Prediction of HPLC Retention Index Using Artificial Neural Networks and IGroup E-State Indices

Daniel R. Albaugh,[⊥,†] L. Mark Hall,[⊥,‡] Dennis W. Hill,[†] Tzipporah M. Kertesz,[†] Marc Parham,[§] Lowell H. Hall,[||] and David F. Grant*[,†]

Department of Pharmaceutical Sciences, School of Pharmacy, University of Connecticut, 69 North Eagleville Road, Storrs, Connecticut 06269-3092, Hall Associates Consulting, 2 Davis Street, Quincy, Massachusetts 02170, Interactive Analysis, 6 Ruben Duren Way, Bedford, Massachusetts 01730-1666, and Eastern Nazarene College, 23 East Elm Avenue, Quincy, Massachusetts 02170

Received January 13, 2009

A back-propagation artificial neural network (ANN) was used to create a 10-fold leave-10%-out cross-validated ensemble model of high performance liquid chromatography retention index (HPLC-RI) for a data set of 498 diverse druglike compounds. A 10-fold multiple linear regression (MLR) ensemble model of the same data was developed for comparison. Molecular structure was described using IGroup E-state indices, a novel set of structure-information representation (SIR) descriptors, along with molecular connectivity chi and kappa indices and other SIR descriptors previously reported. The same input descriptors were used to develop models by both learning algorithms. The MLR model yielded marginally acceptable statistics with training correlation $r^2 = 0.65$, mean absolute error (MAE) = 83 RI units. External validation of 104 compounds not used for model development yielded validation $v^2 = 0.49$ and MAE = 73 RI units. The distribution of residuals for the fit and validate data sets suggest a nonlinear relationship between retention index and molecular structure as described by the SIR indices. Not surprisingly, the ANN model was significantly more accurate for both training and validation with training set $r^2 = 0.93$, MAE = 30 RI units and validation $v^2 = 0.84$, MAE = 41 RI units. For the ANN model, a total of 91% of validation predictions were within 100 RI units of the experimental value.

INTRODUCTION

The study of quantitative structure-retention relationships (QSRR) of solutes is an important topic in chromatographic thermodynamics. QSRR are the statistically derived relationships between the chromatographic parameters determined for a structurally diverse series of analytes in a given separation system and the descriptors accounting for the structural differences among the analytes studied. QSRR provides a promising method for the estimation of retention times based on descriptors derived solely from the molecular structure to fit experimental data.[1] This investigation details the development of two QSRR models for the prediction of high performance liquid chromatography retention index (HPLC-RI) from nonexperimental structural parameters. Although many studies[2−8] have been performed modeling retention index values, the QSRR models for this set of data are unique for four reasons: 1) the entire set of retention index values was determined in one laboratory, 2) the data set is very diverse compared to many published studies in which a related series of compounds was modeled, 3) the data set is highly druglike, and 4) a set of novel structure descriptors was used. The aim of this study was to derive a model to describe the chromatographic retention of druglike substances on a given chromatographic system, which can then be used for future retention predictions of new solutes.

Separation scientists use conventional relative measures of solute migration in individual chromatographic techniques. These are the well-known Kováts indices in gas chromatography (GC), logarithms of capacity factors in column liquid chromatography, R_F values in thin layer chromatography, and electrophoretic mobilities in capillary zone electrophoresis. Thin-layer chromatography and GC have long been used to screen biological samples for the presence of drugs. Identification of compounds by these methods relied primarily on the comparison of the retention characteristics of the compound to retention information found in previously collected databases. Recently, HPLC-RI has been used for drug screening. HPLC-RI has the advantage in that solute compounds need not be volatile, and band broadening and asymmetric elution can be circumvented by the use of appropriate desorption agents in the mobile phase.[9] One of the major advantages of HPLC-RI is the ease of transferability between laboratories through the use of homologous compounds such as n-nitroalkanes. An HPLC-RI is generated using a homologous series of compounds with increasing lipophilicity (such as a series of n-nitroalkanes), and the RI value is based on relative (rather than absolute) retention time when compared with compounds within the series eluting just prior to and just after the solute of interest. Generally small changes in structure result in measurable changes in the retention index. This is due to the ability of

* Corresponding author phone: (860)486-4265; fax: (860)486-5792; e-mail: david.grant@uconn.edu.
† University of Connecticut.
‡ Hall Associates Consulting.
§ Interactive Analysis.
⊥ Both authors contributed equally to this work.
|| Eastern Nazarene College.

10.1021/ci9000162 CCC: $40.75
Published on Web 03/24/2009

each of the atoms in a molecule to effect the relative distribution of the molecule between the stationary phase and mobile phase. For this reason, retention indices are frequently used to aid in characterization of compounds.

E-state indices are structure-information representation (SIR) descriptors that encode the electron accessibility of each atom in a molecule. The E-state is based on an intrinsic state for each atom that is modified by all other atoms in the molecule. This formalism for atoms has been extended to a bond-level index that encodes information about the electron accessibility of the bond between each pair of atoms in the molecule. The E-state indices encode both the electron distribution and local topology at each atom or bond in the atom-level or bond-level indices, respectively. The atom and bond level E-state indices have been grouped in various ways to form atom-type, hydrogen atom-type, functional-group type, group-type, internal hydrogen bonding, bond-type, and bond-group indices. A novel set of inputs was developed for this study using the SIR method of molecular description[10,11] based on the work of Hall and Kier.[12] A new grouping system was developed for this study called the interaction group (IGroup), which is a variation of the functional group type E-state indices.

The molecular connectivity indices are derived from representation of a molecule as a sigma-bonded network. As previously reported,[13,14] simple and valence delta values are used to label individual atoms according to relationships with every other atom in the molecule. A connectivity based topological profile can then be developed by deconstructing the molecule into sets of fragments of a given length of consecutive bonds, rings, or other subgraph branching patterns. In this way, the ramifications of skeletal variation are encoded in the molecular connectivity indices extending from simple branching patterns that differentiate among isomers, to the description of complex ring systems. The low order molecular connectivity indices are highly correlated to bulk properties such as volume, mass, and surface area, while the higher order indices tend to encode information about specific subgraphs within the sigma-network. The kappa shape indices are molecular connectivity descriptors that differentiate between molecules on the basis of varying shape[15] (e.g., globular, elongated, symmetrical).

The elementary structure information indices provide simple parameters such as formula weight, counts of atoms, rings and circuits, counts of bonds, and counts for elements that occur in organic molecules. Descriptors may also be categorized as global[16] or feature-based.[17,18] Global descriptors tend to describe the molecule as a whole and relate to properties of the molecule that are not generally associated with a specific region of the molecule or a specific interaction. Examples of global descriptors include molecular weight, volume, flexibility, branching, ring structure, and lipophilicity metrics. Global descriptors tend to have a nonzero value for every molecule in a data set. Feature based descriptors encode the presence of specific molecular substructures. These substructure features can be localized to a specific region of the molecule and are associated with specific interactions. Feature-based descriptors generally have a much lower population than global descriptors where the zero value is an empty-set zero representing the absence of the feature.

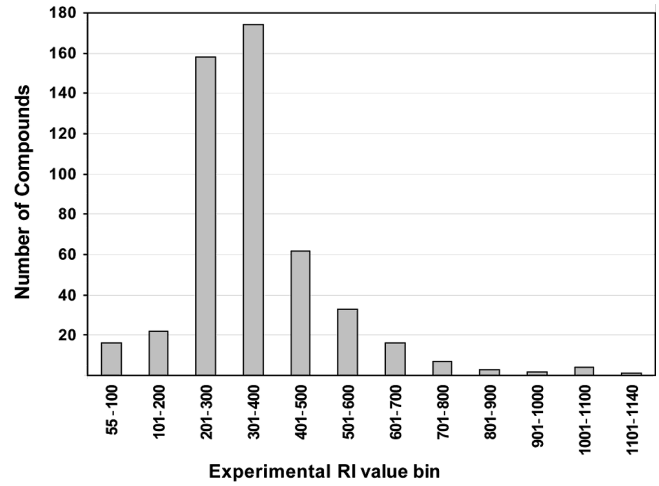

Figure 1. Distribution of retention index values for data set.

The models investigated in this study were 10-fold LGO 10%[19,20] (leave-10%-out) cross- validated ensembles generated using MLR and ANN learning algorithms. The ensemble approach[21,22] involves dividing available data into 10 subsets and training a separate model with data from each of the resulting data partitions. The final prediction of the ensemble is comprised of the average value of the 10 predictions for each compound.

