A	x	−	N/A	+*
249	3,4	Lancaster	N/A	x	−	N/A	−
250	1,2,3,4	Chem Service	2.21	2.32	−	2.43	−
251	3,4	Chem Service	N/A	x	−	N/A	−
252	5,6	Chem Service	N/A	x	−x	N/A	+x
253	5,6	Chem Service	N/A	x	−x	N/A	+x
254	5,6	Chem Service	N/A	x	−x	N/A	−x
255	5,6	Chem Service	N/A	x	−x	N/A	+x
256	2,3,4	Chem Service	N/A	x	−	2.48	−
257	1,2,3,4	Chem Service	2.50	2.47	−	2.59	−
258	3,4	Aldrich	N/A	x	−	N/A	−
259	3,4	Aldrich	N/A	x	−	N/A	−
260	1,2,3,4	Aldrich	−0.33	−0.01	+	−0.60	+
261	3,4	Aldrich	N/A	x	−	N/A	−
262	3,4	Aldrich	N/A	x	−	N/A	−
263	3,4	Aldrich	N/A	x	−	N/A	−
264	3,4	Aldrich	N/A	x	−	N/A	−
265	3,4	Aldrich	N/A	x	−	N/A	−
266	3,4	Aldrich	N/A	x	−	N/A	−
267	1,2,3,4	Aldrich	1.71	0.80	+	1.21	+
268	3,4	Aldrich	N/A	x	−	N/A	−

Table 3. (Continued)

no.	set[a]	synthesis[b]	human log IC_{50} (μM) obsd[c]	human log IC_{50} (μM) calcd[d]	human cls. calcd[e]	murine log IC_{50} (μM) obsd[c]	murine cls. calcd[e]
269	2,4,5	Aldrich	N/A	x	−x	1.85	−*
270	3,4	Aldrich	N/A	x	−	N/A	−
271	3,4	Aldrich	N/A	x	−	N/A	−
272	3,4	Aldrich	N/A	x	−	N/A	+*
273	1,2,3,4	ref 16	1.30	1.69	+	2.00	+
274	3,4	Aldrich	N/A	x	+*	N/A	+*
275	3,4	Aldrich	N/A	x	−	N/A	−
276	3,4	Aldrich	N/A	x	−	N/A	−
277	3,4	Aldrich	N/A	x	−	N/A	−
278	3,4	Aldrich	N/A	x	−	N/A	−
279	3,4	new	N/A	x	−	N/A	−
280	5,6	new	N/A	x	+x	N/A	+x
281	1,3,4	Chem Service	2.38	1.93	−	2.88	−
282	1,2,3,4	Chem Service	2.28	2.14	−	2.62	−
283	1,2,3,4	Chem Service	2.11	1.42	−	2.36	−
284	1,2,3,4	Chem Service	1.91	2.45	−*	2.29	−
285	1,3,4	Chem Service	2.50	2.45	−	N/A	−
286	1,2,3,4	Chem Service	1.60	2.21	−*	2.27	+
287	1,2,3,4	new	0.52	0.64	+	0.87	+
288	1,2,3,4	ref 16	0.88	0.33	+	0.58	+
289	1,2,3,4	new	1.09	0.98	+	1.30	+
290	3,4	Aldrich	N/A	x	−	N/A	−
291	3,4	Aldrich	N/A	x	−	N/A	+*
292	3,4	Aldrich	N/A	x	−	N/A	−
293	3,4	Aldrich	N/A	x	−	N/A	−
294	3,4	Aldrich	N/A	x	−	N/A	−
295	3,4	Aldrich	N/A	x	−	N/A	−
296	3,4	Aldrich	N/A	x	+*	N/A	+*
297	3,4	Aldrich	N/A	x	−	N/A	−
298	3,4	Aldrich	N/A	x	−	N/A	−
299	3,4	Aldrich	N/A	x	−	N/A	−
300	3,4	Aldrich	N/A	x	−	N/A	−
301	1,2,3,4	new	−0.32	0.84	+	−0.96	+
302	1,3,4	ref 16	−0.89	−0.56	+	−1.22	+
303	1,2,3,4	new	1.71	2.35	−*	1.78	+
304	3,4	Chem Service	N/A	x	−	N/A	−
305	1,2,3,4	Chem Service	2.45	1.95	−	2.32	−
306	1,2,3,4	Cayman Chem.	1.69	0.41	−*	1.78	+
307	1,2,3,4	Aldrich	2.48	2.28	−	2.03	+
308	3,4	Aldrich	N/A	x	−	N/A	−
309	3,4	Aldrich	N/A	x	−	N/A	−
310	3,4	Aldrich	N/A	x	−	N/A	−
311	3,4	Aldrich	N/A	x	−	N/A	−
312	3,4	Aldrich	N/A	x	−	N/A	−
313	3,4	Aldrich	N/A	x	−	N/A	−
314	3,4	Aldrich	N/A	x	−	N/A	−
315	5,6	new	N/A	x	−x	N/A	−x
316	1,2,3,4	Aldrich	2.64	2.28	−	2.63	+*
317	1,2,3,4	Aldrich	2.61	2.46	−	2.63	−
318	1,2,3,4	new	0.49	0.69	−*	−0.42	+
319	1,2,3,4	new	1.34	0.80	+	0.00	+
320	1,2,3,4	ref 28	1.62	0.14	+	0.52	+
321	2,3,4	ref 28	N/A	x	+*	−0.49	+
322	1,2,3,4	new	1.18	0.34	−*	1.28	+
323	1,2,3,4	ref 28	0.00	−0.04	+	−0.47	+
324	1,2,3,4	ref 28	−0.38	−0.05	+	0.71	+
325	1,2,3,4	new	1.60	0.75	+	1.54	+
326	1,3,4	ref 28	−0.60	0.05	+	−1.30	+
327	1,3,4	ref 28	−0.96	−0.26	+	−1.30	+
328	1,2,3,4	ref 28	2.36	−0.64	+*	−0.89	+
329	1,2,3,4	new	−0.26	−0.25	+	−1.00	+
330	3,4	new	N/A	x	−	N/A	+*
331	1,2,3,4	new	0.85	1.61	+	0.32	+
332	1,2,3,4	new	−0.55	0.65	+	−0.60	+
333	1,2,3,4	new	1.89	1.36	+	0.40	+
334	1,3,4	new	−0.82	0.78	−*	−1.30	+
335	1,3,4	ref 27	1.13	0.72	+	−1.30	+

