JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Development of $Ecom_{50}$ and Retention Index Models for Nontargeted Metabolomics: Identification of 1,3-Dicyclohexylurea in Human Serum by HPLC/Mass Spectrometry

L. Mark Hall,[†] Lowell H. Hall,[‡] Tzipporah M. Kertesz,[§] Dennis W. Hill,[§] Thomas R. Sharp,[§] Edward Z. Oblak,[§] Ying W. Dong,[∥] David S. Wishart,[∥] Ming-Hui Chen,[⊥] and David F. Grant*[,§]

[†]Hall Associates Consulting, Quincy, Massachusetts, United States

[‡]Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts, United States

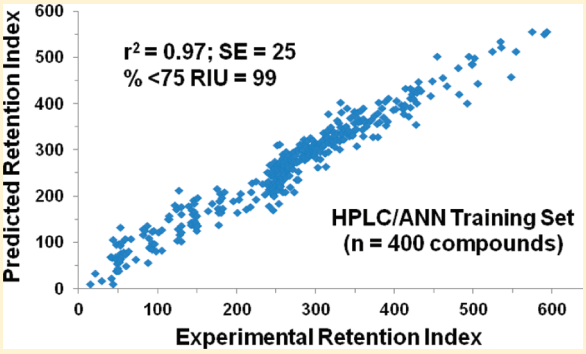
[§]Department of Pharmaceutical Sciences, University of Connecticut, Storrs, Connecticut, United States

[∥]Department of Biological Sciences, University of Alberta, Edmonton Alberta, Canada

[⊥]Department of Statistics, University of Connecticut, Storrs, Connecticut, United States

Supporting Information

ABSTRACT: The goal of many metabolomic studies is to identify the molecular structure of endogenous molecules that are differentially expressed among sampled or treatment groups. The identified compounds can then be used to gain an understanding of disease mechanisms. Unfortunately, despite recent advances in a variety of analytical techniques, small molecule (<1000 Da) identification remains difficult. Rarely can a chemical structure be determined from experimental "features" such as retention time, exact mass, and collision induced dissociation spectra. Thus, without knowing structure, biological significance remains obscure. In this study, we explore an identification method in which the measured exact mass of an unknown is used to query available chemical databases to compile a list of candidate compounds. Predictions are made for the candidates using models of experimental features that have been measured for the unknown. The predicted values are used to filter the candidate list by eliminating compounds with predicted values substantially different from the unknown. The intent is to reduce the list of candidates to a reasonable number that can be obtained and measured for confirmation. To facilitate this exploration, we measured data and created models for two experimental features; MS $Ecom_{50}$ (the energy in electronvolts required to fragment 50% of a selected precursor ion) and HPLC retention index. Using a data set of 52 compounds, $Ecom_{50}$ models were developed based on both Molconn and CODESSA structural descriptors. These models gave r^2 values of 0.89 to 0.94 depending on the number of inputs, the modeling algorithm chosen, and whether neutral or protonated structures were used. The retention index model was developed with 400 compounds using a back-propagation artificial neural network and 33 Molconn structure descriptors. External validation gave a v^2 = 0.87 and standard error of 38 retention index units. As a test of the validity of the filtering approach, the $Ecom_{50}$ and retention index models, along with exact mass and collision induced dissociation spectra matching, were used to identify 1,3-dicyclohexylurea in human plasma. This compound was not previously known to exist in human biofluids and its elemental formula was identical to 315 other candidate compounds downloaded from PubChem. These results suggest that the use of $Ecom_{50}$ and retention index predictive models can improve nontargeted metabolite structure identification using HPLC/MS derived structural features.



INTRODUCTION

Over the past decade, the field of metabolomics[1] has expanded to encompass a wide range of applications in genetics, environmental sciences, human health, and preclinical toxicity studies. Improvements in mass spectrometry and liquid chromatography have enabled high-throughput detection of a very large number of experimental peaks from complex biological samples. Measurements can be made in a short period of time and metabolites can be detected even when present at low concentration. Likewise, improvements in chemometric tools have enabled the high throughput extraction and comparison of raw data to determine which experimental "features" are significantly different between sample groups. In this context a feature (or "target feature") refers to a single observed experimental end point, or peak, which corresponds to a specific molecular structure. When a complex biologic sample is processed by HPLC and mass spectrometry, a very

Received: February 15, 2012
Published: April 10, 2012

large number of features are observed in the selected ion chromatograms. The ultimate goal is to match a specific molecular structure to each observed feature. Unfortunately, the process of structure elucidation for these features has remained time-consuming, costly, and is frequently unsuccessful.[2]

Currently, structure identification for mass spectrometry based metabolomics relies heavily on database searching. To facilitate the identification process, a number of metabolite databases have been compiled. These often contain detailed structural and chemical information of known metabolites[3−5] (i.e., molecular weight, elemental formula, logP, mass spectral fragmentation patterns, retention index). For example, the combined use of exact mass, retention index, and MS fragmentation pattern has been previously demonstrated to be a robust technique for identifying metabolites using an in-house database of these descriptors.[3] Because these metabolite databases are incomplete, their use precludes identification of metabolites or metabolite classes not contained in the database. Therefore, there are clear advantages in using chemical databases that contain a much larger number and greater diversity of compounds, including compounds that have not been confirmed to exist in biological systems. One such database, PubChem, consists of molecules compiled from more than 125 smaller databases[6,7] and currently contains approximately 30 million compounds with masses between 50 and 1000 Da. Obviously, it is not feasible for any single laboratory to measure experimental features (i.e., fragmentation patterns and retention indices) for such a large number of compounds. However, as previously described,[2] large databases such as PubChem can be annotated with chemical and structural information predicted by computational techniques, thus increasing the utility of the database for identification purposes. For example, computational predictions of logP are now included for most compounds in PubChem.

Previous reports have demonstrated the validity of identifying an unknown compound by matching its experimental collision induced dissociation (CID) spectra with predicted CID spectra of candidate compounds contained in databases.[8,9] In these cases, the CID spectra of the database-derived compounds were predicted using commercial and noncommercial software. The initial list of candidate compounds was obtained from PubChem based on their having the same exact mass as the unknown. All compounds with the exact mass of the unknown compound were extracted from the database, and their CID spectra were predicted. The predicted CID spectra were compared to the observed spectra and database compounds are ranked based on number of matches. This combination of accurate mass and accurate mass fragmentation prediction was sufficient for identifying a majority of the 102 compounds tested. Some structures were not correctly identified due to the inability of the software to accurately predict their fragmentation mechanisms. In addition, Hill et al. found no a priori way to determine which compounds or compound classes were poorly predicted by the software.[8] Current CID prediction software is not sufficiently accurate to discriminate unambiguously among all potential candidates contained in large databases. Thus, the prediction of additional orthogonal experimental features could be of tremendous value to aid in the identification of unknown compounds using HPLC/MS. Recently, Kumari et al. combined exact mass, CID fragmentation prediction, and retention index prediction for identifying metabolites using accurate mass gas-chromatography/time-of-flight mass spectrometry.[10] Although analyte derivitization required for gas chromatography can complicate data analysis, their results also suggest that chemical feature prediction is a paradigm worth pursuing for nontargeted metabolomics.

In the present study, $Ecom_{50}$ and retention index (RI) were evaluated as additional experimental features that are applicable to electrospray-based MS analysis and could augment exact mass and predicted CID spectra in structure identification. These features can be measured using HPLC/MS, and if the end points are successfully modeled using readily calculated structure descriptors, use of predicted values could improve database filtering efficiency. We reasoned that by comparing exact mass, along with experimental vs predicted values for $Ecom_{50}$, retention index, and CID spectra, we could better discriminate among candidate compounds from large databases such as PubChem and reduce the number of potential matches to a target unknown.

MATERIALS AND METHODS

Chemicals. Ethisterone was purchased from Steroloids (Newport, RI). HPLC grade methanol (Chromasol), trifluoroacetic acid, and 1,3-dicyclohexylurea (DCHU) were purchased from Sigma-Aldrich (St. Louis, MO). *N*-Cyclohexyl-2-methylpiperidine-1-carboxamide was purchased from Chembridge (San Diego, CA). HPLC grade acetonitrile was purchased from Fisher Scientific (Hampton, NH). Nitromethane, nitroethane, *n*-nitropropane, *n*-nitrobutane, *n*-nitropentane, and *n*-nitrohexane were purchased from Aldrich (St. Louis, MO). *n*-Nitroheptane, *n*-nitrooctane, *n*-nitrononane, and *n*-nitrodecane were synthesized as described by Aderjan and Bogusz (1988). HPLC grade heptafluorobutyric acid was purchased from Thermo Scientific (Rockford, IL). Reagent grade water was generated by a Barnstead Diamond reverse-osmosis water purification system. 1,3-Bis(cyclopentylmethyl)urea was synthesized and characterized as described in Supporting Information SI10. The sources of all other chemical compounds for which $Ecom_{50}$ and RI were measured are listed in Supporting Information SI1−SI3.

Biological Samples. Human serum and cerebral spinal fluid (CSF) were obtained from the Rocky Mountain Multiple Sclerosis Center (Englewood, CO). Samples were used in accordance with the University of Connecticut Institutional Review Board.

Detection of Unknown Compound in Human Serum Sample MSC205s. During an HPLC/MS analysis of human CSF taken from control patients and those with multiple sclerosis, an unknown compound with a retention time of approximately 13 min and an apparent protonated mass of m/z 225 (exact protonated mass = 225.1958 Da) was detected in many of the control and multiple sclerosis samples. The m/z 225 unknown compound was targeted for additional investigation because it was a significant peak in these samples. More importantly, a search of the 2009 Metlin,[4] HMDB,[5] DrugBank,[11] CSF-metabolome,[12] and KEGG[13] databases of known biological and bioactive compounds returned no matches at that exact mass. Thus, this unknown compound represented an ideal opportunity to test our approach. Further analysis of CSF samples was precluded by insufficient volume; however, a sample of human serum (labeled MSC205s) from a control patient was obtained and this sample was extracted and analyzed by HPLC/MS for presence of the m/z 225 compound. Here, 100 μL of the serum sample was pipetted

into a 1.5 mL polypropylene centrifuge tube (Eppendrof, Hauppauge, NY). Cold methanol (400 μL) was added to the tube, and the vortex was mixed for 1 min and allowed to incubate at −20 °C for 1 h. The sample was centrifuged at 14 000 rcf for 10 min, and an aliquot (450 μL) of the supernatant phase was transferred to a clean 1.5 mL polypropylene tube. The solvent was evaporated to dryness under reduced pressure in a Vacufuge (Eppendrof, Hauppauge, NY) evaporator at ambient temperature. The sample residue was dissolved in 25 μL of methanol by vortex mixing for 30 s and then diluted with 25 μL of reagent grade water with vortex mixing for 5 s. Then, 8 μL of the sample was analyzed on the described HPLC/UV system. The $Ecom_{50}$ value of a peak in the sample (detected at ∼13 min with a nominal m/z value of 225) was determined by reanalyzing the sample extract on the HPLC system in the MS/MS mode, isolating the m/z 225 ion, and collecting the CID spectrum at 17.0, 19.0, and 21.0 eV at 0.5 s intervals. The CID mass spectral profile was obtained on the m/z 225 ion at 25 eV. The $Ecom_{50}$ value of the m/z 225 compound was calculated as described by eq 2, and the retention index was calculated as described in the retention index measurement section. A mouse serum sample was processed by the same methods to ensure that the m/z 225 peak was not the result of a contaminant associated with our methodology. The mass of the m/z 225.1958 Da ion was assumed to be the protonated molecular ion of the unknown compound. The monoisotopic molecular weight (MIMW) of the unknown was therefore calculated as

$$\text{MIMW} = \text{protonated monoisotopic molecular weight} - 1.0073. \quad (1)$$

The MIMW for the m/z 225 ion was determined to be 224.1885 Da. In order to determine if the structure of this unknown could be identified using our database filtering method, predictive models were created for the $Ecom_{50}$ and RI experimental features and were used in conjunction with exact mass and CID fragment predictions.

