articles

molecular pharmaceutics

In Silico Prediction of Ionization Constants of Drugs

Pil H. Lee,*,† Saravanaraj N. Ayyampalayam,‡ Lionel A. Carreira,*,‡ Marina Shalaeva,§ Shobha Bhattachar,|| Rachel Coselmon,|| Salwa Poole,⊥ Eric Gifford,# and Franco Lombardo§,▽

Computer Assisted Drug Discovery, Research Formulation, Pharmaceutical Science, Chemical Technology, Lead Discovery Group, Pfizer Global Research and Development, 2800 Plymouth Road, Ann Arbor, Michigan 48105, Computational Chemistry and Molecular Property Group, Eastern Point Road, Groton, Connecticut 06340, and Department of Chemistry, University of Georgia, Athens, Georgia 30602

Received February 12, 2007; Revised Manuscript Received June 13, 2007; Accepted June 14, 2007

Abstract: Most pharmacologically active molecules contain one or more ionizing groups, and it is well-known that knowledge of the ionization state of a drug, indicated by the pK_a value, is critical for understanding many properties important to the drug discovery and development process. The ionization state of a compound directly influences such important pharmaceutical characteristics as aqueous solubility, permeability, crystal structure, etc. Tremendous advances have been made in the field of experimental determination of pK_a, in terms of both quantity/speed and quality/accuracy. However, there still remains a need for accurate in silico predictions of pK_a both to estimate this parameter for virtual compounds and to focus screening efforts of real compounds. The computer program SPARC (SPARC Performs Automated Reasoning in Chemistry) was used to predict the ionization state of a drug. This program has been developed based on the solid physical chemistry of reactivity models and applied to successfully predict numerous physical properties as well as chemical reactivity parameters. SPARC predicts both macroscopic and microscopic pK_a values strictly from molecular structure. In this paper, we describe the details of the SPARC reactivity computational methods and its performance on predicting the pK_a values of known drugs as well as Pfizer internal discovery/development compounds. A high correlation ($r^2 = 0.92$) between experimental and the SPARC calculated pK_a values was obtained with root-mean-square error (RMSE) of 0.78 log unit for a set of 123 compounds including many known drugs. For a set of 537 compounds from the Pfizer internal dataset, correlation coefficient $r^2 = 0.80$ and RMSE = 1.05 were obtained.

Keywords: pK_a; in silico prediction; SPARC; macroscopic (microscopic) ionization constants; drugs; tautomer model; prediction error

Introduction

Most drug molecules contain one or more sites that can reversibly disassociate or associate a proton (a hydrogen ion) to form a negatively charged anion or a positively charged cation. The reversibility means that a compound is always in an equilibrium state with some fraction protonated and the rest deprotonated.[1]

The ionization state of a compound, indicated by the pK_a value, greatly influences many biopharmaceutical properties

* Authors to whom correspondence should be addressed. P.H.L.: Pfizer, Inc., Computer Assisted Drug Discovery, 2800 Plymouth Road, Ann Arbor, MI 48105; tel, 734-433-0218; fax, 734-622-2782; e-mail, pilhlee@gmail.com. L.A.C.: tel, 706-542-2050; e-mail, butch@sunlc3.chem.uga.edu.

† Computer Assisted Drug Discovery, Pfizer Global Research and Development.

‡ University of Georgia.

§ Computational Chemistry and Molecular Property Group.

|| Research Formulation, Pharmaceutical Science, Pfizer Global Research and Development.

⊥ Chemical Technology, Pfizer Global Research and Development.

Lead Discovery Group, Pfizer Global Research and Development.

▽ Current address: Novartis Institutes for Biomedical Research, Cambridge, Massachusetts.

10.1021/mp070019+ CCC: $37.00
Published on Web 07/13/2007

such as partition coefficient, aqueous solubility as a function of pH, and pharmacokinetic properties, such as blood−brain barrier (BBB) and permeability.[2] In the preformulation area, pK_a is exploited for forming salts of compounds in order to achieve desirable biopharmaceutical properties and solid-state characteristics that may be lacking in the free form of the compound.[3] There are also a number of studies reporting a correlation between pK_a and various ADMET and biological properties. One example is the apparent volume of distribution, a parameter used to predict the half-life of a drug in the body. Prediction of the volume of distribution in humans could be achieved via the Oie−Tozer equation.[4] This model requires the fraction of drug unbound in tissue to be determined. It has been shown[5] that the fraction unbound in tissue (f_{ut}) could be calculated using the fraction unbound in plasma (f_u), a value fairly easy to obtain, and physicochemical parameters: lipophilicity (E log D) and degree of ionization (f_i):

$$\log f_{ut} = 0.0080 - 0.2294(E \log D) - 0.9311 f_i(7.4) + 0.8885(\log f_u)$$

where

f_{ut} = fraction of drug unbound in tissue

f_u = fraction of drug unbound in plasma

$f_i(7.4)$ = fraction of drug ionized at pH = 7.4

The coefficient for the fraction ionized ($f_i(7.4)$) is the largest in the equation, thus underlining the significance of the state and degree of ionization. An accurate pK_a is needed to calculate the $f_i(7.4)$.

The experimental methods used to measure pK_a values have been well established and used extensively in drug discovery and development stages in the pharmaceutical industry. A comprehensive list of publications about pK_a determinations can be found in the review article.[6] However, accurate in silico prediction methods are used in many cases to estimate pK_a of compounds for which one has no physical sample, i.e., virtual compounds, and as an important guide when setting out to measure pK_a using experimental methods.

Since the theoretical prediction methods only need a chemical structure, the advantage of the prediction methods is in not requiring the physical samples of compounds. Also, fast, in silico prediction of pK_a for a large virtual combinatorial library of compounds is especially useful to evaluate compounds before synthesis.

A variety of prediction methods have been developed over the years including linear free energy relation[7] (LFER), CoMFA,[8] semiempirical,[9] and ab initio quantum mechanical calculations.[10] The LFER methods have been very successful and are based on a wealth of empirical data. While being fast, errors can be encountered if the molecule in question contains fragments not found in the training set. The LFER methods are implemented in widely used commercial programs.

More recently, many interesting and novel methods have also been developed by various groups: molecular tree structured fingerprints and PLS,[11] the COSMO-RS method, a combination of the quantum chemical dielectric continuum solvation model COSMO with a statistical thermodynamics treatment for more realistic solvation (RS) simulations,[12] a method based on semiempirical and information-based descriptors,[13] and an approach using group philicity.[14]

The current status and recent progress in computational approaches to pK_a prediction in terms of the accuracy limits are discussed in a recent review article.[15]

The computer program SPARC[16] (SPARC Performs Automated Reasoning in Chemistry) was developed to

(1) Wells, J. I. *Pharmaceutical preformulation*, 1st ed.; Ellis Horwood Ltd: London, 1998; p 25.

(2) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T. Estimation of pK_a using semiempirical molecular orbital methods. Part1: Application to phenols and carboxylic acids. *Quant. Struct.−Act. Relat.* **2002**, *21*, 457−472.

(3) Bhattachar, S. N.; Deschenes, L. A.; Wesley, J. A. Solubility: it's not just for physical chemists. *Drug Discovery Today* **2006**, *11* (21 and 22), 1012−1018.

(4) Oie, S.; Tozer, T. N. Effect of altered plasma protein binding on apparent volume of distribution. *J. Pharm. Sci*. **1979**, *68*, 1203−1205.

(5) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J. Med. Chem.* **2004**, *47*, 1242−1250.

(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**, *1037*, 445−454.

(7) Perrin, D. D.; Dempsey. B.; Serjeant, E. P. *pK$_a$ prediction for organic acids and bases*; Chapman and Hall: New York, 1981.

(8) Kim, K. H.; Martin, Y. C. Direct prediction of dissociation constants (pK_as) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted-imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. *J. Med. Chem.* **1991**, *34*, 2056−2060.

(9) Citra, M. J. Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere* **1999**, *38*, 191−206.

(10) Jang, Y. H.; Sowers, L. C.; Cagin, T.; Goddard, W. A., III. First principles calculation of pK_a values for 5-substituted uracils. *J. Phys. Chem. A* **2001**, *105*, 274−280.

(11) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870−879.

(12) Klamt, A.; Eckert, F.; Diedenhofen, M. First principle calculations of aqueous pK_a values for organic and inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pK_a scale. *J. Phys. Chem. A* **2003**, *107*, 9380−9386.

(13) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for drug like compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.*, in press.

(14) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a prediction using group philicity. *J. Phys. Chem. A* **2006**, *110*, 6540−6544.

(15) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Exp. Opin. Drug Metab. Toxicol.* **2006**, *2*, 139−155.

predict numerous physical properties such as vapor pressure, distribution coefficient, and GC retention time as well as chemical reactivity parameters such as pK_a and electron affinity. SPARC predicts both macroscopic and microscopic pK_a values strictly from molecular structure using relatively simple reactivity models.

In this paper, we describe the details of the SPARC reactivity computational methods and its performance on predicting the pK_a values of various organic compounds including many known drugs in comparison with experimental values. Previously, SPARC was used to calculate the pK_a of 4338 IUPAC screened pK_a measurements with the RMSE of 0.38. In this paper the SPARC system is run against a set of much more complicated and less well screened data.

Experimental Section

Experimental pK_a Measurements. The pK_a values were determined by capillary electrophoresis (CE), which is based on observation of the effective mobility of an ionizable compound in a series of electrolyte solutions of constant ionic strength and different pH. The pK_a values are obtained by fitting the effective mobility as a function of pH to a suitable model for the number of ionizable groups. Two very similar but slightly different experimental protocols have been used in two separate labs. The first protocol [protocol 1] is the pressure-mediated capillary electrophoresis method using the Combisep cePRO 9600, which has a 96 capillary array, enabling it to perform 96 separations simultaneously. Buffers are used in determining pK_a values based on log P values. For compounds with log $P < 3$, a set of 24 aqueous buffers (pH range 1.8 to 11.2) are used. For compounds with log $P > 3$, four sets of cosolvent buffers containing 30%, 40%, 50%, and 60% of methanol and ranging in pH from 2.1 to 10.8 are used. From the pK_a measurements at the different methanol concentrations, a plot of pK_a vs methanol concentration is created. The aqueous pK_a (0% methanol) is obtained by extrapolating the plot to zero methanol concentration. The average standard deviation for the measurements from this protocol is $\pm$0.1 pK_a unit. The detailed methodology can be found in ref 17. The second protocol [protocol 2] is using Beckman P/ACE System MDQ with a set of 12 aqueous buffers (pH range 2.0 to 11.5, see Table 2 in ref 6) are used. The average standard deviation for the measurements from this protocol is $\pm$0.2 pK_a unit. A detailed description of the method is illustrated in ref 6.

