Evaluation of pK_a Estimation Methods on 211 Druglike Compounds

John Manchester,* Grant Walkup, Olga Rivin, and Zhiping You

Infection Discovery, AstraZeneca R&D Boston, 35 Gatehouse Drive, Waltham, Massachusetts 02451

Received January 11, 2010

The pK_a values of 211 discovery (druglike) compounds were determined experimentally using capillary electrophoresis coupled with ultraviolet spectroscopy and a novel fitting algorithm. These values were compared to those predicted by five different commercially available pK_a estimation packages: ACDLabs/pK_a, Marvin (ChemAxon), MoKa (Molecular Discovery), Epik (Schrodinger), and Pipeline Pilot (Accelrys). Even though the topological method MoKa was noticeably faster than ACD, the accuracy of those two methods and Marvin was statistically indistinguishable, with a root-mean-squared error of about 1 pK_a unit compared to experiment. Pipeline Pilot and EpiK both produced pK_a estimates in significantly worse agreement with the experiment. Interestingly, on a number of compounds, the predictions due to ACD v12 were in poorer agreement with the experiment than ACD v10. Microscopic and "apparent" pK_a predictions were also compared using ACD v10. Microscopic pK_as gave significantly worse agreement with the experiment than the "apparent" values. In all cases, the errors appeared to be randomly distributed across chemical series.

INTRODUCTION

In the process of optimizing compounds within drug discovery campaigns, it is commonly desirable to know the pK_a values of any ionizable groups present within a given test compound. This knowledge can be valuable for a large variety of reasons including improving the potency against the primary drug target, reducing potency toward undesired targets (for example, hERG), impacting physical properties such as solubility, and for modulating pharmacological properties such as bioavailability, distribution, plasma protein binding, and clearance. For example, the polarity of small molecules has long been associated with these properties,[1,2] and polarity is profoundly influenced by whether a compound exists at least partially in an ionized state at physiologic pH. Additionally, more than two-thirds of small-molecule pharmaceuticals brought to market since 1900 contain at least one ionizable group (Prous Integrity search). Whether these compounds ionize at physiologic pH is most readily assessed using pK_a values.

In efforts to take greater control over the design-make-test cycle typically implemented in modern drug discovery efforts, considerable attention has been given to providing accurate pK_a measurements with good throughput. As summarized in a recent review, increasing attention given to pK_a during drug discovery is evidenced by the development of high-throughput methods for rapid pK_a determination.[3] While experimental methods continue to become more sophisticated and refined, it is often desirable to predict pK_as for "virtual compounds," i.e., those that have been described by a compound designer (chemist or modeler) but that have not yet been synthesized. Many different algorithms for predicting pK_as have been developed, and a few have been packaged into commercial computer software applications. With a host of pK_a estimation tools now available and with the potentially far-reaching impacts that the pK_a makes to the final properties of investigational compounds, questions naturally arise as to which prediction package performs best and what types of errors or bias might occur. Toward this end, we have measured the pK_a values for 211 druglike compounds and calculated their pK_as using a number of available software packages that implement differing estimation strategies.

METHODS

pK_a Determination. Experimental pK_a values were determined by a pressure-assisted capillary electrophoresis/ultraviolet absorbance spectroscopy technique as described previously.[4] Samples to be analyzed were prepared from 10 mM dimethylsulfoxide (DMSO) stocks and diluted into electrophoresis buffers to a typical assay concentration of 100 μM compound. Capillary migration was performed in ten buffers at an ionic strength fixed at 0.05, with the pH ranging from 2.50 to 11.0 in approximately single unit steps. Together with molecular weight and collected effective mobility, a model is chosen to be fitted for the pK_a values.[5] There are a total of nine different candidate equations that model the particular cases of monoacid, monobase, diacid, dibase, monoacid with monobase, tribase, triacid, diacid with monobase, and finally monoacid with dibase. The derivation and application of these equations was described previously in ref 6. Resulting pK_a values were returned after the curve fitting routine performed a nonlinear least-squares minimization to the best fitting model. Performance of the apparatus was routinely checked with standard samples of lidocaine (pK_a 7.90), ibuprofen (pK_a 4.55), and benzoic acid (pK_a 4.20) to ensure data quality. The reproducibility of results with this apparatus (95% confidence) is <0.1 pK_a units.

