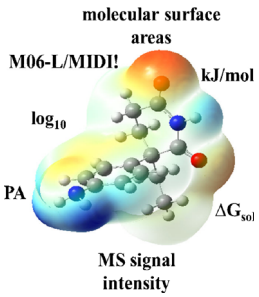
J. Am. Soc. Mass Spectrom. (2017) 28:278–285
DOI: 10.1007/s13361-016-1536-4

RESEARCH ARTICLE

Prediction of Mass Spectral Response Factors from Predicted Chemometric Data for Druglike Molecules

Christopher J. Cramer,[1] Joshua L. Johnson,[2,3] Amin M. Kamel[2,3]

[1]Department of Chemistry, Supercomputing Institute, and Chemical Theory Center, University of Minnesota, 207 Pleasant St. SE, Minneapolis, MN 55455, USA
[2]Novartis Institutes for BioMedical Research, Metabolism, and Pharmacokinetics (MAP), 250 Massachusetts Ave., Cambridge, MA 02139, USA
[3]*Present Address:* Biogen, Drug Metabolism and Pharmacokinetics (DMPK), 125 Broadway, Cambridge, MA 02142, USA



Abstract. A method is developed for the prediction of mass spectral ion counts of drug-like molecules using in silico calculated chemometric data. Various chemometric data, including polar and molecular surface areas, aqueous solvation free energies, and gas-phase and aqueous proton affinities were computed, and a statistically significant relationship between measured mass spectral ion counts and the combination of aqueous proton affinity and total molecular surface area was identified. In particular, through multilinear regression of ion counts on predicted chemometric data, we find that $\log_{10}$(MS ion counts) = –4.824 + c_1•PA + c_2•SA, where PA is the aqueous proton affinity of the molecule computed at the SMD(aq)/M06-L/MIDI!//M06-L/MIDI! level of electronic structure theory, SA is the total surface area of the molecule in its conjugate base form, and c_1 and c_2 have values of $–3.912 \times 10^{-2}$ mol kcal^{-1} and 3.682×10^{-3} Å^{-2}. On a 66-molecule training set, this regression exhibits a multiple R value of 0.791 with p values for the intercept, c_1, and c_2 of 1.4×10^{-3}, 4.3×10^{-10}, and 2.5×10^{-6}, respectively. Application of this regression to an 11-molecule test set provides a good correlation of prediction with experiment (R = 0.905) albeit with a systematic underestimation of about 0.2 log units. This method may prove useful for semiquantitative analysis of drug metabolites for which MS response factors or authentic standards are not readily available.

Keywords: Chemometric data, Aqueous proton affinity, Total surface area, Mass spectral ion counts of drug-like molecules

Received: 5 May 2016/Revised: 17 October 2016/Accepted: 18 October 2016/Published Online: 10 November 2016

Introduction

Metabolite identification studies play a critical part of the early pharmaceutical discovery process and are required to speed up lead optimization and candidate selection for development. However, a major limitation of these studies is that neither radiolabeled material nor analytical standards of drug metabolites are usually available for quantification of the amounts of the various metabolites formed. The ionization efficiency of even closely related molecules within the ion-source of a mass spectrometer may vary, and simple metabolic modifications to a drug can drastically alter the mass spectral response, prohibiting comparison of MS peak areas to estimate abundance of the identified metabolites. Furthermore, it is known that both molar absorptivity (ε) and the maximum absorption wavelength (λ_{max}) may be altered as a drug is metabolically modified, making the comparison of UV-derived peak areas without correction factors ambiguous. Therefore, development of a sensitive and specific technique for the quantitation of drug metabolites without the use of synthetic analytical standards or radiolabel would represent a major advance in preliminary metabolism screening in drug discovery.

Efforts toward obtaining a universal quantitative detection have been focused upon several detectors such as ultraviolet (UV) [1], evaporative light scattering detector (ELSD) [1, 2], chemiluminescence nitrogen detector (CLND) [1, 3], and charged aerosol detector (CAD) [4], among others. All of these detectors, however, have advantages as well as limitations. Proton nuclear magnetic resonance (^{1}H NMR) [1, 5] and Electronic REference To access In-vivo Concentrations (ERETIC) method [1, 6] has also been utilized as a "Gold Standard" to enable quantification by integrating proton

Correspondence to: Amin M. Kamel; *e-mail:* amin.kamel@biogen.com

signals. However, the ERETIC method is time-consuming and cost-ineffective [1].

The objective of this study is to develop a novel mass spectrometric quantitative detection strategy without a requirement for specific analyte standards to achieve a rapid, cost-effective, and readily automated quantification method. The major challenging aspect of this work is the correct identification of the most important parameters that influence analyte response during the electrospray ionization (ESI) process. Various solution-phase factors such as mobile-phase additives, solution pH, pK_a, analyte concentration and solvent composition [7–13] and gas-phase reactions including proton affinity [14, 15] have been reported to have a significant effect on ESI response.

In order to establish bounds on concentrations of trace analytes, it would be substantially more convenient to rely on validated relationships between *computed* (i.e., in silico) molecular properties and ionization response factors, since many computed properties can be determined efficiently and quickly. Various workers have employed uni- or multivariate statistical models to identify correlations between electrospray ion intensities and various computed (or measured) molecular properties for alkaloid cations [16], polar metabolites [17], alcohols [15], small molecule pharmaceuticals [13], and tripeptides [18]. Key quantities identified as having potential utility in multivariate relationships include polar and nonpolar surface areas, gas-phase basicities/proton affinities, solvation free energies [13, 18], molecular volume, octanol-water distribution coefficient (log *D*), and absolute mobility [17].