SPARC Computational Methods. The SPARC system addresses the calculation of physicochemical properties strictly from molecular structure. The query being addressed will normally fall into one of two categories. The first category is best described as a "whole molecule" problem, where the whole molecule interacts with itself or other molecules (solvents). Here the mechanistic interactions used to calculate these properties involve dispersion, induction, dipole, and hydrogen bonding interactions. This approach is the one used for calculating vapor pressure, boiling point, diffusion coefficient, electron affinity, activity coefficient, solubility, Henry's constant, and distribution coefficients in SPARC's physical properties calculator. The other category of calculations involves "reaction at a center" where the chemistry at a center is changed. These include ionization pK_a, hydrolysis, hydration, and tautomeric equilibria. The combination of the two types of calculation can lead to more exotic calculations such as gas-phase pK_a, nonaqueous pK_a, and reduction potentials. Even more complex are reactions that involve the coupling of several mechanisms such as ionization pK_a where there can be simultaneous tautomeric and hydration equilibria involved in the reaction.

In this paper we will concentrate on ionization pK_a and a description of the SPARC reactivity computational methods are presented. SPARC seeks to analyze chemical structure relative to a specific reactivity query in much the same manner as an expert chemist would do so. Molecular structures are broken into functional units with known chemical reactivity called reaction centers. The intrinsic behavior of the reaction center then is "perturbed" for the compound in question by describing mechanistically the effects on the basic reactivity of molecular structure appended to the reaction center using perturbation theory. SPARC utilizes a classification scheme that defines the role of structural constituents in effecting or modifying reactivity of the center and quantifies the various "mechanistic" descriptions commonly utilized in physical organic chemistry such as resonance, field effects (both direct and indirect (mesomeric)), sigma induction, intramolecular hydrogen bonding, and steric effects as they pertain to resonance and differential solvation of the reaction center.

A "toolbox" of mechanistic perturbation models has been developed that can be implemented where needed for a specific reactivity query. Resonance models were developed and calibrated on light absorption spectra,[18] whereas electrostatic models (direct/indirect field and sigma induction effects) were developed on ionization pK_a.[18,19] The SPARC computational approach is based on blending well-known, established methods such as SAR,[20,21] LFER,[22] and PMO theory.[23,24] SPARC uses SAR for structure−activity analysis,

(16) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. A rigorous test for SPARC's chemical reactivity models: estimation of more than 4300 ionization pK$_a$s. *Quant. Struct.−Act. Relat.* **1995**, *14*, 348−355.

(17) Zhou, C.; Jin, Y.; Kenseth, J. R.; Stella, M.; Wehmeyer, K. R.; Heineman, W. R. Rapid pK$_a$ estimation using vacuum-assisted multiplexed capillary electrophoresis (VAMCE) with ultraviolet detection. *J. Pharm. Sci.* **2005**, *94*, 576−589.

(18) Karickhoff, S. W.; McDaniel, V. K.; Melton, C. M.; Vellino, A. N.; Nute, D. E.; Carreira, L. A. Predicting chemical reactivity by computer. *Environ. Toxicol. Chem.* **1991**, *10*, 1405.

(19) Hilal, S. H.; Carreira, L. A.; Melton, C. M.; Baughman, G. L.; Karickhoff, S. W. Estimation of ionization constants of azo dyes and related aromatic amines: environmental implication. *J. Phys. Org. Chem.* **1994**, *7*, 122−141.

(20) Lowry, T. H.; Richardson , K. S. *Mechanism and Theory in Organic Chemistry*, 3rd ed.; Harper & Row: New York, 1987.

LFER to estimate thermodynamic or thermal properties, and PMO theory to describe quantum effects such as charge distribution, delocalization energy, and polarizability of a π electron network.

Our approach to predict chemical reactivity involves the location of primary reactive units within the molecule. These reactive sites, which are termed reaction centers (C), are in general the smallest subunits to which the reactivity of interest can be ascribed. Any molecular structure appended to C is viewed as a "perturber" (termed perturber structures, P). For ionization pK_a,

$$P{-}C_i \overset{K_a}{\longleftrightarrow} P{-}C_f$$

where i and f denote the initial and final states of the reaction center, P is the perturber structure presumed to be unchanged by the reaction, and pK_a is the ionization equilibrium constant. The pK_a is expressed in terms of contributions of the components P and C,

$$pK_a = (pK_a)_C + \delta_P(pK_a)_C$$

where $(pK_a)_C$ is the ionization constant of the unperturbed reaction center, and $\delta_P(pK_a)_C$ is the pK_a perturbation brought about by the perturber structure. All perturbations are to the overall pK_a reaction and are therefore all differential in nature. E.g., the resonance perturbation is the differential of the resonance stabilization of the initial and final states.

The pK_a perturbation is factored into mechanistic components,

$$\delta_P(pK_a)_C = \delta_{elec}(pK_a)_C + \delta_{res}(pK_a)_C + \delta_{solv}(pK_a)_C + \delta_{HB}(pK_a)_{C-S}$$

where $\delta_{res}(pK_a)_C$, $\delta_{elec}(pK_a)_C$, $\delta_{solv}(pK_a)_C$, and $\delta_{HB}(pK_a)_{C-S}$ describe the differential resonance, electrostatic, solvation, and hydrogen bonding of P with the protonated and unprotonated states of C, respectively. Electrostatic interactions are derived from local dipoles or charges (monopoles) in P interacting with charges or dipoles in C. δ_{res} describes the change in the delocalization of π electrons of the two states due to P. Hydrogen bonding and solvation effects are derived from interactions of the structural elements of P that are contiguous to C with the two states through hydrogen bonding or steric blockage of solvent access to C, respectively.

Currently, the SPARC is confined to elements H, B, C, N, O, halogens, P, S, some As, and Se. Allenic groups are not fully implemented.

(21) Taft, R. W. *Progress in Organic Chemistry*; John Wiley & Sons: New York, 1987; Vol. 6.
(22) Hammett, L. P. *Physical Organic Chemistry*, 2nd ed.; McGraw Hill: New York, 1970.
(23) Dewar, M. J. S.; Dougherty, R. C. *The PMO Theory of Organic Chemistry*; Plenum Press: New York, 1975.
(24) Dewar, M. J. S. *The Molecular Orbital Theory of Organic Chemistry*; McGraw Hill: New York, 1969.

SPARC's Chemical Reactivity Models. The modeling of the perturber effects for pK_a relates to the reactivity center, C. S denotes substituent groups that "instigate" this perturbation. For electrostatic effects, S contains electric dipole and/or monopole fields; for resonance, S donates or receives electrons from the reaction center. R links the substituent, S, and reaction center, C, and serves as a conductor of the perturbation (for field the R acts as a spacer and modifies the distance from S to C, for resonance R may be an intervening π network that can conduct electrons). A given substituent, however, may be part of the structure, R, connecting another substituent to C, and thus functions as a conductor for the second substituent.

The modeling of the perturber effects for chemical reactivity relates to the structural representation:

$$S{-}_iR_j{-}C$$

where $S{-}_iR_j$ represents the perturber P appended to the reaction center C. The *i* and *j* represent anchor atoms in R that connect to S and C, respectively. Perturbations are factored into three independent components for the structural components C, S, and R: (1) substituent strength, which describes the potential of a particular S to "exert" a given effect (e.g., for electrostatic interactions this would be the magnitude of the dipole and/or monopole strength), (2) molecular network conduction, which describes the "conduction" properties of the molecular structure R, connecting S to C with regard to a given effect, and (3) reaction center susceptibility, which rates the response of C to the effect in question.

For each reaction center and substituent, SPARC catalogs appropriate characteristic parameters. Substituents include all non-carbon atoms and aliphatic carbon atoms contiguous to either the reaction center or a π unit. Some heteroatom substituents containing π groups are treated collectively as substituents (e.g., $-NO_2$, $-C{\equiv}N$, $-C(=O)O$, etc.). The only requisites are that they be structurally and electronically well-defined. Also, these units must be terminal with regard to resonance interactions (no pass-through resonance).

The contributions of the structural components C, S, and R are quantified independently. For example, the strength of a substituent in creating an electrostatic field effect depends only on the substituent regardless of the C, R, or the reactivity property of interest. This allows substituent strength to be tabulated and used in developing other models such as hydrolysis and hydration. The goal of SPARC is to develop this type of mechanistic toolbox using reactivities such as pK_a that are abundant and well-measured and then use these tools in areas where there is a paucity of data. Likewise, the molecular network conductor R is modeled so as to be independent of the identities of S, C, or the property being estimated. The susceptibility of a reaction center to an electrostatic effect quantifies only the differential interaction of the initial state versus the final state with the electrical field. The susceptibility of the reaction center to this perturbation gauges only the reaction $C_{initial} - C_{final}$ and

is completely independent of both R and S. *This factoring and quantifying of each structural component independently provides parameter "portability" and, hence, permits model portability to all structures and, in principle, to all types of reactivity.* SPARC's chemical reactivity models have been designed and parametrized to be portable to any chemical reactivity property and any chemical structure. For example, chemical reactivity models are used to estimate macroscopic/microscopic ionization pK_a for both simple and complex organic compounds,[18,19,22] and the same reactivity models have been used to calculate electron affinity,[25] hydrolysis,[26,27] and hydration.[28]

Electrostatic Effects Models. Electrostatic effects on reactivity are derived from charges or electric dipoles in the substituents (S) interacting through space (R) with charges or dipoles in the reaction center (C). These effects include direct electrostatic effects (field effect), indirect electrostatic effects (mesomeric effect), and sigma induction effects.

Field Effect Model. The field effect of a given substituent is given by a multipole expansion

$$\delta_{\text{field}}(\text{p}K_a)_C = \frac{\delta q_C q_S}{r_{CS} D_e} + \frac{\delta q_C \mu_S \cos\theta_{CS}}{r^2 D_e}$$

where q_S is the charge on the substituent, approximated as a point charge located at a point in S; μ_S is the substituent dipole located at a point in S; δq_C is the change in charge of the reaction center accompanying the ionization reaction, presumed to be located at point C; cos θ_{CS} gives the orientation of the S dipole relative to C, the r's are the appropriate distances of separation, and D_e is the effective dielectric constant for the intervening molecular conduction medium.