Measurement of $Ecom_{50}$ Values. The $Ecom_{50}$ experimental feature was chosen as a compliment to exact mass and CID fragment spectra matching because recent work has shown that isomeric structures can be differentiated using this experimental feature.[14] In addition, $Ecom_{50}$ data is easily obtained during the collection of CID spectra data.

In preparation for the development of an $Ecom_{50}$ model, measurements were taken on a diverse set of 54 compounds listed in the Supporting Information SI1. Collision energy at 50% survival yield (CE_{50}) values were measured as described by Kertesz et al.[15] and converted to center of mass energy at 50% survival yield ($Ecom_{50}$) using the following formula:

$$Ecom_{50} = CE_{50}M_{rg}/(M_{rg} + M_x) \quad (2)$$

where M_{rg} is the natural isotope weighted mass of the reagent gas (argon in this case, 39.945) and M_x is the monoisotopic mass of the test compound's molecular ion. $Ecom_{50}$ values were used (rather than CE_{50} values) because $Ecom_{50}$ values are independent of the type of collision gas used and are thus more compatible with data obtained in different laboratories. The range of experimental $Ecom_{50}$ values was 0.70−7.82 eV. Structures of the neutral compounds for the $Ecom_{50}$ data set were downloaded from PubChem[16] as computer-readable three-dimensional structure data (SD) files and verified by cross-referencing with the Merck Index.[17]

Development of $Ecom_{50}$ Models. $Ecom_{50}$ models were developed using multiple structure description and modeling methods for comparison. Version 1.4 of winMolconn[18] was used to generate structure descriptors, which were used as input to SAS[19] for the generation of both multiple linear regression (MLR) and partial least-squares (PLS) models. Also, the CODESSA[20] software package (in conjunction with AMPAC[21]) was used to generate both structure descriptors and MLR models based on the CODESSA input selection algorithm. The $Ecom_{50}$ model data set and these two approaches are described in the following sections.

$Ecom_{50}$ Data Set Structure Profile. The compounds used to generate the $Ecom_{50}$ model contain from 1 to 5 nitrogen atoms in several functional group types. The number of oxygen atoms in a given molecule ranges from 0 to 11. Forty-eight molecules contain at least one oxygen atom. Oxygen is present in several functional group types including alcohol, ether, ketone, ester, acid, amide, and phosphate. Four compounds have permanently charged quaternary nitrogens, and one N-oxide is present. Molecular weight ranges from 88.2 to 607.7 Da with an average of 263.2 Da and standard deviation of 140.5 Da. Acyclic, cyclic, and multiple cyclic structures are present. The data set contains nitrogen atoms in 15 different chemical environments. This data set contains a useful overall variety of nitrogen atom types, but the groups are heavily loaded as alkyl amines as well as pyridines and amides. Ureas are under-represented and others such as imines, oximes, and nitriles are absent. In this sense, the set is not well-balanced in overall chemical space with respect to nitrogen functional groups. However, the data are sufficient to provide an initial evaluation for using $Ecom_{50}$ as a filtering option for nontargeted metabolomics that is approximately orthogonal to retention index. Two compounds that were initially considered for inclusion in the models (gallamine, a trication, and ATP, a triphosphate) were subsequently considered outside the chemical space of the 52 other structures and were removed from the model data set. A histogram of the measured $Ecom_{50}$ values for the 52 remaining compounds is given in Figure 1.

Protonation Site Selection. Models of $Ecom_{50}$ were generated from descriptors calculated for both the neutral and monoprotonated forms of the structures. Protonated structures were considered because positive ion electrospray produces

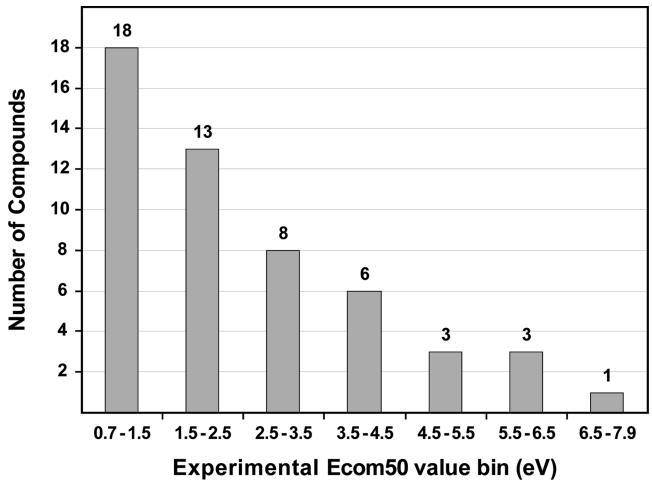


Figure 1. Distribution of measured experimental values for 52 compound $Ecom_{50}$ data set.

monoprotonated ions. Protonation alters the internal energy distribution of a molecule, so it was reasonable to infer that it could be useful to calculate descriptors based on the monoprotonated form. The actual protonation site is not known with certainty. Indeed, evidence exists that protonation site can be influenced by prevailing conditions in an LC-MS experiment such as mobile phase pH, aqueous−organic ratio, and buffer concentration.[22] However, the most likely protonation site for molecules in this data set is a nitrogen atom. For molecules that contain a single nitrogen atom, the proton was assigned to that atom. For structures that contain more than one nitrogen atom, it was necessary to evaluate protonation site options. Two methods were examined in this study: one based on partial charge and a second based on a combination of experimental and quantum chemically computed proton affinity.

To assign the protonation site according to partial charge, the SD file of the neutral structure was used to compute semiempirical quantum chemical geometry optimizations using the AMPAC[21] program and the AM1 (Austin Model 1) semiempirical model, originally developed by Dewar et al.[23−26] Results were verified by frequency calculation to ensure that a stationary state geometry (i.e., a conformation with at least a local energy minimum) was achieved. The electrostatic charge distribution of the geometry-optimized neutral molecule was examined. The atom showing the highest partial negative charge was chosen as the site to attach the H^+. Following protonation at that site, the geometry of the resulting cationic structure was reoptimized using the same frequency calculation criterion. Most structures in this test set contain an obvious single highly negatively charged atom. Several compounds show multiple atoms with negative partial electrostatic charges comparable in magnitude. The single site with highest partial negative charge was consistently chosen.

A second method assigned the site of protonation based on proton affinity. For those structures which contained more than one nitrogen atom, a combination of experimental and calculated proton affinity values was used to make the assignment. Proton affinity data available from NIST consists largely of molecules with a single nitrogen atom.[27] This data does, however, suggest the relative proton affinity of nitrogen in different chemical environments. For example, proton affinity values for alkyl amines are in the order tertiary > secondary > primary; also pyridine > pyrrole. However, when a nitrogen atom in a particular group occurs in varied electronic surroundings, the picture is less clear.

Quantum mechanical computations were carried out to compare computed proton affinities for the structures with multiple nitrogen atoms. Density functional computations at both the 6-31G* and the 6-311+G* level were used in addition to a semiempirical method (RM1).[28] All computations were done with Spartan-08 software.[29] Of the 52 compounds, 32 have nitrogen atoms in more than one type of electronic setting. For 22 of these 32 structures, it was reasonable to assign a protonation site on the basis of simple rules derived from experimental proton affinity data. An example is the protonation preference for alkyl amines in the order tertiary > secondary > primary and also for pyridines and ureas > pyrroles. Since there were combinations of nitrogen functional groups for which there was no experimental data, and confirmation of the simple rule assignments was useful, density functional theory DF(B3LYP 6-31G*) was used to compute the proton affinity for all 32 structures as:

$$\text{proton affinity} = E(\text{protonated}) - [E(\text{neutral}) + E(H^+)] \quad (3)$$

The energy of all possible monoprotonated nitrogen forms was calculated, along with the energy of the neutral form. For molecules which contain more than one nitrogen atom, the protonated species with the greatest proton affinity was taken to be the most stable. For nine compounds (isoniazid, allantoin, 3-methyl-L-histidine, *N*-methyl-L-histidine, 3-hydroxy-2-naphthone hydrazide, nalidixic acid, 8-hydroxy-2-deoxyguanosine, 4-dimethylaminopyridine, and ormetoprim), assignments were determined entirely on the basis of the greatest computed proton affinity. These computations confirmed the simple rule assignment for the other 22 structures, as the rule based assignment agreed with the calculated proton affinity in all cases. It is of note that, except for *N*-methyl-L-histidine and 4-aminohippuric acid, the semiempirical method RM1 gave the same protonation site indication as the DF 6-31G* computations.

Molconn Modeling of $Ecom_{50}$. Structure descriptors for the $Ecom_{50}$ compounds were generated from an SD file using winMolconn.[18] The representations of molecular structure calculated by winMolconn are generally referred to as topological structure descriptors and are divided into two broad categories, electrotopological state (E-State indices) and molecular connectivity (chi indices). E-State descriptors represent the electron accessibility at each atom, a combination of electron distribution and local atom topology. In addition to individual atom descriptors, the basic atom-level E-State information is developed into several types of structure descriptors including atom types, bond types, functional group descriptors, minimum and maximum atom-level values in a structure, and internal hydrogen bonding descriptors. The molecular connectivity (chi) descriptors represent features of molecular skeletal structure including degree of branching, ring structure, and overall skeletal ramification. For this modeling, the difference valence chi indices are used because they are independent of molecular size and represent skeletal variation, including the influence of heteroatoms. In contrast to the largely electronic information in E-State descriptors, chi indices primarily encode the topology of the molecular skeleton. This overall method of structure descriptor development has become known as Structure Information Representation (SIR).[30−36]

Molconn structure descriptors most closely related to the training set structure features were used to create a pool of descriptors available to the modeling algorithms. Descriptors were additionally filtered so that no pair shows a pairwise correlation greater than 0.75. For modeling on protonated structures, descriptors included the protonated form of the E-State index for all protonated nitrogen and quaternary nitrogen N^+ atoms. The names and descriptions of indices in the descriptor pools for Molconn modeling are given in Supporting Information SI4−SI7.

Model building and prediction were done with the SAS Statistical System 9.1.[19] To facilitate variable selection for MLR models, the SAS RSQUARE procedure was used. This method determines models for all possible combinations of a specified number of descriptors. RSQUARE does not use a stepwise procedure for descriptor selection. SAS REG was used for creation of multiple linear regression (MLR) models and associated statistics. SAS was also used to develop partial least-squares (PLS) models. In the PLS method,[37,38] the procedure

constructs linear combinations (latent factors) of all the descriptors in the pool such that each successively created variable (latent factor) improves the correlation with the dependent variable. The resulting latent factors are created as an orthogonal (independent) set. Up to eight latent factors were considered for PLS models and up to 8 descriptors were considered for MLR models. The best MLR and PLS models for the protonated and nonprotonated forms of the data set were kept for further analysis.

CODESSA Modeling of $Ecom_{50}$. Katritzky and co-workers explored structure−property and structure−activity relations using a multidimensional approach.[39,40] They introduced the CODESSA method, in which a large number of molecular descriptors can be calculated, based upon a quantum chemical optimization of a molecular geometry, and can be correlated with experimental results to find a structure−property correlation.