Table 3. (Continued)

no.	set[a]	synthesis[b]	human log IC_{50} (μM) obsd[c]	human log IC_{50} (μM) calcd[d]	human cls. calcd[e]	murine log IC_{50} (μM) obsd[c]	murine cls. calcd[e]
337	1,3,4	ref 27	−0.43	0.13	+	−1.30	+
338	1,3,4	ref 27	−0.92	0.11	+	−1.30	+
339	1,3,4	ref 27	−0.82	−0.32	+	−1.30	+
340	3,4	ref 27	−1.00	x	+	−1.30	+
341	1,3,4	new	−0.19	−0.05	+	−1.30	+
342	1,3,4	new	−0.16	0.13	+	−1.30	+
343	3,4	new	−1.00	x	+	−1.30	+
344	3,4	new	−1.00	x	+	−1.30	+
345	1,2,3,4	new	0.88	0.76	+	−0.19	+
346	3,4	ref 27	−1.00	x	+	−1.30	+
347	3,4	new	−1.00	x	+	−1.30	+
348	1,2,3,4	new	2.69	2.29	−	2.51	−

[a] Denotes set memberships, where 1 is human quantitation, 2 is murine quantitation, 3 is human classification, 4 is murine classification, 5 is human external prediction, and 6 is murine external prediction. [b] Denotes the reference for the synthesis or commercial origin. "New" indicates that the compound was not described before, but was synthesized in a manner consistent with the examples in this and other papers. [c] IC_{50}s of compounds denoted by N/A could not be determined because these compounds were above the highest concentration (500 μM) of inhibitor tested. [d] Predicted log IC_{50} (μM) values from the Type III five-descriptor model. An x denotes compounds that were not predicted with the quantitative model. [e] Classification results are denoted by + (active) or − (inactive), with an asterisk (*) denoting a misclassification. An x after the classification result denotes a prediction of a compound that was not considered in training or validation because of missing data.

Objective Feature Selection. When dealing with a large pool of descriptors, there is a chance that many of them will offer little or no important information, or that several descriptors may contain highly correlated or even identical information. For the model building techniques discussed below, a genetic algorithm or simulated annealing method was employed to search the descriptor space to find the combinations of descriptors that would best correlate molecular structure with sEH inhibition. It was therefore desirable to trim the overall descriptor pool of useless or redundant information. Objective feature selection, which does not utilize the dependent variable, was performed using only TSET compounds to create a reduced descriptor pool that maximized the amount of information within the reduced pool space. The maximum size of the final reduced pool of descriptors was constrained only by the following condition: the number of reduced pool descriptors could be no more than 60% the number of training set compounds. This limitation has been shown to reduce the possibility of chance correlations when building models.[57]

Two methods were used to remove less informative descriptors from the pool: identical tests and pairwise correlations. If a descriptor held zero or identical information across more than 70−85% of its range, then it was removed from consideration. The remaining descriptors were subjected to pairwise correlations with all other descriptors. If two descriptors had a pairwise correlation of greater than 0.80−0.93, then one of those descriptors was removed at random. Because of the different sets and distribution of compounds in the training, cross-validation, and prediction sets, the above procedures were carried out separately for each data set using separate identical test and correlation cutoff values. As a result, the number and types of descriptors in each reduced descriptor pool were different for the four data sets.

Quantitative Model Formation and Validation. Three methods were used to create quantitative models for prediction of sEH inhibition values (log IC_{50} [μM]). Type 1 models used linear feature selection and linear model development, type 2 models employed the descriptors chosen in type 1 feature selection in a nonlinear CNN, and type 3 models used nonlinear CNNs for both feature selection and model construction. For type 1, TSET and CVSET compounds were combined to form a larger training set. In nonlinear modeling, these compounds were kept separate − TSET members for training and CVSET members to prevent CNN overtraining. In all modeling, PSET compounds were used only at the end of the procedure to validate the predictive ability of each model.

Linear Regression Models. A simulated annealing optimization routine[35] was used to screen the reduced descriptor pool to find the smallest subsets (models) of descriptors that would accurately predict log IC_{50} (μM) values based on multiple linear regression. Three-descriptor subsets were chosen as a starting point, and the root-mean-square error (RMSE) was calculated for TSET compounds for several models. Models were assessed by descriptor *T*-values to ensure that the magnitude of errors was no greater than 25% of the descriptor coefficient, and only models with descriptor *T*-values $>|4|$ were examined further. A variance of inflation factor (VIF) were calculated to check for multicollinearities by regressing each descriptor against all others in the model. VIFs were calculated as $[1/(1 - R^2)]$, where R is the multiple correlation coefficient. Models were considered free of multicollinearities if the VIF values for all model descriptors were less than 10. Finally, the following statistical values were calculated to check for outliers in each model: residuals, standardized residuals, studentized residuals, leverage points, DFFITS values, and Cook's distance.[58,59] Model sizes were then increased sequentially and evaluated as above until no significant improvement (decrease) in RMSE was observed by adding another descriptor. Once the best model was found, PSET compounds were used to validate the models. Results were plotted for visual inspection.

Nonlinear CNN Models. Once a valid type 1 model was found, its descriptors were used as inputs for a nonlinear CNN. A three-layer, fully connected, feed-forward CNN, described in detail previously,[37,60] was used for nonlinear model formation. The number of input neurons equaled the number of descriptors in the type 1 model, while a single output neuron generated the calculated log IC_{50} values. The number of neurons in the hidden layer was increased sequentially until no marked improvement was seen in TSET RMSE. One restriction placed on this process was to keep the ratio of TSET observations to CNN adjustable parameters (weight and bias terms) greater than two. Having too many CNN adjustable parameters can lead to chance correlations during model development.[61]

Network training was optimized using the BFGS (Broyden-Fletcher-Goldfarb-Shanno) quasi-Newton method.[62−65] TSET compounds were used to adjust weights and biases in the CNN to minimize the RMSEs, and overtraining of the networks was prevented by use of the CVSET. TSET RMSEs continually improve during training, but at a certain point the network starts to memorize idiosyncrasies of the individual compounds in the TSET and loses its ability to generalize. Therefore, the RMSEs of the CVSET compounds were periodically checked throughout training. The weights and biases that produced the minimum CVSET error were considered optimal. It was at this minimum that the network was losing its ability to generalize and so training was halted.