To facilitate model portability, each term in this equation is resolved into contributions of the structural component S, R, and C,

$$\delta(\text{p}K_a)_{\text{field}} = \rho_{\text{elec}} \sigma_R F_S$$

where ρ_{elec} is the susceptibility of a given reaction center to electric field effects, F characterizes the magnitude of the field component, charge or dipole, on the substituent (D_e is subsumed into this number), and σ_R is the appropriate $1/r_{CS}$ or $1/r_{CS}^2$ term. An uncharged substituent has one field strength parameter (F_i, dipole field strength), whereas a charged substituent has two, F_q and F_i. F_i describes the effective substituent dipole inclusive of the anchor atom i, which is assumed to be a carbon atom. For cases involving a non-carbon anchor atom, F_i is adjusted based on the electronegativity of the anchor atom relative to carbon. F_i incorporates the effective dielectric constant for the molecular cavity and any unit conversion factors for charges, distances, etc. F_q describes the effective S charged field strength. F_q incorporates effective charge on S as well as D_e and any unit conversion factors.

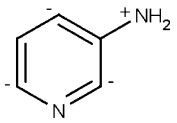
For all reaction centers, electrostatic interactions are calculated relative to a fixed geometric reference point, c, which was chosen to approximate the center of charge for the carboxylate anion, $r_{Cj} = 1.3$, where the length unit is the aromatic carbon−carbon length. The ρ_{elec} for the reaction center reflects electric field changes for these reactions gauged relative to the carboxylic acid reference, but also subsumes any differences in charge distribution relative to the reference point, c.

The distances between the reaction center and the substituent are estimated as a summation of the respective distance contributions of C, R, and S as

$$r_{CS} = r_{Cj} + r_{ij} + r_{iS}$$

r_{ij} is calculated by summation over delineated units on the shortest molecular path from i to j. All aliphatic bonds contribute 1.1 units; double and triple bonds contribute 0.9 and 0.8 unit, respectively. For aromatic ring systems SPARC uses templates for benzene, naphthalene, anthracene, and phenanthrene that contain distances between each constituent atom pair. For other polyaromatic hydrocarbons the system creates the template on the fly to get the distances. The dipole orientation factors, cos θ_{CS}, are ignored except in those cases where S and C are attached to the same rigid R_π unit. In these situations, they are assumed to depend solely on the point(s) of attachment, (i,j), and are precalculated and stored in SPARC databases. Conventional bond angles are assumed except for ortho configurations where substituent bond angles are expanded slightly.

Mesomeric Field Effect. The mesomeric field effect was first proposed to explain the very large perturbations on pK_a values of amino pyridines and guanidines.



The substituent, NH_2 group (S) can "induce" electric fields in the R (the aromatic ring) that can interact electrostatically with C (the pyridine nitrogen). This indirect interaction is called the "mesomeric field effect." The amino group in this structure should exert a positive *direct field effect* and lower the pK_a. Here, however, the observed effect is exactly the opposite of that predicted. The pK_a of *m*-aminopyridine is 6.1, which is greater than the pK_a of pyridine (5.2). In this case, the NH_2 induced negative charges are ortho and para to the in-ring N. The loss of charge from the eminus N to

(25) Hilal, S. H.; Carreira, L. A.; Karickhoff, S. W.; Melton, C. M. Estimation of electron affinity based on structure activity relationships. *Quant. Struct.−Act. Relat.* **1993**, *12* (4), 389−396.

(26) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A.; Shrestha, B. P. Estimation of carboxylic acid ester hydrolysis rate constants. *QSAR Comb. Sci.* **2003**, *22* (9−10), 917−925.

(27) Whiteside, T. S.; Hilal, S. H.; Carreira, L. A. Estimation of phosphate ester hydrolysis rate constants. I. Alkaline hydrolysis. *QSAR Comb. Sci.* **2006**, *25* (2), 123−133.

(28) Hilal, S. H.; Bornander, L. L.; Carreira, L. A. Hydration equilibrium constants of aldehydes, ketones and quinazolines. *QSAR Comb. Sci.* **2005**, *24* (5), 631−637.

the ring induces a more positive charge on the N (S). The positive induced N charge and the direct field effect lower the pK_a of the in-ring N whereas the negative charges raise the pK_a of the in-ring N. The proximity of the negative charges ortho to the in-ring N leads to an overall increase in pK_a.

In SPARC, this mesomeric field effect is treated as a collection of discrete charges, q_R, with the contribution of each described by the equation below. As was the case for the direct field effects, it is desirable to resolve, and to parametrize independently, the contributions of structural units S, R and C,

$$\delta_{M_F}(\mathrm{p}K_a)_C = \rho_{elec} M_F \sum_k \frac{q_{ik}}{r_{kC}}$$

where M_F gauges the ability or strength of a given S to induce a field in R_π. It describes the π-induction ability of a particular substituent relative to a surrogate CH_2^-. q_{ik} is the charge induced at atom k, with the reference probe attached at atom i calculated using PMO theory. r_{kC} is the through-cavity distance to C as described in the direct field effect. The magnitude of a given M_F parameter describes the relative field strength, and the sign of M_F specifies the positive or negative character of the induced charge in the R_π.

Sigma Induction Model. Sigma induction derives from electronegativity differences between the reaction center and the substituent. This effect is transmitted progressively through a chain of σ-bonds among atoms. This is a short-range interaction that is strong when the two atoms are bonded to each other and any effect beyond the second atom is negligible and ignored. The σ contribution is expressed as

$$\delta_{sigma}(\mathrm{p}K_a)_C = \rho_{elec} d\chi_{CS}$$

where ρ_{elec} is the susceptibility of a given reaction center to electric field effects and $d\chi_{cs}$ is the difference in the effective electronegativity of C and S. The electronegativities of the reaction centers and the substituents are estimated based on the electronegativity of the methyl group that was chosen to be the reference group.

Resonance Model. Resonance involves the delocalization of π electrons into or out of the reaction center. Resonance stabilization energy in SPARC is a differential quantity, related directly to the extent of electron delocalization in the initial state versus the final state of the reaction center. The source or sink of electrons in P may be the substituents and/or R-π units contiguous to the reaction center. Substituents that withdraw electrons from a reference point (e.g., $-NO_2$, $-C{=}O$, etc.) are designated S+, and those that donate electrons (e.g., $-NR_2$, $-OH$, etc.) are designated S−. The R-π units withdraw or donate electrons or may serve as "conductors" of π-electrons between resonance units. Reaction centers are likewise classified as C+ and C− denoting withdrawing and donating of electrons, respectively. To model this effect, the reaction center is replaced by a surrogate electron donor (reference source), CH_2^-. The distribution of nonbonded molecular orbital (NBMO) charge from this surrogate donor is used to quantify the acceptor potential for the perturber structure P. The reactivity perturbation is given by

$$\delta_{res}(\mathrm{p}K_a)_C = \rho_{res}(\Delta q)_C$$

where $(\Delta q)_C$ is the fraction loss of NBMO charge from the surrogate reaction center, and the susceptibility, ρ_{res}, of a given reaction center to resonance quantifies the differential "donor" ability of the two states of the reaction center relative to the reference donor CH_2^-. In parametrization of resonance effects, resonance strength, E_r, is defined for all the substituents (i.e., the ability to donate or receive electrons). Resonance susceptibility is defined for all the reaction centers. Resonance "conduction" in R_π networks is modeled so as to be portable to any array of R_π units or to linking any resonant source or sink groups.

Solvation Effects Model. C_i and C_f frequently differ substantially in degree of solvation, with the more highly charged moiety solvating more strongly. Thus, steric blockage of the reaction center is distinguished from the steric-induced twisting of the reaction center incorporated in electron delocalization interactions. Differential solvation is a significant effect in the protonation of organic bases (e.g., $-NH_2$, in-ring N, $=$N) but is less important for acidic compounds except for highly branched aliphatic alcohols.

In SPARC's reactivity models, differential solvation of the reaction center is incorporated in $(\mathrm{p}K_a)_C$, ρ_{res}, and ρ_{elec}. If the reaction center is bonded directly to more than one hydrophobic group or if the reaction center is *ortho* or *para* to hydrophobic substituent, then $\delta_{sol}(\mathrm{p}K_a)_C$ must be calculated. The $\delta_{sol}(\mathrm{p}K_a)_C$ contributions for each reaction center bonded directly to more than one hydrophobic group are quantified based on the sizes and the numbers of hydrophobic groups attached to the reaction center and/or to the number of the aromatic bridges that are *approximate* to the reaction center as

$$\delta_{solv}(\mathrm{p}K_a)_C = \rho_{solv}(\nu_i + \nu_j + \nu_k)$$

where ρ_{solv} is the susceptibility of the reaction center to differential solvation due to steric blockage of the solvent, and ν are the solid angles occluded by the hydrophobic P that is bonded directly (i), *ortho* (j), or *perri* (k) to the reaction center.

Intramolecular Hydrogen-Bonding Effects Model. Intramolecular hydrogen bonding is a direct site coupling of a proton-donating (α) site with a proton accepting (β) site within the molecule. The reaction center might interact with S through intramolecular hydrogen bonding and thus impact the pK_a. The C_i and C_f frequently differ substantially in degree of hydrogen-bonding strength with a S. In aromatic, π-ring or π-aliphatic (e.g., diguanide) systems, where the reaction center is contiguous to the substituent and where a stable 4-, 5-, or 6-member ring may be formed, $\delta_{HB}(\mathrm{p}K_a)_{C-S}$ must be estimated. $\delta_{HB}(\mathrm{p}K_a)_{C-S}$ is a differential quantity that

describes the H-bonding differences of the C_i vs C_f with S as

$$\delta_{HB}(pK_a)_{C-S} = HB_C S_i F$$

where the HB_C is the differential H-bond strength for C−S where C and S approximate to each other, S_i is a reduction factor for steric-induced twisting of C, and F is 1 or 0.6 for aromatic and π-ring systems, respectively. For C that might H-bond with more than one substituent, the H-bonding contribution for each S is calculated and the stronger contributor is selected.

Temperature Dependence. For processes that can be modeled in terms of some equilibrium (or pseudo equilibrium component), the temperature dependence can be expressed in the van't Hoff representation,

$$f(\Delta pK_a) = A_C + \delta_S(\Delta pK_a)_C + [B_C + \delta_H(\Delta pK_a)_C]/T$$

where A_C and B_C are the entropic and the enthalpic van't Hoff coefficients for the reaction center and δ_H and δ_S are enthalpic and entropic perturbations, respectively. To date, all perturbations have been assumed to be predominantly enthalpic and δ_S has been assumed to be zero. The van't Hoff factors (A and B) can be derived from temperature data for the reaction center or inferred from simple structures with minimal perturbational contributions. $\delta_H(\Delta pK_a)_C$ is a sum of all the effects described above. When the enthalpic perturbation cancels the B parameter as in the *p*-nitroaniline, little or no temperature dependence is observed. Some systems may have perturbations large enough to change the sign of the slope of the pK_a temperature dependence (e.g., the third pK_a of phosphoric acid).