Advances in electronic structure theory [19] and quantum mechanical continuum solvation models [20, 21] have made the computation of proton affinities [22–24] and solvation free energies [25, 26] increasingly accurate and affordable (and molecular surface areas are trivially computed, of course). Importantly, insofar as the goal is to determine a statistically relevant relationship between ESI MS response factors and molecular properties, it is not necessarily critical to find computational models that are accurate in an absolute sense—all that is required is that they be systematically accurate in a relative sense (that is, trends in a computed property must be predicted accurately in order to identify correlations with experimental properties, but the absolute quantities themselves may be off so long as they are off by systematic amounts). This reduced demand on computational accuracy permits a wide range of efficient theoretical models to have potential utility in developing useful multivariate predictive models.

In this work, we report the assembly of a training set of 66 drugs and drug-like molecules for which we computed various chemometric quantities, including polar and molecular surface areas, aqueous solvation free energies, and gas-phase and aqueous proton affinities. We identify a statistically significant relationship between measured mass spectral ion counts at a specific concentration and the combination of aqueous proton affinity and total molecular surface area. When applied to a test set of 11 drug-like molecules, the derived protocol provides a good correlation of prediction with experiment, suggesting its potential utility for broader application.

Experimental

Chemicals and Materials

Unless otherwise indicated, reagents used were purchased from Sigma-Aldrich (St. Louis, MO, USA). *N*-acetyl mesalazine was purchased from Enamine (Monmouth Junction, NJ, USA). Novartis compounds were synthesized in-house. LC-MS grade acetonitrile and water were purchased through VWR International (Radnor, PA, USA).

Mass Spectrometry

A linear ion trap hybrid Orbitrap Elite (Thermo Scientific, San Jose, CA, USA) mass spectrometer interfaced with a Dionex UltiMate 3000 (Thermo Scientific) HPLC and a CTC PAL autosampler was used in this study. All *m/z* measurements were performed within the Orbitrap. Data was collected in full scan profile mode between *m/z* 100 and 1000 in positive ionization mode. The resolution was set at 30,000 at *m/z* 400 with an AGC target of 1.00e6. The following source parameters were used for all samples analyzed; sheath gas flow rate 40, aux gas flow rate 30, sweep gas flow rate 2 (all units arbitrary), spray voltage 4.00 kV, capillary temperature 275 °C, and S-Lens rf level of 60%.

Sample Preparation and Introduction

Stock solutions (1–10 mM corrected for salt, purity ≥99%) of 66-molecule training set in LC-MS grade acetonitrile were diluted to 10 μM in 1:1 acetonitrile:water with 0.1% formic acid. For flow injection analysis the HPLC was controlled using Chromeleon Xpress (Thermo Scientific), the mass spectrometer was controlled using Tune Plus (Thermo Scientific), and the autosampler using Xcalibur. Fifteen μL was injected with a flow rate of 700 μL/min with a mobile-phase composition of 50% acetonitrile, 50% water, and 0.1% formic acid. For flow injection analysis, each compound was injected three times on three different days. Data was analyzed using Xcalibur Qual browser (Thermo Scientific) and Microsoft Excel. Extracted ion chromatograms were generated using the theoretical monoisotopic masses, including any observed adducts or neutral losses ±5 ppm. Peak areas were used for all calculations. The intra- and inter-day variability was ≤5% (zonisamide) (intra-day) and ≤4% (acetaminophen) (inter-day) percent relative standard deviation (%RSD).

To examine data reproducibility, six commercially available compounds (acetaminophen, traxoprodil, *N*-acetyl mesalazine, zonisamide, verapamil, and dexamethasone) from the 66-molecule training set were carefully selected based on the diversity of their chemical structures and to cover a wide range of molecular weight (150–450 Da), aqueous proton affinity (–250 to –300 kcal/mol), and total molecular surface area (~203 to 600 $Å^2$). To account for the effect of mobile composition on the MS ionization efficiency, all six compounds were well separated and spread out at the entire HPLC gradient system used for the standard biotransformation protocol. Compounds

(10 μM) were spiked into 0.1M phosphate buffer pH 7.4 and vortex mixed. The sample was then diluted with three volumes of LC-MS grade acetonitrile. The samples were vortex mixed and centrifuged at 10,000 rpm for 10 min. Two hundred μL of the supernatant was transferred to a 96-well plate and the sample was dried to completion under nitrogen. The sample was redissolved into 200 μL of 95% water, 5% acetonitrile with 0.1% formic acid; 20 μL was injected onto a Waters (Milford, MA, USA) Symmetry C18 column (2.1 × 150 mm, 5 μm beads) and analyzed by LC-MS. Buffer A was LC-MS grade water with 0.1% formic acid and buffer B was LC-MS grade acetonitrile with 0.1% formic acid. The following gradient was used: 0–5 min 5% B, 15 min 40% B, 25 min 95% B, 25–30 min 95% B, 31 min 5% B, 31–36 min 5% B. The column eluent from the first 5 min was diverted to waste. To make sure the results were reproducible, a total of three injections were made. Data was analyzed using Xcaliber Qual browser and Excel. Extracted ion chromatograms were generated using the theoretical monoisotopic masses, including any observed adducts or neutral losses ±5 ppm. Peaks were manually picked and peak areas were used for all calculations.