Compounds. Compounds were selected from discovery programs and represent 11 different chemotypes and a variety

* Corresponding author e-mail: john.manchester@astrazeneca.com; tel: (781) 839-4844; fax: (781) 839-4640.

10.1021/ci100019p
Published on Web 03/12/2010

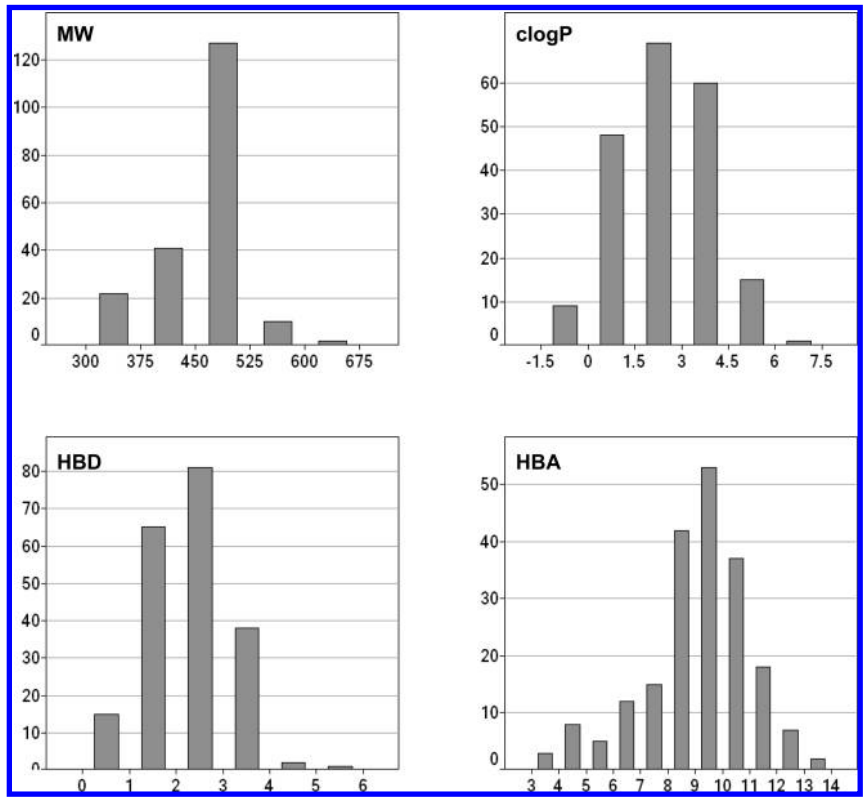


Figure 1. Physicochemical property distributions of the 211 compounds used in this study. Starting in the upper left are molecular weight (MW), computed octanol−water partition coefficient (clogP), hydrogen-bond donors (HBD), and hydrogen-bond acceptors (HBA). The *y*-axes are the numbers of compounds in each bin.

Table 1. Numbers and Types of Titratable Groups in the Data Set

acids		bases	
carboxylic acids	8	secondary amines	58
heterocyclic acids	8	tertiary amines	105
		oxazinones	155
		benzodiazepines	13
		heterocycles	31
total	16		362

of titratable groups. The physicochemical property distributions of the compound set are summarized in Figure 1, and structures for 85 of the compounds which have been previously disclosed along with the measured pK_as are available in the Supporting Information. Table 1 lists the numbers and types of titratable groups represented in the full set. The physical property distributions of the subset in the Supporting Information parallel those of the full set, as do the relative proportions of types of titratable groups, with the exception of tertiary amines (only 19 are represented) and benzodiazepines (11 of the total 13 are contained in the Supporting Information). Compounds that were tested had passed routine purity quality control standards. They were generally known to have both ^{1}H NMR and high-resolution mass spectra consistent with the desired structure as well as reverse-phase HPLC based purity measurements that indicate >90% purity.

Calculations. Computer programs for pK_a estimation from five different vendors were evaluated: ACDLabs/pK_a v10.0 and v12.0,[7,8] Marvin from ChemAxon,[9] MoKa from Molecular Discovery,[10] Pipeline Pilot from Accelrys,[11] and EpiK from Schrödinger.[12,13] However, differences between the apparent and microscopic pK_as were noted, and both sets of results are presented here.