DATA SET

Compounds and Source. The data set used in the study[9] contained 498 compounds with a distribution of retention indices indicative of a diverse set of chemicals as shown in Figure 1. Reagents were obtained from commercial sources and were of analytical purity or better.

Data Set Structural Profile. A list of 571 compound names and RI values were obtained from the study of Hill and Kind with private communication with the author.[9] Molecular structures were then obtained from the PubChem Project[23] Database by searching on the compound name. A total of 73 compounds were eliminated from the study because compound names could not be matched to confirmed structures in the PubChem Database. Data for the remaining 498 compounds were used for model building and validation. The structures used to generate the model descriptors were rendered in the neutral form with four compounds having a permanent positive charge on a nitrogen atom (e.g., quaternary amine). Table 1 gives an outline of the structural characteristics of the data set, including functional groups present. More than 99% of structures contain at least one ring, while nearly 82% contain an aromatic ring. The compounds contained an average of 4 hydrogen bond acceptors, 1.4 hydrogen bond donors, and 5.7 rotatable bonds. Nearly 40% of the structures contained a molecular configuration amenable to the formation of an internal hydrogen bond. The average values for formula weight, number of hydrogen bond donors, and hydrogen bond acceptors are consistent with druglike compounds.

Structural diversity in this data set is considerable. The formula weights ranged from 99 to 926 Da, and compounds vary from simple aromatic hydrocarbons, such as biphenyl and naphthalene, to alpha-amanitin, a cyclic polyamide, with dual fused 25-membered ring systems and 25 heteroatoms.

Table 1. HPLC-RI Data Set Structural Attributes

structure attribute	average[a]	number[b]	percent[c]
FW[d]	262.7	498	100%
ring[e]	2.38	494	99.2%
aromatic ring	1.08	408	81.9%
heteroaromatic ring	0.23	106	21.3%
nonaromatic ring	1.06	266	53.4%
fused ring system	0.44	221	44.3%
rotatable bonds	5.7	477	95.8%
heteroatom	4.3	495	99.4%
number H bond acceptor	4.01	495	99.3%
number H bond donor	1.4	388	77.9%
internal hydrogen bond	0.95	197	39.6%
TPSA[f]	58.9	495	99.4%
amine	0.42	202	40.6%
aniline	0.08	38	7.6%
pyridine N	0.20	77	15.5%
pyrole N	0.09	47	09.4%
ether	0.38	109	21.9%
ester	0.11	52	10.4%
ketone	0.17	62	12.4%
alcohol	0.35	108	21.7%
carboxylic acid	0.12	56	11.2%
phenol	0.24	93	18.7%
amide	0.25	108	21.7%
urea	0.09	44	8.8%
sulfonamide	0.09	42	8.4%
chlorine	0.15	62	12.4%
fluorine	0.09	24	4.8%

[a] Average count in the data set for the specified group. [b] Number compounds with specified attribute. [c] Percent of compounds in the data set with at least one example of the specified attribute. [d] FW = formula weight. [e] Compound contains a ring structure. [f] TPSA = static surface area of O, N, P, and S along with associated hydrogen atoms calculated by the Etrl method.

Table 2. Underpopulated Structure Features

structure attribute	number[a]	percent[b]
No Population Management[c]		
nitro	14	2.8%
vinyl CH	13	2.6%
Managed with IGroups[d]		
carbamate	9	1.8%
guanadine	4	0.8%
N-oxide	4	0.8%
benzamidine	3	0.6%
thiourea	3	0.6%
acetylene CH	2	0.4%
thioamide	2	0.4%
sulfuric acid	1	0.2%
Managed with IGroups and Explicit Description[e]		
furan	3	0.6%
oxazole	5	1.0%
thiophene	2	0.4%
thiazole	4	0.8%
thiol	1	0.2%
cyano	4	0.8%
Managed with Explicit Description, Manual Test-Set Assignment[f]		
N+ (permanent charge)	4	0.8%
bromine	4	0.8%

[a] Number of compounds with at least one example of the structure feature. [b] Percent of the data set with at least one example of the structure feature. [c] Low population features with enough examples to be included in an explicit descriptor. [d] Underpopulated features included in an IGroup descriptor. [e] Some atoms included in IGroups, nonincluded atoms encoded as explicit descriptors. [f] Underpopulated features included with explicit descriptor, these compounds were manually assigned to test sets to ensure that only one example was left out in any given test set.

Structural diversity and the presence of underpopulated features present significant challenges for creating a model of this data set. Because of the nature of the RI end point, it is optimal to have information about each atom explicitly described in the model indices. This is especially true for features containing heteroatoms. Statistical limitations on the number of input descriptors made it unlikely that every feature could be included in an independent fragmentlike index. It is also common for the list of input descriptors to be filtered in order to eliminate indices that have a nonzero value for less than a given percentage of the data set, commonly 3%−5%. Even when such low population indices are left in the descriptor pool, they are often ignored by input selection algorithms because they impact a small number of rows and have minimal or negligible leverage on the statistical outcome. Table 2 shows the underpopulated structure features present in the data set. The population is small for each of the listed features, but a total of 78 compounds, or approximately 16% of the data set, contain at least one of these underpopulated features.

A common method for managing the combination of diversity, underpopulation, and limited data is to eliminate the unusual structures from the study and constrain the applicability domain of the resulting model accordingly. This study will, in part, explore the use of the IGroup E-state indices as an alternative method for managing these issues. The IGroup E-state indices are a novel set of descriptors that combine information from similar structure features, thus reducing the total number of indices needed to encode information about a given group of structure features. The IGroup E-state indices are described in greater detail in the Methods section.

The structure features in Table 2 are given in four groups. The features in the first group have a low population but sufficient members for a specific descriptor to be included in the model without the need for special treatment in assigning the compounds to folds for the model test sets. The next two sets are not sufficiently populated to allow a specific descriptor to be included as some features have as few as one example in the data set. The inability to include explicit descriptors for these features arises, in part, from the use of a 10-fold ensemble model. In the creation of the 10 training folds, it is necessary to avoid the situation where a descriptor has the value of zero for all training rows, so caution must be used in including descriptors with so few examples. Two related methods were employed to encode information about these feature sets. The features listed in the set, "Managed with IGroups" were included by combining atom-level information from underpopulated functional group with atom-level information from similar functional groups of greater population. This was done for the features in the entitled "Managed with IGroups and explicit description" as well, but these features contained certain atoms that could not be reasonably combined with any feature of greater population. Information from these atoms was encoded in explicit descriptors, and compounds containing the features were "hand-assigned" to the 10 training folds to prevent the occurrence of a training fold with all zero values for any index. Information about the fourth group of features was included using explicit descrip-

tors and "hand-assignment" as there was no relevant IGroup related to these functional groups.

METHODS

Measurement Procedure. The HPLC retention characteristic of compounds used in this study was determined on a Zorbax, C-8, 4.6 × 250 column using a mobile phase consisting of 0.15 M phosphoric acid and 0.05 M triethylamine in water (mobile phase A) and 0.15 M phosphoric acid, 0.05 M triethylamine in 80% acetonitrile in water (mobile phase B). The flow rate was 2.0 mL/min, and compounds were analyzed by either a 20 or 30 min gradient. A homologous series of n-nitroalkanes (C1 through C10) was analyzed at the beginning of each batch analysis of compounds to establish the retention index scale. Acetophenone was coanalyzed with the nitroalkane calibration solution as well as the solutions containing the test compounds. A column test solution of representative compounds (morphine, amphetamine, methamphetamine, ethylmorphine, salicylic acid, desipramine, imipramine, phenylbutazone, and mefenamic acid) was used daily to check reproducibility of retention index determinations.

Retention indices were calculated as described,[9] in which the relative retention times were determined for each test compound and the nitroalkanes by dividing the retention time of the compound by the retention time of the acetophenone analyzed in the same system. The retention index was calculated by the formula given in eq 1

$$RI = \left(\frac{(RR_{tx} - RR_{tz})\Delta z 100}{RR_{tz+\Delta z} - RR_{tz}}\right) + 100z \qquad (1)$$

where RR_{tx} = relative retention time of test compound, RR_{tz} = relative retention time of 1-nitroalkane eluting just before test compound, $RR_{tz+\Delta z}$ = relative retention time of 1-nitroalkane eluting just after test compound, z = number of carbons in 1-nitroalkane eluting just before test compound, and Δz = carbon difference between 1-nitroalkane reference standards.