Initial weights and biases was assigned randomly at the onset of network training, and the final network results were dependent upon those weights and biases. To reduce the dependence of results on initial weights and biases, and to ensure that CNN results were reliable, a committee of five randomly initialized networks was used. Their outputs were averaged to produce a final predicted log IC_{50} for a compound, and these average values were then used to calculate the final RMSE and *R* values.

Fully Nonlinear CNN Models. Fully nonlinear quantitative models used a CNN for both feature selection and model building. The descriptor sets found using linear feature selection are not necessarily the best subsets when considering a nonlinear relationship between molecular structure and IC_{50} (μM) values. Therefore, a genetic algorithm routine with a CNN fitness evaluator was used to determine the best subsets of descriptors from the reduced descriptor pools. Once the best subsets of descriptors were found, they were trained and tested by the same type 2 procedures outlined above.

Classification Models. The *k*-Nearest Neighbor (*k*NN) algorithm is a fast, supervised learning method that assigns a class to a compound based on the distances to its *k* nearest neighbors in descriptor space. Euclidean distances from a compound of interest to all other TSET compounds are measured, and the shortest *k* distances are used to assign the class of that compound. In this study, $k = 3$ was used to prevent ties in a binary problem. A genetic algorithm employing the *k*NN as a fitness function was used to search the descriptor space to find small subsets that would minimize the number of incorrect classifications. Starting with three descriptors, the genetic algorithm produced several models that were evaluated based on their classification rates for the TSET compounds. Subset sizes were increased until no significant improvement was seen in TSET classification. Models were then validated using the PSET compounds.

Two less successful classification methods were applied to data sets 3 and 4 and will not be covered in detail. Linear discriminant analysis (LDA)[66,67] maximizes the distance between class means relative to their variances, while a radial basis function neural network[67,68] is a nonlinear classification approach. Both schemes were utilized as fitness evaluators in a genetic algorithm search of the reduced descriptor pool. As with the *k*NN approach, the search began with three descriptors and several models were analyzed, then model sizes were increased until no marked improvement in TSET classification rates was observed.

Results and Discussion

Data Set 1. A simulated annealing routine was used to search the reduced pool of 95 descriptors for subsets ranging from three to eight descriptors. A five-descriptor model was chosen as optimal. The TSET RMSE was 0.802 log units ($r^2 = 0.44$) and PSET RMSE was 0.776 log units ($r^2 = 0.47$). Pairwise correlations among the descriptors ranged from −0.27 to 0.68 (mean $|r| = 0.20$), and descriptor *T*-values were $>|4|$.

The five descriptors from the best type 1 model were passed to a CNN, and network architectures from 5−2−1 to 5−5−1 were trained using a committee approach of five networks. The best results were found using a network architecture of 5−5−1 that provided a TSET RMSE of 0.695 log units ($r^2 = 0.56$), CVSET RMSE of 0.732 log units ($r^2 = 0.54$), and PSET RMSE of 0.668 log units ($r^2 = 0.60$). Using nonlinear modeling on descriptors chosen by linear feature selection improved the RMS errors for both TSET (13.3%) and PSET (13.9%) compounds.

Type 3 fully nonlinear models usually provide the best modeling results in a QSAR, but due to their greater computational cost it is desirable to find suitable models sizes and network architectures using linear and hybrid modeling. A genetic algorithm routine using a CNN fitness evaluator was applied to a 5−5−1 network architecture, and several five-descriptor models found by this method were trained and validated using the committee approach outlined above. The descriptors of this model are listed in Table 4. WTPT-5 is the sum of all path weights starting from nitrogen atoms, revealing atom-specific branching information.[42] 2SP3 is a count of sp^3-hybridized carbons that are attached to two other carbons. This denotes the amount of branching and relative size of the molecule when considering carbon-based side groups on the R_1−R_4 positions. MDE−44 is the molecular distance edge term between all quaternary carbons, which gives an indication of the number of quaternary carbons and the path lengths between them.[47] For example, compound **87** and **91** each have three quaternary carbons, **87** has a higher MDE−44 value than **91** because the quaternary carbons are closer together. If two compounds had different number of quaternary carbons but their path length products were equal, the compound with fewer carbons would have the higher value. BCUT-58 is the negative second Burden eigenvalue weighted by the atomic polarizability that relates distribution of charge information throughout the molecule.[53] SAAA-2 is the average surface area of hydrogen bond acceptor atoms (oxygen, nitrogen, sulfur, fluorine) in the molecule.[51]

Table 4. Descriptors for the Quantitative Type 3 Model of Human Soluble Epoxide Hydrolase

descriptor[a]	type	ranges	average	standard deviation
WTPT-5	topo	0−11.50	5.77	1.54
2SP3	topo	0−17.00	5.17	3.70
MDE−44	topo	0−34.80	1.32	2.93
BCUT-58	topo	0.98−1.95	1.81	0.11
SAAA-2	hyb	−1.32−31.90	13.78	4.97

[a] WTPT-5, the sum of all path weights starting from nitrogen atoms;[42] 2SP3, the count of sp^3-hybridized carbons attached to two carbons; MDE−44, the distance edge term between all quaternary carbons;[47] BCUT-58, the negative second Burden eigenvalue weighted by atomic polarizability;[53] SAAA-2, the average acceptor atom surface area [$\Sigma(SA)_{acc}$/no. of acceptor atoms].[51]

TSET and CVSET RMS errors were improved over type 2 results (11.4% and 7.9%, respectively) and with the exception of one outlier, PSET RMSE also improved. TSET RMSE was 0.616 log units ($r^2 = 0.66$), CVSET RMSE was 0.674 log units ($r^2 = 0.61$), and PSET RMSE was 0.914 log units ($r^2 = 0.33$). Removing one severe outlier, **328**, from the 21-member prediction set resulted in a PSET RMSE error = 0.653 log units ($r^2 = 0.58$). It is not clear why this compound was predicted so poorly, as none of its descriptor values fell outside the range of TSET compound descriptors, and it appears structurally similar to many other compounds in the data set. It is interesting to note that the predicted value of −0.644 log units for this compound is much closer to the observed value of −0.886 log units for the murine data set than the observed human set value of 2.362 log units. See Table 3 for calculated log IC_{50} (μM) values.