RESULTS AND DISCUSSION

The estimation and calculation of pK_a values has a long and rich history within organic and medicinal chemistry. In the 1930s, Louis Hammett systematized the observation that the rates of reactions undergoing acid−base catalysis could be correlated with the ionization constants of the acids or bases participating in those reactions.[14] Using the pK_as of substituted benzoic acids as a baseline, he further showed that the relative effects of substituents on rates were transferable across a range of reactions and depended primarily on the nature of the ionizing species involved.[15] Thus, the pK_a of a known ionizing center (such as a carboxylic acid) in a novel environment can be estimated using the Hammett equation,

$$\mathrm{p}K_a = \mathrm{p}K_a^0 - \rho \sum \sigma_i \qquad (1)$$

in which the reference value pK_a^0 is perturbed by nearby chemical substituents *i* to varying degrees according to the substituent effects σ_i. The reactivity factor ρ is specific to the type of ionizing center and describes its sensitivity to substituent effects, which initially were confined to electronic effects. Taft later extended this approach to accommodate steric effects,[15−18] and eq 1 is sometimes referred to as the Hammett−Taft equation.

Several modern pK_a estimation methods are based on the Hammett−Taft approach. Probably the best known is the ACDLabs tool,[7] which has set the industry standard for accuracy, speed, and ease of use.[19] ACDLabs has adopted a very fine-grained approach, with distinct Hammett-type equations for more than 650 ionizing centers. The method is also extensively parametrized with more than 3000

substituent constants and incorporates methods for estimating substituent effects for species outside the training set. In addition, an algorithm is implemented for estimating apparent pK_as for multiprotic compounds, which can cause discrepancies between the predicted and measured values when the pK_as of different ionizing groups on the compound are within 1 to 2 units. The total training set size of about 16 000 compounds is typical among pK_a calculators of this type. The training set can be supplemented with user-supplied compounds using a convenient graphical interface, and prediction accuracy can also be improved by encoding additional Hammmett equations for novel ionizing centers or using more fine-grained descriptions of those centers.

EpiK,[12] a Hammett−Taft-type method developed by Schrödinger, follows a similar fine-grained approach to ACD, with over 800 ionizing centers and over 650 substituents encoded. However, a somewhat smaller set of approximately 3300 compounds was used to train the method. Algorithms are implemented to handle tautomerization and multiple ionizing centers, but the method is not presently able to estimate macroscopic pK_as. ChemAxon's Marvin[9] is another tool based on the Hammett−Taft approach[20] and apparently makes additional use of calculated atomic charges[21] in estimating pK_as. However, no further information on implementation or the size of the training set was available at the time of this writing.

Quantitative structure−activity relationships (QSARs) have also been developed to represent perturbations to an ionizing center, due to neighboring chemical groups, without explicitly constraining the form of the model as in the Hammett−Taft formalism. A few approaches have employed fast molecular orbital methods too,[21–24] and others have used topological descriptors to capture those affects.[25,26] However, extensibility beyond the original training/test sets seems to be a challenge among these methods.

An interesting development in the application of QSAR to pK_a estimation was the introduction of molecular tree structured fingerprints.[27] In this approach, the ionizing center and its chemical environment are represented with fingerprints, which are partitioned into blocks that represent chemical features discovered as the algorithms march outward from the ionizing center one bond at a time. The speed inherent in this sort of approach makes it amenable to training on large numbers of compounds, and in at least two cases, it allowed the exploitation of corporate databases to produce in-house pK_a models at two major pharmaceutical companies.[28,29] A third method, which makes use of GRID interaction fields in calculating fingerprints,[30] has been developed into the commercial application MoKa.[10]

Another topological method, available in Accelrys Pipeline Pilot,[11] uses path fingerprints to capture the environment of ionizing centers. Six models, trained on a total of 12 000 compounds, are used to independently describe aliphatic and aromatic acids and bases, phenols, and heterocyclic amines. If a test compound cannot be classified into one of those six groups, then a generic model based on either all acids or all bases is used. The models correlate the presence of specific paths of up to 6 atoms from an ionizing center with the difference in the pK_a of that center relative to the mean for its respective category.

The difference between topological methods and those based on the Hammett/Taft formalism is particularly interesting because in principle the two approaches capture the same information about the environment of the ionizing center. Does the theoretical formalism provide greater sensitivity and, thus, better accuracy or is the use of a large training set sufficient to represent chemical environments significantly different from the average, adding robustness to the generally greater speed of the informatics approach?

