J Comput Aided Mol Des (2007) 21:681–691
DOI 10.1007/s10822-007-9133-z

Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules

John C. Shelley · Anuradha Cholleti · Leah L. Frye · Jeremy R. Greenwood · Mathew R. Timlin · Makoto Uchimaya

Received: 15 March 2007 / Accepted: 28 August 2007 / Published online: 27 September 2007

Abstract Epik is a computer program for predicting pK_a values for drug-like molecules. Epik can use this capability in combination with technology for tautomerization to adjust the protonation state of small drug-like molecules to automatically generate one or more of the most probable forms for use in further molecular modeling studies. Many medicinal chemicals can exchange protons with their environment, resulting in various ionization and tautomeric states, collectively known as protonation states. The protonation state of a drug can affect its solubility and membrane permeability. In modeling, the protonation state of a ligand will also affect which conformations are predicted for the molecule, as well as predictions for binding modes and ligand affinities based upon protein–ligand interactions. Despite the importance of the protonation state, many databases of candidate molecules used in drug development do not store reliable information on the most probable protonation states. Epik is sufficiently rapid and accurate to process large databases of drug-like molecules to provide this information. Several new technologies are employed. Extensions to the well-established Hammett and Taft approaches are used for pK_a prediction, namely, mesomer standardization, charge cancellation, and charge spreading to make the predicted results reflect the nature of the molecule itself rather just for the particular Lewis structure used on input. In addition, a new iterative technology for generating, ranking and culling the generated protonation states is employed.

Keywords Hammett and Taft (HT) equations · Ionization · Mesomers · pK_a · Protonation state · Tautomerization · Tautomers

Introduction

Many drug-like molecules can gain or lose protons in solution—a behavior we will collectively refer to as ionization. In addition, they can also lose a proton from one location and gain a proton at another location or vice versa, effectively an intramolecular transfer of a proton, referred to as a tautomerization. The ionization and tautomerization processes can be quite active in aqueous solutions and thus transform a ligand into a number of related ionization and tautomeric forms, collectively referred to as protonation states, on physiologically relevant time scales. While there are often many potential protonation states for a medicinal chemical, usually only a small number are present at significant levels under physiological conditions. Which protonation states predominate can be affected by pH and the local molecular environment of the compound. The solubility and membrane permeability of a drug can vary significantly, often in obvious ways with protonation state. Thus, the protonation state of a molecule is of inherent

Electronic Supplementary Material The online version of this article (doi:10.1007/s10822-007-9133-z) contains supplementary material, which is available to authorized users.

J. C. Shelley (✉) · L. L. Frye · M. R. Timlin · M. Uchimaya
Schrödinger, LLC, 101 SW Main St., Suite 1300, Portland OR 97204, USA
e-mail: jshelley@schrodinger.com

A. Cholleti
D. E. Shaw India Software, PL (DESIS), 3rd Floor, G. Pulla Reddy Building, 6-3-879 Begumpet, Hyderabad, Andhra Pradesh 500 016, India

J. R. Greenwood
Schrödinger, LLC, 120 West 45th Street, 29th Floor, New York, NY 10036, USA

interest, particularly in a biological context. In addition, the pK_a values for the relevant protonation states may be of interest to synthetic chemists. The answers obtained from other computational applications can be quite sensitive to the protonation state of the input structure. For instance, the predicted conformations of a drug-like molecule, as well as the binding mode and binding affinity of ligands with proteins, are all expected to be influenced by the protonation state of a ligand.

Epik [1] is a program for pK_a value prediction and protonation state generation for drug-like molecules. For the purposes of Epik, the acid/base chemistry considered is limited to that of the Brønsted/Lowry type, where the reaction consists of the simple removal or addition of a proton to a small organic molecule. Thus, other types of acidity such as that of Lewis acids are not considered. Nor is any reaction involving a change in the connectivity of heavy atoms, such as covalent hydration or dehydration, certain ring condensations or openings, and so on. The emphasis is on the kind of acid/base reactions of most importance to the behavior and structure of drug-like molecules in solution. Epik relies on the well-established and proven Hammett and Taft [2] (HT) linear free energy approaches for predicting pK_a values for aliphatic and aromatic acids or bases.

Based upon these pK_a values, a proton can be added or removed from the molecule at various sites. In the resulting protonation state, subsequent ionizations may be viable, however their pK_a values will usually have changed significantly as compared to those calculated before the molecule was ionized. As a result, Epik uses an iterative approach to generate protonation states in which pK_a values are estimated, the protonation state is adjusted and the whole process is repeated. Along the way unimportant protonation states are culled to ensure that the process is efficient and the output is focused on the relevant structures.