Computational Procedures

For each studied molecule, the structure was first drawn in GaussView [27] (a visualization tool that accompanies the electronic structure suite *Gaussian 09* [28], used for all density functional calculations described below) and saved as a Gaussian input file. That structure was then imported to the molecular mechanics software program PC Model [29]. Next, the GMMX search algorithm within the PC Model program was employed to search for the lowest energy conformer (typically 1000 steps of search with otherwise default selection of parameters; this search covers rotation about single bonds (including amide and ester bonds) and ring conformations) using the MMX force field that is part of the PC Model program. The lowest energy conformer from the MMX search was then optimized (gas phase) at the M06-L [30] level of density functional theory, using the MIDI! basis set [31, 32] and a density fitting basis set that improves efficiency when employed with a local density functional [19] (this level of theory is denoted M06-L/MIDI!), The optimized structure was re-imported into PC Model and its polar, saturated nonpolar, unsaturated nonpolar, and total molecular surface areas were computed using the standard van der Waals radii employed by the package (we note that the definitions of "polar," "saturated nonpolar," and "unsaturated nonpolar" surface areas are not universal, however, as we found no significance for these variables in our analyses, we do not consider their details further here).

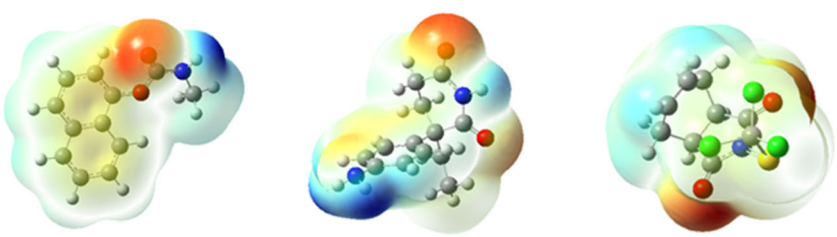
The most basic site of the molecule was then identified (either through chemical intuition or, when necessary, through examination of multiple possibilities), and the molecule was protonated at that site. Again, the structure was optimized (gas phase) at the M06-L/MIDI! level of theory (see Figure 1).

Next, for the gas-phase geometries of both the conjugate base and acid forms (i.e., as single points), aqueous solvation free energies were computed using the SMD aqueous solvation model [33] (this level of theory is denoted SMD(aq)/M06-L/MIDI!//M06-L/MIDI!). The aqueous proton affinity is computed as the difference in the single-point energies in solution of the protonated species and the neutral species (typically in the range of –250 to –300 kcal/mol; note that this quantity is consistent with the typical ion convention in solution, which takes the free energy of the solvated proton to be zero [34]). In preliminary work, we examined whether reoptimization of structures in aqueous solution led to improved results. Such reoptimizations were found to have negligible statistical significance. As reoptimization in solution roughly doubles the computational time per molecule, we relied on single-point calculations alone to compute aqueous proton affinities (see also Results and Discussion section).

Results and Discussion

Measurement of Target Property

A total of 66 drug-like molecules including eight Novartis compounds (compounds A, B, C, D, E, F, G, and H) and commercially available drugs (training set) were chosen based on their availability and diverseness of molecular weight, structure (phase I and phase II metabolism), and physical chemical properties. Flow injection analysis demonstrated



Carbaryl Aminoglutethimide Captan

Figure 1. Representative electrostatic potentials (ESPs) for the optimized structures of carbaryl, aminoglutethimide, and captan, three compounds chosen for this study, showing a ball-and-stick representation of the optimized M06-L/MIDI! geometries with gray balls representing carbon, white hydrogen, blue nitrogen, red oxygen, green chlorine, and yellow sulfur. The most red site on the ESP represents the most negative electrostatic potential (attractive to a positive test charge) and was selected for protonation. Blue regions are repulsive while intermediate colors of the spectrum reflect the quantitative change from attractive to repulsive

concentration-dependent peak areas between 10 and 100 μM (data not shown). A concentration of 10 μM was chosen because at that concentration the detector was not saturated and that concentration is the substrate concentration used in our standard biotransformation assays. Compounds used for data reproducibility (acetaminophen, traxoprodil, *N*-acetyl mesalazine, zonisamide, verapamil, and dexamethasone) eluted between 7 and 18 min and peak areas for all compounds were reproducible with a % RSD ≤ 7.4%.

Chemometric and Statistical Analysis

Beginning with a training set of data for 40 molecules (Table 1), we considered a number of different quantities, including the aqueous free energy of solvation of the neutral form of the solute, computed either from a single-point energy calculation at the gas-phase optimized geometry or including relaxation in aqueous solution (columns 2 and 3 in Table 1), the solute proton affinity, computed either in the gas phase or including solvation free energies from either single-point calculations at gas-phase geometries or including relaxation effects in solution (columns 4–6 in Table 1), as well as the saturated nonpolar, unsaturated nonpolar, polar, and total molecular surface areas for the M06-L/MIDI! optimized structures (columns 7–11 in Table 1).