Model data for the CODESSA study was the same as for the Molconn study described above. SD files for both the neutral and protonated structures were used to compute semiempirical quantum chemical geometry optimizations using AMPAC with AM1. Results were verified by frequency calculation to ensure that a stationary state geometry (a conformation with at least a local energy minimum) was achieved. While doing the geometry optimization of the neutral and protonated structures, AMPAC was set to generate output results for use by CODESSA.[20] Quantum mechanical calculation results were then used to calculate CODESSA models for $Ecom_{50}$. CODESSA calculates upward of 500 molecular descriptors for each structure submitted for consideration. Modeling then consists of systematically examining each of the descriptors available for all structures to find the best empirical model(s), which are based on multiple linear regression (MLR) of the selected descriptors. The best eight variable MLR models for the protonated and nonprotonated forms of the data set were kept for further analysis.

Development of a Retention Index (RI) Model. Predictive models for RI have been previously developed for specific classes of compounds.[41−47] A model suitable for nontargeted metabolite structure identification would require the ability to predict RI values for a diverse set of biologically relevant small molecules. In this regard, we previously developed a robust HPLC RI predictive model demonstrating the potential for predicting RI for a structurally diverse group of compounds based on chemical structure.[30] The mobile phase used to measure this previously published data set is not appropriate for the currently proposed LC/MS metabolomic method because it is not compatible with electrospray ionization mass spectrometry. This necessitated the measurement of new data and the corresponding development of a new model.

RI Model Compounds and Source. In order to build a robust predictor of chromatograph RI for small molecules, it was first necessary to develop a training set of experimentally measured values. RI values were experimentally determined for a structurally diverse group of 411 small molecules consisting of endogenous compounds, endogenous metabolites, drugs, and drug metabolites. Compounds chosen for the study were limited to the set of biologic elements (C, H, N, O, S, and P) and contained at least one protonatable atom to facilitate detection by mass spectrometry in positive ion mode. For the majority of compounds, the protonatable atom was nitrogen. A total of 12 steroid-like compounds with alpha-beta unsaturated ketone groups were also measured because of the prevalence of this scaffold in biologic systems. The ketone group has been observed to protonate at electrospray pH levels allowing these compounds to be detected in positive ion mode.[30]

Retention Index Measurements. Retention index was determined for the 411 compounds on a Zorbax Stable Bond C-18, 1 mm × 150 mm (3.5 μm particle size) column (Agilent, Santa Clara, CA) with a mobile phase consisting of 0.01% (v/v) heptafluorobutyric acid (HFBA) in water (solvent A) and 0.01% HFBA in 90% acetonitrile (solvent B) using an Agilent 1100 capillary HPLC system (Agilent, Santa Clara, CA). A linear gradient from 0% solvent B to 100% solvent B at a flow rate of 75 μL/min was used to elute the compounds from the column. A 1 μL portion of 0.34 mM mixtures of approximately 20 test compounds per solution were analyzed in duplicate using gradient times of 17 and 35 min. At the beginning and end of each batch of sample analyses, 1 μL of a homologous mixture of C_1 through C_{10} *n*-nitroalkanes (1.12 μM mobile solvent phase B) was analyzed with the effluent column tubing attached to an Agilent 1100 diode array detector (Agilent, Santa Clara, CA) monitoring 210 nm.

For the analysis of test samples the tubing from the end of the column was attached to the electrospray source of a Micromass Q-TOF-2 (Waters Associates, Beverly, MA) mass spectrometer operated in positive ion mode. The electrospray source of the mass spectrometer was operated at a cone potential of 20 V, a capillary potential of 3.0 kV, and a source temperature of 120 °C. Ethisterone (0.3 mM) was coanalyzed with the nitroalkanes and the test samples as a retention reference to determine the void time difference between the diode array detector and the mass spectrometer. The retention times of test compounds were determined from their respective reconstructed molecular ion chromatographic peak.

The retention times of the respective *n*-nitroalkanes analyzed before and after the batch analysis of a set of test samples were averaged and used as the calibration references for calculating the RI of the test compounds. The offset in the retention time between the diode array and the mass spectrometer was calculated as the difference between the average retention time of ethisterone determined using the diode array detector and the retention time of ethisterone obtained on the mass spectrometer during a respective set of test compound analyses. This value was added to the retention time of test compounds analyzed on the mass spectrometer to correct for the void time difference between the two detectors. During solvent gradient analysis, the composition of the mobile phase remains constant at the starting composition until the mixing of the two mobile phases transverses the dwell volume of the system. During this period, compounds are eluted under isocratic conditions.

We have determined that a linear relationship exists between the log of the retention times of the homologous series of *n*-nitroalkanes and the number of carbon atoms in the structure of each individual compound in an isocratic reverse phase HPLC system. The retention index of each *n*-nitroalkane was defined as 100 times the number of carbon atoms. Therefore the retention indices of test compounds that eluted during the isocratic portion of the mobile phase were calculated by the following formula:

$$\mathrm{RI} = (\log T_x - \log T_z)100/(\log T_{z+1} - \log T_z) + 100z \quad (4)$$

where, T_x is the corrected retention time of the analyte, T_z is the retention time of the nitroalkane eluting just before the

analyte, T_{z+1} is the retention time of the nitroalkane eluting just after the analyte, and z is the number of carbons in the nitroalkane eluting just before the analyte.

The dwell time for the system used in this study was determined as the greatest retention time from the set of 411 compounds in which the retention time in the 17 min gradient, and the 34 min gradient was the same value. For this system, 1-methylpyrrolidine had a 17 min solvent gradient retention time of 5.54 min and a 35 min solvent gradient retention time of 5.53 min. The respective retention times for the next highest retained compound, 2-aminopyrimidine, were 5.83 and 6.01 min. Therefore a dwell time of 5.54 min was used as the end of the isocratic portion of the mobile phase, and this time corresponded to a retention index of 221.8. The C_3 through C_{10} *n*-nitroalkanes eluted during the gradient portion of the mobile phase. The trend function in an Excel spreadsheet was used to determine the fourth-order polynomial equation between the retention time and the retention index of the dwell time and the C_3 through C_{10} *n*-nitroalkanes. This equation was then used to calculate the retention index of each test compound from the corrected retention time. The assigned retention index for each compound used in the development of the retention model was the average value for duplicate analysis of each compound in the 17 min gradient system.

Experimental reproducibility of the RI measurements was calculated by performing 12 replicate measurements of 15 test compounds and generating a two-factors mixed effects model[48] from the resulting data. Using this method, the total measurement error for a given compound was calculated to be 2.24 RIU. This suggests an experimental reproducibility of ±6.7 RIU where, on a large number of replicate measurements on the same compound, 99.7% of measurements could be expected to fall within ±6.7 RIU of the mean. Repeatability was calculated to be 1.07, and reproducibility is 3.93. Total measurement error for a given compound is equal to repeatability + reproducibility, so it is calculated that reproducibility contributes 78.6% of the variability of measurement error.

Data Set Retention Index Profile. A histogram of the measured RI values is given in Figure 2. The measurement range was 578 RIU. Phosphoserine had the minimum value of 15 RIU, while carbazole had the maximum at 593 RIU. The mean is at 264 RIU with a standard deviation of 114 RIU. The distribution of measurement values is not normal, but the majority of the data points cluster around the mean. The largest RI value is 2.9 standard deviations above the mean, and the smallest is 2.2 standard deviations below the mean. The median value of 268 RIU is very close to the mean. It was not necessary to remove compounds from the study as experimental outliers since no values were more than three standard deviations from the mean value.

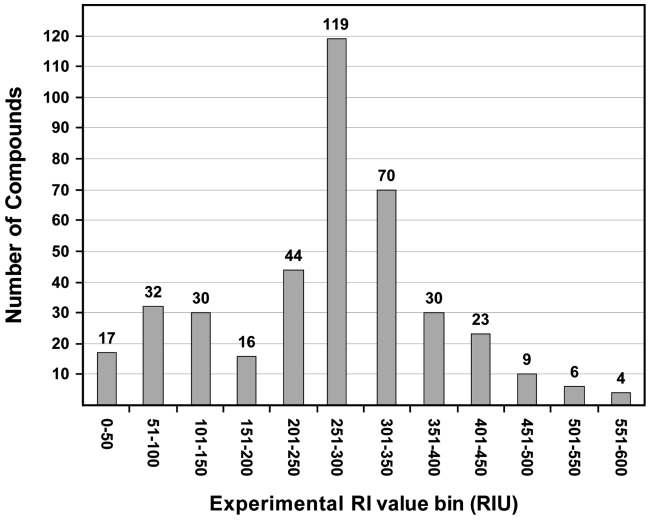


Figure 2. Distribution of measured experimental values for the retention index data set. RIU = 100z where z = the number of carbons in the nitroalkane series (see eq 4).

Data Set Structural Profile and Diversity. Compounds measured for the RI study have an average molecular weight of 218 Da with a minimum of 79 Da, maximum of 609 Da, and standard deviation of 94 Da. More than 40 common organic functional groups are observed in the data set. Additionally, there are many analogues of the common functional groups and a variety of carbon skeletal features such as phenyl, napthyl, cyclohexyl, and *t*-butyl groups (Table 1). There also exists in the data lower population functional subgraphs that are related to more highly populated functional groups such as amide. Some of these are "amide-like" features that contain at least one amide subgraph that is part of an extended conjugated system of additional carbon and nitrogen atoms. Examples in the data set include urea, imide, guanine, xanthine, cytosine, uracil, barbiturate, uric acid, carbamte, hydantoin, pyrazolidinone, and flavin (Table 2).

Table 1. HPLC-RI Data Set High Population Features

feature	count[a]	percent[b]	average[c]
MW[d]	400		218.2
ring[e]	319	79.8%	1.7
aromatic ring	186	46.5%	0.6
heteroaromatic ring	95	23.8%	0.3
nonaromatic ring	180	45.0%	0.77
fused ring system	128	32.0%	0.33
rotatable bonds	381	95.3%	6.1
heteroatom	400	100%	4.4
hydrogen bond acceptor	396	99.0%	4.1
hydrogen bond donor	343	85.8%	1.9
internal hydrogen bond	232	58.0%	1.8
TPSA[f]	399	99.8%	68.2
amine	212	53.0%	0.57
anilline	44	11.0%	0.12
pyridine	40	10.0%	0.11
pyrole	33	8.3%	0.09
ether	80	20.0%	0.32
ester	27	6.8%	0.08
ketone	33	8.3%	0.11
alcohol	107	26.8%	0.42
carboxylic acid	113	28.3%	0.32
phenol	41	10.3%	0.12
amide	73	18.3%	0.20

[a]Number compounds with specified attribute. [b]Percent of compounds in the data set with at least one example of the specified feature. [c]Average value in the data set for the specified feature (rings average = 1.7 indicates that compounds in the data set have an average or 1.7 rings). [d]MW = molecular 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[49] method.

Table 2. HPLC-RI Low Population Structure Features

feature	count[a]	%[b]	IGroup[c]
aldehyde	3	0.8%	aliphatic O
diazole	12	3.0%	pyridine, pyrrole
purine	12	3.0%	pyridine, aniline, pyrrole
guanadine	6	1.5%	aliphatic N
pyrimidine	4	1.0%	pyridine, aniline
diphenylamine	1	0.3%	pyrrole
carbamate	8	2.0%	amide, aliphatic O
oxime	1	0.3%	aliphatic N, aliphatic O
guanine	9	2.3%	amide, pyridine, pyrrole
xanthine	9	2.3%	amide, pyridine, pyrrole
cytosine	8	2.0%	amide
uracil	6	1.5%	amide
urea	6	1.5%	amide
barbiturate	4	1.0%	amide
uric acid	4	1.0%	amide
imide	3	0.8%	amide
hydantoin	3	0.8%	amide
flavin	2	0.5%	amide
pyrazolidinone	2	0.5%	amide
phosphate	16	4.0%	acid O
sulfonic acid	7	1.8%	acid O
sulfinic acid	1	0.3%	acid O
phosphonate	1	0.3%	acid O
thioether	9	2.3%	aliphatic S
thioanisole	6	1.5%	aliphatic S
disulfide	3	0.8%	aliphatic S
sulfoxide	1	0.3%	aliphatic S, aliphatic O
vinyl	13	3.3%	acidic CH
quaternary N+	9	2.3%	protonated N
pyridinium N+	5	1.3%	protonated N
thiazole	3	0.8%	aromatic S, pyridine
thiophene	2	0.5%	aromatic S

[a]Count of compounds with at least one example of the feature. [b]Percent of compounds with at least one example of the feature. [c]IGroup descriptor or descriptors where structure information for atoms in the group was assigned.