Effective tautomerizations via sequential protonation and deprotonation can, in principle, generate all tautomers for the molecule. However, there are many such tautomerizations in which the first ionization would generate a very low probability intermediate form which would be eliminated from further processing. To help alleviate this problem, Epik also uses a rapid, template-based tautomerization facility developed for another Schrödinger product, LigPrep [3]. There are more than 200 types of tautomerizations in the database. The iterative approach for generating protonation states described in the previous paragraph also includes a tautomerization step in each cycle.

One challenge for HT approaches is that most databases record molecules as a single Lewis structure, with specific patterns of single bonds, double bonds and formal charges. In many molecules, particularly those with aromatic ring systems, the state of the molecule is better thought of as being a weighted combination of a number of distinct resonance contributors each corresponding to a distinct Lewis structure. So the Lewis structure used in the database provides some of the information regarding the charge distribution in the molecule but the Lewis structure used for a particular functionality is not standardized and may vary from database to database. In this article we will use the term mesomer to refer to the Lewis structure corresponding to particular resonance contributor [4]. An additional complication is that many methods for recognizing molecular fragments utilize particular Lewis structures. Recognizing all of the various forms for the molecular fragments presents a practical problem. Thus programs using the HT approach must generalize from single Lewis structures and specific fragment patterns in order to make reliable pK_a predictions.

To address the problem of the dependence of the pK_a estimates on the Lewis structure provided, Epik 1.5 uses an approach in which the mesomeric representation the input molecule is standardized for fragment identification and then novel formal charge cancellation and delocalization mechanisms are used to make the predictions more representative of the nature of the molecule beyond the limitations of a particular mesomeric representation.

Since Epik's calculations make extensive use of stored data and in a limited sense chemical knowledge, it is accurate and fast enough to be useful in processing large databases of drug-like molecules. Epik can make pK_a predictions with water or DMSO as the solvent. For this article we are focusing on the parameterization for water because it is more important in drug development. This article describes the technology used Epik 1.5 which was released in May 2007.

Epik is one of a number of programs available that also predict pK_a values and in some cases protonations states, including ACD's pK_a and Tautomers [5] modules, Marvin from ChemAxon [6], Sparc from the EPA [7], Jaguar from Schrodinger [8] and Chemsilico's CSpKa™ [9] to name a few. It is beyond the scope of this article to fairly compare these approaches with each other and Epik. We look forward to seeing such comparisons conducted by interested but neutral third parties.

In the following section, methods, we will review the technology employed by Epik for pK_a prediction and protonation state adjustment including tautomerization. In the Training and performance section we will present results obtained from the application of Epik to a broad range of medicinal compounds. In the last section we will summarize Epik's capabilities, limitations and prospects for further development.

Methods

In this section, the key technologies employed by Epik will be described. Epik's implementation of HT technology along with charge spreading extensions will be outlined in the pK_a value estimation section. The *Protonation and deprotonation*, and *Tautomerization* sections describe how Epik introduces individual changes in protonation state. Procedure for generating protonation states delineates the overall iterative process for generating and prioritizing the most important protonation states. The last section, *Procedure for pK_a scanning*, outlines how Epik estimates the pK_a values in a manner that facilitates comparison with experimental pK_a values.

pK_a value estimation

Epik uses the well-established and complimentary HT [2] methodologies that organic chemists have successfully used for decades to estimate pK_a values for small aromatic and aliphatic molecules, respectively. Both approaches use the same formal equation to estimate the pK_a of an acidic or basic functional group (ABG), i,

$$\mathrm{pKa}_i = \mathrm{pKa}_i^0 + s_i - \rho_i \sum_j \sigma_j \quad (1)$$

Each type of ABG has two parameters: an unperturbed pK_a value, pKa^0, and a response parameter, ρ, which indicates how sensitive the pK_a for this ABG is to the groups of atoms (substituents) attached to it. For each substituent, j, there are σ parameters that reflect how strongly it perturbs ABGs. The pK_a^0, ρ, and σ parameters are almost always fit to experimental pK_a data. In only a few cases where experimental data was not available and experience suggested that the pK_a values given by Jaguar's pKa module were very accurate, pK_a^0 values were set using Jaguar's predictions. There are different σ parameters depending on whether the substitution is on an aliphatic portion of the molecule (σ^*) or on an aromatic ring system where three parameters are used depending on whether the substituent is ortho (σ_o), meta (σ_m) or para (σ_p) to the ABG. s is a statistical factor which accounts for differences in the number of hydrogen atoms equivalent to the one involved in the ionization of this ABG in its acidic (n_a), and basic (n_b) forms and is given by:

$$s = -\log(n_a/n_b) \quad (2)$$

Note that equivalent hydrogen atoms on other equivalent ABGs within the molecule should be included in these counts. Each type of ABG in Epik's database has an additional value associated with it that gives a rough estimate of the uncertainty in the pK_a value predictions for that type of site. Epik uses the SMARTS® [10] language to identify the various chemical features within molecules and then associate HT parameters with them. One non-obvious problem is that often multiple SMARTS patterns, each with different associated parameters, match a particular ionizable group. Each of the SMARTS patterns in Epik's pattern has numeric priority, usually calculated from the SMARTS pattern itself, which determines which pattern is selected when multiple matches are detected. The appendix outlines how these priorities were determined.

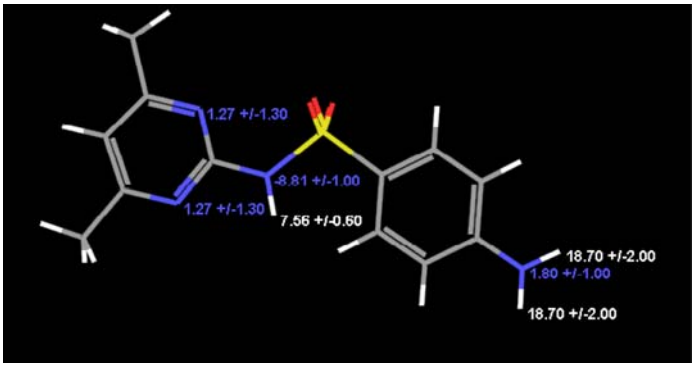
Epik also utilizes a number of extensions to HT technologies, including:

- transmission with attenuation of substituent effects through intervening atoms [2]
- use of some special σ terms for particular combinations of ABGs and substituents [2, 11]
- use of σ terms to correct for the presence of additional heteroatoms in aromatic rings [2, 11]
- corrections for amine groups in one or two rings [2]
- use of σ terms to correct for ring fusion in aromatic ring systems [2, 12]
- methods for estimating σ values from σ_o, σ_m and σ_p values in fused aromatic rings [2, 13, 14]
- virtual conversion of five membered rings into six membered rings when selecting amongst σ_o, σ_m and σ_p [2]

Epik also uses alternate ρ values for ABGs where the influence of substitution locations is not covered by the pattern that best matched the ABG. Figure 1 illustrates the estimation of the pK_a value for the sulfonamide hydrogen using the Hammett parameters for the drug sulfamethazine.

Since Epik uses SMARTS patterns for matching, it necessarily has to represent molecules with localized integral bonds and formal charges, i.e., individual formal Lewis structures, even when a single Lewis structure by itself does not capture the delocalization of the species. In those cases the molecule is usually better represented as a combination of Lewis structures, resonance contributors which we will refer to as mesomers. This situation presents challenges. One issue is that having more than one way to represent various ABGs and substituents can greatly expand the number of SMARTS patterns that need to be encoded. Another challenge that mesomers pose is that they often represent artificial localizations of the formal charge, thus basing the calculation on the nature of one mesomeric representation can provide an inaccurate description. As such, making predictions based a single mesomeric form limits the range of accurate applicability for patterns for ABGs influenced by mesomeric behavior.

We have implemented an empirical approach for handling molecules where a single mesomeric representation is inappropriate, involving standardization of the mesomeric form processed, as well as the partial cancellation and distribution of formal charges, to produce spread formal



Experimental pKa: 7.38
Epik pKa: 7.56

Breakdown:	
pKa^0 (H-N(-Ar)-SO_2-Ar):	8.31
Substituent contributions:	
para amine	0.992
meta methyl	0.104
meta methyl	0.104
Aromatic heteroatoms:	
ortho N	-0.975
othro N	-0.975
pKa	7.56

Fig. 1 Illustration of pK_a estimation using the Hammett equation for the sulfonamide proton in Sulfamethazine. A sulfonamide between two aromatic rings has a unperturbed pK_a (pK_a^0) of 8.31. The contributions ($\rho\sigma$) from the substituents (an amino and two methyl groups) and the ring heteroatoms are added to this value to provide Epik's estimate for the pK_a. The Maestro graphical user interface can display these estimated pK_a values along with their uncertainties

charges. The spread formal charges are used to generate parameters which are interpolations between those for the charged and uncharged versions for substituents and heteroaromatic atoms. Groups capable of participating in mesomerization are encoded in Epik's database along with a charge bias factor, w, which is a real number whose sign indicates the sign of the charge this site will tend to attain and whose magnitude is proportional to how strongly it competes with other sites for charge. We should emphasize that such spread formal charges as calculated internally by Epik, like any partial charge representation, are not real, however these charges never-the-less do provide a useful means to estimate the perturbing influence of various chemical features on the ABG in the context of HT technology.