Measurement Reproducibility. Measurement reproducibility was determined by testing 66 drugs on both a Waters (Milford, MA) and Hewlett-Packard (Palo Alto, CA) HPLC systems. Analysis of the selected drugs were performed at different times (over a period of three years), on different columns using solvent systems prepared at different times. The retention indexes of the individual drugs varied between the two systems by an average absolute difference of 5.7 RI units (0 to 27 RI units).

Structure Descriptors. Structure descriptors used as modeling inputs were generated with the winMolconn[24] software package, which calculates a large number of molecular descriptors from the connection table of the submitted structure without the use of an optimized three-dimensional geometry. The calculated descriptors fall into several categories that include the electrotopological state (E-state), molecular connectivity, kappa shape, and elementary structure information indices. Examples of each class of descriptors were included in the MLR and ANN models.

When modeling a diverse data set, it is common practice to subject the data to a descriptor selection algorithm through which a large pool of descriptors is pruned down to an optimal subset. This common procedure presents a problem when modeling an end point such as retention index where every atom in the molecule contributes to the target value and many structure features may be underpopulated. The concept of selecting specific structure features to be included while excluding others is counterintuitive for modeling an end point of this kind. Since the retention index value can be used for compound identification, the underlying assumption exists that each atom in the molecule has a measurable influence over the retention time. If an atom in a molecule is changed, a different RI value would be expected from the modified structure. An optimal characterization of structure relative to RI value would necessarily contain information about each atom.

Two separate aspects of molecular structure must be described in order to encode the information relevant for modeling a chromatographic retention index end point. The first aspect consists of electronic features that determine the extent of noncovalent interactions between the solute and either the mobile or stationary phase. These interactions include electrostatic dipole, hydrogen bonding, and van der Waals dispersion. The second aspect consists of global (whole-molecule) bulk property characteristics such as molecular size, shape, flexibility, skeletal branching patterns, lipophilicity, and hydrophilicity metrics. Global indices tend to have a nonzero value for every molecule. Information encoded in the global indices is necessary to address specific aspects of the chromatographic measurement. Since isomers have different RI values, a description of branching patterns and ring structure is a necessary complement to the electronic feature information. In general, molecular shape and volume effect retention time independently of other electronic influences. Both the global and feature specific aspects of molecular structures were described using structure-information representation (SIR) indices.[10,11] SIR indices have been successfully used to model multiple end points ranging from ADME parameters and measures of toxicity to physical properties and receptor interaction strengths.[25−30] The IGroup E-state indices were used to describe the feature specific electronic information, while global structure parameters were encoded using a combination of molecular connectivity chi and kappa indices, graph counts, and some specific E-state indices.

The Interaction Group E-State Indices. Feature specific electronic information was encoded using a novel set of SIR indices called Interaction Group E-state Indices (IGroups). The IGroups were used, in part, to address the desirability of encoding electronic information about every atom in each molecule in order to adequately model the retention index. The goal of the IGroups system is to take the valence state electron accessibility for each atom, as explicitly encoded in the atom-level E-state, and then combine atom-level indices into group indices where the valence electrons could reasonably be expected to participate in similar noncovalent interactions in solution. As an example case, amide-like features include a carbonyl carbon with an alpha sp^2 oxygen and an alpha sp^3 nitrogen. The nitrogen may have up to two attached hydrogen atoms in its neutral form. The oxygen atom is part of a strong dipole and is a very strong hydrogen bond acceptor. The nitrogen is part of a weaker dipole and is a weaker acceptor. Any hydrogen present on the nitrogen may act as a hydrogen bond donor. Any of the atoms in the

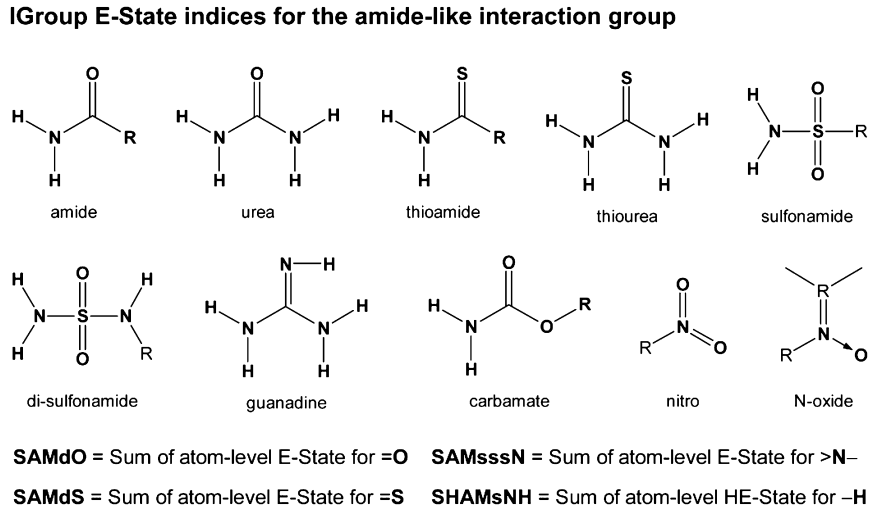

Figure 2. Amide-like functional groups, generally nitrogen with alpha sp^2 carbon and a beta withdrawing heteroatom, are reduced to four IGroup E-state descriptors. Oxygen in nitro and N-oxide is included in SAMdO, while the nitrogen atom is left out of the IGroup descriptors. If all combinations of hydrogen atoms on nitrogen are counted (1° amide, 2° amide, etc.) at least 39 fragments are required to encode the functional groups covered by these 4 descriptors. This does not include extended amide-like systems such as imides, pyrimidines, and barbiturates which would require many additional fragments.

group may participate in dispersion interactions. In the IGroup system, three indices are used to characterize this structure feature. The oxygen atom-level E-state is assigned to the SAMdO index (E-state, amide-like monovalent sp^2 oxygen). The nitrogen atom-level E-state is assigned to the SAMsssN index (E-state, amide-like trivalent sp^3 nitrogen), and the hydrogen atom-level HE-state for any hydrogen atoms is assigned to the SHAMsNH index (HE-state, amide-like −N-H hydrogen). These three indices form the amide-like IGroup.

If this approach was taken for all features similar to amide groups that are found in drugs, (amide, urea, thioamide, thiourea, sulfonamide, guanadine, carbamate) a total of 20 indices, in 7 IGroups, would be necessary to fully describe every atom in these groups. As an alternative, an assumption is made that the oxygen of a urea group will participate in noncovalent interactions similar to those of the amide oxygen. It follows that it may be reasonable for the amide and urea sp^2 oxygen indices to be merged into a single descriptor. This is a weighted sum in that the atom-level E-state value of the oxygen will not be the same in the amide and urea because the E-state is a reflection of variation in valence electron density at the oxygen resulting from the different local environments.

Extending the concept, the atom-level E-state values from sp^2 oxygen in amide, urea, sulfonamide, and carbamate are combined into the SAMdO descriptor. The atom-level E-state values from sp^3 nitrogen in amide, urea, thioamide, thiourea, sulfonamide, guanadine, and carbamate are combined into the SAMsssN descriptor. The hydrogen atom-level HE-state values from sp^3 hydrogen on the nitrogen in amide, urea, thioamide, thiourea, sulfonamide, guanadine, and carbamate are combined into the SHAMsNH descriptor. A separate index, SAMdS, is created for the sp^2 sulfur atom in thioamide and thiourea, and the atom-level E-state for the carbonyl carbon in all of these groups is added to a sp^2 carbon IGroup index.

Using this system, every atom in all of these groups, except for the tetravalent sulfur in sulfonamide, is explicitly encoded in one of 5 indices. The tetravalent sulfur was left out because it has minimal valence electron density and is stericaly buried. The sp^2 oxygen in nitro and N-oxide groups was also included in the amide-like SAMdO IGroup descriptor. As a result of implementing the IGroup method, a model-learning algorithm may weight the contribution of the different IGroup elements according to interactions that specific atoms may undergo instead of being limited to the application of a coefficient to the feature as a whole. The members of the amide-like IGroup are illustrated in Figure 2.