A plot of calculated vs observed log IC_{50} values is shown in Figure 1. The 21 PSET compounds are represented by solid triangles. The plot shows substantial spread, which is not unusual for biological activity studies. The RMSE for the 21 PSET compounds is 0.914

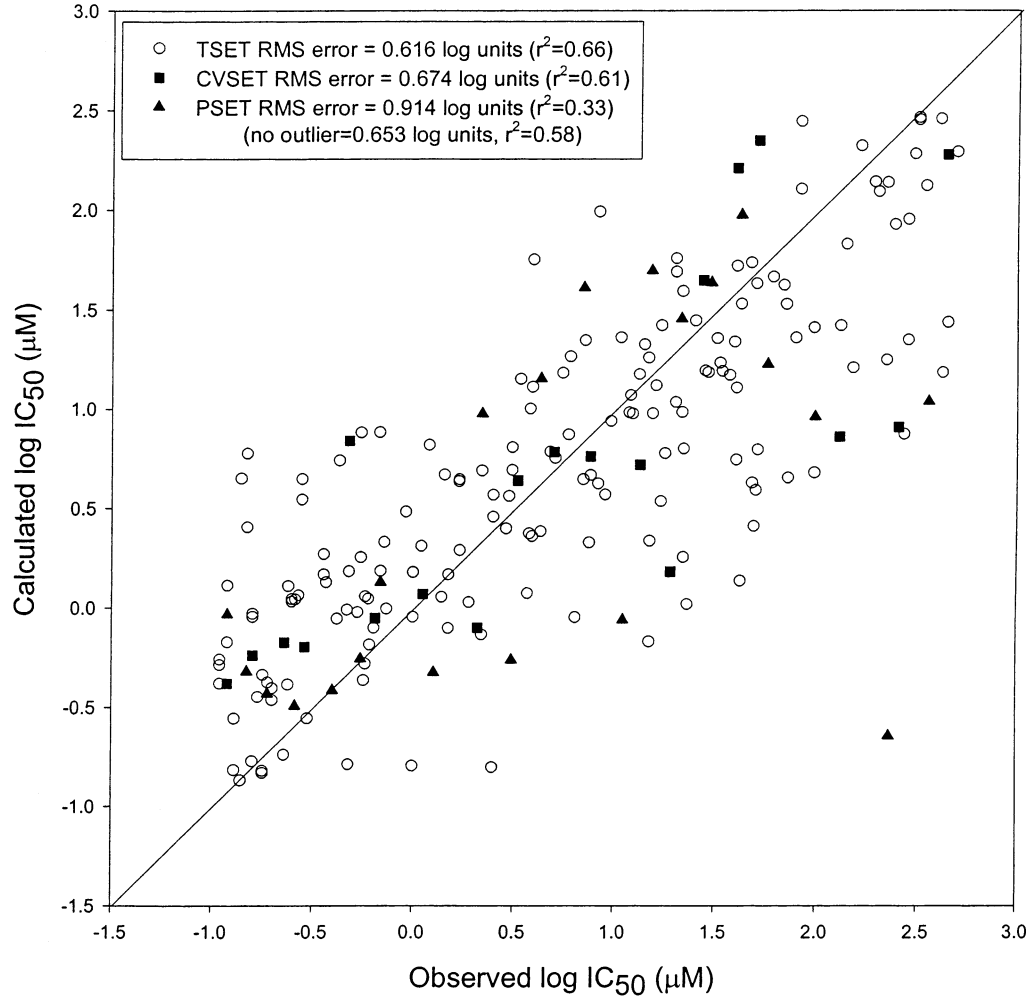


Figure 1. Plot of calculated versus observed log IC_{50} (μM) of the data set 1 type 3 quantitative model (human soluble epoxide hydrolase inhibitors). TSET (n = 167), CVSET (n = 19), PSET (n = 21).

log units, and the RMSE for 20 PSET compounds (neglecting one outlier 328) is 0.653 log units. Of the 21 PSET compounds, 14 have predicted log IC_{50} values within 0.653 log units of their observed values, and seven have log IC_{50} values differing from their observed value by more than this. The best calculated log IC_{50} value is for compound 329, and the worst outlier is 328. These two compounds have very similar structures, but somehow the five descriptors chosen as the best overall set is able to describe 329 very well but 328 very poorly.

After obtaining these results, a second assay was performed on 328. The original IC_{50} reported in this study was 230 ± 20 μM, and the IC_{50} value obtained from the second assay was 190 ± 30 μM. It was noted that compound 328 dissolved much slower in DMF than 329, suggesting the formation of microcrystals that could affect solubility. Likewise, the assay curve for 328 showed a distinct break, whereas assay curves for similar compounds showed a smooth transitional curve. This further supports the evidence of crystal formation at higher concentrations. For several compounds, the IC_{50} values for human sEH inhibition were on the order of 10 times the IC_{50} values for murine sEH inhibition. Extrapolating the IC_{50} value on the lower end of the curve shows this to be true also for 328, and it is believed that the higher observed value reported is due largely to the poorer solubility. This may also explain the poorly predicted value from the model. The *tert*-octyl group on 328 may also contribute the higher observed value because of steric hindrance in humans, but the crystal structure has yet to be determined.

Data Set 2. Both linear and nonlinear modeling were performed on data set 2 using the above procedures, but results were not as robust as those found with data set 1.

Randomization Experiments. Scrambling experiments were performed on the data to ensure that the best linear and nonlinear models presented here were the result of a structure−activity relationship and not due to random effects. The dependent variables were randomly scrambled, and models were trained and validated using the above methodologies. An increase in RMS errors and decrease in r^2 was desired to show that scrambled inhibition values had little correlation to their structures. For data set 1, the scrambled five-descriptor type 1 model had a TSET RMSE of 1.039 log units (r^2 = 0.07) and PSET RMSE of 1.174 log units (r^2 = 0.06). The type 3 model with 5−5−1 architecture using scrambled inhibition values resulted in a TSET RMSE of 0.935 log units (r^2 = 0.23), CVSET RMSE of 1.115 log units (r^2 = 0.08), and PSET RMSE of 1.082 log units (r^2 = 0.01). Visual inspection of a calculated vs observed plot revealed the data clustered about the network average value. From this, we concluded that chance correlation had little or no effect in driving model development.