Table 2. RMSE of pK_a Models on 211 Multiprotic Druglike Compounds

	acids	bases	overall	*N* acids[a]	*N* bases[b]	*N* overall
ACD v10	0.6	0.8	0.8	13	348	361
ACD v12	0.8	0.8	0.8	12	343	365
ACD v10 micro	0.9	1.3	1.3	13	347	360
MoKa	1.6	0.9	1.0	14	357	370
EpiK	2.3	3.0	3.0	12	334	345
Marvin	0.8	0.9	0.9	14	355	370
Pipeline Pilot	1.0	2.7	2.6	11	357	368

[a] Total number of acidic groups detected by each method (out of a total of 14 determined experimentally). [b] Total number of basic groups detected by each method (358 were determined experimentally).

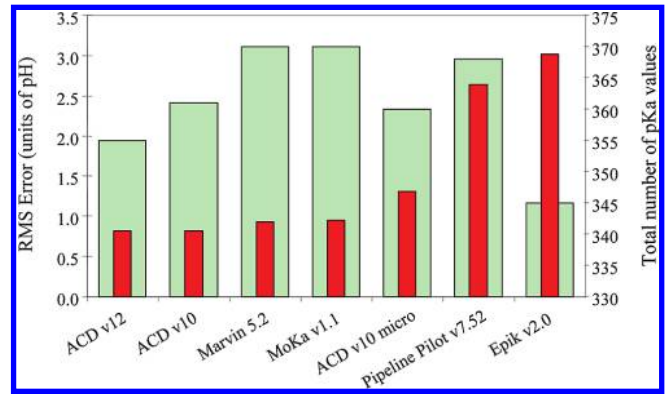


Figure 2. RMSE for the 7 different approaches examined (red/dark bars), superimposed on the total number of pK_a values predicted by each approach (green/light bars).

Table 2 shows the root-mean-squared error (RMSE) for the 7 different approaches tested. There are no statistically significant differences among the overall RMSE obtained among ACD v10 and v12 (in "apparent" pK_a prediction mode), MoKa, or Marvin ($p < 0.01$). These approaches all gave an RMSE of approximately 1 pK_a unit. This is similar to the RMSE of 0.90 obtained in the original validation of MoKa performed by its developers,[31] as well as the average absolute residual of 0.92 on druglike compounds obtained by the creators of EpiK,[12] but significantly larger than the mean quadratic errors of prediction observed for ACD and Marvin by others.[19] Pipeline Pilot and EpiK both gave significantly larger RMSEs of 2.6 and 3.0, respectively ($p < 0.01$). The difference in RMSE observed between "apparent" and "microscopic" pK_as calculated using ACD v10 is significant ($p < 0.05$).

Except for Pipeline Pilot, errors were similar between acidic and basic groups. This lower RMSE for Pipeline Pilot among acids is in addition to missing prediction values for several of the measured acids. Only MoKa and Marvin returned predicted values for all 14 measured pK_as for acidic groups. None of the methods identified all 358 basic groups; MoKa and Pipeline Pilot predicted values for 357, and Marvin followed closely with 355. These relationships can be more easily observed in Figure 2.