Structure Descriptors for RI Compounds. On the basis of the structure features of the compounds for which the RI was determined, a group of 33 structure information representation (SIR) descriptors[35] was selected for use in creating a predictive model. All descriptors values were calculated using the winMolconn software, v2.1.[50] An automated feature selection algorithm was not used to pick descriptors. Descriptors were chosen in an attempt to explicitly describe every feature of every molecule that would likely influence the retention time under the known measurement conditions.

Certain challenges exist when modeling a data set that is small compared to the level of diversity. The most significant issue involves the attempt to encode chemical information relative to the modeled end point in a statistically acceptable number of inputs. The Interaction Group E-State (IGroup) family of descriptors was utilized as an alternative to counting each possible variation of every functional group as a separate descriptor, which would have resulted in a very large number of low population inputs. The IGroups methodology[30] attempts to explicitly represent every atom in the molecule by combining atom level contributions from atoms that would be expected to participate in similar noncovalent interactions in solution. Functional subgraphs are categorized into groups such as the amide-like features described above. There are several other groups including acids (carboxylate, phosphate, phosphonate, sulfonate, sulfinate), anilines (aniline, benzamidine), aromatic nitrogen (pyridine, quinoline, pyrimidine), permanent charge nitrogen, and others. For each group, a separate descriptor is created for each noncarbon atom in the group. The amide-like IGroup descriptors consist of three indices: one for double bond oxygen, one for nitrogen atoms, and one for hydrogen bonded to nitrogen. The separate indices consist of the sum of the atom-level E-State value for all atoms of the type in the molecule. A list of the IGroups indices and corresponding atom types is given in Supporting Information SI8.

This method allows for the description of all subgraphs similar to the parent functional group in a small number of indices. Additionally, when a model is used for screening, subgraphs similar to those in the training data can be accounted for in a reasonable manner, even if the exact subgraph is not present in training data. Including amide, urea, imide, and other similar NC(=O) subgraphs, the amide-like group of features requires at least 100 different feature counts to enumerate all the possible combinations of nitrogen atom types that could be present in screening data. The IGroups method condenses this information into to three inputs that can be used to encode information for any subgraph of the amide-like type.

Along with the IGroup indices, a number of descriptors were added to encode information about size, shape, branch points, rings, flexibility, hydrophilicity, lipophilicity, and internal hydrogen bonding. These indices encode bulk structure features that likely influence partitioning characteristics during liquid chromatography. These descriptors and their definitions are listed in Supporting Information SI9. Two of these indices are new descriptors developed to characterize hydrophilicity and lipophilicity. The ratio valence hydrophilic index (rvalHyd) is the sum of the atom level E-State values of all hydrophilic atoms divided by the sum of atom level E-State values for every atom in the molecule. Since the atom level E-State value is a measure of the build up of valence electron density at an atom, this index describes the proportion of valence electron density that is associated with hydrophilic atoms in a size independent index. The sumLip index (total lipophilic estate) is the sum of the atom level E-State value for all lipophilic atoms in the molecule and is size dependent.

Input Normalization and Scaling. In preparation for modeling, descriptor values for the 33 inputs were normalized and scaled from 0.1 to 0.6. The maximum value of 0.6 leaves headroom in the event that a compound to be predicted has a larger value than was observed in the model data set. For feature descriptors, including the IGroups, a value of zero indicates the absence of the feature in the molecule. The 0.0 input values were preserved through scaling and normalization; also the mean and standard deviation used for normalization was taken on rows with a nonzero value for the input. After normalization, any compound with a normal value of 5 or greater for any input was excluded. These are compounds with a descriptor input value greater than 5 standard deviations above or below the mean value for other compounds with a nonzero value for the index. A total of 11 compounds were excluded on the basis of large normal values leaving a total of 400 compounds for modeling.

Data Set Partitioning. The remaining 400 compounds were partitioned into a 4 × 10 × 10 ensemble configuration by creating four validation subsets, each containing approximately 25% of the compounds. This procedure results in four splits of

the data, each with approximately 75% of the data for model fitting, and a unique 25% reserved for validation. The validation subsets were chosen so that the structural diversity and experimental activity were as similar as possible across all four sets. Each compound was used in exactly one validation set. The 75% fit set for each split was further subdivided into 10 train-test folds. Each training fold has a unique 10% of the 75% set aside for cross validation and the remainder used for neural net training.

Unlike the 1 × 10 × 10 ensemble described previously,[30] the use of a 4 × 10 × 10 ensemble allows each compound to be included in model training, validation, and cross validation. One of the four ensemble models gives a validation prediction for each compound, and each compound is used for model fitting in the other three models. Since the four validation sets are created by the same method, it is unlikely that any single set will yield a picture of model validation capability that is intrinsically more representative than the others. It is suggested that the four validation sets together give a more complete picture of prediction accuracy and limit the likelihood of bias associated with validation compound selection.

Neural Network Methodology. The Emergent ANN software package[51,52] was used to create a separate ensemble model for each of the four data splits. For each split, a model was developed for each of the 10 individual train-test folds. A three layer back-propagation net with 33 input neurons and 16 hidden neurons was used with online weight updates, 0.25 learning rate, and 0.9 momentum. For each fold, learning was conducted from a total of 50 different random initial weight starting points and the network was allowed to iterate through learning epochs until the mean absolute error (MAE) for the validation set reached a minimum. Training was stopped at this point, and statistics were calculated for train, test, and validate subsets. After learning from all 50 initial weight sets was completed, the run that produced the smallest validation MAE was kept as the model for that fold. The final model used for database filtering consists of the average predicted value of the individual models from the 40 total training sets.

Filtering of Candidate List of Possible Matches to the *m/z* 225 Unknown. A search of possible matches for the *m/z* 225 unknown (candidates with similar chromatographic and mass spectral characteristics) was performed by retrieving all compounds from the PubChem database (July 2009) with a MIMW of 224.1885 Da ± 10 ppm. The search was limited to compounds composed of carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorus, with the additional stipulation that the structure contain at least one protonatable atom. Redundant compounds and multiple diasteriomers and enantiomers were eliminated. The RI values and the $Ecom_{50}$ values of candidate compounds were predicted using the respective models described above. Additionally, the predicted protonated CID fragment ions for each candidate compound were calculated using Mass Frontier[53] in the rules prediction mode as described previously.[8] The list of candidates was filtered using the $Ecom_{50}$ and RI models with a filter range of ± three standard errors of prediction from the respective measured values. The ±3 SE range corresponds to a 99.8% confidence interval and gives a high probability that a compound matching the unknown will survive the filter process if present in the candidate list. On the basis of experimental values for the *m/z* 225 unknown ($Ecom_{50}$ = 2.76 eV, RI = 499 RIU), the resulting filter ranges were 2.76 ± 1.59 eV for $Ecom_{50}$ and 499 ± 114 RIU for RI. Candidates with predicted values outside the specified ranges were removed from the list and surviving candidate compounds were rank ordered on the number of predicted fragment ions that matched the measured CID fragment ions of the unknown.

RESULTS

$Ecom_{50}$ Modeling Based on Molconn Descriptors. For models based on Molconn descriptors, the use of PLS resulted in significantly better statistics than MLR (Table 3). The probable reason for this is that the set of structures in the training set covers a wide range of structure types and functional groups. On the basis of the typically accepted ratio of observations to statistical variables, the number of statistical variables should not be greater than eight for a data set of 52 compounds. Using eight descriptors in an MLR model places a limitation on the amount of structure information that can be included. On the other hand, PLS includes a larger fraction of the available structure information in the full pool of descriptors by making use of orthogonal linear combinations (latent factors) of all 75 descriptors in the pool. Our preliminary work indicates that up to seven latent factors could be used while maintaining statistical significance; beyond seven, little statistically significant improvement is observed.

Table 3. $Ecom_{50}$ Molconn PLS and MLR Modeling Results (Protonated and Neutral Structures)

model	n^a	$r^{2\,b}$	s^c	F^d	$q^{2\,e}$	$s_{press}{}^f$	inputs^g
protonated structures							
PLS	52	0.931	0.46	84.3	0.910	0.53	7
MLR	52	0.862	0.65	34.3	0.812	0.76	8
neutral structures							
PLS	52	0.905	0.54	59.9	0.855	0.67	7
MLR	52	0.848	0.68	30.7	0.762	0.90	8

[a]Number of compounds. [b]Correlation coefficient for training set calculations. [c]Standard error (RMSE) for training set calculations. [d]Fisher *F*-test for training set calculations. [e]Correlation coefficient for leave-one-out cross validation. [f]Standard error for leave-one-out cross validation predications. [g]Number of descriptor inputs.

For MLR modeling, the pool of descriptors was deceased to 68 by removing descriptors which are highly intercorrelated with other descriptors. The SAS RSQUARE procedure was run to select descriptors up to a maximum of 8. With these settings, the procedure performs MLR regressions on all possible combinations of one to eight variables. The statistically best models are reported and the model with the highest r^2 for eight variables was selected.

On the basis of these modeling results (Table 3), we focused primarily on the PLS model for protonated structures since the statistics for that model are significantly better than for the neutral structures. The direct statistics for the PLS model on protonated structures yields r^2 = 0.931, s = 0.46, F = 84.3, and n = 52. The leave-one-out method of cross-validation yields q^2 = 0.910 and s_{press} = 0.53. A plot of calculated versus measured $Ecom_{50}$ values (Figure 3) reveals a reasonable spread of predicted values about the line that represents the measured values. In the statstical sense, there are no outliers (residual >3 sigma) resulting from the PLS model for protonated structures. Also this plot shows the sparseness of data with $Ecom_{50}$ values above 3.50 eV (25% of compounds covering 68% of data range) and the much higher density of values below 2.00 eV (47% of compounds covering 18% of data range).

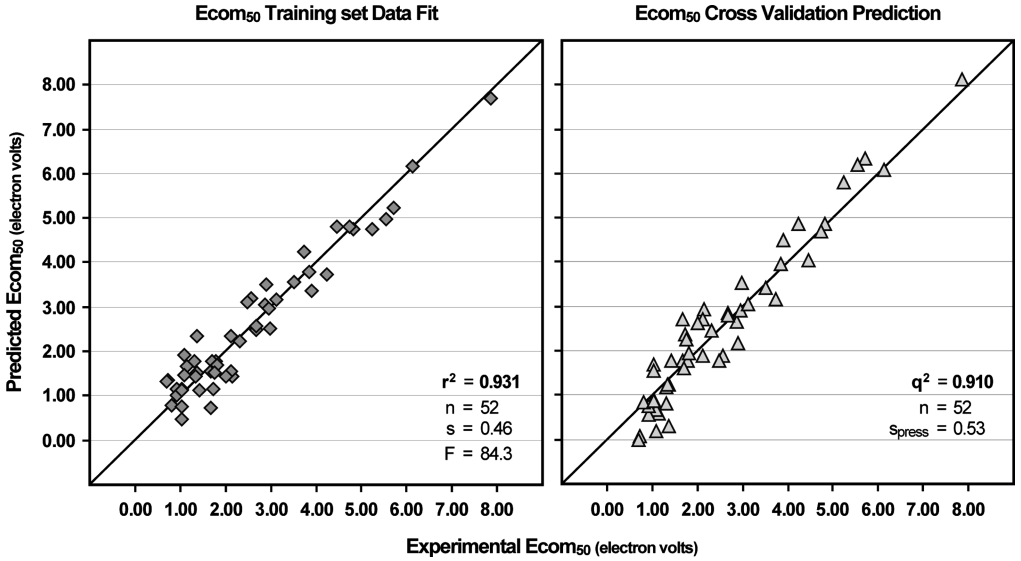


Figure 3. Molconn modeling results for $Ecom_{50}$ data giving experimental versus both calculated and predicted values from the PLS model (seven latent factors) using protonated structures. Model fitting calculation data points are given on the left and leave-one-out cross validation prediction data points are given on the right.