The first stage in Epik's treatment of mesomeric systems involves adjusting the representation of the input structure to produce a zwitterion-like mesomeric representation (as illustrated in Fig. 2). In most cases this process involves transferring a negative charge from a neutral nitrogen atom bonded to three other atoms from within an aromatic ring or amino group directly bonded to an aromatic ring, to an oxygen or sulfur atom bonded only to the same aromatic ring system or an imine nitrogen atom bonded to the same aromatic ring system.

After producing the standard mesomeric representation for the molecule as a whole, charge spreading calculations are carried out separately for each ABG. Groups are identified that can participate in mesomerization and share an atom with the ABG or are in an aromatic system that shares atoms with the ABG. The net formal charge on each mesomeric group, m, is given by F_m. ABGs can contain mesomerizable sites too. If so, the formal charge and charge bias factor for that site should be adjusted so that they lie halfway between the acidic and basic forms for the ABG, ensuring that the pK_a calculations will give the same answer when starting from either form. This involves adding 0.5 to or subtracting 0.5 from F_m for this site for acidic or basic forms of the ABG, respectively, and using the average of

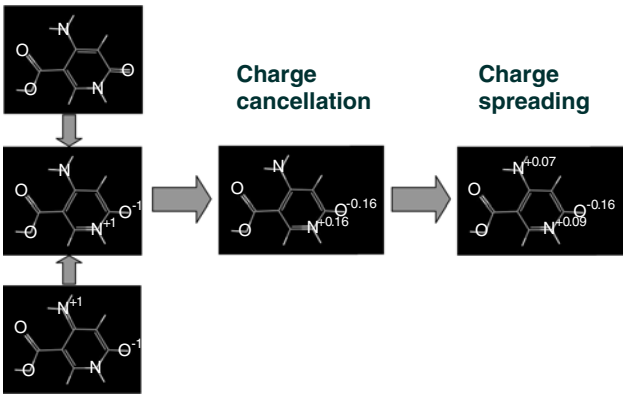


Fig. 2 An illustration of Epik's mesomer processing technology. At left are three resonance contributors, or mesomers, for the same molecule. If the input structure has either the top or the bottom form it is transformed into the middle, zwitterionic form. If the sums of the positive and negative formal charges are both non-zero charge cancellation is carried out. In the charge spreading stage, the remaining fractional formal charges are redistributed amongst the heteroatom sites in the mesomers

the w_m values for the acidic and basic forms of the ABG. The following sums run over all mesomeric sites for each ABG:

$$
\begin{aligned}
Q_+ &= \sum_m F_m H(F_m) \\
Q_- &= \sum_m F_m H(-F_m) \\
Q_T &= Q_+ + Q_- \\
W_+ &= \sum_m w_m H(w_m) \\
W_- &= \sum_m w_m H(-w_m) \\
n_+ &= \sum_m H(w_m) \\
n_- &= \sum_m H(-w_m) \\
H(x) &= \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \le 0 \end{cases}
\end{aligned}
\tag{3}
$$

If the totals for both of the positive and negative formal charges, Q_+, and Q_- , are non-zero charge cancellation is

used to reduce their magnitudes. The cancellation charge C is calculated using the following scheme:

$$C = \begin{array}{rlll} -Q_- & \text{if} & Q_T \geq 0 \quad \text{and} & n_- = 0 \\ -c_f Q_- & \text{if} & Q_T \geq 0 \quad \text{and} & n_- > 0 \\ Q_+ & \text{if} & Q_T < 0 \quad \text{and} & n_+ = 0 \\ c_f Q_+ & \text{if} & Q_T < 0 \quad \text{and} & n_+ > 0 \end{array} \tag{4}$$

where the charge cancellation factor, c_f, is assigned a value of 0.84 based upon empirical experience. The cancellation charge is then used to adjust the totals of the formal charges:

$$\begin{aligned} Q'_+ &= Q_+ - C \\ Q'_- &= Q_- + C \end{aligned} \tag{5}$$

which are in turn used to calculate the spread formal charges, f_m, on each of the mesomerization sites.