Considering that urea can have from zero to four hydrogen atoms on the associated nitrogen atoms, five fragment descriptors are necessary to differentiate among all possible urea configurations. Extending this to amide, urea, thioamide, thiourea, sulfonamide, disulfonamide, guanadine, and carbamate, 34 fragment descriptors are necessary to characterize all possible combinations of nitrogen and hydrogen atoms. Even if sufficient rows of data were available to allow for the use of a large number of inputs, there is a strong probability that the majority of these fragment descriptors would be underpopulated and difficult to force into a model. This situation highlights a specific problem that the IGroups were intended to address. Using traditional fragment counts, options are limited to the traditional approach of selecting some features to be in a model while leaving out others, or combining fragment counts, even though the fragments have different numbers of hydrogen, nitrogen atoms, and possible oxygen atoms. Use of the IGroup system allows all of the relevant information to be included in a reasonable number of descriptors. The IGroups also allow for the inclusion of extended amide-like systems like imides, barbiturates, and pyrimidines that have an extended conjugated system of nitrogen and carbonyl/thione groups. Oxygen and nitrogen atoms in such extended systems were included in the amide-like IGroup when connected to the feature by a sp^2 carbon or if an sp^3 nitrogen was alpha to a nitrogen already part of the feature.

In addition to the amide-like features, IGroups were formed for acids, anilines, aromatic nitrogen, amines, aliphatic oxygen, aliphatic sulfur, sp nitrogen, permanently charged nitrogens, and a group for bromine−iodine. IGroups

Table 3. HPLC RI Model Indices

index name	sum of atom level E-state or HE-state for
	IGroup E-State[a]
SAMdO	=O in amide like groups, nitro, n-oxide
SAMdS	=S in thioamide like groups
SAMsssN	>N- in amide like groups, guanadine
SHAMsNH	-H from amide like -NH
ACDdO	=O in acid groups
SACDsO	-O- in acid groups
SAniN	>N- in aniline groups
mSHAniNH	-H from aniline -NH
SAromN	>N- in pyridine/pyrrole groups
SHAromNH	-H from pyridine/pyrrole -NH
SalphN	>N- or =N- amine
SHalphNH	-H from amine -NH or -NH_2
SAlphO	O- aliphatic (ether, ester, alcohol, ketone)
SHAlphOH	-H from aliphatic -OH
SalphS	-S or =S aliphatic (thioether, thione, thiol)
SCspN	≡N (cyano, azide)
SallNp	permanent charge nitrogen
SsBrI	-Br bromine and -I iodine
Ssp3C	all sp3 carbon atoms
Ssp2C	all sp2 carbon atoms
SspC	all sp carbon atoms
SHacCH	-H in acidic CH
SaromC	all aromatic carbon atoms
	Atom-type E-State[b]
SaaO	aromatic oxygen
SaaS	aromatic sulfur
SsF	-F flourine
SsCl	-C chlorine
THB345	total internal hydrogen bonding path 3,4,5

index name	description
	Global Bulk Property Indices[c]
xv0	chi valence 0
xch5	chi simple chain 5
xch6	chi simple chain 6
xch7	chi simple chain 7
dx2	difference simple chi 2
ka3	kappa alpha 3
nrbond	number of rotatable bonds
ncirc	number of circuits
EPSA	E-state polar surface area
Hmax	largest hydrogen atom-level E-state

[a] A total of 23 IGroup indices were used. [b] A total of 5 atom-type E-state indices were used. [c] A total of 10 global descriptors were used including 5 molecular connectivity indices, 1 kappa shape index, 2 feature counts, 1 group-type E-state, and 1 single-atom hydrogen E-state.

for sp^3 carbon, sp^2 carbon, sp carbon, aromatic carbon, and acidic carbon hydrogen (allylic, acetylenic CH) were also used. The list of IGroups and their definitions are given in Table 3. It was necessary to include atom-type E-states for 4 individual structure features because they could not be reasonably included in an IGroup group. These included aromatic oxygen, aromatic sulfur, fluorine, and chlorine. A total of 27 indices were used to encode feature specific electronic information. A very small number of atoms were left out of this description. The hydrogen on one thiol was not assigned to a descriptor because there was only one in the data set, although the corresponding sulfur atom was assigned to the aliphatic sulfur IGroup. The sulfur of sulfonamide groups was also not assigned for reasons stated earlier as were the nitrogen atoms in nitro and N-oxide. The E-state internal hydrogen bonding descriptor THB345 was also included to encode the strength of a hydrogen bond in configurations where a 5-, 6-, or 7-member internal hydrogen bond may form, as the presence of an internal hydrogen bond would affect the capacity of a donor or acceptor to interact with the hydrophilic mobile phase. A list of all the IGroup indices is given in Table 3.

Global Descriptors. A total of 10 global descriptors were chosen for this project to describe the physical characteristics of a solute interacting with the stationary phase of a chromatography column. Molecular volume is correlated with RI, so the molecular connectivity index chi valence zero (xv0) was used because of its high correlation to molecular volume and surface area.[31] The connectivity index difference simple chi 2 (dx2) was used to characterize branching independent of size in order to differentiate among straight chain and branched skeletons such as are common in isomers. Since ring structures tend to have shorter retention times than their straight chain analogs, molecular connectivity indices simple chi chain 5, 6, and 7 (xch5, xch6, xch7) were used to characterize 5-, 6-, and 7-member rings. The count of circuits was also used to differentiate fused ring systems from individual rings. Globular structures also tend to have shorter retention times compared to linear structure of similar composition, so the shape index kappa alpha three (ka3) was also used. The ka3 index becomes larger as a structure becomes more linear (extended) and less highly branched, for the same number of atoms.[32] The count of rotatable bonds (nrbond) was used as a measure of molecular flexibility. Two E-state indices were also used as global descriptors. The EPSA (E-state Polar Surface Area) gives the summed atom level E-state values for nonhalogen heteroatoms. The Hmax index gives the largest single hydrogen atom-level E-state index in the molecule. This is generally the most acidic hydrogen in the structure. The E-state global indices may serve to categorize or classify molecules in the modeling process in a manner similar to a lipophilicity metric. A list of all the global indices is given in Table 3.

A total of 38 descriptors were used in the models generated in this study. There was no automated descriptor selection utilized for the project, as part of the purpose of the investigation was to examine the suitability of these indices to model a retention index end point. The valence chi versions of the chi chain indices (xch5, xch6, xch7) were evaluated as an alternative to the simple chi versions, but the simple chi were retained because no improvement in performance was observed.

Data Set Partitioning. A 10-fold, leave-10%-out (LGO, 10%) cross-validated ensemble model requires the available data to be partitioned into three subsets. A portion of the data is set aside for the purpose of model validation, and the remainder of the data, which will be used to fit the model, is partitioned into 10 training folds. Each training fold is a variation of the fit data with approximately 10% of the data left out as a test set. The training data from each fold are used to create a distinct model, the predictive capacity of which is evaluated using the 10% test set (leave-out set). After training, each of the 10 resulting models is used to predict the validation set, and the average of the 10 predictions is used as the final predicted value. It has been common practice to generate these subsets with a random