Table 5. Descriptor Ranges and Averages for Active and Inactive Compounds for Seven-Descriptor Data Set 3 *k*NN Classification Model: Human SEH Inhibition

		range		average (std dev)	
descriptor[a]	type	active	inactive	active	inactive
MOLC-9	topo	1.10−3.46	1.47−4.56	2.05 (0.51)	2.67 (0.65)
1SP3	topo	0−5	0−10	0.79 (1.07)	1.17 (1.67)
2SP3	topo	0−17	0−16	6.19 (3.93)	3.39 (4.62)
PND-5	topo	0−312	0−92.10	17.60 (32.10)	13.84 (19.29)
SAAA-3	hyb	0.002−0.75	0−0.38	0.10 (0.07)	0.13 (0.07)
CHAA-1	hyb	−2.19 to −0.42	−2.00 to −0.29	−1.26 (0.28)	−1.11 (0.39)
WHIM-42	geo	0.37−0.69	0.39−0.70	0.58 (0.05)	0.55 (0.05)

[a] MOLC-9, topological index **J**;[46] 1SP3, count of sp^3-hybridized carbons attached to one heteroatom; 2SP3, count of sp^3-hybridized carbons attached to two carbons; PND-5, superpendentic index calculated from pendent oxygen atoms;[69] SAAA-3, $\Sigma(SA)_{acc}/(SA)_{tot}$; CHAA-1, $\Sigma(Q)_{acc}$;[51] WHIM-42, 1st component accessibility weighted by atomic Sanderson electronegativities.[56]

Data Set 3. The best model for the human classification data set 3 was found using *k*NN, which produced the seven-descriptor model listed in Table 5. MOLC-9 is the topological index **J** or the averaged distance sum connectivity,[46] which reveals molecular branching information. On average, inactive compounds had higher values for this descriptor than active compounds. The next two descriptors account for the carbon types in the hydrophobic side chains of the inhibitors. 1SP3 and 2SP3 are counts of sp^3-hybridized carbons bound to one and two carbons, respectively. Both descriptors give some degree of branching information but also give insight into the amount of hydrophobicity on the nitrogen-bonded substituent groups. On average, active inhibitors had fewer 1SP3 carbons and more 2SP3 carbons than inactive compounds. Inhibition may be increased by longer, less-branched side groups attached to the nitrogens, allowing for greater flexibility while maintaining hydrophobic character. PND-5 is the superpendentic index[69] from pendant oxygen atoms. Calculated from the pendent matrix, this descriptor relays branching information of the molecule with respect to terminal oxygens, in the majority of cases, the carbonyl oxygen. SAAA-3 and CHAA-1 are both hydrogen-bonding descriptors,[51] calculated by assuming a mixture of solute and solvent. SAAA-3 is the ratio of acceptor atom surface area to total molecular surface area, and CHAA-1 is the summation of charges on all acceptor atoms. These descriptors shed light on the relative available acceptor atom area on the inhibitor and the associated charges of the acceptor atoms, oxygen nitrogen, sulfur, and fluorine, on the surface of the inhibitor that may interact with the active site on the sEH enzyme. Active compounds had slightly less relative acceptor atom surface area than inactive compounds but with greater negative charges on average. WHIM-42 is related to atomic distribution and molecular density around the Cartesian coordinate origin and first principal axis determined by principal component analysis, weighted by atomic electronegativity.[56] The interpretability of this descriptor is poor but gives some sense of the electronic interactions occurring near the molecules center of mass. Very little difference was seen for average WHIM values of active vs inactive compounds, yet removing this descriptor from consideration resulted in poorer classification.

This model correctly classified 89.1% of the TSET compounds and 91.4% of the prediction set compounds (89.4% overall). The confusion matrix of classification results is shown in Table 6. Individual predicted class values are shown in Table 3. For both TSET and PSET, there were higher percentages of false positives than false negatives, due in part to a higher number of active compounds in the entire data set. Overall, active compounds were correctly classified at 92.8% and inactive compounds at 82.9%, with 20 false positives and 16 false negatives. Average descriptor values for misclassified compounds predominantly fell between the descriptor averages of active and inactive compounds. For example, the false-active SAAA-3 descriptor average was 0.11 and the false-inactive SAAA-3 average was 0.13. These values corresponded well to the active and inactive descriptor averages of 0.10 and 0.13, respectively. This trend was observed for the majority of misclassified compound descriptor averages. Log IC_{50} values of the misclassified compounds were distributed throughout the range of inhibition data and not centered around the active-inactive cutoff value. The most poorly predicted compound in the human quantitative model, 328, was also misclassified as an active compound in this model.

Table 6. Confusion Matrix of Training and Prediction Set Compounds for Seven-Descriptor *k*NN Classification Model: Human SEH Inhibition

	training set = 89.1%[a]		prediction set = 91.4%[b]	
actual class	active	inactive	active	inactive
active	183	15	23	1
inactive	18	88	2	9

[a] Active TSET compounds correct classification = 92.4%; inactive TSET compounds correct classification = 83.0%. [b] Active PSET compounds correct classification = 95.8%; inactive PSET compounds correct classification = 81.8%.

LDA produced a model of seven descriptors with TSET, CVSET, and PSET classification rates of 87.0%, 85.7%, and 88.6%, respectively. Using a RBF−NN also resulted in a seven-descriptor model with a five-run average classification rate of 81.6 ± 2.2%, 77.7 ± 5.1%, and 81.1 ± 7.7% for the TSET, CVSET, and PSET, respectively.