CODESSA Modeling. We developed a preliminary $Ecom_{50}$ model using CODESSA descriptors based on 3D structure representations. CODESSA modeling calculations were initially limited to finding the best model using five molecular descriptors. For neutral structures, several models were found with the best five-descriptor model, giving an r^2 correlation of 0.874. To obtain a model with an R^2 greater than 0.9, and to be consistent with the Molconn modeling, the number of allowed descriptors was increased to eight. The best eight-descriptor CODESSA model obtained had an r^2 = 0.920 (Table 4).

Table 4. $Ecom_{50}$ CODESSA MLR Modeling Results (Protonated and Neutral Structures)

model	n^a	r^{2b}	s^c	F^d	q^{2e}	inputs^f
protonated structures						
MLR	52	0.943	0.38	89.6	0.917	8
neutral structures						
MLR	52	0.920	0.45	63.2	0.882	8

[a]Number of compounds. [b]Correlation coefficient for training set calculations. [c]Sigma for training set calculations. [d]Fisher *F*-test statistic for training set calculations. [e]Correlation coefficient for leave-one-out cross validation. [f]Number of descriptor inputs.

In addition to modeling neutral structures, we also used CODESSA for developing an $Ecom_{50}$ model using protonated structures. Protonation sites based on both partial charge and proton affinity were assigned as described above. Appropriate structures files were created for compounds under the two criteria and CODESSA models built for $Ecom_{50}$ with both sets of structures. For the criterion whereby a proton is attached to the most electrostatically negative site on the molecule, the best eight-descriptor CODESSA model had an r^2 of 0.917 and q^2 of 0.882. For protonation at a site assigned by highest proton affinity, the best eight-descriptor model gave an r^2 of 0.943 and q^2 of 0.917 (Table 4).

A comparison of the statistics of the various models suggests that the use of protonated structures was of marginal value for the CODESSA modeling procedure. The training set standard error of the model based on protonated structures is 15% smaller than for the model based on neutral structures and the correlations coefficients are different by 0.03. For the purpose of creating CODESSA based models, it is questionable whether the additional computational effort to compute a minimized structure, examine its partial negative charge distribution or assign a charge site by proton affinity, and then recompute optimized geometries is warranted to obtain statistical improvements of that magnitude. Plots of the neutral and protonated models are given in Figure 4.

Retention Index Model. An artificial neural network (ANN) ensemble was used to generate the retention index model because previous work showed linear modeling by MLR to produce substantially inferior results.[30] Statistics were generated for training set fit, cross-validation prediction of the training data by 10-fold leave-10%-out, and external validation prediction on a set of compounds not used for modeling.

For each of the 4 data splits described in the Materials and Methods section, each compound was used for training in 9 of the 10 train-test folds. This resulted in 36 training fit values for each compound, each calculated by a different neural net model. All 36 values were averaged to generate the calculated 4 × 10 × 10 ensemble RI value from training. Training set calculations show a high correlation with measured data with r^2 = 0.95, mean absolute error (MAE) = 19 RIU, and standard error (SE) = 25 RIU. Each compound was left out of exactly one of the four data splits for use as validation data. Validation compounds were predicted by all 10 models from the split in which they were left out, and all 10 values were averaged to generate the predicted RI value from validation. External validation predictions are also well-correlated with experimental values with v^2 = 0.87, MAE = 30 RIU, and SE = 38 RIU. A total of 93% of validation predictions are within 75 RIU of the measured value (Table 5). The largest validation error was observed for bitolterol at 161 RIU.

Each compound was used for cross validation in one of the 10 folds of each of the 3 splits where the compound was part of the fit set used for training. This resulted in three cross-validation predictions for each compound which were averaged to generate the predicted RI value from cross-validation. Being

dx.doi.org/10.1021/ci300092s | J. Chem. Inf. Model. 2012, 52, 1222−1237

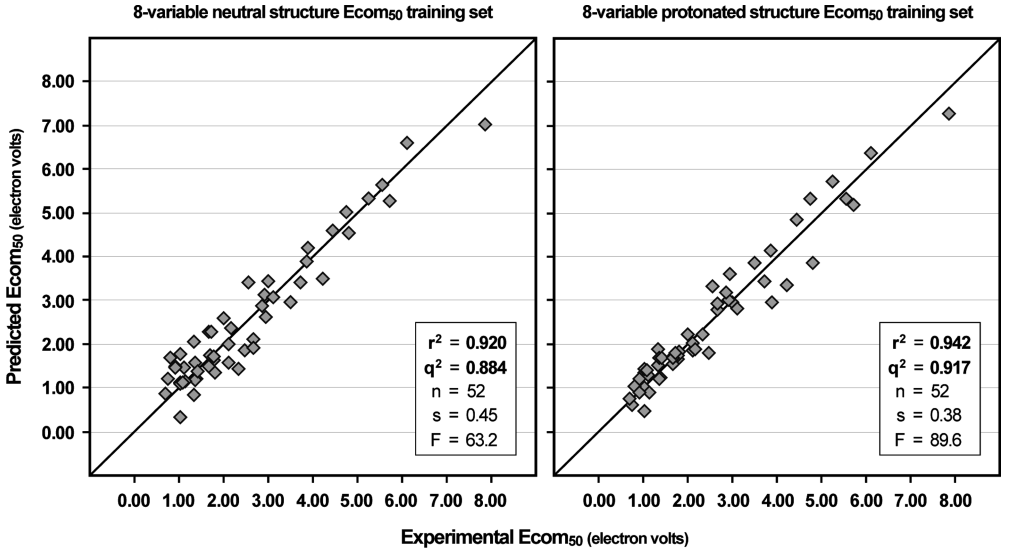


Figure 4. CODESSA training set results for $Ecom_{50}$ data. Plots show experimental versus calculated values from two eight-variable models. The left plot shows the model based on neutral structures and the right plot shows the model based on protonated structures. The q^2 leave-one-out cross validation correlation coefficient is also given.

Table 5. Retention Index ANN Modeling Results

subset	n^a	r^{2b}	MAE^c	SE^d	$<75_{RIU}{}^e$
train	400	0.95	19	25	99%
external validation	400	0.87	30	38	93%
cross validation	400	0.83	36	46	90%

[a]Number of compounds. [b]Square of the correlation coefficient. [c]Mean absolute error $=1/n\sum(|RI_{pred} - RI_{exp}|)$. [d]Standard error (root-mean-square error). [e]Percent of predicted values $\pm 75_{RIU}$ of the measured value.

an average of only three predicted values, the cross validation predictions show the lowest correlation with $q^2 = 0.83$, MAE = 36 RIU, and SE = 46 RIU. A total of 90% of cross-validation predictions are within 75 RIU of the measured value. These statistics are still quite reasonable given the small number of predictions that are averaged for this value.

Examination of the corresponding plots (Figure 5) shows that performance is similar across the RI range for both training and validation. Validation performance, however, is somewhat diminished above 400 RIU (MAE of 40 RIU as compared to 30 RIU for the overall data set), and the majority of predictions above 500 RIU are low. This could be the result of limited data in this experimental range, insufficient nonlinearity in the neural net architecture, or a deficit in the description of lipophilic features likely responsible for longer retention times.

Given the large number of features of low population, an analysis was performed to tabulate the performance of the

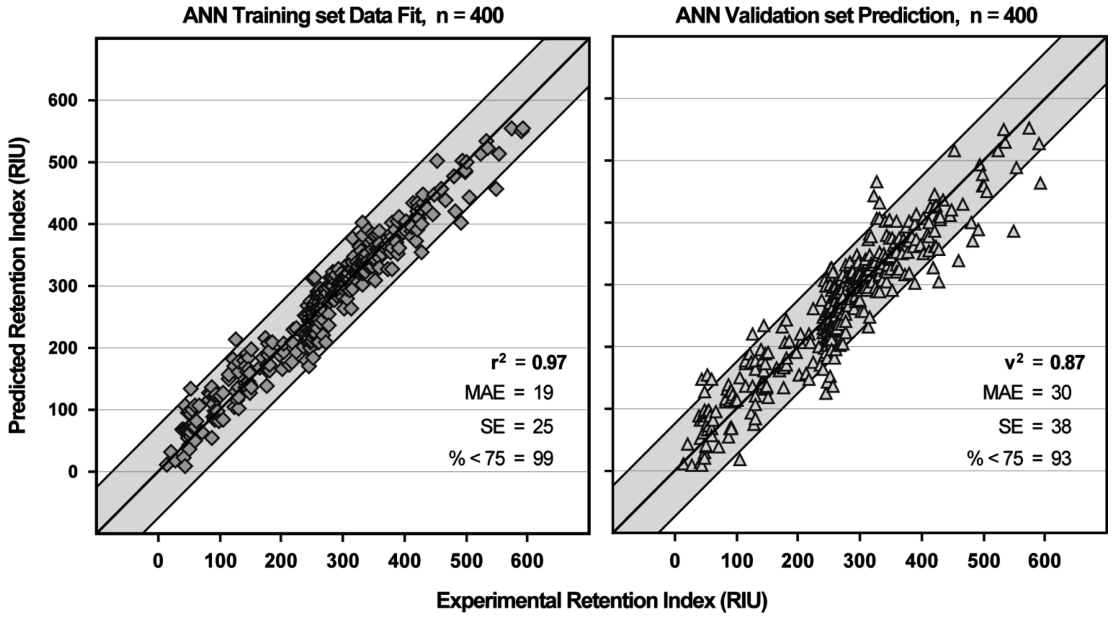


Figure 5. Artificial neural network results for training set and validation set are shown with the squared correlation coefficient, mean absolute error, standard error, and percent of compounds with residuals less than 75 RIU. The shaded region on the validation plot represents ±75 RIU from the experimental value.

dx.doi.org/10.1021/ci300092s | J. Chem. Inf. Model. 2012, 52, 1222−1237

model when predicting compounds with at least one low population feature (Table 6). The MAE for compounds with at least one low population feature is 32 RIU, which is only slightly higher than the overall MAE. The number of low population features in a compound does not appear to be a significant factor as predictions for compounds with multiple low population features appear to be of similar quality to the rest of the data. The largest MAE values occur in subsets where only one or two compounds with the feature are present in the data. These values are still well within two standard errors. The overall performance of the model in predicting compounds with low population features implies that the IGroup method of combining features with similar solution interaction characteristics into groups is at least partially successful in mitigating the issue of low population features.