Insofar as the chemometric quantities in Table 1 are either free energies, or surface areas, which are likely to be correlated linearly with free energy quantities associated with surface

Table 1. Chemometric Data for 40 Molecules

	$\log_{10}$	Neutral ΔG^o_S[a]		Proton affinity[a]			Surface area[b]			
Molecule	Ions	SP[c]	Opt[d]	Gas	Aq. SP[c]	Aq. opt[d]	Saturated nonpolar	Unsaturated nonpolar	Polar	Total
Aminoglutethimide	7.51	–10.6	–11.0	–219.0	–267.3	–268.3	150.5	38.06	97.05	286.5
Bromacil	7.06	–7.03	–7.08	–226.7	–267.9	–267.9	194.2	12.04	63.04	269.9
Caffeine	6.09	–7.06	–8.00	–221.4	–267.7	–267.9	134.8	25.05	73.03	233.7
Carbaryl	6.00	–6.05	–7.03	–223.4	-267.0	–268.0	139.2	71.00	49.06	259.8
Carbofuran	7.49	–8.01	–8.06	–228.4	–267.9	–268.4	187.3	45.01	58.04	290.8
Celecoxib	7.11	–9.00	–9.07	–214.7	–258.6	–259.8	207.2	107.1	84.09	399.2
Citalopram	8.51	–5.05	–6.02	–252.3	–299.1	–299.2	264.0	98.03	37.04	399.7
Dapsone	7.37	–17.4	–17.9	–234.1	–266.8	–266.8	97.05	96.09	98.06	293.0
Diazinon	8.49	–3.04	–3.09	–237.3	–281.9	–283.1	249.5	28.00	52.03	329.7
Diclofenac	6.97	–8.09	–9.02	–213.4	–259.7	–262.5	170.8	88.06	55.01	314.5
Linezolid	8.07	–14.4	–15.6	–243.2	–287.3	–288.1	257.9	42.05	88.08	389.2
Methomyl	5.64	–5.01	–5.06	–226.1	–271.6	–271.7	155.3	6.005	60.03	222.1
Mirtazapine	8.16	–8.05	–9.01	–248.4	–294.8	–295.4	234.7	93.07	8.00	336.3
Naproxen	5.03	–7.08	–8.02	–205.8	–261.3	–262.6	171.8	61.08	59.08	293.3
Nefazadone	8.75	–18.0	–19.2	–221.5	–269.3	–270.3	396.5	96.06	54.03	547.4
Traxoprodil	8.32	–14.6	–16.8	–260.5	–300.2	–300.4	243.8	97.08	58.09	400.5
Valdecoxib	6.98	–11.1	–11.8	–212.3	–258.7	–259.3	114.9	116.2	88.00	319.0
Zaleplon	7.76	–10.7	–11.3	–211.8	–253.2	–253.6	178.2	106.0	78.09	363.1
Zileuton	6.98	–12.5	–12.8	–234.6	–273.5	–274.8	145.3	51.09	78.02	275.4
Zomepirac	7.21	–9.04	–9.08	–233.1	–273.9	–274.6	157.3	73.02	85.05	315.9
Zonisamide	5.81	–10.5	–10.9	–205.6	–258.0	–259.2	79.02	55.03	96.05	231.0
Acyclovir	7.09	–24.7	–26.4	–247.8	–288.7	–288.9	95.03	34.03	140.0	269.6
Acecainide	8.01	–14.6	–16.8	–259.4	–298.1	–298.2	235.9	49.08	78.08	364.5
Atropine	8.34	–10.0	–10.8	–244.5	–295.6	–296.1	285.5	48.02	56.04	390.1
Azelastine	8.06	–10.0	–10.0	–241.3	–297.5	–297.5	285.2	98.05	51.02	434.9
Clindamycin	8.26	–16.6	–17.7	–237.1	–286.3	–288.1	347.6	0.00	93.02	440.7
Fleroxacin	8.31	–16.8	–18.0	–234.3	–293.8	–294.7	213.0	54.08	98.06	366.3
Galanthamine	8.09	–10.2	–10.6	–250.7	–294.0	–294.8	249.4	48.07	45.04	343.5
Pirmenol	8.47	–8.00	–8.06	–264.2	–303.0	–302.4	327.6	86.05	31.00	445.2
Quinine	8.32	–10.8	–11.2	–249.6	–293.3	–293.8	264.9	71.05	46.01	382.5
Tramadol	8.08	–6.05	–7.03	–246.3	–295.8	–295.8	277.1	41.07	32.01	350.9
Venlafaxine	8.14	–5.07	–7.00	–253.1	–297.8	–297.5	307.2	36.08	30.00	374.0
Verapamil	8.74	–9.07	–9.07	–248.3	–295.9	–295.9	449.3	82.06	69.00	600.9
Vildagliptin	8.15	–14.4	–15.1	–250.3	–296.2	–295.5	281.9	11.01	72.09	365.8
9-Methylguanine	7.27	–18.7	–20.3	–225.0	–278.9	–279.3	57.03	33.03	109.8	200.3
APAP	6.45	–11.1	–11.4	–223.7	–273.9	–274.1	88.03	51.08	63.08	203.9
APAP sulfate	6.49	–13.4	–14.1	–195.2	–244.0	–245.2	95.08	45.04	114.6	255.8
APAP glucuronide	5.43	–22.7	–23.6	–227.5	–273.3	–274.0	139.3	49.00	168.9	357.2
Mesalazine	6.82	–10.1	–10.4	–223.3	–270.0	–269.9	49.04	32.04	99.01	181.0
N-acetyl mesalazine	6.03	–10.6	–11.0	–218.5	–266.1	–266.5	89.00	31.05	108.9	229.5
R with $\log_{10}$ ions	1.00	0.015	0.007	0.669	0.676	0.679	0.748	0.217	0.451	0.702