Data Set 4. The best classification model for the murine classification data set 4 was found using *k*NN, which produced the five-descriptor model listed in Table 7. MOLC-9, the topological index **J**, is described above. As with the human classification model, inactive compounds had higher average MOLC-9 values than active compounds. 2SP3 is described above, and again active compounds had a higher count of 2SP3 carbons than inactive compounds. CHAA-2 is the average charge on all acceptor atoms of the molecule, revealing the potential for hydrogen bonding at the active site.[51] There was very little difference between the average values for

Table 7. Descriptor Ranges and Averages for Active and Inactive Compounds for Five-Descriptor Data Set 4 *k*NN Classification Model: Murine SEH Inhibition

		range		average (std dev)	
descriptor[a]	type	active	inactive	active	inactive
MOLC-9	topo	1.10−3.47	1.47−4.56	2.07 (0.52)	2.72 (0.64)
2SP3	topo	0−17	0−16	6.26 (3.91)	2.80 (4.48)
CHAA-2	hyb	−0.41 to −0.12	−0.44 to −0.09	−0.35 (0.05)	−0.32 (0.09)
WHIM-42	geo	0.37−0.69	0.39−0.70	0.58 (0.05)	0.55 (0.05)
WHIM-84	geo	0.34−0.98	0.32−0.98	0.84 (0.09)	0.74 (0.18)

[a] MOLC-9, topological index **J**;[46] 2SP2, count of sp^2-hybridized carbons attached to two carbons; CHAA-2, $\Sigma(Q)_{acc}/count_{acc}$;[51] WHIM-42, 1st component accessibility weighted by atomic Sanderson electronegativities;[56] WHIM-84, K global shape index weighted by van der Waals volume.[56]

Table 8. Confusion Matrix of Training and Prediction Set Compounds for Five-Descriptor *k*NN Classification Model: Murine SEH Inhibition

	training set = 91.5%[a]		prediction set = 88.6%[b]	
actual class	active	inactive	active	inactive
active	200	10	24	1
inactive	16	78	3	7

[a] Active TSET compounds correct classification = 95.2%; inactive TSET compounds correct classification = 83.0%. [b] Active PSET compounds correct classification = 96.0%; inactive PSET compounds correct classification = 70.0%.

active and inactive compounds, but removing this descriptor from the model resulted in poorer classification results. WHIM-42 is described above, and again there was little difference in average values of active and inactive compounds. WHIM-84 encodes an overall shape of the molecule weighted by van der Waals volume. The overall shape value, *K*, used in the calculation of WHIM-84 is equal to 1 for linear molecules and 0 for ideal spherical molecules, and ranges between 0.5 and 1 for planar molecules.[55] Active compounds had slightly higher values than inactive compounds.

This model correctly classified 91.5% of the TSET compounds and 88.6% of the PSET compounds (91.2% overall). The confusion matrix of classification rates is shown in Table 8. Individual predicted class values are shown in Table 3. For both the TSET and PSET, there were higher percentages of false-actives than false-inactives. Overall, active compounds had a correct classification rate of 95.3% and inactive compounds a rate of 81.7%, with 11 false-negatives and 19 false-positives. As with the human classification results, misclassified compounds had descriptor values that fell between the active/inactive average values, and log IC_{50} (μM) values of misclassified compounds were distributed throughout the range of inhibition data.

LDA produced a model of six descriptors with TSET, CVSET, and PSET classification rates of 88.5%, 85.7%, and 94.3%, respectively. Using a RBF−NN resulted in a seven-descriptor model with a five-run average classification rate of 81.5 ± 2.1%, 86.3 ± 2.4%, and 83.4 ± 7.1% for the TSET, CVSET, and PSET, respectively.

Randomization Experiments. As with the quantitative models, it was important to ensure that the best classification models found were not due to chance. Therefore, a scrambling experiment was also carried out for classification models. The dependent variables, in this case the binary classifiers, were randomly scrambled 10 times. Classification models were built using *k*NN, LDA, and RBF−NN algorithms to find the best models of the same size as those listed above. Poor results, or PSET classification rates within one standard deviation of random class assignment in the TSET, were desired to show that the best models were indeed classifying compounds based on their molecular structure features rather than chance.

Scrambling experiment results for the data set 3 seven-descriptor *k*NN model were 71.6 ± 1.1% for the TSET and 54.3 ± 10.2% for the PSET. The uneven distribution of active and inactive compounds in the training set allowed for a random class assignment of 54.6% in the PSET; therefore, it was concluded that chance effects played little or no role in the formation of the above model.

Scrambling experiment results for the data set 4 five-descriptor *k*NN model were 72.4 ± 1.6% for the TSET and 56.3 ± 10.0% for the PSET. The uneven distribution of active and inactive compounds in the TSET allowed for a random class assignment of 57.7%; thus, chance effects again played little or no role in the formation of the above model.

Scrambling experiments for the LDA and RBF−NN models for human and murine data sets also revealed little or no chance effect in those models.

Conclusions

Successful models that quantitatively predict or classify human and murine soluble epoxide hydrolase inhibition by urea-like compounds have been presented. Both linear and nonlinear quantitative models were created for data set 1 (human sEH), and the fully nonlinear model gave the best results. Classification models using three different algorithms had successful classification rates, with *k*NN providing the best overall TSET and PSET classification rates for data sets 3 (human sEH) and 4 (murine sEH).

Similar structural descriptor types were found in all the models (i.e., hydrogen bonding, carbon types, branching), supporting a relationship between structure and inhibitor activity. Inhibition is affected by the position of the urea moiety at the active site, which is in turn dependent upon the size and shape of the alkyl groups attached to the urea nitrogens. In general, inhibition of murine sEH was achieved with lower inhibitor concentrations than needed for human sEH inhibition. In the few instances where a lower concentration was needed for human sEH inhibition compared to murine sEH inhibition, it was not clear from a structural standpoint why this was so. The differences in effective IC_{50} concentrations are most likely due to subtle differences in the configuration of human and murine sEH, and the manner in which inhibitors interact in those systems. Compounds with very low IC_{50} concentrations (<1.00 μM) seem to favor a structure with hydrogens

in the R2 and R3 positions, and a host of substituents on the R1 and R4 positions. The exact ratio between bulkiness and length on either R1 or R4 is difficult to ascertain, but is encoded in the models through size and branching descriptors such as 2SP3, MOLC-9, WTPT-5, MDE−44, etc., mentioned above. The sizes and shapes of substituents on R1 and R4 then determine how well the compound will sit within the active site pocket to allow hydrogen bonding or other electronic interactions as encoded by descriptors such as SAAA-x, CHAA-x, etc.