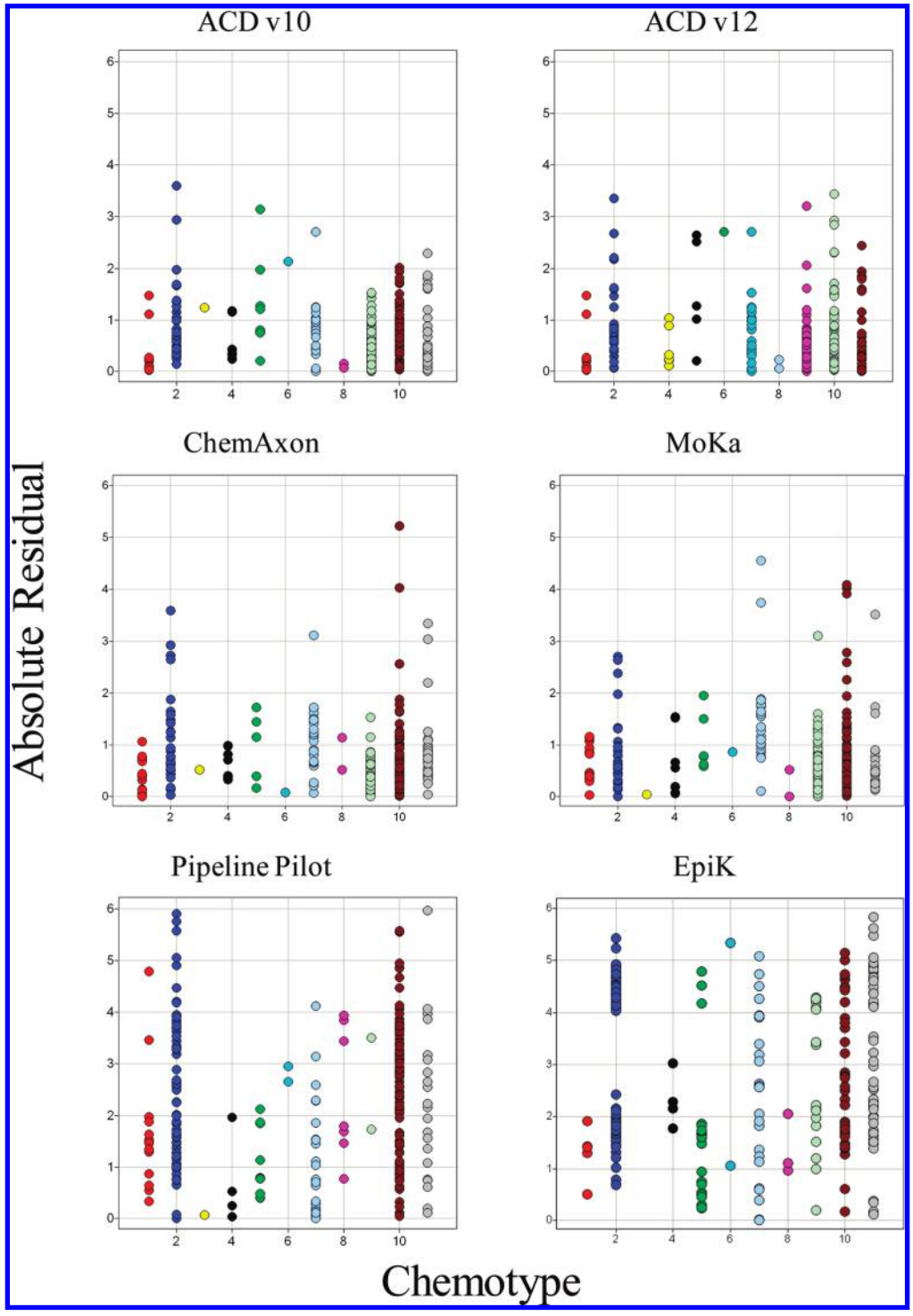


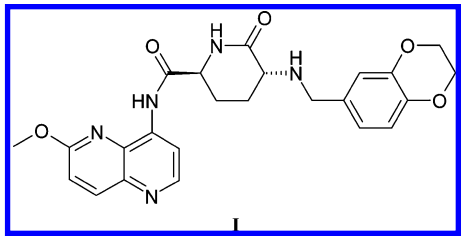
Figure 3. Absolute (unsigned) residual for each of the 211 compounds organized by chemotype across the methods tested. There does not appear to be any obvious "problem scaffold" for any of the methods.

These results differ from another recent evaluation of nine different pK_a estimation programs on about 200 druglike compounds in that the mean absolute errors for all methods in common between that study and this are here larger (for example, an error of about 0.5 was observed vs an error of 0.8 in the present analysis for ACD).[32] The authors point out, however, that the compounds and data used in that evaluation were taken from the literature and that it is likely that some or all of those data were used in training the various programs examined. pK_as for the compounds in the present study have not been previously disclosed, and even though the capability exists in ACD, Marvin, MoKa, and Pipeline Pilot, no attempt was made to train or bias those methods toward the data set.

Errors appear to be randomly distributed across the 11 chemotypes and a variety of titratable groups represented among the 211 compounds in the test set. Figure 3 shows the absolute residual for each pK_a value compared across the chemotypes (numbered 1 through 11).