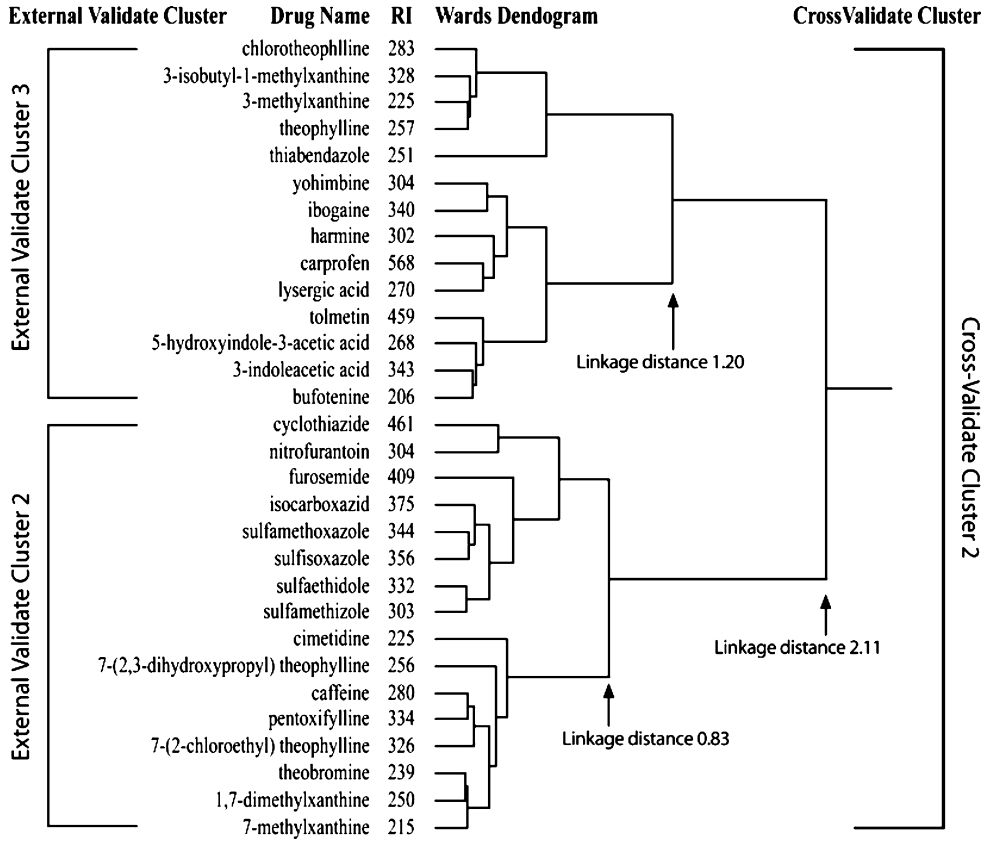

Figure 3. Ward's clustering dendrogram with drug name and experimental RI value for a 30 compound cross-validation cluster and associated external validation subclusters (16 and 14 compounds, respectively). External validation clusters were created by working from the left of the dendrogram until approximately 15 compounds are grouped. Cross-validation clusters are created by combining related external validation clusters. Small linkage distances indicate that the compounds in each cluster are highly related.

split. This study will examine the use of a clustering technique to balance the structure−activity space of the three subsets.

Wards Clustering. The HPLC-RI data set was analyzed by Ward's hierarchical clustering using the MDL QSAR software.[33] Clusters were selected from the output dendrogram by working back from the last level of output where each compound is in its own cluster. Related small clusters were grouped into larger clusters by looking at the experimental values of the related compounds. Figure 3 shows an example of the output dendrogram with the selection of a cluster of related compounds.

The target cluster size was 15−30 compounds. The use of 15 compound clusters allows for an example of the high, medium, and low RI range within a cluster to be set aside for validation without removing more than 20% of the data from a given cluster. Smaller clusters were allowed in cases where the activity range was small across all members of the cluster. A total of 27 clusters were produced, ranging in size from 11 to 35 compounds with an average of 21. Compounds containing an underpopulated structure feature were removed to a separate cluster for each feature. These clusters included thiazole/thiophene, permanently charged nitrogen, cyano, bromine, furan/oxazole, and thioamide/thiourea. The resulting 33 clusters were used to partition the data into a subset for model fitting and a subset for model validation.

Fit/Test and Validate Partition. The 498 compound data set was partitioned into two sets for model building. This was done by setting aside approximately 20% of the data (102 compounds) for use as an external validation set, with the remaining 80% (396 compounds) to be used as a fit set (for model fitting). The 20% validation subset was selected by rank ordering each of the 33 Ward's clusters on the experimental RI value and then taking every fifth compound from each cluster. The compounds with the lowest and highest activity values were not chosen for the validate set. Table 4 shows the partition of the compounds from cluster number three into "fit" and "validate" subsets. This process results in a validation set that is sampled from across the combined structure activity space of the available data. Figure 4 shows a histogram of the activity values in the entire data set as compared to the fit and validate subsets.

The RI values shown in Figure 4 have been subdivided into bins of 200 RI units. It is clear that the RI profile of the fit and validate sets are very similar to each other and to the overall data set. Further, the practice of leaving the lowest and highest activity value of each cluster out of the validate set has not distorted the RI profiles. This is largely because the second highest and lowest value is often very similar to the extreme values, and there are a number of clusters with a limited RI range across all members.

The fit subset of the data was further partitioned into 10-folds for use in building the ensemble model. This was done by removing the validation rows and then assigning each compound to one of 10 cross-validation test sets. The 10 cross-validation test sets are mutually exclusive; each fit compound is left out of exactly one training set and is present in 9 others. Table 5 shows the assignment of the fit set compounds in cluster three to the test sets of 10 training

Table 4. Fit/Validation Partition of Validate Cluster 2 Data

name	RIexp	cluster	group
17-methylxanthine	215	2	F
cimetidine	225	2	**V**
theobromine	239	2	F
1,7-dimethylxanthine	250	2	F
7-(2,3-dihydroxypropyl) theoph.	256	2	F
caffeine	280	2	F
sulfamethizole	303	2	F
nitrofurantoin	304	2	**V**
7-(2-chloroethyl) theophylline	326	2	F
sulfaethidole	332	2	F
pentoxifylline	334	2	F
sulfamethoxazole	344	2	F
sulfisoxazole	356	2	F
isocarboxazid	375	2	F
furosemide	409	2	**V**
cyclothiazide	461	2	F

[a] Compounds within each cluster are rank-ordered on the experimental RI value. Approximately one compound is assigned to the validate set (V) for every 5 compounds in the cluster. Remaining compounds are assigned for model fitting (F). The compounds with the lowest and highest value are not assigned to validate, but the difference between the lowest/highest value and the next value is often not large. The validate compounds are spread as evenly as possible over the activity range for the cluster. This method samples ~20% of the data for validation and helps to insure that the activity space and structure space are as broadly represented as possible.

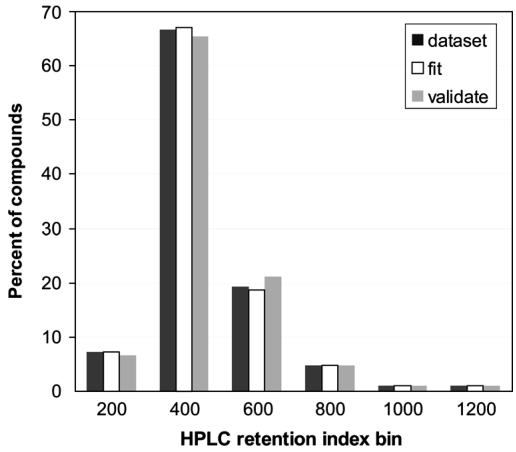

Figure 4. Histogram of activity values for entire data set, fit and validate subsets.

folds. The pattern in which the test set compounds are assigned is designed to prevent multiple compounds from the same activity range from being assigned to the same test set.

Multiple Linear Regression (MLR). The MLR model was generated using the JMP[34] software package to create a MLR of each training fold, and then the created model was used to test the validation set. This was done for each of the 10 training folds resulting in 10 predicted values for each validation row and nine fit calculations for each training row because each row is left out of exactly one training set. Summary training statistics for the ensemble MLR were calculated based on the average of the 9 fit calculations, and validation statistics were based on the average of the 10 predicted values.

Artificial Neural Networks (ANN). In a manner similar to the MLR model, the ANN ensemble model was created by training each fold separately and then taking average statistics across the ensemble. Modeling was done using the Emergent ANN software.[35] A standard back-propagation neural network with online weight update and sequential data loop order was used for all 10 models. The best model utilized a total of 38 input neurons, with a single hidden layer of 20 neurons, and an output layer of 1 neuron. The network was fully connected resulting in a total of 780 connections. Each connection has a single associated weight that is modified during model training. A fold of data was trained with a learning rate of 0.25 until a minimum in the mean absolute error (MAE) for the test set was reached. The learning rate was then lowered to 0.01, and training continued as long as the test set error continued to improve. The validation data was predicted after the final test set absolute error was reached. This process was repeated with 50 different sets of random starting weights for each fold, and the example with the best test set MAE was chosen. Average statistics were calculated across all 10-folds to generate the statistics for the ensemble model.

RESULTS

Statistical Information on the QSAR Models. Two statistical learning algorithms were utilized in this investigation. MLR was used as a baseline method. In addition to the MLR, ANN was used to investigate the possibility that the structure activity relationship was nonlinear. Statistics were generated for training set fit, cross-validation of the training data by 10-fold leave-10%-out, and external validation on a set of compounds not used for modeling. The statistical results obtained using a training set of 396 compounds and a validation set 102 compounds are summarized in Table 6.