[a] Kcal mol^{-1}.
[b] $Å^2$.
[c] Single-point.
[d] Optimized in solution.
APAP = acetyl-*para*-aminophenol, acetaminophen

tensions, volatilities, etc., their variations should be associated with log variations in concentration measures. In the case of ion counts, the observed quantity may be related to a concentration (while all analytes are at a common *total* concentration, their conjugate acid/base speciation may vary, which may influence ionization counts), but is also related to efficiency of ionization and other factors associated with the electrospray experiment. As the measured data may be expected to have some concentration-like character, and as they span nearly three orders of magnitude in ion counts (and no computed molecular quantity is likely to span so similarly large a range in linear space), we chose to seek a correlation with the common logarithm of the ion counts measured from experiments with 10 μM analyte concentrations (column 1 of Table 1). The last row of Table 1 indicates the single variable correlation coefficient R between the various individual tabulated quantities and the $\log_{10}$ ion counts. It is immediately apparent that the solvation free energy of the neutral solute is completely uncorrelated with the ion count data ($R < 0.02$), whereas strong correlation is observed with the gas-phase and aqueous proton affinities computed with or without solvent relaxation ($R > 0.66$), as well as with the saturated nonpolar and total surface areas ($R > 0.70$).

The observation of a good correlation with proton affinity (basicity) is consistent with prior multivariate analyses along similar lines [7, 8]. Considering the various proton affinity computational protocols, they provide data that are highly correlated ($R > 0.94$ between methods). As single-point solvated calculations include the important physical effect of aqueous solvation at very small cost, we elected to continue to explore further with this descriptor. With respect to the surface areas, the saturated nonpolar and total surface areas show the highest correlation with the target ion count data. Those two descriptors are also highly correlated with one another ($R = 0.915$). As the total surface area is much more unambiguously defined, we elected to carry this descriptor forward for further analysis as well.

Considering multilinear regressions that included, along with the single-point aqueous proton affinity and the total surface area, other descriptors not already strongly correlated with those two (i.e., neutral molecule solvation free energy, unsaturated nonpolar surface area, and polar surface area) failed to provide meaningful statistical improvements. However, the correlation coefficient between the proton affinity and total surface area descriptors themselves over the 40 molecules in Table 1 is $R = 0.482$, suggesting that the training set could be improved through addition of molecules reducing this cross-correlation.

To that end, we added an additional 26 molecules, computing only the necessary two descriptors (Table 2). When taken together over the 66 molecules in the expanded training set, the cross-correlation between the proton affinity and total surface area descriptors is reduced to $R = 0.228$, which we deem acceptable.

Table 2. Chemometric Data for an Additional 26 Molecules

Molecule	$\log_{10}$ Ions	Proton affinity[a] Aq. SP[c]	Surface area[b] Total
N-methyl piperdine	7.18	–296.8	178.8
Dexamethasone	7.81	–279.6	413.4
Meropenem	8.01	–292.7	379.5
Phencyclidine	7.93	–298.3	329.8
Procainamide	8.03	–303.4	323.8
Compound A	7.74	–271.4	527.9
Compound B	7.68	–271.1	670.7
Coumarin	5.52	–267.8	185.6
Umbelliferone	5.64	–269.7	192.1
Umbelliferone Gluc	5.25	–268.2	348.4
Hymecromone	6.09	–271.1	212.7
Hymecromone Glucuronide	6.03	–269.8	369.9
Nefamostat	8.31	–291.9	405.4
4-Guanidinobenzoic acid	7.54	–291.6	217.8
6-Amidino-2-napthol	7.77	–298.4	233.4
Phenolphthalein	7.06	–274.4	357.9
Phenolphthalein Glucuronide	7.18	–271.9	516.4
Compound C	7.76	–278.3	418.1
Compound D	8.48	–293.2	579.8
Compound E	8.39	–293.2	560.6
Clozapine	8.49	–287.3	381.1
Clozapine *N*-oxide	8.39	–286.3	381.2
Clozapine *N*-desmethyl	8.32	–287.4	358.2
Compound F	6.95	–265.3	478.6
Compound G	6.61	–265.9	439.5
Compound H	7.62	–278.6	393.7

[a] Kcal mol^{-1}.
[b] Å^2.
[c] Single-point.

Over the 66 molecules in the training set, bilinear regression of the $\log_{10}$ ion counts (IC) on the aqueous proton affinity (PA) and total molecular surface area (SA) provides

$$\log_{10}(\mathrm{IC}) = -4.824 - \left(3.912 \times 10^{-2}\ \mathrm{mol\ kcal^{-1}}\right) \cdot \mathrm{PA} + \left(3.682 \times 10^{-3}\ \mathrm{Å^{-2}}\right) \cdot \mathrm{SA} \quad (1)$$

with multiple R value 0.791, multilinear regression F ratio of 52.5, and p values for the intercept, c_1 and c_2 of 1.4×10^{-3}, 4.3×10^{-10}, and 2.5×10^{-6}, respectively. The standard errors on the intercept, c_1 and c_2 are 1.447, 0.530×10^{-2} mol kcal^{-1}, and 0.711×10^{-3} Å^{-2}, respectively. Note that qualitatively, Equation 1 predicts that molecules having proton affinities (or, more loosely, basicities) that are larger in magnitude—recalling that the quantity itself is negative—will show increased ion counts, as will molecules that are larger in overall surface area. Thus, many of the molecules in Tables 1 and 2 with larger observed ion counts include basic amine-type functionality of one sort or another.