Table 6. Mean Absolute Error of Prediction for Compounds with Low Population Features

group		count[a]	MAE[b]
overall data set		**400**	**30.4**
1−4 low population features		165	32.3
1 low population feature		140	32.2
2 low population features		23	33.4
3−4 low population features		2	26.5
feature	count	MAE	feature	count	MAE
phosphate	16	30.7	pyrimidine	4	35.9
vinyl	13	22.9	methyluric acid	4	24.0
diazole	12	48.9	aldehyde	3	15.4
purine	12	36.3	disulfide	3	21.9
xanthine	9	20.9	guanine	3	37.8
thioether	9	40.7	hydantoin	3	38.0
quaternary N^+	9	33.6	imide	3	19.8
cytosine	8	22.3	thiazole	3	31.3
carbamate	8	33.6	flavin	2	26.6
sulfonic acid	7	33.4	pyrazolidine	2	56.5
guanosine	6	31.9	thiophene	2	35.2
guanadine	6	33.7	phosphonate	1	19.0
uracil	6	28.9	diphenylamine	1	40.1
urea	6	52.9	oxime	1	65.1
thioanisole	6	24.9	sulfinic acid	1	49.3
pyridinium N^+	5	25.2	sulfoxide	1	8.4
barbiturate	4	25.4			

[a]Count of compounds with at least one example of the feature. [b]Mean absolute error for all compounds with at least one example of the indicated feature.

Filtering List of Possible Matches to Unknown *m/z* 225 in Human Serum Sample MSC205s. As stated previously, a compound was detected in human serum sample MSC205s with a measured protonated molecular mass of 225.1958 Da (MIMW 224.1885 Da), $Ecom_{50}$ value of 2.76 eV, and RI value of 499. The $Ecom_{50}$ and RI models developed in the study were used to filter 315 compounds with MIMW within ±10 ppm of 224.1885 Da that were retrieved from the PubChem database as possible matches to the unknown. The goal of filtering was to eliminate structures that were unlikely to be a match because their predicted $Ecom_{50}$ and RI values were greater than three standard errors of prediction from the measured values of the unknown.

Both the CODESSA and Molconn models were evaluated for $Ecom_{50}$ filtering. The Molconn PLS model based on protonated structures was chosen for final filtering because it eliminated a larger number of the 315 candidate structures as compared to the CODESSA models, or Molconn model based on neutral structures. The improved filtering efficiency was likely a consequence of the Molconn protonated model predicting a larger range of $Ecom_{50}$ values for the 315 compounds compared to the best CODESSA model (5.7 eV compared to 3.12 eV). Since the model statistics and corresponding three standard error filter ranges were similar, the smaller range of predicted values for the CODESSA model resulted in fewer compounds outside the boundaries of the filter range. At this point, however, the size of the $Ecom_{50}$ data set is too small to draw meaningful conclusions about the potential suitability of either method as they relate to predictive models of the $Ecom_{50}$ experimental feature.

Of the initial 315 compound candidate set, a total of 280 compounds passed the $Ecom_{50}$ filter, having predicted values within ±1.59 eV of the 2.76 eV measured value for the unknown (Table 7). Of these 280 compounds, 65 had predicted RI values within ±114 RIU of the measured RI value for the unknown. These final 65 compounds were rank ordered based on the number of matching predicted CID fragments. A match was defined as a predicted CID fragment ion mass within ±10 ppm of the mass of an experimentally observed fragment ion from the unknown. Of these 65 compounds, 11 were found to have 3 predicted fragments matches to the unknown spectrum. The same three matching predicted fragments were found in all 11 final candidates.

Table 7. Model Filtering Results on 315 PubChem Candidates for Match to *m/z* 225 Unknown

filter	range[a]	pass[b]
exact mass (from PubChem)	224.1885 Da ± 10 ppm	315
$Ecom_{50}$ filter	2.76 ± 1.59 eV	280
RI filter	499 ± 114 RIU	65
predicted CID fragment	3 matches	11

[a]Range of values used for filter, compounds outside of range do not pass. [b]Count of compounds passing filter.

The final list of 11 compounds was rank ordered on the absolute value of the difference between the predicted RI value and the *m/z* 255 unknown measured RI value. The predicted RI was chosen for final candidate ranking because of the larger data set for the RI model as compared to the $Ecom_{50}$ model. The rank order of the 11 final candidate compounds is given in Table 8 along with the PubChem CID number, name and filter predicted values. None of the top four final candidates was available for commercial purchase. The top ranking final candidate, 1,3-bis(cyclopentylmethyl)urea was synthesized according to the method described in Supporting Information SI10. Samples of both 1,3-dicyclohexylurea and *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide were commercially available and were also obtained. These three compounds (Figure 6) were analyzed experimentally to determine their $Ecom_{50}$ values, HPLC retention times, and CID mass spectral profiles for direct comparison with those of the unknown compound.

Confirmation of 1,3-Dicyclohexylurea in Human Serum. The MSC205s sample extract containing the unknown *m/z* 225 ion peak and solutions of standards for the three candidate compounds were sequentially analyzed on the HPLC system in MS/MS mode isolating the *m/z* 225 as a precurson ion and using a collision energy of 25 eV. For $Ecom_{50}$ determinations, samples and reference standards were analyzed

Table 8. PubChem Compounds of Mass 224.1885 ± 10 ppm with the Closest Match of Predicted $Ecom_{50}$, Predicted RI, and Predicted CID Fragments to the Experimental Data for the Unknown Compound *m/z* 225 Found in Human Serum[a]

		predicted[c]		predicted[e]			
PubChem CID[b]	compound name	RI	a-res[d]	$Ecom_{50}$	a-res[f]	predicted fragments[g]	fragment matches[h]
19071309	1,3-bis(cyclopentylmethyl)urea	492	7	2.77	0.01	20	3
23592100	5,6-ditert-butyl-2-methyl-2,3-dihydro-1H-pyrimidin-4-one	437	62	3.11	0.35	58	3
18384398	4-butyl-1-(2-ethylbutyl)-3*H*-imidazol-2-one	428	71	2.72	0.04	65	3
18384329	4-butyl-1-(3,3-dimethylbutyl)-3*H*-imidazol-2-one	423	76	2.73	0.03	56	3
4277	1,3-dicyclohexylurea	413	86	3.29	0.53	24	3
19906651	1,1-dicyclohexylurea	411	88	3.11	0.35	31	3
4207227	*N*-cyclohexyl-4-methyl-piperidine-1-carboxamide	410	89	3.46	0.70	25	3
2903296	*N*-cyclohexyl-2-methyl-piperidine-1-carboxamide	410	89	3.48	0.72	46	3
896432	(3S)-*N*-cyclohexyl-3-methyl-piperidine-1-carboxamide	409	90	3.48	0.72	31	3
21027155	1-cyclohexyl-3-[(*E*)-4-methylpent-2-enyl]urea	396	103	2.63	0.13	26	3
20497885	1-allyl-1-cyclohexyl-3-propylurea	385	114	2.77	0.01	66	3

[a]All compounds have the same molecular formula $C_{13}H_{24}N_2O$. [b]PubChem CID number. [c]RI model prediction. [d]Absolute residual, absolute value of difference between RI prediction, and measured RI value for unknown. [e]$Ecom_{50}$ model prediction. [f]Absolute residual, absolute value of difference between $Ecom_{50}$ prediction and measured $Ecom_{50}$ value for unknown [g]Number of fragment ions predicted by Mass Frontier software. [h]Number of predicted CID fragment ions matching measure fragments from unknown. The following matching fragments were predicted in all 11 candidates [predicted(experimental)]: 83.0855(83.0860), 100.1121(100.1117), 143.1179(143.1173).

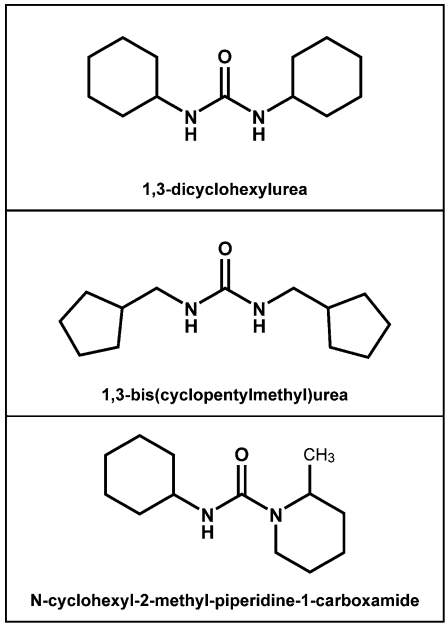


Figure 6. Three $C_{13}H_{24}N_2O$ isomer candidate compounds resulting from the PubChem database filtering process that were acquired and tested as possible matches to the unknown.

in duplicate on the same system collecting the CID spectrum at 17.0, 19.0, and 21.0 eV at 0.5 s intervals. The retention times, calculated $Ecom_{50}$ values, and CID fragmentation profile of the unknown were compared to those of each standard. A linear regression correlation coefficient between the unknown CID spectrum and that of each standard spectrum was calculated and used as a measure of the similarity CID spectral profiles. The HPLC retention and mass spectral data for the unknown compound in sample MSC205s and for candidate compounds 1,3-bis(cyclopentylmethyl)urea, 1,3-dicyclohexylurea, and *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide are shown in Table 9 and Figure 7.

All three candidate compounds are isomers and thus have the same theoretical protonated MIMW of 225.1961 Da. During the confirmation analysis, the measured protonated MIMW of

Table 9. HPLC Retention, $Ecom_{50}$, and Molecular Ion Mass for *m/z* 225 Ion in Test Sample MSC205s and Candidate Compounds[a]

sample ID	RT[b]	$Ecom_{50}$[c]	$(M + H^+)$[d]
MSC205s	12.947 ± 0.008	2.68 ± 0.00	225.1958
4277	12.946 ± 0.000	2.67 ± 0.01	225.1953
19071309	13.480 ± 0.008	2.75 ± 0.03	225.1952
2903296	13.111 ± 0.003	1.99 ± 0.03	225.1966

[a]MSC205s = human serum sample with *m/z* 225 unknown. 4277 = 1,3-dicyclohexylurea. 19071309 = 1,3-bis(cyclopentylmethyl)urea. 2903296 = *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide. [b]Measured HPLC retention time in minutes. [c]Measured $Ecom_{50}$ value in electronvolts. [d]Molecular ion mass (for monoprotonated ion).

1,3 dicyclohexylurea was 6 ppm of the theoretical value; whereas, the measured protonated MIMW of 1,3-bis-(cyclopentylmethyl)urea and *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide differered from this value by 2 and 11 ppm, respectively. The measured protonated MIMW of the unknown compound differed from the theoretical mass of the reference compounds by 2 ppm. The standard deviation of the mass accuracy of the mass spectrometer used in this study was determined in our laboratory to be −4.4 ± 8.9 ppm.

Table 9 lists the average retention times obtained for the unknown and three reference compounds. A *t*-test comparison of the average retention time of the unkown to that of the standards indicated a significant ($p < 0.005$) difference between the retention times of the unknown and 1,3- bis-(cyclopenylymethyl)urea and *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide but no significant ($p < 0.005$) difference between the retention time of the unknown and 1,3-dicyclohexylurea.