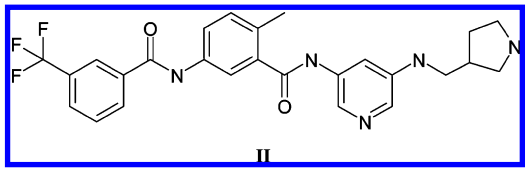
For example, consider compound **I**, for which the measured pK_a for the secondary amine is 6.5. ACD v10 and v12 predicted values of 6.7 and 6.9, in excellent agreement with the experiment, whereas MoKa and Marvin both predict values of 7.8. The difference in prediction between ACD

versions reflects differences in the parametrization of Hammett equations between the versions and is probably not significant, but it is interesting to note that v12 was able to predict exact apparent pK_as for this compound (as opposed to approximate apparent values otherwise used by default), and the value for the secondary amine is in exact agreement with the experiment at 6.5.



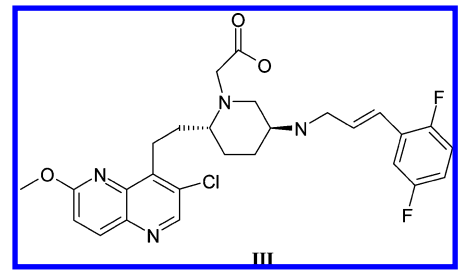
I

Compound **II** serves as a counter example, where MoKa and Marvin do better than ACD. Experimentally, the pK_a of the strongest base on **II** is 8.2. (In this case, assignment can be made relatively unambiguously to the pyrrolidine.) ACD v10 predicts a pK_a of 6.07, and ACD v12 predicts a value of 5.49, giving relatively large residuals. MoKa predicts a value of 8.97, and Marvin predicts a value of 8.11, in good agreement with the experiment.



II

As a final example, compound **III** has a measured pK_a for the strongest base of 8.7. MoKa predicts a value of 8.97 for the strongest base of **III**, in excellent agreement with the experiment (residual of 0.3), which it assigns to the amine. ChemAxon predicts a value of 9.33, also for the amine, giving a residual of 0.6. ACD v12 predicts a value of 10.87 for the strongest base, which it identifies as the piperidine, with a large residual of 2.9. Interestingly, ACD v10 predicts a value of 9.75, in much better agreement with the experiment, with a residual of 1.1.



III

These examples illustrate that ACD, MoKa, and Marvin are indistinguishable in that each can sometimes give large errors. There are a number of cases where the predictions from ACD12 are worse than those from ACD10. All the methods tested (except EpiK) can incorporate "local" data in order to bias predictions. However, these facilities were not evaluated since, to do so properly, it seems necessary to include some assessment of the models' ability to generalize beyond the "local" data, which was beyond the scope of the current effort.

Microscopic vs "Apparent" pK_a. A particular challenge for pK_a prediction methods are compounds that contain multiple titratable groups in which corresponding pK_as are similar in magnitude. Nicotinic acid is an example of such a compound. The pK_a of the isolated pyridine is ~4, and that of the carboxylic acid is about 4.5. However, pK_as of 2 and 6 are measured experimentally because the protonated microstate of the pyridine stabilizes the deprotonated microstate of the carboxylic acid and vice versa. The situation becomes much more complex with >2 groups with similar pK_as. This phenomenon drastically complicates the interpretation of experimentally determined pK_as, and often, the measured values represent an average of several simultaneous protonation "events". This ambiguity complicates pK_a prediction, as the models are generally trained to predict the microscopic pK_as of those isolated centers. It can also be troublesome for medicinal chemists who typically think of pK_a values as corresponding to particular ionizing centers. More to the point, it can undermine the medicinal chemists' confidence in (and willingness to use) pK_a predictions.

ACDLabs and ChemAxon have built into their pK_a prediction algorithms facilities which estimate the distribution of species among potential microstates and, then, estimate