The MLR model showed minimally acceptable correlation statistics with a training $r^2 = 0.65$, cross-validation $q^2 = 0.44$, and external validation $v^2 = 0.49$. The mean absolute error (MAE) for the MLR model was 83.6 RI units for the train, 83.5 RI units for the cross-validate, and 79.5 RI units for the external validate. This magnitude of error constitutes approximately 7−8% of the data range. While the model indicates a definite correlation between structure and the RI value, the fit and predicted values are only within 100 RI units of the experimental value 65%−75% of the time. The distribution of residuals for both the train and validate data sets suggests the possibility of a nonlinear relationship[36] between the RI end point and molecular structure as described by this set of structure indices. This is evidenced by the poor performance at both ends of the activity range. A plot of the training and cross-validation data points for the MLR model is given in Figure 5.

The ANN model shows a much higher correlation with a training $r^2 = 0.93$, cross-validation $q^2 = 0.76$, and external validation $v^2 = 0.83$. The MAE for the ANN model was 30.3 RI units for the train, 53.7 RI units for the cross-validate, and 40.8 RI units for the external validate. This magnitude of error constitutes 2.8% of the data range for the training set, 4.9% for the cross-validate, and 3.7% for the validate set and approximately half the relative error of the MLR model. A total of 91% of the ANN validation predictions were within 100 RI units of the experimental value. A plot of the training and cross-validation data points for the ANN model is given in Figure 6. An examination of the plot of

Table 5. Partition of Cross-Validate Cluster 2 (EV 2&3) into Train and Test (Cross-Validate) for 10 Folds

name	RI	EVC[a]	G[b]	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9
bufotenine	206	3	F	**S**	T	T	T	T	T	T	T	T	T
17-methylxanthine	215	2	F	T	**S**	T	T	T	T	T	T	T	T
3-methylxanthine	225	3	F	T	T	**S**	T	T	T	T	T	T	T
theobromine	239	2	F	T	T	T	**S**	T	T	T	T	T	T
1,7-dimethylxanthine	250	2	F	T	T	T	T	**S**	T	T	T	T	T
7-(2,3-dihydroxypropyl) theophylline	256	2	F	T	T	T	T	T	T	T	T	T	S
theophylline	257	3	F	T	T	T	T	T	T	T	T	**S**	T
5-hydroxyindole-3-acetic acid	268	3	F	T	T	T	T	T	T	T	**S**	T	T
lysergic acid	270	3	F	T	T	T	T	T	T	**S**	T	T	T
caffeine	280	2	F	T	T	T	T	T	**S**	T	T	T	T
harmine	302	3	F	**S**	T	T	T	T	T	T	T	T	T
sulfamethizole	303	2	F	T	**S**	T	T	T	T	T	T	T	T
yohimbine	304	3	F	T	T	**S**	T	T	T	T	T	T	T
7-(2-chloroethyl) theophylline	326	2	F	T	T	T	**S**	T	T	T	T	T	T
3-isobutyl-1-methylxanthine	328	3	F	T	T	T	T	**S**	T	T	T	T	T
sulfaethidole	332	2	F	T	T	T	T	T	T	T	T	T	S
pentoxifylline	334	2	F	T	T	T	T	T	T	T	T	**S**	T
ibogaine	340	3	F	T	T	T	T	T	T	T	**S**	T	T
3-indoleacetic acid	343	3	F	T	T	T	T	T	T	**S**	T	T	T
sulfamethoxazole	344	2	F	T	T	T	T	T	**S**	T	T	T	T
sulfisoxazole	356	2	F	**S**	T	T	T	T	T	T	T	T	T
isocarboxazid	375	2	F	T	**S**	T	T	T	T	T	T	T	T
cyclothiazide	461	3	F	T	T	**S**	T	T	T	T	T	T	T
carprofen	568	3	F	T	T	T	**S**	T	T	T	T	T	T
cimetidine	225	2	**V**	V	V	V	V	V	V	V	V	V	V
thiabendazole	251	3	**V**	V	V	V	V	V	V	V	V	V	V
chlorotheophlline	283	3	**V**	V	V	V	V	V	V	V	V	V	V
nitrofurantoin	304	2	**V**	V	V	V	V	V	V	V	V	V	V
furosemide	409	2	**V**	V	V	V	V	V	V	V	V	V	V
tolmetin	459	3	**V**	V	V	V	V	V	V	V	V	V	V

[a] External validate cluster. [b] Group (fit set or validate set). Cross-validate cluster 2 is made up of 2 related validate clusters from the same Wards tree. The validate clusters are combined and rank-ordered on experimental RI, and the validate compounds are set aside. The remaining compounds (fit set group) are partitioned into 10-folds of training "T" and test "S" subsets. Assignment to the test set is made according to the rank-ordered cluster list so that the activity range across each cluster is well sampled. This process results in training folds that are different but balanced in structure−activity space. This results in 10 pairs of data sets, each containing ~90% of the data for train and ~10% of the data for test. Each fit set compound is assigned to exactly one test and is included in the train set for the other 9 folds.

Table 6. Statistical Results for MLR and ANN Models

statistic	MLR	ANN	DIFF[a]
Training Set			
n[b]	396	396	−
r^{2}[c]	0.65	0.93	+0.28
MAE (RI units)[d]	83.6	30.3	−53.3
SE (RI units)[e]	86.5	38.9	−47.6
% < 100 (RI units)[f]	65%	98%	+33%
Cross-Validation Set			
n	396	396	−
q^{2}	0.44	0.76	+0.32
MAE (RI units)	83.5	53.7	−29.8
SE (RI units)	94.4	74.6	−19.8
% < 100 (RI units)	68%	86%	+20%
External-Validation Set			
n	102	102	−
v^{2}	0.49	0.83	+0.34
MAE (RI units)	79.5	40.8	−38.7
SE (RI units)	80.8	61.0	−19.8
% < 100 (RI units)	73%	91%	+18%

[a] Difference between MLR and ANN models. [b] Number of compounds in the indicated data set. [c] Square of the correlation coefficient. [d] Mean absolute error (RIU). [e] Standard error (RIU). [f] Percent of compounds with an absolute residual of less than 100 s.

the ANN model shows that performance is similar across the entire RI range for both the training and validation compounds. Although the two validation compounds with the largest RI values are both underpredicted, the magnitude of the residuals is not as large as the magnitude seen in the MLR model, and performance at the low end of the RI range is similar to the bulk of the data. The worst validation prediction for the ANN model was for the RI of papaverine (an opium alkaloid antispasmodic), which was overpredicted by 295 RI units.

DISCUSSION

Forensic toxicological analysis often deals with the identification of unknown substances in biological matrices. Analytical techniques, such as mass spectrometry, assist the toxicologist in the identification of these unknown substances. It has been shown that mass spectral comparison facilitates the identification of unknown substances,[37] and although mass spectral data are an indispensable tool for substance identification, some compounds show indiscriminate or very similar mass spectra. In this case, other tools such as retention index are necessary to adequately characterize an unknown substance. However, many databases do not contain HPLC retention index data. Therefore, it would be beneficial to have an automated means of predicting HPLC retention index from chemical structures. By designing this model to predict HPLC retention index from chemical structures, it has the ability to fill gaps left by existing prediction tools. However, it is important to keep in mind that the model did focus on drug molecules, and therefore the predicting capability may be limited to drug or druglike substances. Keeping in mind