A plot of predicted versus experimental log data for a training set of 66 molecules (purple circles) and a test set of 11 molecules (brown squares, see below for further discussion) at nominal 10 μM concentration is provided in Figure 2. Note that the most significant outliers are at lower left, that is, compounds with very low *measured* ion counts but reasonably high *predicted* ion counts. The three lowest points are naproxen and two glucuronides.

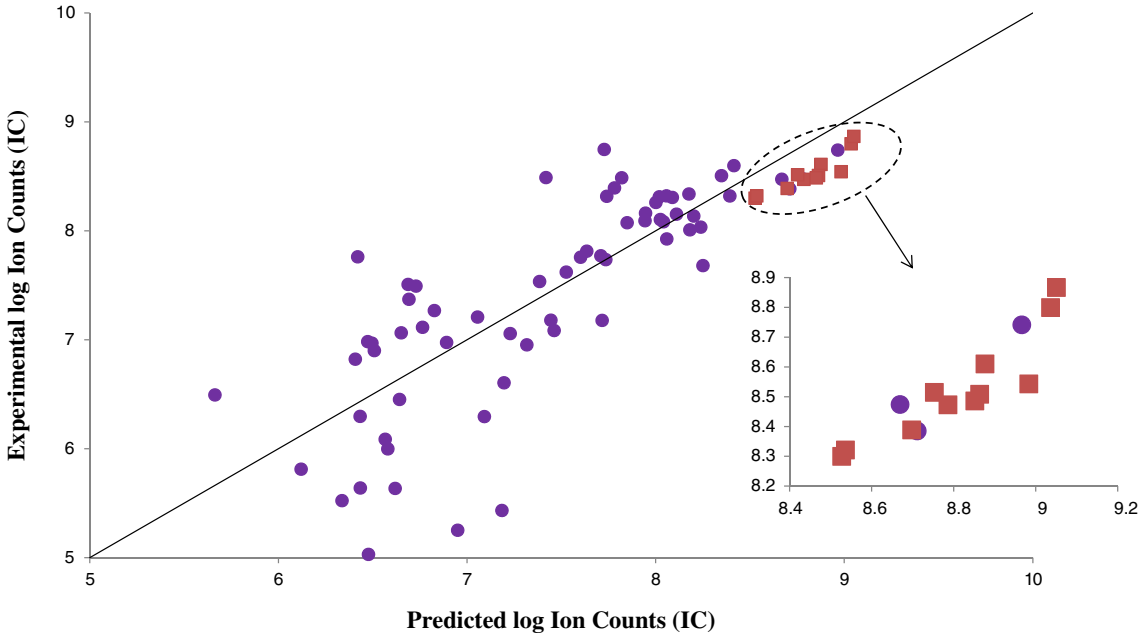


Figure 2. Log/log plot of predicted versus observed mass spectral ion counts for a training set of 66 molecules (purple circles) and a test set of 11 molecules (brown squares) at nominal 10 mM concentration. Expanded area of graph includes the 11 molecule test data set

If one treats these three compounds as outliers, regression on the remaining data leads to a significantly improved F ratio (67.5). However, the regression itself is affected primarily in the intercept, which is expected in the event of removing such outliers (that are over-predicted). On the whole, regression on the full data set seems likely to be the most general and useful in the absence of further expansion of the training set. While one could consider making the presence of a glucuronide group a variable (a constant), with only four in the training set, this seems statistically dubious. Over the full 66-molecule training set, the mean absolute deviation between experiment and prediction from Equation 1 is 0.44 log units.

To assess the robustness of our fit, we next considered a test set of 11 molecules, chosen from a proprietary list of Novartis compounds and having limited similarity to molecules used in the training set. Employing Equation 1, we predicted log ion count data that compare very well to experimental measurement, as illustrated in Figure 2. There is an excellent correlation of predicted values with the variation in the data, and the mean absolute deviation between the predicted and experimental values is 0.29 log units. This deviation is smaller than that observed over the full training set. It is also systematic (i.e., every prediction somewhat overestimates the observed value). We reconsidered whether a regression equation removing the lowest outliers mentioned above would have improved predictive performance on the 11-molecule test set, but it was essentially unaffected.

We further examined the robustness of our predictive protocol by pooling all 77 data and randomly assigning each point to one of seven 11-member subsets in order to

Table 3. Results from Leave-11-Out Cross-Validation Studies on Combined Data

Subset	Intercept[a]	c_1[a]	c_2[a]	F_{train}[b]	R_{train}[c]	MUE[d]	R_{test}[e]
1	–4.178	-3.728×10^{-2}	3.298×10^{-3}	80.2	0.847	0.569	0.681
2	–4.244	-3.724×10^{-2}	3.428×10^{-3}	62.7	0.816	0.294	0.886
3	–3.978	-3.670×10^{-2}	3.193×10^{-3}	67.6	0.826	0.378	0.82
4	–3.501	-3.512×10^{-2}	3.164×10^{-3}	59.9	0.81	0.478	0.868
5	–4.471	-3.870×10^{-2}	2.939×10^{-3}	58.7	0.807	0.281	0.854
6	–5.031	-3.999×10^{-2}	3.405×10^{-3}	69.6	0.83	0.479	0.836
7	–4.687	-3.870×10^{-2}	3.431×10^{-3}	70.7	0.832	0.383	0.84
Mean	–4.299	-3.768×10^{-2}	3.265×10^{-3}				
StdDev	0.496	0.160×10^{-2}	0.181×10^{-3}				