Calculating Macroscopic Ionization Constants. The methods described above allow SPARC to calculate microscopic pK_a ionization constants. Microscopic pK_a ionization constants describe the equilibrium that exists between the two species that are related by the loss of a proton. Molecules that contain only one ionizable group need only this micro constant to describe the state of ionization and the distribution of the two species as a function of pH. The problem of describing the state of ionization and the distribution of species as a function of pH grows exponentially with the number of ionizable sites. For a molecule that has N ionizable sites, there are N macroscopic ionization constants which can be measured. There are, however, $2^{N-1} \times N$ microscopic ionization constants and 2^N microscopically different species or states. For example, tyrosine contains 3 ionizable sites: the carboxyl, aromatic hydroxyl, and ammonium groups. Since each of the three groups may exist in either of two states, tyrosine may exist in 8 (2^3) microscopically different forms. The most positive of these 8 states is the cation, with net charge $Z = 1$; the most negative is the divalent anion, with $Z = -2$. Each of the two intermediate states of net charge $Z = 0$ and $Z = -1$, respectively, may have three microscopically different forms. Each of the ionizable groups in tyrosine is characterized by four micro constants, since the tendency of each group to accept or donate a proton depends on the ionization state of the other two groups. Hence, there are 12 (3×2^2) microscopic ionization constants connecting the 8 species. A microscopic ionization constant governing the reaction $A \leftrightarrow B + H^+$ can be expressed as $K = [B][H^+]/[A]$ and is a nonlinear constant. However, if the pH is held constant then $P = K/[H^+] = [B]/[A]$ is a linear constant relating the initial and final species. At a given pH all of the possible species are connected through a coupled linear network. At a given pH (constant H^+) the fraction of any species that exists as a result of ionization can be expressed as $D_{ij...k}/D$ where D can be expressed as

$$D = \frac{1}{0!} + \frac{\sum_i k_i[H]^{L_i}}{1!} + \frac{\sum_i \sum_{j \neq i} k_i k_{ij}[H]^{L_{ij}}}{2!} + \ldots + \frac{\sum_i \sum_{j \neq i} \cdots \sum_{k \neq i,j\ldots} k_i k_{ij} \ldots k_{ij\ldots k}[H]^{L_{ij\ldots k}}}{N!}$$

and $L_{ij...k}$ is the charge of the final state ($ij...k$ state). The factorial is the number of different thermodynamic paths that lead to the $ij...k$ state, and $D_{ij...k}$ is one of terms in the denominator (D). For example, the fraction of neutral species would be $1/D$ and the fraction of a singly ionized species would be $k_i \times [H]L_i/D$. The macroscopic pK_a values are determined by generating the fraction of each species as a function of pH. The curves of species having the same charge are summed. The pH at crossing of the summed curves for the species having a net charge n with the summed curves for the species having net charge $n + 1$ represents a macroscopic pK_a.

Table 1 illustrates the detailed contribution from each term using several examples.

Tautomer Equilibrium Constants. Tautomers are rapidly converting isomers of a structure. Tautomerism is a chemical process in which the double bonds in a molecule are rearranged with synchronized hydrogen atom shift to form an isomer. Tautomeric isomers play a very important role in determining the physicochemical properties of a molecule.

Tautomerism is a two step process; first the molecule is attacked by an acid or base for the gain or loss of a proton respectively and a double bond rearrangement. The second step is the loss or the gain of the proton from the first step. Hence it is catalyzed by acid or base. This makes the process dependent on the solvent properties.

To model the tautomeric equilibrium constant (K_T), the energy difference between the two isomeric structures, the ketone form and the enol form, has to be modeled. SPARC operates as a perturbation calculator to a reference structure; hence, it cannot directly model the tautomeric equilibrium constant. Rather SPARC uses an indirect thermodynamic loop to calculate K_T.

SPARC Tautomer Model. SPARC does not calculate absolute energies of molecular structures. To calculate tautomeric equilibrium constants, SPARC uses its microscopic pK_a and Henry's constant models to calculate K_T using

Table 1. Example Calculations and Mechanisms

SMILES	ref[a]	stat[b]	res[c]	field[d]	MF[e]	sigma[f]	HB[g]	solv[h]	calcd[i]	obsd[j]
CO	14.3	0	0	0	0	1.2	0	0	15.5	15.1−15.5
Oc1ccccc1	14.3	0	−4.3	0	0	0	0	0	10.0	9.9−10.0
Oc1ccc(N(=)=O)cc1	14.3	0	−5.95	−1.09	−0.52	0	0	0	6.74	6.9−7.1
Oc1ccc(N(=)=O)ccc1	14.3	0	−4.3	−1.42	−0.21	0	0	0	8.38	8.2−8.4
Oc1c(N(=)=O)cccc1	14.3	0	−5.61	−2.67	−0.04	0	1.09	0	7.06	7.2
CN	9.83	−0.48	0	0	0	1.06	0	0	10.42	10.6
CNc1ccccc1	9.83	−0.3	−4.35	0	0	1.01	0	−1.08	5.11	4.8
Nc1ccc(N(=)=O)cc1	9.83	−0.48	−6.41	−1.46	−0.72	0	0	0	0.76	1.0
Nc1ccc(N(=)=O)ccc1	9.83	−0.48	−4.45	−1.91	−0.27	0	0	0	2.72	2.5−2.6
Nc1c(N(=)=O)cccc1	9.83	−0.48	−6.04	−3.58	−0.06	0	−0.14	0	−0.47	−0.3
n1ccccc1	2.36	0	2.59	0	0	0	0	0	4.96	5.2
n1ccc(N)cc1	2.36	0	5.20	−0.40	1.76	0	0	0	8.91	9.1
n1ccc(N)ccc1	2.36	0	2.59	−0.59	1.29	0	0	0	5.65	6.0
n1c(N)cccc1	2.36	0	5.20	−1.18	0.32	0	0.02	0	6.72	6.7−6.8

[a] pK_a of reaction center. [b] Statistical factor. [c] Resonance effect. [d] Direct field effect. [e] Mesomeric field effect. [f] Sigma induction effect. [g] Intramolecular hydrogen-bonding effect. [h] Solvation effect. [i] Calculated pK_a. [j] Experimentally observed pK_a.

a thermodynamic loop. This thermodynamic loop consists of five distinct steps.

After identifying the atoms involved in the tautomeric process (atoms 8, 9, and10 of molecule 1), the pK_a for the deprotonation of the sp^3 atom (atom 10) is calculated using SPARC's pK_a calculator. This pK_a gives us the free energy to deprotonate the molecule.

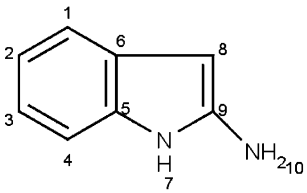
In the next step the ionized molecule will be transferred to vacuum. The energy required to complete this step is calculated using SPARC's Henry's constant calculator.

Once the ionized molecule is moved into space, it can be rearranged to the ionized form of the other tautomer with a synchronized switching of the double bond and the lone pair. Both the structures are resonance structures of the same ion. The energy cost for this rearrangement is zero.

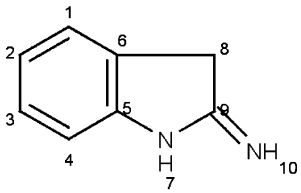
Next the rearranged molecule is transferred back into solution. The energy required for this transfer is calculated using the Henry's constant model.

Finally the rearranged molecule is now protonated and the energy is calculated by determining the pK_a of the sp^3 atom (atom 8 of molecule 2).

Using this approach SPARC can also calculate the equilibrium constant of all types of tautomers. SPARC can calculate pK_a in various solvents and makes it possible for the tautomer model to calculate the pK_T in various solvents,



Molecule 1



Molecule 2

The model is illustrated in Figure 1, where the tautomeric equilibrium is calculated for 2-amino-3-methyl-1*H*-indole.

Limitations and Workarounds. In the process of modeling tautomer equilibrium constants using experimentally observed data, it was realized that the solvation energy

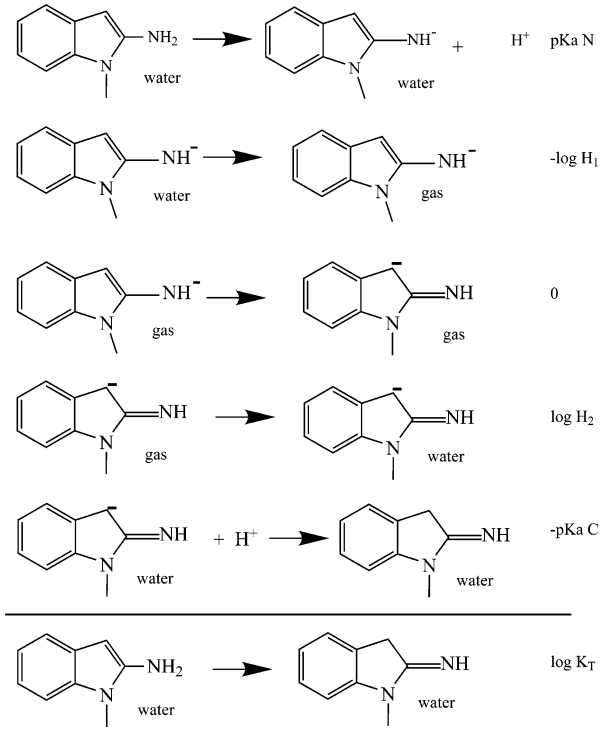


Figure 1. SPARC thermodynamic loop to calculate the tautomeric equilibrium constant.

difference between the two ionic forms is negligible and is within the noise range of the system. So, under normal operation circumstances the differential Henry's energy term would be ignored.

The SPARC tautomer model relies on the pK_a model to perform the necessary ionization calculations. Most of the ionization pK_a calculations performed for the tautomer model are of the type carbon acid, nitrogen acid, and hydroxy acid ionizations. The perturbation effects in the case of carbon acid and nitrogen acid pK_a values are very large. This large perturbation effect leads to a large error of estimation for these calculations, which in turn reflects on the error of estimation for the tautomer equilibrium calculation. The other

reason for the poorer performance of the carbon acid model is the scarce availability of well-measured carbon ionization constants. In order to keep the estimation error low during the tautomeric network calculations, we devised a reliability assignment to all tautomer calculations. These reliabilities are the product of the reliability score of the two pK_a calculations involved. In this system all carbon acid pK_a calculations are assigned a score of 0.5 and other pK_a calculations are assigned a score of 1. The reliability scores calculated are used to eliminate less reliable calculations when averaging the K_Ts for a node.