Scrambling experiments resulted in poorer models for classification and quantitation, providing evidence that results were not due to chance. The models that best predicted log IC_{50} values and correctly classified compound activity contained descriptors that may offer insight into the role those features play in inhibiting the sEH enzyme in human and murine systems. These models may be used as a tool for screening similar compounds whose activity or IC_{50} values are unknown, provided that those new compounds have structures similar to those used in training these models. One strategy may be to apply the classification models to an unknown to determine its activity. Those compounds deemed active by the classification could be tested further through the quantitative model for an estimate of inhibitor concentration.

Acknowledgment. The authors thank Drs. Deanna L. Dowdy, Marvin H. Goodrow, Bruce G. Hammock, James R. Sanborn, and Craig E. Wheelock, from U. C. D., for the synthesis of some of the structures used in this study. Electrospray and MALDI mass spectral analyses were performed at U. C. D. by Dr. John W. Newman. This work was supported in part by NIEHS grant no. R37-ES02710, NIEHS Center for Environmental Health Sciences grant no. P30-ESO5707, UC Systemwide Biotechnology Research and Education Training Grant no. 2001-07, and by the NIH/NIEHS Superfund Basic Research program no. P42-ES04699.

Supporting Information Available: A table of structures and observed IC_{50} values for the 348 compounds considered in this study. Histograms of data distribution for the active-inactive cutoff. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hammock, B. D.; Storms, D. H.; Grant, D. F. Epoxide Hydrolases. *Comprehensive Toxicology*: *Biotransformation*; Elsevier: New York, 1997; pp 283−305.
(2) Oesch, F. Mammalian Epoxide Hydrolases: Inducible Enzymes Catalysing the Inactivation of Carcinogenic and Cytotoxic Metabolites Derived from Aromatic and Olefinic Compounds. *Xenobiotica* **1972**, *3*, 305−340.
(3) Argiriadi, M. A.; Morisseau, C.; Hammock, B. D.; Christianson, D. W. Detoxification of Environmental Mutagens and Carcinogens: Structure, Mechanism, and Evolution of Liver Epoxide Hydrolase. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10637−10642.
(4) Greene, J. F.; Newman, J. W.; Williamson, K. C.; Hammock, B. D. Toxicity of Epoxy Fatty Acids and Related Compounds to Cells Expressing Human Soluble Epoxide Hydrolase. *Chem. Res. Toxicol.* **2000**, *13*, 217−226.
(5) Kosaka, K.; Suzuki, K.; Hayakawa, M.; Sugiyama, S.; Ozawa, T. Leukotoxin, A Linoleate Epoxide: Its Implication in the Late Death of Patients with Extensive Burns. *Mol. Cell. Biochem.* **1994**, *139*, 141−148.
(6) Moghaddam, M. F.; Grant, D. F.; Cheek, J. M.; Greene, J. F.; Williamson, K. C. et al. Bioactivation of Leukotoxins to Their Toxic Diols by Epoxide Hydrolase. *Nat. Med.* **1997**, *3*, 562−566.
(7) Sinal, C. J.; Miyata, M.; Tohkin, M.; Nagata, K.; Bend, J. R. et al. Targeted Disruption of Soluble Epoxide Hydrolase Reveals a Role in Blood Pressure Regulation. *J. Biol. Chem.* **2000**, *275*, 40504−40510.
(8) Yu, Z.; Xu, F.; Huse, L. M.; Morisseau, C.; Draper, A. J. et al. Soluble Epoxide Hydrolase Regulates Hydrolysis of Vasoactive Epoxyeicosatrienoic Acids. *Circ. Res.* **2000**, *87*, 992−998.
(9) Fang, X.; Kaduce, T. L.; Weintraub, N. L.; Harmon, S.; Teesch, L. M. et al. Pathways of Epoxyeicosatrienoic Acid Metabolism in Endothelial Cells. *J. Biol. Chem.* **2001**, *296*, 14867−14874.
(10) Argiriadi, M. A.; Morisseau, C.; Goodrow, M. H.; Dowdy, D. L.; Hammock, B. D. et al. Binding of Alkylurea Inhibitors to Epoxide Hydrolase Implicates Active Site Tyrosines in Substrate Activation. *J. Biol. Chem.* **2000**, *275*, 15265−15270.
(11) Borhan, B.; Jones, A. D.; Pinot, F.; Grant, D. F.; Kurth, M. J. et al. Mechanism of Soluble Epoxide Hydrolase. *J. Biol. Chem.* **1995**, *270*, 26923−26930.
(12) Nardini, M.; Ridder, I. S.; Rozeboom, H. J.; Kalk, K. H.; Rink, R. et al. The X-ray Structure of Epoxide Hydrolase from *Agrobacterium radiobacter* AD1. *J. Biol. Chem.* **1999**, *274*, 14579−14586.
(13) Mullin, C. A.; Hammock, B. D. Chalcone Oxides − Potent Selective Inhibitors of Cytosolic Epoxide Hydrolase. *Arch. Biochem. Biophys.* **1982**, *216*, 423−439.
(14) Morisseau, C.; Du, G.; Newman, J. W.; Hammock, B. D. Mechanism of Mammalian Soluble Epoxide Hydrolase Inhibition by Chalcone Oxide Derivatives. *Arch. Biochem. Biophys.* **1998**, *356*, 214−228.
(15) Draper, A. J.; Hammock, B. D. Inhibition of Soluble and Microsomal Epoxide Hydrolase by Zinc and Other Metals. *Toxicol. Sci.* **1999**, *52*, 26−32.
(16) Morisseau, C.; Goodrow, M. H.; Dowdy, D.; Zheng, J.; Greene, J. F. et al. Potent Urea and Carbamate Inhibitors of Soluble Epoxide Hydrolases. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 8849−8854.
(17) Morisseau, C.; Newman, J. W.; Dowdy, D. L.; Goodrow, M. H.; Hammock, B. D. Inhibition of Microsomal Epoxide Hydrolases by Ureas, Amides, and Amines. *Chem. Res. Toxicol.* **2001**, *14*, 409−415.
(18) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.
(19) Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci* **1998**, *38*, 489−496.
(20) Mattioni, B. E.; Jurs, P. C. Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232−240.
(21) Bakken, G. A.; Jurs, P. C. Classification of Multidrug- Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.
(22) Mattioni, B. E.; Jurs, P. C. Development of Quantitative Structure−Activity Relationship and Classification Models for a Set of Carbonic Anhydrase Inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 94−102.
(23) Kauffman, G. W.; Jurs, P. C. QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.
(24) Bakken, G. A.; Jurs, P. C. QSARs for 6-Azasteroids as Inhibitors of Human Type 1 5a-Reductase: Prediction of Binding Affinity and Selectivity Relative to 3-BHSD. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1255−1265.
(25) Nakagawa, Y.; Wheelock, C. E.; Morisseau, C.; Goodrow, M. H.; Hammock, B. G. et al. 3-D QSAR Analysis of Inhibition of Murine Soluble Epoxide Hydrolase (MsEH) by Benzoylureas, Arylureas, and Their Analogues. *Bioorg. Med. Chem.* **2000**, *8*, 2663−2673.
(26) Newman, J. W.; Denton, D. L.; Morisseau, C.; Koger, C. S.; Wheelock, C. E. et al. Evaluation of Fish Models of Soluble Epoxide Hydrolase Inhibition. *Environ. Health Perspect.* **2001**, *109*, 61−66.
(27) Morisseau, C.; Goodrow, M. H.; Newman, J. W.; Wheelock, C. E.; Dowdy, D. L. et al. Structural Refinement of Urea Based Soluble Epoxide Hydrolases Inhibitors. *Biochem. Pharmacol.* **2002**, *63*, 1599−1608.
(28) Severson, T. F.; Goodrow, M. H.; Morisseau, C.; Dowdy, D. L.; Hammock, B. D. Urea and Amide-based Inhibitors of the Juvenile Hormone Epoxide Hydrolase of the Tobacco Hornworm. *Insect Biochem. Mol. Biol.* **2002**, *32*, 1741−1756.
(29) Grant, D. E.; Storms, D. H.; Hammock, B. D. Molecular Cloning and Expression of Murine Liver Soluble Epoxide Hydrolase. *J. Biol. Chem.* **1993**, *268*, 17628−17633.
(30) Beetham, J. K.; Tian, T.; Hammock, B. D. cDNA Cloning and Expression of a Soluble Epoxide Hydrolase from Human Liver. *Arch. Biochem. Biophys.* **1993**, *305*, 197−201.
(31) Wixtrom, R. N.; Silva, M. H.; Hammock, B. D. Affinity Purification of Cytosolic Epoxide Hydrolase Using Derivatized Epoxy-activated Sepharose Gels. *Anal. Biochem.* **1988**, *169*, 71−80.