Figure 7 shows that the molecular ion peak and 4 fragment ion peaks in the 25 eV CID spectrum of the unknown are also present in the spectra of 1,3-dicyclohexylurea. The CID spectrum of 1,3-bis(cyclopentylmethyl)urea had 3 fragment ion peaks and the moleclar ion peak in common with that of the unknown. This compound also had an additional low intenisty fragment ion peak at *m/z* 55.1 that was not detected in the unknown spectrum. The CID spectrum of *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide shows a very low intensity

dx.doi.org/10.1021/ci300092s | J. Chem. Inf. Model. 2012, 52, 1222−1237

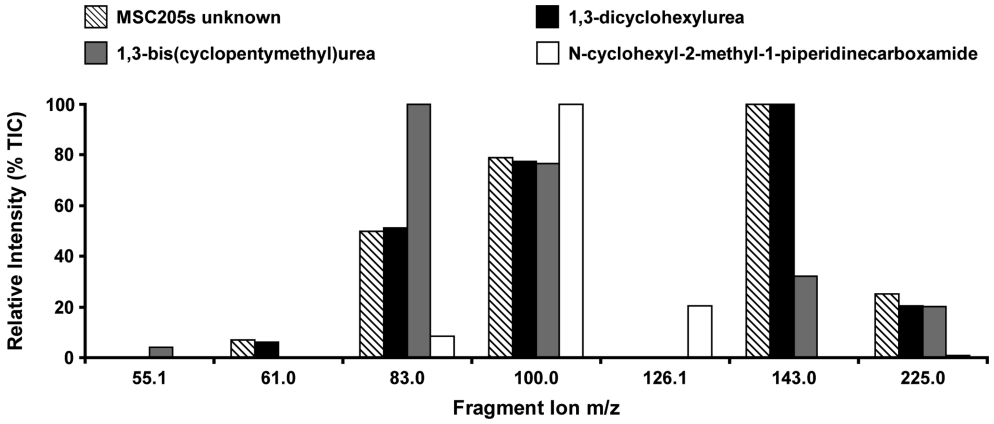


Figure 7. Comparison of the CID fragmentation profile of unknown m/z 255 compound in MSC205s to those of the 1,3-bis(cyclopentymethyl)urea, 1,3-dicyclohexylurea, and *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide standards.

ion peak for the moleuclar ion, two ion peaks in common with the CID spectrum of the unknown, and one ion peak that is not present in the other two standards or the unknown CID spectra. The measured masses of all common fragment ion peaks between spectra differed by less than 17 ppm. For a complete qualitative comparison of the mass spectral profiles, the correlation coefficient of the relative intensities of the matching ion peaks between the CID spectrum of the unknown and those of the three standards was calculated. The cross correlation value approaches 1 as the relative intensities of matched ion peaks between two spectra approach the same values, thus indicating a close match between the two profiles. The cross correlation between the unknown and 1,3-dicyclohexylurea is 0.9977, the cross correlation between the unknown and 1,3-bis(cyclopentylmethyl)urea is 0.4033, and the cross correlation between *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide is 0.1563.

The $Ecom_{50}$ value for the unknown (Table 9) was not significantly (t test, $p < 0.005$) different from that of 1,3 dicyclohexylurea or 1,3 bis(cyclopenylmethyl)urea, but was significantly (t-test, $p < 0.005$) different from the $Ecom_{50}$ value of *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide.

These data suggest that the overall chromatographic and mass spectral characteristics of the unknown are consistent with those of 1,3-dicyclohexylurea but not with 1,3 bis-(cyclopenylmethyl)urea or *N*-cyclohexyl-2-methyl-piperidine-1-carboxamide.

DISCUSSION

In this study we obtained new experimental data and developed two new models designed to filter large chemical databases for possible matches to an experimentally observed unknown. This work also served as a test of the overall filtering methodology in the identification and confirmation of 1,3-dicyclohexylurea as an unknown compound in a human serum sample observed during HPLC/MS analysis.

A significant aspect of the $Ecom_{50}$ modeling described here is the use of input variables based largely on molecular structure but not specifically designed to describe bond breaking. Both the Molconn and CODESSA structure descriptors are representations of molecular structure information, rather than quantities related to bond energy. The encoded information includes atom and bond arrangement, electron distribution, skeletal ramification, and geometric quantities, encoded at atom, bond, and whole molecule levels. Both MLR and PLS methods were used to develop models with correlation coefficients >0.93 for a truly eclectic data set. This outcome could be considered surprising given the absence of descriptor information specifically related to bond energy. The modeling results suggest that use of the molecular structure features described herein is a workable basis for model development of the novel molecular characteristic, $Ecom_{50}$. Furthermore, because the modeling descriptors are a representation of molecular structure, future models based on a much larger data set with better distribution in chemical data space may permit elucidation of structure features most strongly related to fragmentation. Molconn structure descriptors, based on molecular topology, do not require geometry optimization and CODESSA descriptors use a semiempirical level quantum chemical determination of geometry. Because the statistics of the CODESSA models were not significantly different from the Molconn PLS model, and the CODESSA models were not as effective in filtering the candidate database, the use of semiempirical optimized geometry did not appear to provide a significant advantage for modeling this data set.

Molconn descriptor calculations are not computationally intensive. This permits the development of high speed methods for filtering and compound identification as demonstrated in this work. However, since the best Molconn model made use of protonated input structures, additional work is required to more completely develop a computationally expedient algorithm for assigning the site of protonation. The CODESSA model does not have the limitation of requiring an assigned protonation site, but is computationally more intensive and did not prove to be as effective in filtering. Of the 315 structures taken from the PubChem database with an identical monoisotopic molecular weight to the unknown compound, the Molconn PLS protonated $Ecom_{50}$ model was able to filter out 35, or ~11%. This is not an especially large percentage, but given that this is a linear model based on very limited data, there is reason to believe that a nonlinear model based on a much larger data set will provide significantly better filtering. Neither the Molconn or CODESSA method provided optimum results, but we assert that, with limited data available for modeling, it is far too early in the investigation to draw any meaningful conclusions about which method might finally prove more effective. It remains to be seen if the findings of this study hold true when modeling a larger database that is more normally populated in both structure and activity space.

The model of retention index was found to show reasonable proficiency in differentiating among a group of structures with identical monoisotopic molecular weight and molecular formula. Of the 280 structures from the PubChem database that pass the $Ecom_{50}$ filter, the RI model was able to filter out an additional 212 compounds (based on a 99.8% confidence interval). Used in combination with the $Ecom_{50}$ model, approximately 80% of the candidate structures from the PubChem database were eliminated.

As was found in our previous RI modeling study,[30] good validation statistics were obtained despite an 8 to 1 ratio between the number of training rows and input descriptors. A 10 to 1 ratio is one of the least conservative ratios commonly deemed acceptable. The question arises as to whether or not maintaining a small number of descriptors relative to the number of training rows is actually necessary in order for a model to have acceptable validation statistics or if some other effect is involved. It is possible that a ratio of less than 10 to 1 is sufficient in some cases, but it is also possible that the population of input descriptors should be considered along with the number of descriptors. Looking at the nonzero population, on average, compounds have a nonzero input value for only 18.6 of the 33 descriptors (σ = 2.98, min = 10, max = 26). A ratio of much closer to 15 to 1 is maintained if only descriptors with a nonzero values are counted. This 15 to 1 ratio would pass the much more conservative $(\sqrt{n}) - 1$ rule. Eriksson suggests that models do not have a significant number of irrelevant inputs when the difference between train and validation correlation coefficients does not exceed 0.2,[54] which is well satisfied in this case. The quality validation results of the RI study suggest that the population of input descriptors should be taken into account when evaluating the ratio of inputs to training rows or if the traditionally accepted ratios are overly conservative.

The data used for the RI model is still lacking in structures with long retention times. Retention times of over 1000 RIU have been observed for some endogenous metabolites, yet the largest retention index in the data used was 593 RIU. The addition of compounds with longer retention times would likely give access to a larger domain of applicability for the final model. The data set is also lacking a significant number of simple structures which may provide useful information to the model about more complex structures, which can be considered as various combinations of simple compounds. As stated in the Materials and Methods section, there are a large number of features of low population in the data set. Although we suggest that the good validation statistics of the RI model indicate that the IGroups method was reasonably successful in mitigating the commonly seen detrimental effects of underpopulated features, it is also likely that increasing the population of these features would enhance the predictive performance of the model. Measurement of additional compounds will also provide invaluable independent validation data for the current model. Such data would help define the applicability domain by unambiguously identifying compounds for which the model makes accurate predictions. Revealing the strengths and weakness of current models will assist in tuning both the input descriptors and neural net modeling process. Though we believe the validation criteria used in this study were well-constructed, validation with data on which the model is partially based will never enumerate predictive capability in the way that newly measured data can.

As mentioned previously, 1,3-dicyclohexylurea has not been previously detected in human serum, so our finding is rather surprising. 1,3-dicyclohexylurea is a known byproduct formed during automated solid-phase peptide synthesis, and thus, the potential for contamination in our samples should be considered. However, we feel this is unlikely for several reasons. First, this compound was not found in all samples, even those coming from the same source. Second, in samples where it was found, the amounts were highly variable, ranging from barely detectible to approximately 10 μg/L. Third, we found this compound in human serum and CSF but not in mouse serum using the same methodology. Interestingly, 1,3-dicyclohexylurea is a potent inhibitor of soluble epoxide hydrolase with a K_i of approximately 30 nM.[55] Soluble epoxide hydrolase has been implicated in cardiovascular disease[56] and inflammation in mammals,[57] and thus soluble epoxide hydroase inhibitors have been considered for various cardiovascular indications.[58] On the other hand, soluble epoxide hydrolase inhibitors have recently been shown to dramatically enhance tumor metastasis in rodent models.[59] We estimate that in the human samples used in our study, the concentration of 1,3-dicyclohexylurea was as high as 10 μg/L or 45 nM, which is approximately equivalent to its K_i for soluble epoxide hydrolase. Although our findings will require verification in larger studies, it is temping to speculate that by regulating soluble epoxide hydrolase activity in vivo, endogenously produced 1,3-dicyclohexylurea may be physiologically significant. The endogenous source of 1,3-dicyclohexylurea remains unknown. However, it has been previously identified in two species of plants, *Portulaca oleracea*[60] and *Toddalia asiatica*,[61] suggesting the existence of a biochemical pathway for its synthesis in vivo.

This study illustrates the use of predictive algorithms for $Ecom_{50}$, retention index, and mass spectral fragmentation in discriminating among compounds retrieved from structural databases to aid in the identification of compounds detected in biological samples. A unique feature of these predictive algorithms is that they can be applied to any chemical database that provides structure files. Therefore, structure identification is not limited to the contents of small biological databases containing premeasured experimental variables. Thus, as demonstrated here, unknown endogenous compounds can be successfully identified. This would not be possible using experimental libraries of endogenous compounds or searching databases comprised of known biological compounds.

Although this approach is promising, there remain several limitations. The three predictive models used ($Ecom_{50}$, retention index, and CID fragmentation) will require considerable improvements; $Ecom_{50}$ by using larger numbers of compounds and neural network modeling, RI by increasing the number and diversity of compounds used for developing the neural net model, and CID fragmentation by including bond energy considerations. Our goal is to increase model predictive accuracy so that the three standard error filter ranges are approximately $\pm$0.75 eV for the $Ecom_{50}$ model and $\pm$75 RIU for the RI model. In addition, this approach still relies on the assumption that the unknown compound is contained in the chosen database. Although the presence of matches to an unknown seems likely for many mammalian metabolites, it is clearly not true for metabolomic studies of plants or other phyla for which potential metabolite lists are far from complete. Thus, the incorporation of structure prediction software that uses an elemental formula as a starting point will allow this procedure to become database-enabled rather than database-dependent.

Clearly, the addition of models for other orthogonal experimental features would be advantageous in this regard. There are two reasons for this. First, no single experimental feature, or even combination of featuers, is unique to a single compound; multiple compound candidates will have the same exact mass and RI value, for example. Second, since models are created using data sets with a limited selection of structures, compared to the databases that are screened with them, it is unlikely that predicted values will be as accurate as the corresponding experimental measurement. Increasing the number of orthogonal chemical descriptors could allow for enhanced discrimination between candidates without the need for models based on an exhaustive and impractical number of compounds. However, we emphasize that even in their current state, the models described here reduced the number of potential matches for the unknown by approximately 80%, and by including CID spectra matching, approximately 96% of the 315 candidate compounds were eliminated. Additionally, the combination of three experimental features used for final verification of the unknown was successful in unambiguously differentiating among three structurally similar isomers in the final candidate list without the need for purifying the unknown compound. Taken together, these results suggest promise of this approach for identifying the structures of unknown compounds in nontargeted metabolomics research.