[a] Bilinear regression coefficients for Equation 1 (cf. text for units).
[b] Multilinear regression ratio on training subset.
[c] Pearson correlation coefficient for bilinear regression on training subset.
[d] Mean unsigned error (absolute residuals) for 11-molecule test set using bilinear regression from 66-molecule training set.
[e] Pearson correlation coefficient for experimental and predicted values in 11-molecule test set using bilinear regression from 66-molecule training set.

perform a cross-validation analysis. We carried out bilinear regression to generate equations analogous to Equation 1 on the remaining 66 molecules associated with each subset, and we examined the utility of those equations for predicting log ion count data treating the left-out 11 molecules in each case as test sets. The results are summarized in Table 3. For the seven different bilinear regressions, we obtained intercepts of –4.299 ± 0.496, coefficients multiplying the aqueous proton affinity of (–3.768 ± 0.160) $\times 10^{-2}$ mol kcal^{-1}, and coefficients multiplying the total surface area of (3.265 ± 0.181) $\times 10^{-2}$ Å^{-2}. These values all fall within the standard error ranges associated with these parameters for Equation 1 itself. The multilinear regression F ratios predicted for the seven fits range from 58.7 to 80.2, whereas the Pearson correlation coefficients R (for the training set) range from 0.807 to 0.847.

Considering the application of each regression to make predictions for its specific test set, the mean unsigned residuals over the various test sets range from 0.281 to 0.569, whereas the associated Pearson correlation coefficients R range from 0.681 to 0.886. In all cases but that of subset 1, $R > 0.8$; this reflects the presence of two of the outliers having very low ion counts in subset 1, which also contributes to this subset exhibiting the largest unsigned error in the residuals.

The good observed performance overall of the various bilinear regressions analogous to Equation 1 on the different randomized 11-molecule test sets is encouraging. We anticipate that this overall model may prove to be useful in the future for estimating concentrations of unknown analytes through the comparison of predicted ion count data to observed data.

Conclusions

We have demonstrated that semiquantitative information on drugs and metabolites can be obtained by an in silico approach and without using radiolabeled compounds and chemically synthesized metabolite standards. We propose that the reported methodology herein would allow a quick semiquantitative assessment of drugs and metabolites much earlier in drugs discovery, hence, identifying metabolism-related liabilities of key compounds and effectively managing the development of a lead series. Additionally, this semiquantitative approach could be applied in early development to assess whether humans produce metabolites that are adequately covered by preclinical species or form unique/disproportionate metabolites that require additional safety testing.

Acknowledgments

The authors thank Professor Burnaby Munson, Dr. Shawn Harriman, and Dr. Franco Lombardo for helpful discussions, and Mr. Kevin Colizza for preliminary lab work. During initial phases of this work, Professor Christopher Cramer received compensation as a consultant to Novartis.

References

1. Lane, S., Boughtflower, B., Mutton, I., Paterson, C., Farrant, D., Taylor, N., Blaxill, Z., Carmody, C., Borman, P.: Toward single-calibrant quantification in HPLC. A comparison of three detection strategies: evaporative light scattering, chemiluminescent nitrogen, and proton NMR. Anal. Chem. **77**, 4354–4365 (2005)
2. Fries, H.E., Evans, C.A., Ward, K.W.: Evaluation of evaporative light-scattering detection for metabolite quantification without authentic analytical standards or radiolabel. J. Chromatogr. B Anal. Technol. Biomed. Life Sci. **819**, 339–344 (2005)
3. Deng, Y., Wu, J.T., Zhang, H., Olah, T.V.: Quantitation of drug metabolites in the absence of pure metabolite standards by high-performance liquid chromatography coupled with a chemiluminescence nitrogen detector and mass spectrometer. Rapid Commun. Mass Spectrom. **18**, 1681–1685 (2004)
4. Poplawska, M., Blazewicz, A., Bukowinska, K., Fijalek, Z.: Application of high-performance liquid chromatography with charged aerosol detection for universal quantitation of undeclared phosphodiesterase-5 inhibitors in herbal dietary supplements. J. Pharm. Biomed. Anal. **84**, 232–243 (2013)
5. Webster, G.K., Marsden, I., Pommerening, C.A., Tyrakowski, C.M., Tobias, B.: Determination of relative response factors for chromatographic investigations using NMR spectrometry. J. Pharm. Biomed. Anal. **49**, 1261–1265 (2009)
6. Nuzzo, G., Gallo, C., d'Ippolito, G., Cutignano, A., Sardo, A., Fontana, A.: Composition and quantitation of microalgal lipids by ERETIC (1)H NMR method. Mar. Drugs **11**, 3742–3753 (2013)
7. Kamel, A.M., Brown, P.R., Munson, B.: Effects of mobile-phase additives, solution pH, ionization constant, and analyte concentration on the sensitivities and electrospray ionization mass spectra of nucleoside antiviral agents. Anal. Chem. **71**, 5481–5492 (1999)
8. Kamel, A.M., Brown, P.R., Munson, B.: Electrospray ionization mass spectrometry of tetracycline, oxytetracycline, chlorotetracycline, minocycline, and methacycline. Anal. Chem. **71**, 968–977 (1999)
9. Constantopoulos, T.L., Jackson, G.S., Enke, C.G.: Effects of salt concentration on analyte response using electrospray ionization mass spectrometry. J. Am. Soc. Mass Spectrom. **10**, 625–634 (1999)
10. Ikonomou, M.G., Blades, A.T., Kebarle, P.I.: Investigations of the electrospray interface for liquid chromatography-mass spectrometry. Anal. Chem. **62**, 957–967 (1990)
11. Loo, J.A., Loo, R.R., Light, K.J., Edmonds, C.G., Smith, R.D.: Multiply charged negative ions by electrospray ionization of polypeptides and proteins. Anal. Chem. **64**, 81–88 (1992)
12. Wang, G., Cole, R.B.: Disparity between solution-phase equilibria and charge state distributions in positive-ion electrospray mass spectrometry. Org. Mass Spectrom. **29**, 419 (1994)
13. Mandra, V.J., Kouskoura, M.G., Markopoulou, C.K.: Using the partial least squares method to model the electrospray ionization response produced by small pharmaceutical molecules in positive mode. Rapid Commun. Mass Spectrom. **29**, 1661–1675 (2015)
14. Kebarle, P., Peschke, M.: On the mechanisms by which the charged droplets produced by electrospray lead to gas phase ions. Anal. Chim. Acta **406**, 11–35 (2000)
15. Amad, M.H., Cech, N.B., Jackson, G.S., Enke, C.G.: Importance of gas-phase proton affinities in determining the electrospray ionization response for analytes and solvents. J. Mass Spectrom. **35**, 784–978 (2000)
16. Tang, L., Kebarle, P.: Dependence of ion intensity in electrospray mass spectrometry on the concentration of the analytes in the electrosprayed solution. Anal. Chem. **65**, 3654–3668 (1993)
17. Chalcraft, K.R., Lee, R., Mills, C., Britz-McKibbin, P.: Virtual quantification of metabolites by capillary electrophoresis-electrospray ionization-mass spectrometry: predicting ionization efficiency without chemical standards. Anal. Chem. **81**, 2506–2515 (2009)
18. Raji, M.A., Frycak, P., Temiyasathit, C., Kim, S.B., Mavromaras, G., Ahn, J.M., Schug, K.A.: Using multivariate statistical methods to model the electrospray ionization response of GXG tripeptides based