A list of SPARC calculated tautomeric equilibrium constants and the observed values[29] are listed in Table 2. A plot of the observed and calculated pK_T values is shown in Figure 2. Based on the very small number of published experimental investigations of the tautomeric properties of chemicals, the r^2 value for the plot at 0.895 is not bad. The effect of other processes involving the compounds of interest brings about a large margin of error in the observed data available in the literature. When such coupled processes are involved, simple experiments fail to measure these equilibrium constants accurately.

Table 2. A List of Observed and SPARC Calculated Tautomeric Equilibrium Constants (pK_T)

reactant	product	obsd[29]	calcd
CC═O	C═CO	4.66	3.57
CCC═O	CC═CO	3.9	2.74
CCCC═O	CCC═CO	5.2	3.10
CC(C)C═O	CC(C)═CO	2.8	2.86
CC(═)C	C═C(O)C	8.22	8.27
CCC(═)CC	CC═C(O)CC	7.44	7.32
CC(C)C(═)C(C)C	CC(C)═C(O)C(C)C	7.52	7.39
CC(═)CC	C═C(O)CC	8.76	8.28
CC(═)CC	CC(O)═CC	7.51	7.32
CC(═)C(C)C	C═C(O)C(C)C	8.61	8.28
CC(═)C(C)C	CC(O)═C(C)C	7.33	7.39
CC(═)C(C)(C)C	C═C(O)C(C)(C)C	8.76	8.28
c1ccc(OC)ccc1C(═)C	c1ccc(OC)ccc1C(O)═C	7.31	6.77
c1ccc(C)ccc1C(═)C	c1ccc(C)ccc1C(O)═C	6.95	6.74
c1ccccc1C(═)C	c1ccccc1C(O)═C	6.63	6.84
c1ccc(Cl)ccc1C(═)C	c1ccc(Cl)ccc1C(O)═C	7.77	7.01
c1ccc(Cl)cc1C(═)C	c1ccc(Cl)cc1C(O)═C	7.57	7.01
c1ccc(C(F)(F)F)cc1C(═)C	c1ccc(C(F)(F)F)cc1C(O)═C	7.55	7.24
c1ccc(N(═)═O)cc1C(═)C	c1ccc(N(═)═O)cc1C(O)═C	7.13	7.23
c1ccc(N(═)═O)ccc1C(═)C	c1ccc(N(═)═O)ccc1C(O)═C	6.95	7.32
c1(C)cc(C)cc(C)c1C(═)C	c1(C)cc(C)cc(C)c1C(O)═C	6.92	6.66
c1ccccc1C(═)C(C)C	c1ccccc1C(O)═C(C)C	6.48	6.14

(29) Toullec, J. In *Keto-enol equilibrium constants: The Chemistry of Enols*; Rappoport, Z., Ed.; John Wiley and Sons Ltd.: New York, 1990.

(30) *The Merck Index*, 13th ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 2001.

(31) Avdeef, A. *Absorption and Drug Development: Solubility, Permeability, and Charge State*; John Wiley & Sons: Hoboken, NJ, 2003.

(32) Box, K.; Bevan, C.; Comer, J.; Hill, A.; Allen, R.; Reynolds, D. High Throughput Measurement of pKa Values in a mixed-buffer linear pH gradient system. *Anal. Chem.* **2003**, *75*, 883−892.

(33) Seiler, P. The simultaneous determinationof partition coefficient and acidity constant of a substance. *Eur. J. Med.Chem*. **1974**, *9* (6), 663−665.

(34) McFarland, J. W.; Berger, C. M.; Froshauer, S. A.; Hayashi, S. F.; Hecker, S. J.; Jaynes, B. H.; Jefson, M. R.; Kamicker, B. J.; Lipinski, C. A.; Lundy, K. M.; Reese, C. P.; Vu, C. B. Quantitative structure-activity relationships among macrolide antibacterial agents: in vitro and in vivo potency against *Pasteurella miltocida*. *J. Med. Chem.* **1997**, *9*, 1340−1346.

(35) Reichard, R. E.; Fernelius, W. C. Formation constants of 6-methyl-2-picolylmethylamine with some common metal ions. *J. Phys. Chem*. **1961**, *65*, 380−381.

(36) Ishihama, Y.; Nakamura, M.; Miwa, T.; Kajima, T.; Asakawa, N. A Rapid Method for pK_a Determination of Drugs Using Pressure-Assisted Capillary Electrophoresis with Photodiode Array Detection in Drug Discovery. *J. Pharm. Sci.* **2002**, *91*, 933−942.

(37) Hong, D. D. Chloroquine phosphate. *Anal. Profiles Drug Subst.* **1976**, *5*, 61−85.

(38) Tariq, M.; Al-Badr, A. A. Chloroquine. *Anal. Profiles Drug Subst.* **1984**, *13*, 95−125.

(39) Szakacs, Z.; Beni, S.; Varga, Z.; Orfi, L.; Keri, G.; Noszal, B. Acid-base profiling of imatinib (Gleevec) and its fragments. *J. Med. Chem*. **2005**, *48*, 249−255.

(40) 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**, *23*, 2833−2841.

(41) Takacs-Novak, K.; Noszal, B.; Hermecz, I.; Kereszturi, G.; Podanyi, B.; Szasz, G. Protonation equilibria of quinolone antibacterials. *J. Pharm. Sci.* **1990**, *79* (11), 1023−1028.

(42) Ross, D. L.; Elkinton, S. K.; Riley, C. M. Physicochemical properties of the fluoroquinolone antimicrobials. III. 1-Octanol/water partition coefficients and their relationships to structure. *Int. J. Pharm.* **1992**, *88* (1−3), 379−839.

(43) Kristl, A.; Vrečer, F. Preformulation investigation of the novel proton pump inhibitor lansoprazole. *Drug Dev. Ind. Pharm.* **2000**, *26* (7), 781−783.

(44) Ruiz, R.; Rafols, C.; Roses, M.; Bosch, E. A potentially simpler approach to measure aqueous pKa of insoluble basic drugs containing amino groups. *J. Pharm. Sci.* **2003**, *92*, 14731481.

(45) Wan, H.; Holmen, A. G.; Wang, Y.; Lindberg, W.; Englund, M.; Nagard, M. B.; Thompson, R. A. High-Throughput screening of pKa values of pharmaceuticals by pressure-assisted capillary electrophoresis and mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2639−2648.

(46) Altomare, C.; Cellamare, S.; Summo, L.; Fossa, P.; Mosti, L.; Carotti, A. Ionization behaviour and tautomerism-dependent lipophilicity of pyridine-2(1H)-one cardiotonic agents. *Bioorg. Med. Chem.* **2000**, *8*, 909−916.

(47) Kaufman, J. J.; Semo, N. M.; Koski, W. S. Microelectrometric titration measurement of the pK_a's and partition and drug distribution coefficients of narcotics and narcotic antagonists and their pH and temperature dependence. *J. Med. Chem*. **1975**, *18*, 647−655.

(48) Ungell, A.-L.; Nylander, S.; Bergstrand, S.; Sjöberg, Å.; Lennernäs, H. Membrane transport of drugs in different regions of the intestinal tract of the rat. *J. Pharm. Sci*. **1998**, *87* (3), 360−366.

(49) Mannhold, R.; Dross, K. P.; Rekker, R. F. *Quant. Struct.−Act. Relat*. **1990**, *9*, 21−28.

(50) Irvin, J. L.; Irvin, E. M. Apparent ionization exponents of homologs of quinacrine; electrostatic effects. *J. Am. Chem. Soc.* **1950**, *72*, 2743−2749.

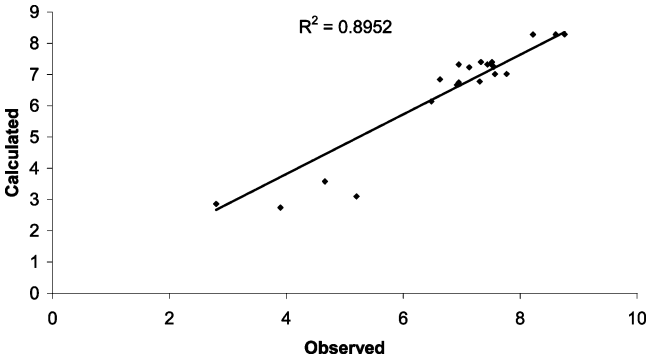


Figure 2. SPARC tautomer model performance. Plot of observed vs SPARC calculated tautomeric equilibrium values.

Building Tautomer Networks. The SPARC tautomer model is used recursively to determine all possible tautomers from an initial tautomeric form. This enables us to find tautomers of tautomers, thereby creating a tautomer network map with the different tautomer forms forming nodes, and the equilibrium constants between these nodes are computed. A doubly recursive algorithm is used to find the tautomers and calculate their cumulative equilibrium constant. The progression down the network is of the depth first search type. This algorithm enables SPARC to identify all tautomer forms and their equilibrium constants irrespective of the initial tautomeric form.

In the process of developing tautomer networks a large number of unproductive pathways are encountered. Unproductive pathways are defined as a particular speciation pathway where the cumulative *K* product for a species/isomer is very small, leading to no viable progression of the path. These unproductive pathways also consume a large amount of calculation cycles. In order to identify and remove these unproductive paths, we designed a filter to examine the results after every cycle of calculation. This filter will check combined *K* for each species in the network. This filter stops the progression of any of the paths if the terminal species is found to have a combined *K* lower than a set threshold. This threshold is set at 0.0002 for the tautomer network model.

The reliability calculated for each tautomer equilibrium constant is used to determine the reliability of each pathway leading to a node in a tautomer network map. This reliability of paths is used to decide whether that particular path's *K* is included in the averaging of *K*'s for a node. The path is ignored if the cumulative reliability of that path is less than 0.15. A sample calculation of the tautomer network for acetylacetone is shown in Figure 3.

Integration of Hydration and Tautomerization. The hydration model is used to generate all the hydrated forms of the starting molecule and determine their hydration constants. These molecules and their respective equilibrium constants are combined and fed into the tautomer network model to determine the possible tautomer and their integrated constants.

Let us use the model to determine the reaction pathways of 2-oxocyclohexanecarbaldehyde. To illustrate the effect of coupling hydration to tautomerization, two different calculations, one without hydration and one with hydration, are reported. Table 3 lists the output of the model without hydration. Table 4 lists the output of the model with hydration turned on. Both calculations are performed in water.

From the results, the conjugated enol−keto forms are found to be the most stable tautomeric forms in water. This is justified as the double bond conjugation stabilizes the molecules better than the dicarbonyl form.