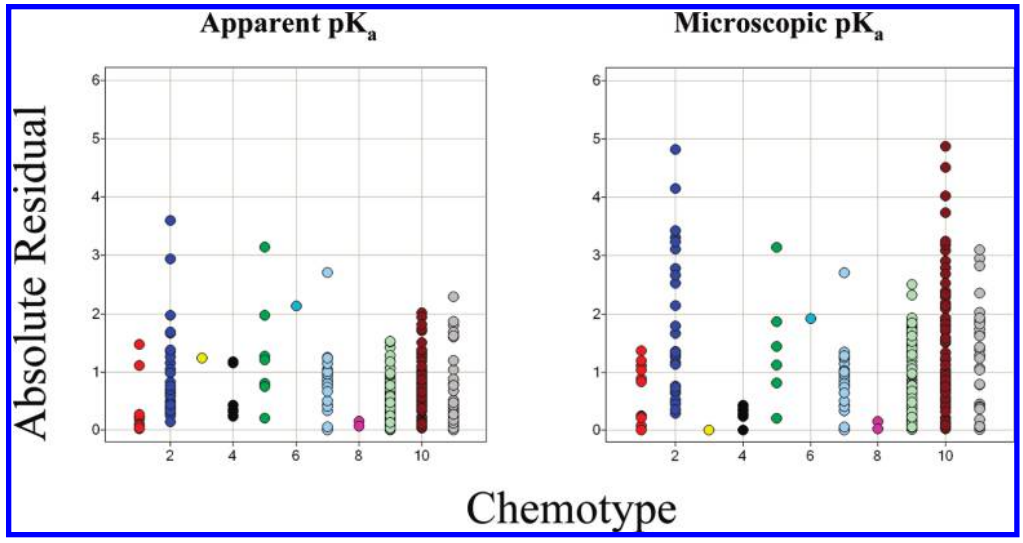


Figure 4. Absolute (unsigned) residuals from ACD v10 for each of the pKas measured for the 211 compounds across chemotypes. Apparent vs microscopic pK_as are compared on the left and right, respectively. The microscopic pK_a calculation tails into larger residuals.

the apparent pK_as resulting from that distribution. The result is (hopefully) better approximations to the experimentally measured pK_as for multiprotic compounds. To test the effect of the "apparent pK_a" perturbation algorithm, we compared the apparent and microscopic pK_as obtained for the 211 compounds using ACDLabs v10. The RMSE for the 211 compounds tested increases to 1.3 vs 0.8 for the "apparent" prediction. This difference is statistically significant ($p < 0.05$). Figure 4 shows that the additional error is not confined to a particular chemotype. Therefore, it appears that the "apparent" pK_a values, while sometimes harder to interpret for the medicinal chemist, do indeed provide overall better agreement with the experiment.

The overall accuracy of the methods examined seem to have more to do with the granularity of the models in describing various ionizing centers than to do with the modeling approach itself. MoKa uses 33 different models to describe a different ionizing center, whereas Pipeline Pilot uses only six; the error for Pipeline Pilot is substantially larger than for MoKa. ACD uses >1500 different Hammett equations to describe ionizing centers of varying complexity. EpiK recognizes 818 ionizing centers but is only trained on 3295 compounds (a total of 4114 pK_a values), indicating an obvious training set dependence. ACD, MoKa, and Pipeline Pilot have all been trained on >10 000 compounds (on the order of 25 000 distinct pK_a values).

It is important to note that these packages are not all intended for the same purposes. ACD is intended for high-resolution work, as in lead optimization, where a relatively small number of compounds are examined, a few at a time. Pipeline Pilot's method, on the other hand, is set up for rapid database screening. EpiK was developed primarily for rapidly assigning protonation states to (large) databases of virtual compounds prior to virtual screening, an application at which it excels. Although our purpose was not to compare CPU times and 211 compounds is probably insufficient to accurately distinguish the performance characteristics of the faster packages, we did qualitatively note that ACD v12 was significantly faster than v10. MoKa and Marvin were both very fast compared to ACD, and both seemed well suited for large database screening.

CONCLUSIONS

The topological method MoKa performed just as well as the Hammett−Taft based methods ACD and Marvin with an RMSE for all three methods of approximately 1 pK_a unit. Pipeline Pilot and EpiK performed significantly worse, highlighting the value of of fine-grained modeling for ionizing centers in combination with large training sets. The best RMSE observed for any of the models was ~1 pK_a unit, achieved by ACD, MoKa, and Marvin. All these packages with the exception of EpiK include the ability to bias predictions using "local" data, but these facilities were not evaluated. Finally, although formal benchmarks were not collected, both MoKa and Marvin seem fast enough for rapid database profiling and seem to offer the best all-around performance.

ACKNOWLEDGMENT

The authors are grateful to Molecular Discovery, ChemAxon, Schrödinger, and ACD for offering trial versions of and support for their softwares, making the work possible. The authors would also like to thank Nicola Colclough, Johan Ulander, Thierry Kogej, Lars Sandberg, and Andrew Leach for helpful discussions.