■ ASSOCIATED CONTENT

Supporting Information

SI1: Measured $Ecom_{50}$ data, material sources, and model values for neutral structures. SI2: Measured $Ecom_{50}$ data, material sources, and model values for protonated structures. SI3: Measured retention index data, material sources, and model values. SI4−7: Tables of Molconn descriptors used in $Ecom_{50}$ PLS model. SI8 and 9: Tables of Molconn descriptors used in RI model. SI10: Description of the procedure used to synthesize and confirm 1,3-bis(cyclopentylmethyl)urea. SI11: Structures of eight final candidates that were not obtained and measured. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Mailing address: University of Connecticut, 69 N Eagleville Road, Storrs, CT 06269. Phone: (860)486-4265. Fax: (860) 486-5792. E-mail: david.grant@uconn.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Funding for this work was provided by NIH (1R01GM087714); Pfizer Inc., Groton, CT; The University of Connecticut Foundation; and by an American Foundation for Pharmaceutical Education Predoctoral Fellowship to T.M.K.

■ REFERENCES

(1) Nicholson, J. K.; Lindon, J. C.; Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, *29* (11), 1181−1189.

(2) Kertesz, T. M.; Hill, D. W.; Albaugh, D. R.; Hall, L. H.; Hall, L. M.; Grant, D. F. Database searching for structural identification of metabolites in complex biofluids for mass spectrometry-based metabonomics. *Bioanalysis* **2009**, *1* (9), 1627−1643.

(3) Evans, A. M.; DeHaven, C. D.; Barrett, T.; Mitchell, M.; Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal. Chem.* **2009**, *81* (16), 6656−6667.

(4) Smith, C. A.; O'Maille, G.; Want, E. J. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27* (6), 747−751.

(5) Wishart, D. S.; Tzur, D.; Knox, C. HMDB: the human metabolome database. *Nucleic Acids Res.* **2007**, *35* (Database issue), D521−D526.

(6) Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **2006**, *7*, 234.

(7) Kind, T.; Scholz, M.; Fiehn, O. How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS One* **2009**, *4* (5), e5440.

(8) 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* (14), 5574−5582.

(9) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *Bioinformatics* **2010**, *11*, 148−160.

(10) Kumari, S.; Stevens, D.; Kind, T.; Denkert, C.; Fiehn, O. Applying in-silico retention retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Anal. Chem.* **2011**, *83*, 5895−5902.

(11) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36* (Database issue), D901−D906.

(12) Wishart, D. S.; Lewis, M. J.; Morrissey, J. A.; Flegel, M. D.; Jeroncic, K.; Xiong, Y.; Cheng, D.; Eisner, R.; Gautam, B.; Tzur, D.; Sawhney, S.; Bamforth, F.; Greiner, R.; Li, L. The human cerebrospinal fluid metabolome. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2008**, *871* (2), 164−73.

(13) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* **2012**, *40*, D109−D114.

(14) Memboeuf, A.; Jullien, L.; Lartia, R.; Brasme, B.; Gimbert, Y. Tandem mass spectrometric analysis of a mixture of isobars using the survival yield technique. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1744−1752.

(15) Kertesz, T. M.; Hall, L. H.; Hill, D. W.; Grant, D. F. CE50: quantifying collision induced dissociation energy for small molecule characterization and identification. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (9), 1759−1767.

(16) *The PubChem Project, Compounds Database*, http://pubchem.ncbi.nlm.nih.gov/ (accessed 2011).

(17) O'Neil M. J., Ed. *The Merck Index*, 14th ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 2006.

(18) *winMolconn*, version 1.1.1.4; Hall Associates Consulting: Quincy, MA, 2008.

(19) *SAS*, 9.1; SAS Institute: Cary, NC, 2004.

(20) *CODESSA*, 2.7.16; Semichem, Inc.: Shawnee, KS, 1995−2011; http://semichem.com.

(21) *AMPAC 9*; Semichem, Inc.: Shawnee, KS, 1992−2011; http://semichem.com.

(22) Wang, J.; Aubry, A.; Bolgar, M. S.; Gu, H.; Olah, T. V.; Arnold, M.; Jemal, M. Effect of mobile phase pH, aqueous-organic ratio, and buffer concentration on electrospray ionization tandem mass spectrometric fragmentation patterns: implications in liquid chromatography/tandem mass spectrometric bioanalysis. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 3221−3229.

(23) Dewar, M. J. S.; Zoebisch, E. G.; Healey, E. F.; Stewart, J. J. P. AM1: a new general-purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(24) Dewar, M. J. S.; Zoebisch, E. G. Extension of AM1 to the halogens. *J. Mol. Struct. (Theochem)* **1988**, *180*, 1−21.
(25) Dewar, M. J. S.; Jie, C. AM1 parameters for phosphorus. *J. Mol. Struct. (Theochem)* **1989**, *187*, 1−13.
(26) Dewar, M. J. S.; Yuan, Y. C. AM1 parameters for sulfur. *Inorg. Chem.* **1990**, *29*, 3881−3890.
(27) http://webbook.nist.gov/chemistry/pa-ser.html (accessed Aug 2011).
(28) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: A reparameterization of AM1 for H, C, N, O. P, S F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101−1111.
(29) *Spartan '08*; WAVEFUNCTION, Inc: Irvine, CA, 2008.
(30) Albaugh, D. R.; Hall, L. M.; Hill, D. W. Prediction of HPLC retention index using artificial neural networks and IGroup E-State indices. *J. Chem. Inf. Model.* **2009**, *49* (4), 788−799.
(31) Hall, L. H.; Kier, L. B.; Hall, L. M. ADME-Tox Approaches, Electrotopological State indices to Assess Molecular Absorption, Distribution, Metabolism, Excretion, and Toxicity. *Comprehensive Medicinal Chemistry II*; Triggle, D. J., Taylor, J. B., Eds., Elsevier: Oxford, UK, 2007; Vol. *5*, pp 555−576.
(32) Hall, L. H.; Kier, L. B.; Hall, L. M. Computer Assisted Drug Design, Topological Quantitative Structure-Activity Relationship Applications: Structure Information in Drug Discovery. *Comprehensive Medicinal Chemistry II*; Triggle, D. J., Taylor, J. B., Eds., Elsevier: Oxford, UK, 2007: Vol. *4*, pp 537−574.
(33) Hall, L. H.; Kier, L. B. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.
(34) Kier, L. B.; Hall, L. H. The Kappa Indices for Modeling Molecular Shape and Flexibility. *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, UK, 1999; pp 455−489.
(35) Hall, L. H.; Structure-Information, A Approach to Prediction of Biological Activities and Properties. *Chem. Biodiversity* **2004**, *1*, 183−201.
(36) Hall, L. H.; Hall, L. M.; Kier, L. B.; Parham, M. E.; Votano, J. R. Interpretation of the role of the Electrotopological State and Molecular Connectivity Indices in the prediction of physical properties and ADME-Tox behavior. Case study: Human Plasma Protein Binding. *Proceedings of the Solvay Conference, Virtual ADMET Assessment in Target Selection and Maturation*, Luzerne, Switzerland; Testa, B., Turski, L., Eds.; IOS Press, 2006; pp 67−100.
(37) Ränner, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. *J. Chemom.* **1994**, *8*, 111−125.
(38) Tobias, R. An Introduction to Partial Least Squares Regression. In *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC, SAS Institute Inc., 1995; pp 1250−1257.
(39) Katritzky, A. R.; Karelson, M.; Lobanov, V. S. SPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure Appl. Chem.* **1997**, *69*, 245−248.
(40) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B. A general treatment of solubility: 1. The QSPR correlation of solvation free energies of single solutes in series of solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794−1805.
(41) Acevedo-Martinez, J.; Escalona-Arranz, J. C.; Villar-Rojas, A.; et al. Quantitative study of the structure-retention index relationship in the imine family. *J. Chromatogr. A* **2006**, *1102* (1−2), 238−244.
(42) Guo, W.; Lu, Y.; Zheng, X. M. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta* **2000**, *51* (3), 479−488.
(43) 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* (7), 400−404.
(44) 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* (4), 1328−1335.
(45) 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* (5), 406−412.
(46) 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* (9), 1550−1563.
(47) 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* (9), 925−933.
(48) Montgomery, D. C. Experiments with Random Factors. In *Design and Analysis of Experiments*, 5th ed., John Wiley &. Sons, Inc.: New York, 2008; pp 511−556.
(49) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714−371.
(50) *winMolconn*, version 1.2.2.1, Hall Associates Consulting: Quincy, MA, 2011.
(51) *Emergent*, version 5.2; University of Colorado: Bolder, CO, 2011.
(52) Aisa, B.; Mingus, B.; O'Reilly, R. The Emergent Neural Modeling System. *Neural Networks* **2008**, *21*, 1045−1212.
(53) Mistrik, R. Determination of molecular structures using tandem mass spectrometry. U.S. Patent no. 7197402 B2, 2007. *HighChem Mass Frontier 5.0*; http://www.highchem.com.
(54) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis, Principles and Applications*; Umetrics AB: Umeå Sweeden, 2001; p 107.
(55) Morisseau, C.; Goodrow, M. H.; Dowdy, D.; Zheng, J.; Greene, J. F.; Sanborn, J. R.; Hammock, B. D. Potent urea and carbamate inhibitors of soluble epoxide hydrolases. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (16), 8849−8854.
(56) Imig, J. D.; Hammock, B. D. Soluble epoxide hydrolase as a therapeutic target for cardiovascular diseases. *Nat. Rev. Drug Discov.* **2009**, *8* (10), 794−805.
(57) Elmarakby, A. A.; Faulkner, J.; Al-Shabrawey, M.; Wang, M. H.; Maddipati, K. R.; Imig, J. D. Deletion of the soluble epoxide hydrolase gene improves renal endothelial function and reduces renal inflammation and injury in streptozotocin-induced type 1 diabetes. *Am. J. Physiol. Regul. Integr. Comput. Physiol.* **2011**, *301* (5), R1307−17.
(58) Imig, J. D. Epoxides and soluble epoxide hydrolase in cardiovascular physiology. *Physiol Rev.* **2012**, *92* (1), 101−30.
(59) Panigrahy, D.; Edin, M. L.; Lee, C. R.; Huang, S.; Bielenberg, D. R.; Butterfield, C. E.; Barnés, C. M.; Mammoto, A.; Mammoto, T.; Luria, A.; Benny, O.; Chaponis, D. M.; Dudley, A. C.; Greene, E. R.; Vergilio, J. A.; Pietramaggiori, G.; Scherer-Pietramaggiori, S. S.; Short, S. M.; Seth, M.; Lih, F. B.; Tomer, K. B.; Yang, J.; Schwendener, R. A.; Hammock, B. D.; Falck, J. R.; Manthati, V. L.; Ingber, D. E.; Kaipainen, A.; D'Amore, P. A.; Kieran, M. W.; Zeldin, D. C. Epoxyeicosanoids stimulate multiorgan metastasis and tumor dormancy escape in mice. *J. Clin. Invest.* **2012**, *122* (1), 178−91.
(60) Rashed, A. N.; Afifi, F. U.; Shaedah, M.; Taha, M. O. Investigation of the active constituents of *Portulaca oleraceae* L. (Portulacaceae) growing in Jordan. *Pakistan J Pharm Sci.* **2004**, *17* (1), 37−45.
(61) Ian-Lih Tsai, I.-L.; Song-Chwan Fang, S.-C.; Tsutomu Ishikawa, T.; Chang, C.-T.; Chen, I.-S. N-cyclohexyl amides and a dimeric coumarin from formosan *Toddalia asiatica*. *Phytochemistry* **1997**, *44* (7), 1383−1386.