Compared to the results without hydration we see more species present in significant quantities. The relative stabilities of two hydrated forms 8 and 9 are explained by the ease

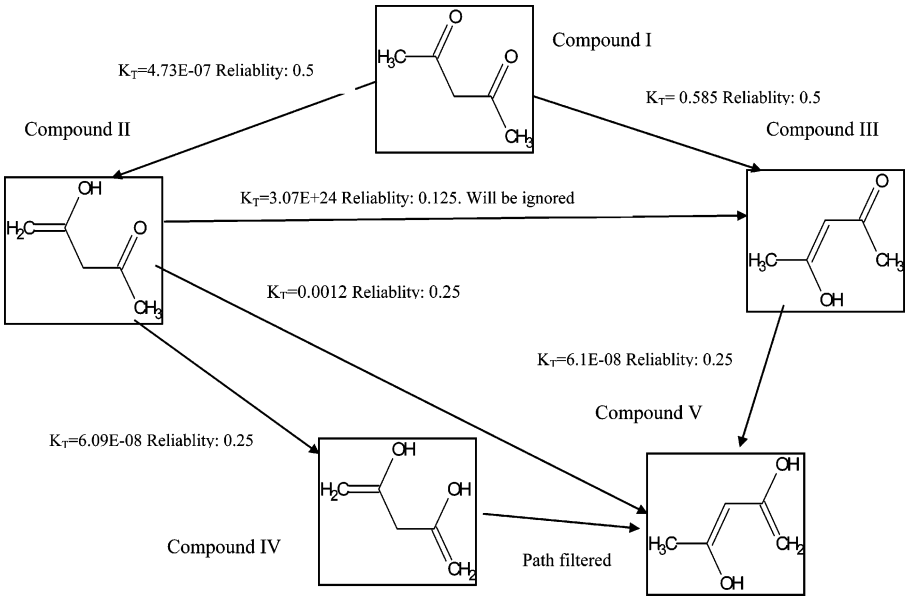
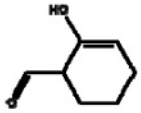
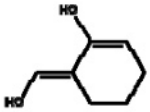
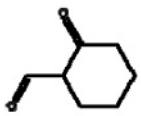
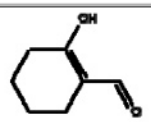
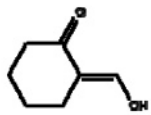


Figure 3. SPARC generated tautomer map for acetyl acetone. The equilibrium constants are indicated next to the arrows with the reliability of the calculation.

Table 3. Results of Integrated Tautomer Network Model without Hydration

No	Tautomer	Relative Abundance	Molecule
1	C1(O)=CCCCC1C=O	7.35E-07	
2	C1(O)=CCCCC1=CO	2.00E-03	
3	C1(CCCCC1C=O)=O	1.00E+00	
4	C1(O)=C(C=O)CCCC1	4.25E+00	
5	C1(C(=CO)CCCC1)=O	4.66E+01	

of hydration of the aldehyde versus that of the ketone. These species are present in significant quantities to affect chemical behavior.

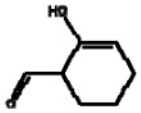
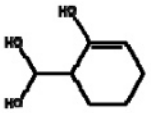
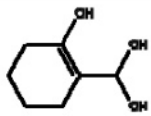
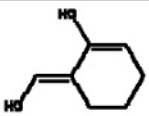
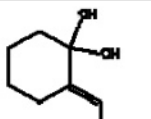
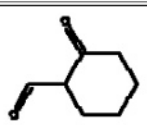
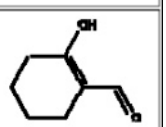
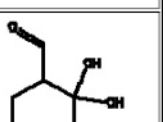
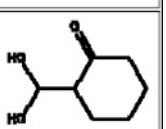
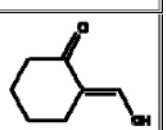
Integration of Speciation, Hydration, and Tautomer Network. Using the hydration coupled tautomer network model the equilibrium constants for the neutral species have been developed. The next step was to determine all the different ionization processes as applied to these molecules and determine their pK_a values.

A full speciation calculation is performed for every species from the tautomer network model. The respective equilibrium constants of the tautomer forms were used to determine the final species fraction as a function of pH. The molecular speciation model is the same model used for pK_a speciation; hence it provides a wealth of information to the user. It calculates the macro constant and micro constants for all chemical processes and determines species fraction as a function of pH.

Let us analyze the results of the fully integrated chemical process model using the same compound. The plot of the species fraction as a function of pH is shown in Figure 4. From the plot we see that below pH 9 the dominant species are the exo−enol form (species 3) and the dihydro form (species 2) with a little bit of the other dihydro form (species 1). At about pH 9 these three species disappear and the ionized form of the dicarbonyl form and the two keto−enol forms dominate (species 6, 7, 8). The species number is the order in which they are listed in the figure.

As the pH increases, the concentration of the dihydro form

Table 4. Results of the Integrated Tautomer Network Model with Hydration

No	Tautomer	Relative Abundance	Molecule
1	C1(O)=CCCCC1C=O	7.35E-07	
2	C1(O)=CCCCC1C(O)O	5.85E-06	
3	C1(O)=C(C(O)O)CCCC1	9.57E-05	
4	C1(O)=CCCCC1=CO	2.00E-03	
5	C1(O)(C(=CO)CCCC1)O	3.53E-03	
6	C1(CCCCC1C=O)=O	1.00E+00	
7	C1(O)=C(C=O)CCCC1	4.25E+00	
8	C1(O)(CCCCC1C=O)O	9.94E+00	
9	C1(CCCCC1C(O)O)=O	2.44E+01	
10	C1(C(=CO)CCCC1)=O	4.66E+01	

is reduced, and the force that drives this reduction is the ionization of the tautomer of the unhydrated form, which in turn drives equilibrium of hydration in favor of the reactant, thus decreasing the concentration of the hydrated forms. This compound is a very good test for coupled reaction models

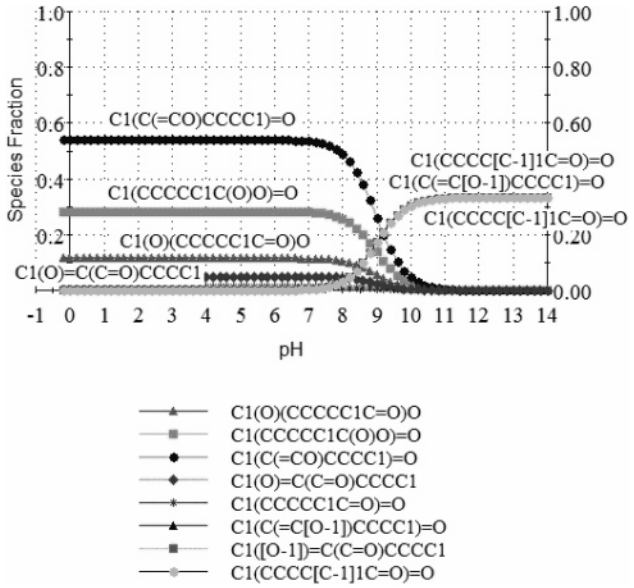


Figure 4. SPARC speciation plot. Plot of species fraction as a function of pH for 2-oxocyclohexanecarbaldehyde.

and is performing very well. A total of 59 micro constants and one macro constant were determined by the speciation calculator.

Estimating Prediction Error. For every predicted value calculated from the computational programs, it will be very useful to provide the users the level of confidence in the prediction. We tried to address this issue for the SPARC program as follows. The SPARC system is based on perturbation models where resonance, field, mesomeric field, sigma induction, differential solvation, and intramolecular hydrogen-bonding effects all perturb the base pK_a of an unperturbed reaction center. For example, all sp^3 N base reactions represent the perturbation of the pK_a of ammonia (corrected for statistical factor). There are cases where the perturbations are very large in one direction and the pK_a is very different from the unperturbed pK_a. There are other cases where the perturbations are very large but have opposite signs that cancel or cases that have very small perturbations that result in a pK_a close to the unperturbed pK_a of ammonia. Since the error in prediction will be related to the magnitude of each perturbation, we need to know the sum of the absolute values of all these perturbations. SPARC error estimation was carried out as follows. A set of well-measured IUPAC pK_a values was constructed for each reaction center type addressed. These centers are OH (acid), SH (acid), sp^3 N (acid and base), sp^2 N (base), C (acid), and oxy-acids (CO_2H, PO_2H, SO_3H, ...). Each set was run and the rms error was calculated. For each batch set, the absolute sum of perturbations was found and averaged. An "error_multiplier" for each type of reaction center was determined as rms(set)/average(absolute sum). The estimated error for an individual micro constant is calculated as error_multiplier(site_type)*absolute sum of the perturbations. The predicted errors and observed errors were very well correlated. For macro pK_a values where several species on the left-hand side of the reaction may couple to several species on the right-hand side of the reaction, the errors for each of the micro constants involved were weighted by their relative abundance in the reaction and summed.

Datasets. For this report, we assembled a set of 123 organic compounds with experimentally measured pK_a values, either Pfizer internal measurements or in the literature. Many of these compounds are known drugs and have been studied previously by various groups using different methods, experimentally as well as computationally. We also applied SPARC to predict the pK_a values for a set of 537 compounds with 735 pK_a values from the Pfizer internal dataset.

To explore the chemistry space for the compounds in the dataset we used, a set of properties were calculated using

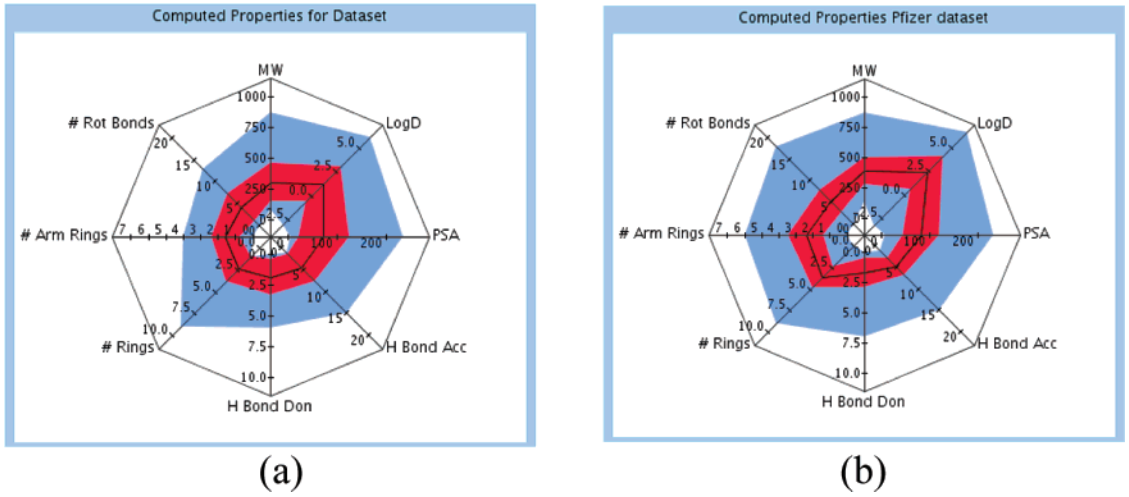


Figure 5. The spider plot for computed properties for the dataset: molecular weight (MW), log *D*, polar surface area (PSA), number of hydrogen bond acceptors (H Bond Acc), number of hydrogen bond donors (H Bond Don), number of rings (# Rings), number of aromatic rings (# Arm Rings), and number of rotatable bonds (# Rot Bonds). In the plot, the blue region represents the range of property, the black line inside the red region is the average value for each property, and the red region is the standard deviation. (a) For a set of 123 known drugs. (b) For a set of 537 Pfizer compounds.