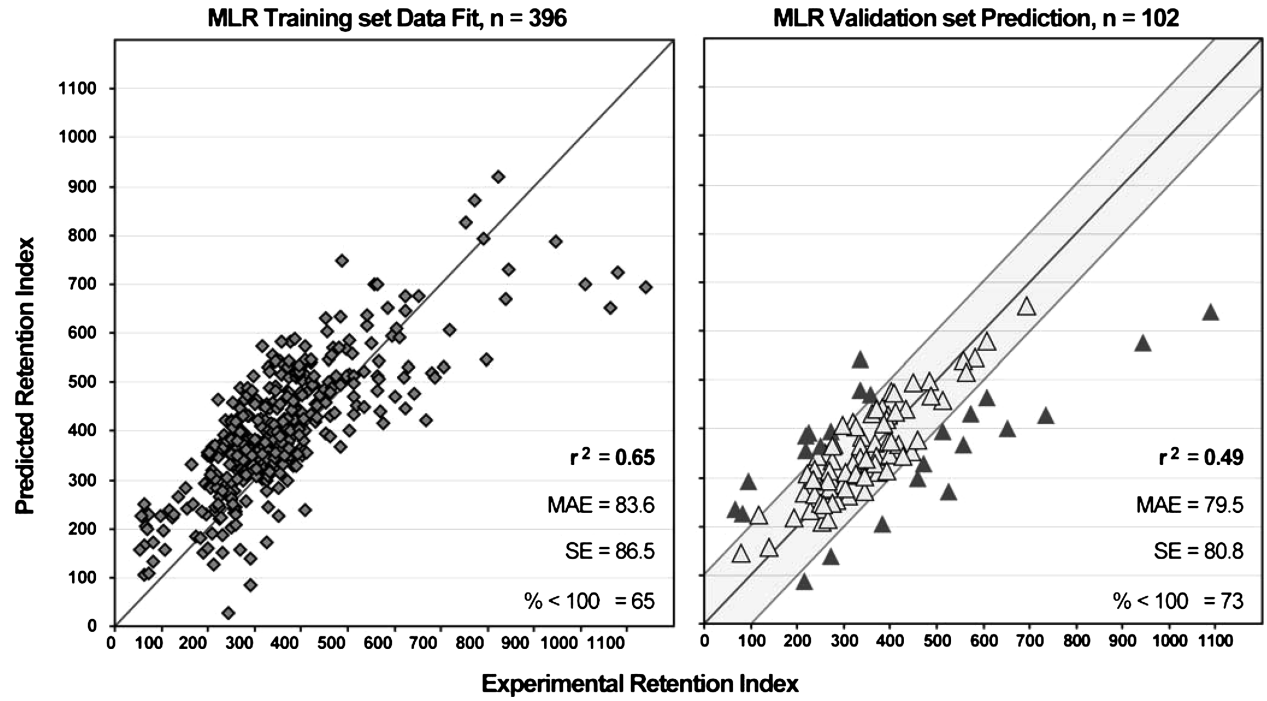

Figure 5. Multiple linear regression results for a training set and a validation set are shown with the squared correlation coefficient, mean absolute error, standard error, and percent of compounds with residuals less than 100 RI units. The shaded area on the validation plot represents ± 100 RI units from the experimental value. The 28 validation compounds with a predicted residual greater than 100 RI units are shown as dark gray triangles.

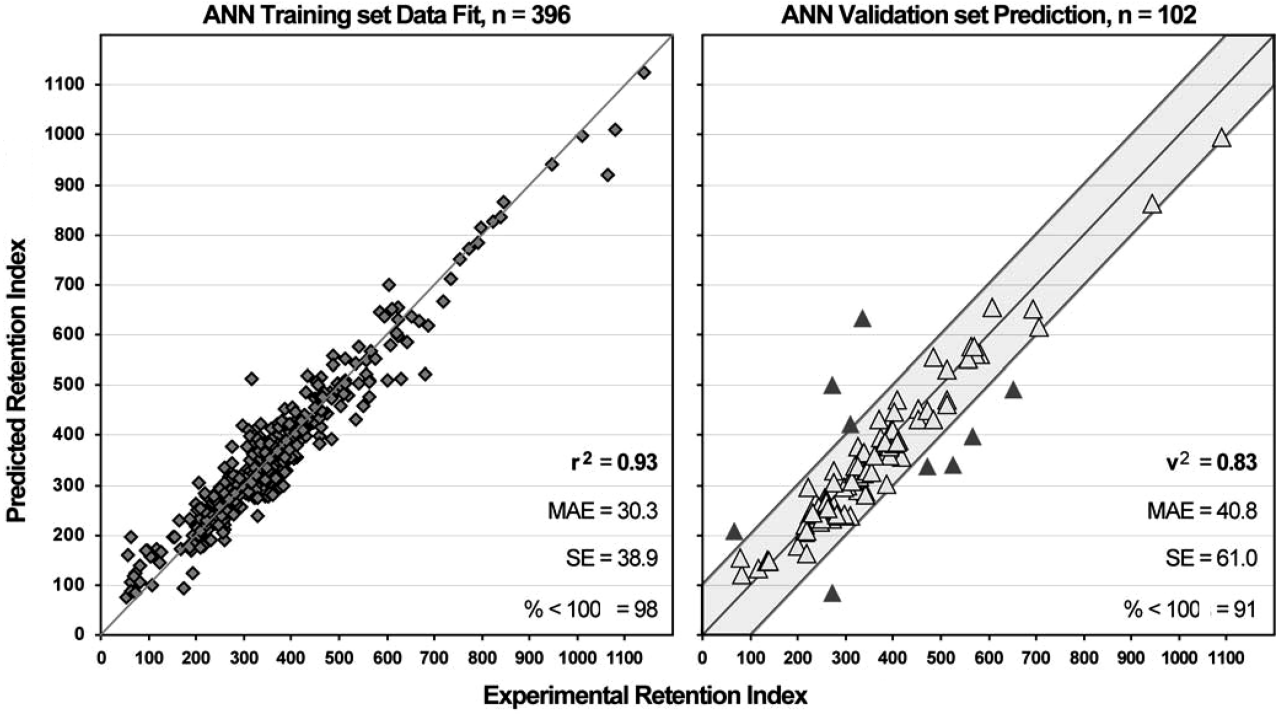

Figure 6. Artificial Neural Network results for a training set and a validation set are shown with the squared correlation coefficient, mean absolute error, standard error, and percent of compounds with residuals less than 100 RI units. The shaded area on the validation plot represents ± 100 RI units from the experimental value. The 9 validation compounds with a predicted residual greater than 100 RI units are shown as dark gray triangles.

the scope of the model, this approach offers a promising tool to support the identification of unknown compounds in both toxicological drug screening and metabonomics field.

One issue of interest arising from the development of the HPLC-RI models is the number of descriptors that were necessary. It is common practice to attempt to limit the number of input columns to a small number relative to the number of training rows. A conservative rule would set the limit at the square root of the number of data points. A less conservative approach is to maintain a minimum 10 to 1 ratio between training rows and input columns. This study utilized 38 input columns for approximately 356 rows of training data (folds vary slightly in number of rows). Within the context of the IGroups description

system, 38 was the minimum number necessary to avoid leaving out relevant structure information for this data set. The number of inputs arises, in part, from the structural diversity of the data and in part from the feature-based description method encoding nearly every atom. Although 38 input descriptors are far fewer than would be required to encode the same information using traditional fragment counts, it is still a sizable number relative to the data set size. The 356:38 (9.4:1) ratio is close to being within the 10 to 1 value recommended by some authors but is more than double the 18 that would be permitted by the much more conservative square root of data points convention. Experience has shown that a model will train well when too many input columns are used, but there will be a dramatic deterioration in validation statistics as a result of overfitting. This is especially the case with neural networks, which are easily overfit under most circumstances.[38] At an early stage of the modeling project, a more traditional descriptor reduction algorithm was used to prune a large descriptor pool. The resulting model utilized 25 inputs but was not as statistically robust at the final model with 38 inputs. The question arises as to why the final MLR model validates reasonable and the final ANN model validates well, independent of the somewhat large number of inputs.

One possibility is that maintaining a ratio near to 10 to 1 is sufficient for reasonable validation. A second possibility is that the population of input columns has a more significant impact than the number of columns. The use of the feature-based approach results in a large number of columns having a value of zero because of the absence of a given feature for a particular row. For the data set as a whole, each row has an average of 23 columns with a value of zero and only 15 columns with a nonzero value. The range for nonzero population runs from 7 column to 21 columns. In a case where a given row has 38 columns, but 23 columns contain a zero, should the number of inputs be evaluated as 38 or 15? In an MLR model, an input value of zero will cause the corresponding term to fall out of the expression. The neural network inputs for this project were scaled such that input values of zero were retained as zero and not scaled to a nonzero number. This causes the neural net to behave in the same way as a MLR equation with respect to empty-set input values. Since the first weight of the network is multiplied by the input value, the input node will not make any contribution to the outcome for that row. Since input values of zero do not have any influence on the outcome of the prediction for a given row, it is reasonable to consider whether the total number of columns may be an overestimate of the quantity of information being utilized for each row. If true, this would lead to the possibility of using a proportionally larger number of inputs relative to the number of data points than may have been previously supposed to lead to highly predictive models. We believe that this is a reasonable view of the modeling process, especially given that the statistical results for both the cross-validation and external validation sets are comparable.