Supporting Information Available: Measured pK_as for 85 of the compounds used in this study are presented along with structures. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *1*−*3*, 3–26.
(2) Waterbeemd, H. v. d.; Gifford, E. ADMET in silico modeling: Towards prediction paradise? *Nature Rev. Drug Disc.* **2003**, 192−204.
(3) Wan, H.; Ulander, J. High-throughput pKa screening and prediction amenable for ADME profiling. *Exp. Opin. Drug. Metab. Toxicol.* **2006**, *1*, 139–155.
(4) Wan, H.; Holmen, A.; Nagard, M.; Lindberg, W. Rapid screening of pKa values of pharmaceuticals by pressure-assisted capillary electrophoresis combined with short-end injection. *J. Chromatogr., A* **2002**, *1*−*2*, 369–377.
(5) Miller, J. M.; Blackburn, A. C.; Shi, Y.; Melzak, A. J.; Ando, H. Y. Semi-empirical relationships between effective mobility, charge, and molecular weight of pharmaceuticals by pressure-assisted capillary electrophoresis: applications in drug discovery. *Electrophoresis* **2002**, *17*, 2833–2841.
(6) Poole, S. K.; Patel, S.; Dehring, K.; Workman, H.; Poole, C. F. Determination of acid dissociation constants by capillary electrophoresis. *J. Chromatogr., A* **2004**, *1*−*2*, 445–454.
(7) *ACDLabs pKa DB*, version 10.0; ACDLabs: Toronto, 2007.
(8) *ACDLabs pKa DB*, version 12.0; ACDLabs: Toronto, 2009.
(9) *Marvin*, version 5.2; ChemAxon: Budapest, 2009.
(10) *MoKa*, version 1.1; Molecular Discovery: Middlesex, UK, 2009.
(11) *Pipeline Pilot*, version 7.5.2; Accelrys: San Diego, CA, 2009.
(12) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, 681–691.
(13) *EpiK*, version 2.0; Schrödinger: New York, 2009.
(14) Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *1*, 125–136.
(15) Hammett, L. P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **1937**, 96–103.
(16) Taft, R. W. Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *J. Am. Chem. Soc.* **1952**, 3120–3128.
(17) Taft, R. W. Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters. *J. Am. Chem. Soc.* **1952**, 2729–2732.
(18) Taft, R. W. Linear steric energy relationships. *J. Am. Chem. Soc.* **1953**, 4538–4539.
(19) Meloun, M.; Bordovska, S. Benchmarking and validating algorithms that estimate pKa values of drugs based on their molecular structures. *Anal. Bioanal. Chem.* **2007**, 1267–1281.
(20) Csizmadia, F.; Tsantili-Kaoulidou, A.; Paderi, I.; Darvas, F. Prediction of distribution coefficient from structure. 1. Estimation method. *J. Pharm. Sci.* **1997**, *7*, 865–871.
(21) Dixon, S. L.; Jurs, P. C. Estimation of pKa for organic oxyacids using calculated atomic charges. *J. Comput. Chem.* **1993**, *12*, 1460–1467.
(22) Citra, M. J. Estimating the pKa of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere* **1999**, *1*, 191–206.
(23) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pKa prediction using group philicity. *J. Phys. Chem. A* **2006**, 6540–6544.
(24) Zhang, J.; Kleinoder, T.; Gasteiger, J. Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.* **2006**, 2256–2266.
(25) Balaban, A. T.; Khadikar, P. V.; Supuran, C. T.; Thakur, A.; Thakur, M. Study on supramolecular complexing ability vis-a-vis estimation fo pKa of substituted sulfonamides: dominating role of Balaban index (J). *Bioorg. Med. Chem. Lett.* **2005**, 3966–3973.
(26) Lee, A. C.; Yu, J.-y.; Crippen, G. M. pKa prediction of monoprotic small molecules the SMARTS way. *J. Chem. Inf. Model.* **2008**, 2042–2053.
(27) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pKa by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *87*, 0–879.

(28) Kogej, T.; Muresan, S. Database mining for pKa prediction. *Curr. Drug Disc. Tech.* **2005**, *4*, 221–229.
(29) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pKa for druglike compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.* **2007**, 450–459.
(30) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pKa prediction method using Grid molecular interaction fields. *J. Chem. Inf. Model.* **2007**, 2172–2181.
(31) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *6*, 2172–2181.
(32) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting pK(a) Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49*, 2801–2812.

CI100019P