(32) Dietze, E. C.; Kuwano, E.; Hammock, B. D. Spectrophotometric Substrates for Cytosolic Epoxide Hydrolase. *Anal. Biochem.* **1994**, *216*, 176−187.
(33) Jurs, P. C.; Chow, J. T.; Yuan, M. Studies of Chemical Structure-Biological Activity Relations Using Pattern Recognition. *Computer-Assisted Drug Design*; American Chemical Society: Washington, D. C., 1979; pp 103−129.
(34) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(35) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(36) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(37) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships For Toxicity Of Phenols Using Regression Analysis And Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(38) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange*; Program 455 ed.; Indiana University: Bloomington, IN.
(39) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(40) Dewar, M. J. S.; Zoebisch, E. G.; Healey, E. F.; Stewart, J. P. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(41) Aleman, C.; Luque, F. J.; Orozco, M. Suitability of the PM3-Derived Molecular Electrostatic Potentials. *J. Comput. Chem.* **1993**, *14*, 799−808.
(42) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(43) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, R. B. Search for All Self-Avoiding Paths on Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5−13.
(44) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley & Sons: New York, 1986; 18−20.
(45) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(46) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(47) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, *λ*. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(48) Pearlman, R. S. Molecular Surface Area and Volumes and Their Use in Structure/Activity Relationships. *Physical Chemical Properties of Drugs*; Marcel Dekker: New York, 1980; pp 321−347.
(49) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; 144−156.
(50) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(51) Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. In *Methods and Principles in Medicinal Chemistry*; 1 ed.; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: Weinheim, 2000; p 667.
(52) Todeschini, R.; Consonni, V. *DRAGON*; 1.11 ed.; Milano Chemometrics and QSAR Research Group: Milano, Italy.
(53) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.-Act. Relat.* **1997**, *16*, 309−314.
(54) Ruecker, G.; Ruecker, C. Counts of All Walks as Atomic and Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683−695.
(55) Todeschini, R.; Gramatica, P. 3D-modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113−119.
(56) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures. Theory. *J. Chemom.* **1994**, *8*, 263−272.
(57) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative-Structure Property Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(58) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; John Wiley & Sons: New York, 1980.
(59) Draper, N. R.; Smith, H. *Applied Regression Analysis*; 2nd ed.; John Wiley & Sons: New York, 1981.
(60) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.
(61) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(62) Shanno, D. F. Conditioning of Quasi-Newton Methods for Function Minimizations. *Math. Comput.* **1970**, *24*, 647−656.
(63) Goldfarb, D. A Family of Variable-Metric Methods Derived by Variational Means. *Math. Comput.* **1970**, *24*, 23−26.
(64) Fletcher, R. A New Approach to Variable Metric Algorithms. *Comput. J.* **1970**, *13*, 317−322.
(65) Broyden, C. G. The Convergence of a Class of Double-Rank Minimization Algorithms. *J. Inst. Math. Its Appl.* **1970**, *6*, 76−90.
(66) Kachigan, S. K. *Statistical Analysis*; Radius Press: New York, 1986.
(67) Bakken, G. A. Prediction of Chemical Properties and Biological Activities of Organic Compounds from Molecular Structure and Use of Pattern Recognition Techniques for the Analysis of Data from an Optical Sensor Array. In *Chemistry*; Pennsylvania State University: University Park, PA, 2001.
(68) Wan, C.; Harrington, P. B. Self-Configuring Radial Basis Function Neural Networks for Chemical Pattern Recognition. *J. Chem. Inf. Comput. Sci* **1999**, *39*, 1049−1056.
(69) Gupta, S.; Singh, M.; Madan, A. K. Superpendentic Index: A Novel Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci* **1999**, *39*, 272−277.

JM020269O