Table 5. Comparison of SPARC pK_a Predictions to Experimental Values

no.	compound name	pK_a expt	pK_a SPARC	residual	prediction error	no.	compound name	pK_a expt	pK_a SPARC	residual	prediction error
1	1-naphthol	9.38[a]	9.38	0.00	0.28	52	erythromycylamine	8.96[h]	6.72	2.24	0.31
2	2,4,6-trimethylpyridine	6.66[a]	7.24	0.58	0.22			9.95[h]	7.95	2.00	0.09
3	2,4,6-trichlorophenol	6.07[a]	6.11	0.04	0.42	53	etoposide	9.96[d]	8.79	1.17	0.33
4	2-aminopyridine	6.63[a]	6.54	0.09	0.33	54	famotidine	6.74,[c] 6.81[d]	6.06	0.59	0.20
5	2-bromoaniline	2.33[a]	2.68	0.35	0.33				8.74		0.41
6	2-ethylaniline	4.26[a]	4.43	0.17	0.25			11.19,[c] 11.24[d]	11.36	0.12	0.18
7	3-aminonaphtoic acid	2.82[a]	2.95	0.13	0.25	55	fenoterol	8.23[g]	8.28	0.05	0.29
		4.61[a]	4.40	0.21	0.19			10.23[g]	9.47	0.76	0.13
8	3-chlorophenol	8.88[a]	9.09	0.21	0.26			11.45[g]	10.68	0.07	0.12
9	3-ethylaniline	4.70[a]	4.95	0.25	0.24				12.08		0.27
10	3-nitrophenol	8.33[a]	8.25	0.08	0.30	56	flumequine	6.27,[c] 6.36[d]	6.66	0.30	0.78
11	4-aminopyridine	9.28[a]	8.69	0.59	0.35	57	flurbiprofen	3.94[d]	4.20	0.26	0.26
12	4-bromoaniline	3.80[a]	3.92	0.12	0.28	58	furosemide	3.52,[c] 3.59[d]	2.55	1.04	0.72
13	4-chloroaniline	3.88[a]	3.94	0.06	0.28			10.63,[c] 10.43[d]	10.02	0.41	0.99
14	4-chlorophenol	9.15[a]	9.35	0.20	0.25	59	homidium	2.47[g]	1.72	0.75	0.43
15	4-nitrophenol	7.10[a]	6.75	0.35	0.38	60	hydrochlorothiazide	8.78[d]	7.69	1.09	1.33
16	abacavir	5.01[b]	4.48	0.53	0.82			10.16[d]	10.29	0.13	1.29
17	acetaminophen	9.63,[c] 9.55[d]	9.60	0.05	0.28	61	ibuprofen	4.45,[c] 4.27[d]	4.46	0.19	0.22
18	acetylsalicylic acid	3.50[c]	3.64	0.14	0.31	62	imatinib	1.71[n]	2.01	0.30	0.12
19	acyclovir	2.34,[c] 2.22[d]	1.35	0.87	0.52			3.10[n]	3.28	0.21	0.11
		9.23,[c] 9.26[d]	8.62	0.64	0.72			3.88[n]			
20	albendazole	3.28[c]	4.75	1.47	1.23			7.70[n]	8.28	0.58	0.17
		9.93[c]	9.74	0.19	0.58	63	imipramine	9.51[c]	9.86	0.35	0.43
21	allopurinol	9.42[d]	9.34	0.08	0.34	64	indomethacin	4.42[c]	4.51	0.09	0.26
22	amifloxacin	5.42[e]	5.93	0.51	0.32	65	isoniazid	3.35[d]	3.98	0.63	0.49
		7.39[f]	8.26	0.87	0.52			10.57[d]	10.63	0.06	0.79
23	amiloride		2.50		0.79	66	ketoconazole	3.15,[o] 3.12[d]	1.64	1.48	0.62
		8.70[g]	8.00	0.70	0.51			6.41,[o] 6.45,[d] 5.89[a]	4.51	1.38	0.56
24	amiodarone	9.06[c]	8.60	0.46	0.38	67	labetalol	7.48[c]	8.00	0.52	0.25
25	amitriptyline	9.49[c]	9.55	0.06	0.31			9.42[c]	10.03	0.61	0.31
26	aniline	4.53[a]	4.72	0.19	0.23	68	lansoprazole	1.33[p]			
27	antazoline	4.41[d]	2.21	2.20	0.63			4.15,[p] 4.11[a]	4.61	0.50	0.36
		10.29[d]	10.51	0.22	0.52			8.84,[p] 9.29[a]	10.50	1.21	1.70
28	atenolol	9.54,[c] 9.56[d]	9.41	0.15	0.35	69	lidocaine	7.95,[c] 7.84[d]	8.38	0.54	0.60
29	atorvastatin	4.50[d]	3.91	0.59	0.23	70	liothyronine	1.88[d]	1.82	0.06	0.54
30	azithromycin	8.74[h]	6.15	2.59	0.37			8.13[a]	7.19	0.94	0.26
		9.45[h]	7.41	2.04	0.12			10.55,[d] 9.16[a]	8.85	0.31	0.26
31	benzoic acid	4.19[a]	3.98	0.21	0.05	71	l-tyrosine	2.20[c]	2.21	0.01	0.46
32	betahistine	4.34[i]	3.72	0.62	0.26			9.06[c]	9.12	0.06	0.17
		9.96[j]	9.86	0.10	0.19			10.12[c]	10.59	0.47	0.18
33	cefadroxil	2.64[d]	3.15	0.51	0.52	72	maprotiline	10.20[q]	10.57	0.37	0.16
		7.17[d]	6.86	0.31	0.24	73	mebendazole	3.43[r]	4.41	0.98	1.22
		9.74[d]	9.53	0.21	0.22			9.93[r]	9.17	0.76	0.59
34	cefazoline	2.20[g]	3.55	1.35	0.67	74	methotrexate	3.31[c]	3.17	0.14	0.09
		12.05[g]	11.80	0.25	0.90			4.00[c]	3.96	0.04	0.11
35	chloroquine	8.25[k]	6.68	1.57	0.34			5.39[c]	4.74	0.65	0.19
		10.37[l]	9.42	0.95	0.43	75	metoprolol	9.56,[c] 9.55[d]	9.47	0.08	0.27
36	chlorthalidone	9.11[g]	8.81	0.30	1.27	76	mexiletine	9.14[c]	9.38	0.24	0.13
		10.98[g]	10.31	0.67	1.23	77	miconazole	6.58,[c,d] 5.52[a]	4.76	0.76	0.50
37	chlorzoxazone	8.24[d]	8.99	0.75	0.51	78	milrinone	5.10[s]	4.20	0.90	0.31
38	cimetidine	6.93,[c] 6.96[d]	5.50	1.46	0.46			9.30[s]	8.40	0.90	0.38
39	cinnamic acid	4.37[a]	4.05	0.32	0.06	79	morphine	8.18[c]	8.84	0.66	0.33
40	clarithromycin	8.99[h]	7.20	1.79	0.33			9.26[c]	10.29	1.03	0.30
41	clomipramine	9.38[m]	9.71	0.33	0.31	80	moxonidine	7.54,[g] 7.50[a]	7.91	0.41	1.09
42	clozapine	4.40[c]	5.44	1.04	0.77	81	nadolol	9.69[c,d]	9.25	0.44	0.28
		7.90[c]	7.58	0.32	0.64	82	nalidixic acid		2.29		0.59
43	codeine	8.22[c]	9.00	0.78	0.33			6.01,[c] 6.19[a]	6.51	0.32	0.66
44	deprenyl	7.48[c]	9.48	2.00	0.31	83	naloxone	7.94[t]	7.22	0.72	0.39
45	desipramine	10.16[c]	10.34	0.18	0.35			9.44[t]	9.89	0.45	0.33
46	dichlorphenamide	8.41[d]	8.16	0.25	1.38	84	nicotine	2.95[a]	3.28	0.33	0.33
		10.27[d]	9.64	0.63	1.33			8.20[a]	7.78	0.42	0.32
47	diclofenac	4.04[d]	4.05	0.01	0.30	85	nifedipine	2.20[d]	1.00	1.20	0.61
48	diltiazem	7.94[d]	8.13	0.19	0.38	86	nitrazepam	3.02[c]	3.31	0.29	0.73
49	diphenhydramine	9.10[c]	9.02	0.08	0.35			10.37[c]	9.63	0.74	0.87
50	enrofloxacin	6.04[g]	6.40	0.36	0.68	87	norfloxacin	6.23,[c] 6.31[d]	6.41	0.10	0.65
		7.83[g]	8.15	0.32	0.53			8.51,[c] 8.59[d]	8.90	0.31	0.44
51	erythromycin	8.88[h]	7.25	1.63	0.32	88	nortriptyline	10.13[c]	10.35	0.22	0.17