Model Applicability Domain. In general, evaluating the potential of a model to make reliable predictions for novel compounds is a nontrivial task. There is no consensus on the optimal way to make such a determination. It is a widely held belief that the future performance of a model is constrained by the structure domain of the training data. According to the work of Jaworska,[39] "Models yield reliable predictions when the models' assumptions are met, and unreliable predictions when these assumptions are violated. The chemical space occupied by a training data set is the basis for estimating where reliable predictions will occur, because, in general, interpolation is more reliable than extrapolation.". This statement would tend to place significant importance on the size and structural diversity of the training data as an indicator of future model performance for use with a broad range of novel compounds.

In addition to diversity in the training data, confidence in future predictions may also be derived from the relative success of external validation predictions, the structure description method, and the model learning algorithm. A large, diverse, training set may tend to indicate a large applicability domain, but it may be equally important to implement an inclusive structure description method that describes (explicitly or implicitly) all relevant structure features. If relevant structure features that may be encountered in novel compounds are represented in the training data, and also encoded in model descriptors, it is reasonable to believe there will be an increased chance of success with future predictions. The use of nonlinear modeling techniques may also be helpful in that extrapolation along a nonlinear parameter may lead to more accurate interpolation at the edges of the training space where linear interpolation are often poor. A nonliner model parameter may also show some success with limited extrapolations for the same reasons.

As described in Tables 1 and 2, structural diversity was significant in the training data used to develop the ANN model for this present study. Since the structure description method was designed to encode all structure characteristics relevant to the RI end point, it is suggested that the ANN RI model could demonstrate a broad applicability domain in future use. The validation set was designed to include examples of every kind of substructure present in the data set. The good overall prediction of the validation structures may be another indication that the model could be successfully applied to an applicability domain of similar diversity to the overall data set. Accurate enumeration of the applicability domain may only be accomplished through independent validation with new data measured subsequent to the development of the model, and thus the specifics of the applicability domain must wait for further study.

CONCLUSIONS

Results in this study clearly demonstrate the utility of SIR IGroup and global SIR descriptors in the prediction of retention indices. The statistical results on the external validation set indicate that the ANN model is superior to the MLR model in making such predictions. The somewhat large number of descriptors relative to the size of the training data did not impair the predictive quality of the model. Over 90% of the predicted values are within 100 RI units of the experimental values. The molecular properties known to be relevant for HPLC-RI, such as molecular size, branching, and polar functional groups, are well represented by the data set. Overall, the developed model may be used to support the identification of unknown substances in cases where experimental HPLC-RI data are not available for candidate structures.

ACKNOWLEDGMENT

This work was partially funded by Grants from Pfizer, Inc., Groton CT, and the University of Connecticut Foundation.

REFERENCES AND NOTES

(1) Li, X.; Luan, F.; Si, H.; Hu, Z.; Liu, M. Prediction of retention times for a large set of pesticides or toxicants based on support vector machine and the heuristic method. *Toxicol. Lett.* **2007**, *175*, 136–144.
(2) Guo, W.; Lu, Y.; Zheng, X. M. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta* **2000**, *51*, 479–488.
(3) Jalali-Heravi, M.; Kyani, A. Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach. *J. Chem. Inf.Comput. Sci.* **2004**, *44*, 1328–1335.
(4) Acevedo-Martínez, J.; Escalona-Arranz, J. C.; Villar-Rojas, A.; Téllez-Palmero, F; Pérez-Rosés, R.; González, L.; Carrasco-Velar, R. Quantitative study of the structure-retention index relationship in the imine family. *J. Chromatogr., A.* **2006**, *1102*, 238–244.
(5) Hadjmohammadi, M. R.; Fatemi, M. H.; Kamel, K. Quantitative structure-property relationship study of retention time of some pesticides in gas chromatography. *J. Chromatogr. Sci.* **2007**, *45*, 400–404.
(6) Rouhollahi, A.; Shafieyan, H.; Ghasemi, J. B. A QSPR study on the GC retention times of a series of fatty, dicarboxylic and amino acids by MLR and ANN. *Ann. Chim.* **2007**, *97*, 925–933.
(7) Konoz, E.; Fatemi, M. H.; Faraji, R. Prediction of Kovats retention indices of some aliphatic aldehydes and ketones on some stationary phases at different temperatures using artificial neural network. *J. Chromatogr. Sci.* **2008**, *46*, 406–412.
(8) Quiming, N. S.; Denola, N. L.; Saito, Y.; Jinno, K. Multiple linear regression and artificial neural network retention prediction models for ginsenosides on a polyamine-bonded stationary phase in hydrophilic interaction chromatography. *J. Sep. Sci.* **2008**, *31*, 1550–1563.
(9) Hill, D. W.; Kind, A. J. Reversed-phase solvent-gradient HPLC retention indexes of drugs. *J. Anal. Toxicol.* **1994**, *18*, 233–242.
(10) Hall, L. H. A structure-information approach to the prediction of biological activities and properties. *Chem. Biodiversity* **2004**, *1*, 183–201.
(11) Hall, L. H.; Hall, L. M. QSAR modeling based on structure-information for properties of interest in human health. *SAR QSAR Environ. Res.* **2005**, *16*, 13–41.
(12) Kier, L. B.; Hall, L. H. *Molecular Structure Description, the Electrotopological State*; Academic Press: San Diego, 1999; pp 13−35.
(13) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(14) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley: New York, 1986.
(15) Hu, Q. N.; Liang, Y. Z.; Yin, H.; Peng, X. L.; Fang, K. T. Structural interpretation of the topological index. 2. The molecular connectivity index, the Kappa index, and the atom-type E-State index. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1193–1201.
(16) Wackermann, J.; Allefeld, C. On the meaning and interpretation of global descriptors of brain electrical activity. Including a reply to X. Pei et al. *Int. J. Psychophysiol.* **2007**, *64*, 199–210.
(17) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A novel shape-feature based approach to virtual library screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1230–1240.
(18) Abolmaali, S. F.; Wegner, J. K.; Zell, A. The compressed feature matrix--a fast method for feature based substructure search. *J. Mol. Model.* **2003**, *9*, 235–241.
(19) Maw, H. H.; Hall, L. H. E-state modeling of dopamine transporter binding. Validation of the model for a small data set. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1270–1275.
(20) Maw, H. H.; Hall, L. H. E-state modeling of corticosteroids binding affinity validation of model for small data set. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248–1254.
(21) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19*, 365–377.
(22) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J. Med. Chem.* **2006**, *49*, 7169–7181.
(23) The Pubchem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed Feb 17, 2009).
(24) *winMolconn*; Hall Associates Consulting: Quincy, MA, 2008.
(25) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
(26) Patankar, S. J.; Jurs, P. C. Prediction of IC50 values for ACAT inhibitors from molecular structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706–723.
(27) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.
(28) Roy, K.; Chakraborty, S.; Saha, A. Exploring selectivity requirements for COX-2 versus COX-1 binding of 3,4-diaryloxazolones using E-state index. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3753–3757.
(29) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19*, 365–377.
(30) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 931–939.
(31) Hall, L. H.; Kier, L. B. The Relation of Molecular Connectivity to Molecular Volume and Biological Activity. *Eur. J. Med. Chem.* **1981**, *16*, 399–407.
(32) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Relations. In *Reviews of Computational Chemistry*; 1991; pp 367−422.
(33) MDL QSAR version 2.2 b365; SymyxTechnology MDL: San Ramon, CA, 2003.
(34) *JMP, version 4.04*; SAS Institute Inc.: Cary, NC.
(35) Aisa, B.; Mingus, B.; O'Reilly, R. The Emergent Neural Modeling System. *Neural Networks* **2008**, *21*, 1045–1212.
(36) Daniel, D.; Wood, F. S. *Fitting Equations to Data*; Wiley-Interscience, John Wiley and Sons: New York, 1980.
(37) Hill, D. W.; Kertesz, T. M.; Fontaine, D.; Friedman, R.; Grant, D. F. Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra. *Anal. Chem.* **2008**, *80*, 5574–5582.
(38) Geman, S.; Bienenstock, E.; Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation* **1992**, *4*, 1–58.
(39) Jaworska, J. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA* **2005**, *3*, 445–459.

CI9000162