on multiple physicochemical parameters. Rapid Commun. Mass Spectrom. **23**, 2221–2232 (2009)
19. Cramer, C.J.: Essentials of Computational Chemistry: Theories and Models. John Wiley and Sons, Chichester (2004)
20. Cramer, C.J., Truhlar, D.G.: Implicit solvation models: equilibria, structure, spectra, and dynamics. Chem. Rev. **99**, 2161–2200 (1999)
21. Tomasi, J., Mennucci, B., Cammi, R.: Quantum mechanical continuum solvation models. Chem. Rev. **105**, 2999–3093 (2005)
22. Liptak, M.D., Shields, G.C.: Comparison of density functional theory predictions of gas-phase deprotonation data. Int. J. Quantum Chem. **105**, 580–587 (2005)
23. Zhao, Y., Truhlar, D.G.: Exploring the limit of accuracy of the global hybrid meta density functional for main-group thermochemistry, kinetics, and noncovalent interactions. J. Chem. Theory Comput. **4**, 1849–1868 (2008)
24. He, X., Fusti-Molnar, L., Merz, K.M.: Accurate benchmark calculations on the gas-phase basicities of small molecules. J. Phys. Chem. A **113**, 10096–10103 (2009)
25. Cramer, C.J.; Truhlar, D.G.: SMx Continuum Models for Condensed Phases. In: Maroulis, G., Simos, T.E. (Eds.) Trends and Perspectives in Modern Computational Science, p 112-140. Brill Academic: Amsterdam (2006)
26. Cramer, C.J., Truhlar, D.G.: A universal approach to solvation modeling. Acc. Chem. Res. **41**, 760–768 (2008)
27. GaussView version 5.0.8, Gaussian Inc, Wallingford, CT (2008)
28. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G.A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H.P., Izmaylov, A.F., Bloino, J., Zheng, G., Sonnenberg, J.L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J.A., Peralta, J.E., Ogliaro, F., Bearpark, M., Heyd, J.J., Brothers, E., Kudin, K.N., Staroverov, V.N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J.C., Iyengar, S.S., Tomasi, J., Cossi, M., Rega, N., Millam, J.M., Klene, M., Knox, J.E., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A.J., Cammi, R., Pomelli, C., Ochterski, J.W., Martin, R.L., Morokuma, K., Zakrzewski, V.G., Voth, G.A., Salvador, P., Dannenberg, J.J., Dapprich, S., Daniels, A.D., Farkas, Ö., Foresman, J.B., Ortiz, J.V., Cioslowski, J.: Gaussian 09, Revision C.01. Gaussian, Inc, Wallingford (2010)
29. PC Model, version 9.2, Serena Software: Bloomington, IN (2010)
30. Zhao, Y., Truhlar, D.G.: The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor. Chem. Accounts **120**, 215–241 (2008)
31. Easton, R.E., Giesen, D.J., Welch, A., Cramer, C.J., Truhlar, D.G.: The MIDI! basis set for quantum mechanical calculations of molecular geometries and partial charges. Theor. Chim. Acta **93**, 281–301 (1996)
32. Li, J., Cramer, C.J., Truhlar, D.G.: MIDI! basis set for silicon, bromine, and iodine. Theor. Chem. Accounts **99**, 192–196 (1998)
33. Marenich, A.V., Cramer, C.J., Truhlar, D.G.: Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J. Phys. Chem. B **113**, 6378–6396 (2009)
34. Kelly, C.P., Cramer, C.J., Truhlar, D.G.: Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton. J. Phys. Chem. B **110**, 16066–16081 (2006)