Table 5 (Continued)

no.	compound name	pK_a expt	pK_a SPARC	residual	prediction error	no.	compound name	pK_a expt	pK_a SPARC	residual	prediction error
89	olsalazine	2.55[g]	2.37	0.18	0.39	107	sulfasalazine	2.65[c]	2.44	0.21	0.38
		2.55[g]	3.01	0.46	0.19			7.95[c]	8.82	0.87	0.30
		11.20[g]	11.67	0.07	0.27			10.51[c]	11.80	1.29	0.76
		12.00[g]				108	sulpiride	9.00[b]	8.71	0.29	0.65
90	omeprazole	4.40[u]	5.22	0.82	0.64			10.19[b]	9.91	0.28	0.60
		8.70[u]	9.96	1.26	0.78	109	tamoxifen	8.48,[c] 8.85[d]	8.98	0.13	0.36
91	papaverine	6.39[c]	6.62	0.23	0.30	110	terbutaline	8.67,[c] 8.7[d]	8.46	0.24	0.29
92	phenobarbital	7.22[g]	7.05	0.17	1.30			9.97,[c] 10.36[d]	10.16	0.20	0.16
		11.68[g]	11.95	0.27	1.07			11.02[c]	11.60	0.58	0.25
93	phenytoin	8.19[d]	8.26	0.07	1.22	111	terfenadine	9.89[d]	9.13	0.76	0.33
94	pheylacetic acid	4.19[a]	4.41	0.22	0.24	112	tetracaine	2.39[c]	2.23	0.16	0.45
95	pindolol	9.54[c]	9.17	0.37	0.28			8.49[c]	8.77	0.28	0.39
96	piroxicam	2.33[c]	1.60	0.73	0.68	113	tetracycline	3.33[g]	4.36	1.03	0.11
		5.07[c]	6.20	1.13	0.40			7.16[g]	8.55	1.39	0.08
97	prazosin	6.83[d]	7.30	0.47	0.90			9.43[g]	9.71	0.28	0.09
98	promazine	9.28[v]	9.59	0.31	0.32	114	theophylline	8.55,[c] 8.65,[d] 9.0[a]	10.00	1.00	0.73
99	propranolol	9.53[c,d]	9.48	0.05	0.26	115	tilmicosin	8.18[h]	6.47	1.71	0.38
100	pyridine	5.33[a]	4.97	0.36	0.11			9.56[h]	8.64	0.92	0.13
101	quinacrine	7.73[w]	8.21	0.48	0.55	116	tolbutamide	5.32[d]	5.26	0.06	1.40
		10.18[w]	10.52	0.34	0.63	117	trazodone	6.69[q]	7.32	0.63	0.47
102	quinine	4.24,[c] 4.16[d]	3.79	0.37	0.43	118	trimethoprim	7.07[c]	6.33	0.74	0.71
		8.55,[c] 8.49[d]	8.16	0.33	0.23	119	trimipramine	9.15[q]	9.80	0.65	0.43
103	ranitidine	2.10[d]	1.80	0.30	0.68	120	trovafloxacin	5.90[c]	5.96	0.06	0.43
		8.62[d]	9.10	0.48	0.82			8.11[c]	7.45	0.66	0.50
104	salicylic acid	2.93[a]	3.04	0.11	0.30	121	verapamil	9.07,[c] 8.76[d]	8.77	0.01	0.35
105	serotonin	9.89[g]	9.75	0.14	0.21	122	vinblastine	5.70[d]	5.98	0.28	0.25
		10.91[g]	11.32	0.41	0.17			7.84[d]	8.43	0.59	0.23
106	sotalol	8.28,[c] 8.25[d]	8.73	0.48	0.88	123	warfarin	4.82,[c] 5.01[d]	5.74	0.73	0.52
		9.72,[c] 9.68[d]	10.25	0.57	0.68						

[a] Protocol 2 (see Experimental Section). [b] Reference 30. [c] Reference 31. [d] Protocol 1 (see Experimental Section). [e] Reference 41. [f] Reference 42. [g] Reference 32. [h] Reference 34. [i] Reference 35. [j] Reference 36. [k] Reference 37. [l] Reference 38. [m] Reference 33. [n] Reference 39. [o] Reference 40. [p] Reference 43. [q] Reference 44. [r] Reference 45. [s] Reference 46. [t] Reference 47. [u] Reference 48. [v] Reference 49. [w] Reference 50.

the Pipeline Pilot program 5.1.0.100: molecular weight (MW), log *D*, polar surface area (PSA), number of hydrogen bond acceptors (H Bond Acc), number of hydrogen bond donors (H Bond Don), number of rings (# Rings), number of aromatic rings (# Arm Rings), and the number of rotatable bonds (# Rot Bonds). The spider plots for the calculated properties are shown in Figure 5. In the plot, the blue region represents the range of each property, the black line inside the red region is the average value for each property, and the red region is the standard deviation from the average value.

Results and Discussion

Comparison of Experimental and Calculated pK_a Values. The calculated pK_a values from the SPARC program are compared with experimentally measured pK_a values for a set of 123 compounds in Table 5. Experimental values are from Pfizer internal measurements using protocols 1 and 2 described in the Experimental Section as well as well-established literature values from many references. For several compounds, multiple pK_a values were observed experimentally since they contain multiple ionizable groups. In total, there are 187 pK_a values from 123 compounds. The experimentally measured pK_a values range from 1.71 (imatinib) to 12.05 (cefazoline). In the table, the residual values, which are the absolute differences between the calculated and experimental values, are presented, as well as the calculated prediction error described in the Estimating Prediction Error section. For each experimental value, we referenced the source in the table.

For several compounds, there was excellent agreement between internal measurements and the literature values, which validated the accuracy of the experimental protocols we employed. A detailed discussion on the internal measurements will be published in the future, but in this paper, we will focus on the comparison of the calculated values from the SPARC program to experimental measurements. The absolute difference between experimental value and SPARC calculated value is calculated using the Pfizer internal measurement if a literature value is also available and the Pfizer measurement using protocol 1 if both internal values are available. For compounds with two calculated values for a single measurement, such as in 54 and 55, and two measurements for a single calculated one, such as in 62, the average value of the two was used to calculate the residual value. The correlation between experimental and calculated values is shown in Figure 6, with the correlation coefficient $r^2 = 0.92$ and the root-mean-square error (RMSE) of 0.78 for 187 pK_a values.

For most of the compounds, the calculated values are in excellent agreement with experimental values, especially for compounds containing a single ionizable group.

Some regular errors were also observed which warrant further discussion. The greatest differences between experi-

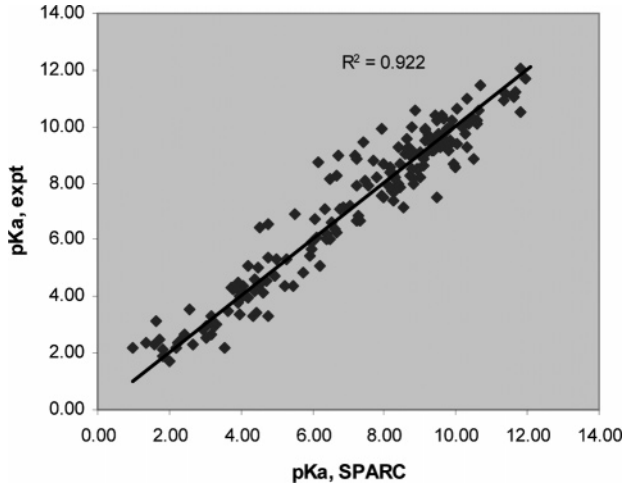


Figure 6. The correlation between experimental pK_a values and the prediction from SPARC. The correlation coefficient $r^2 = 0.92$, and the root-mean-square error (RMSE) = 0.78 for 180 pK_a values.

mental and calculated pK_a values were observed for molecules 30, 40, 51, 52, and 115. From a modeling standpoint the common feature of these molecules is their size. Looking at the observed pK_a for molecule 30 (Azithromycin) at 8.74 and 9.45, it appears that the large number of dipole fields did not affect the pK_a. In SPARC these dipoles produced a perturbation which reduced the pK_a. SPARC's field effect model assumes that the interaction of the dipole/monopole with the reaction center goes through the molecule and uses an effective dielectric constant of 2.4. In large molecules this assumption would not work as many of the interactions would probably go through solvent pockets and will have a much larger dielectric constant. Currently we are working to rectify this problem in SPARC's electrostatic models. Molecules 38, 66, 77 all have imidazole as part of the structure, and it is actively involved in the ionization. Looking at the experimental data the imidazole seems to have a very low susceptibility to field effects from the other dipole and monopoles in the molecule. This can again be a similar problem as discussed above. Molecule 35 (chloroquine) has an observed pK_a value of 8.25 for the quinoline nitrogen. SPARC returned a result of 6.68. We believe that this result can have resulted from the fact that the SPARC steric model forced the amine side chain to rotate further than it should have, which would reduce the resonance effect of the tertiary nitrogen at that position. This will be rectified in future versions. Molecule 44 (deprenyl) has an observed pK_a value of 7.48, compared with the SPARC value of 9.48. While it is unlikely that the beta-phenyl moiety exerts much influence on the pK_a of the tertiary nitrogen center, less is known about the influences of the propargyl side chain. At this time, we have no explanation for the difference in the two values.

Regarding the prediction error, it ranges from 0.05 to 1.7 for 185 pK_a values. There is no strong, direct correlation between the difference in experimental and calculated values and the prediction error. However, if the whole dataset is partitioned into two groups using the median value of 0.35, the average value of the difference is 0.48 for a set of 92 pK_a values with prediction error less than 0.35, while the average value of the difference is 0.63 for a set of 93 pK_a values with prediction error larger than or equal to 0.35.

Overall, SPARC predicts pK_a values of many known drugs in excellent agreement with experimentally measured values. We applied SPARC to predict pK_a values for a Pfizer internal dataset, a set of 537 compounds containing 720 experimentally measured pK_a values. A correlation coefficient of $r^2 = 0.80$ and an RMSE = 1.05 were obtained for 720 calculated pK_a values. As shown in Figures 5a and 5b, the chemistry space for the Pfizer compound set was very similar to that defined by the compounds presented in Table 5, with the exception that the Pfizer compounds had on average slightly more rotatable bonds and a greater proportion of aromatic rings.

Conclusion

The method described herein has been demonstrated to be a reliable predictor of pK_a values for complex drug-like molecules, with a few exceptions as noted. Because SPARC requires only a 2D structure for input, it has been possible to implement this program as part of the predicted properties suite within the Pfizer database. This method offers insight into the effect of structural modification on the ionization state as part of analogue and series design, and along with log $P_{o/w}$ prediction enables the calculation of log D values. Future developments of this software will focus on addressing the overprediction of perturbing effects of large, complex molecules and on the steric factors which can potentially overwhelm important resonance contributions.

The web interface of the SPARC program is freely available in http://sparc.chem.uga.edu. Batch calculations are also provided for academic research by Carreira's group.

Acknowledgment. We would like to thank Dong Li, Jennifer Durkee, and Heather Estrella for their technical support in the program.

Supporting Information Available: A comma separated file with compound names, SMILES, and experimental pK_a values. This material is available free of charge via the Internet at http://pubs.acs.org.

MP070019+