$$f_m = \begin{array}{ll} Q'_+ w_m / W_+ & \textit{if} \quad w_m \geq 0 \\ Q'_- w_m / W_- & \textit{if} \quad w_m < 0 \end{array} \tag{6}$$

The σ value used in the Hammett equation or Taft equation for each mesomeric site, excluding those within the ionizable group itself, is given by:

$$\begin{aligned} \sigma_m &= (1 - r_m)\sigma_{m\alpha} + r_m \sigma_{m\beta} \\ r_m &= f_m - \alpha \\ \beta &= \alpha + 1 \end{aligned} \tag{7}$$

where α is the largest integer less than f_m. $\sigma_{m\alpha}$ and $\sigma_{m\beta}$ are the σ values for the functional group when it possesses formal charges of α and β, respectively.

Every effort is made to ensure that the fitting is not Lewis structure dependent, and that the results are the same regardless of which Lewis structures are input or generated internally. The combination of mesomer standardization, charge cancellation and charge spreading technologies avoid nearly all such issues. In cases where such dependencies are noted they are treated as a bug to be fixed by including additional patterns in the database.

Table 1 summarizes the numbers of the various types of stored data that Epik uses. Ionizable groups are specified using only the SMARTS pattern for the acid, while the SMARTS pattern for the basic form is automatically generated from this.

In molecules with multiple ionization sites the pK_a value for a site depends on the protonation state of each of the other sites. HT technology by its nature predicts pK_a values for the addition or removal of protons to particular site for a particular form of the molecule. Thus Epik's pK_a prediction capability gives pK_a values for the molecule as given. These values are recorded in Epik's log file and in the output structure file. Maestro [15] can display the latter as shown in Fig. 1. In solution at fixed pH different molecules of the compound may participate in different protonation transitions to a significant extent (e.g. when pK_a values for different ionization sites are closer than roughly one pK_a unit). Some experiments such as NMR [16] can be used to give pK_a values for particular ioniziations even under such conditions. However, other types of experiments, such as titrations, measure a net behavior, which is averaged over the different ionization transitions in such mixtures and yields apparent pK_a values. Epik does not predict apparent pK_a values however the distinction is not large if the intrinsic pK_a values differ by more than one unit.

Table 1 The numbers and types of Hammett and Taft parameters, and tautomers used in Epik 1.5

Calculation type	Feature type	Epik 1.5
pK_a prediction	Acid/Base	818
	Substituents	664
	Special σ values	314
	Heteroaromatic	30
	Transmission	3
Tautomerization	Types of tautomerization	217
	Total number of tautomers	1,255

Protonation and deprotonation

Once the ionizable sites in a molecule have been identified one can add or remove protons from these locations to generate additional protonation states. To reduce the potentially significant overhead associated with processing many atypical protonation states, Epik focuses on a target pH range which is specified by a target pH, *tpH*, with a pH range or tolerance, *pHtol*. If the pK_a of an acid is less than *tpH* + *pHtol* then a new form for the molecule in which the acidic proton has been removed will be generated. Similarly, if the pK_a of a base is more than *tpH* – *pHtol*, a new form for the molecule in which the site is protonated is generated. *pHtol* is an intuitive way to specify the range of protonation states that will be generated in terms similar to the uncertainty in the pK_a value estimations and in terms of the potential shifts in the pK_a values due to different environments such as the active site of a protein. Another way to think of *pHtol* is in terms of a minimum population, P_{min} where

$$P_{min} = 10^{-pHtol} \tag{8}$$

such that states with populations less than this value at *tpH* are eliminated. In calculations that generate protonation states *pHtol* is user-controllable and has a default value of 2 (P_{min} = 0.01). This methodology is similar to that used in ionizer stage of LigPrep [3].

Tautomerization

Epik uses a template-based tautomerization facility that has been available in the tautomerizer tool which is part of Schrödinger's LigPrep product [3]. Its goal is to generate the most likely tautomeric forms for drug-like molecules. More than 200 types of tautomerization for aliphatic and aromatic systems are supported. For each type of tautomerization there are two or more tautomeric forms listed. Each tautomer is encoded using a SMARTS pattern along with an estimated population. The estimated populations are obtained from experimental data or derived from Density Functional Theory in Jaguar [17]. For the latter the Jaguar populations are estimated by normalizing the Boltzman weights over the collection of tautomers for that type of tautomerization. To obtain the relative energies, each candidate tautomer is optimized with the B3LYP functional and aug-cc-pVTZ(-f) basis set, in the presence of continuum solvation (Jaguar's Poisson-Boltzmann Self-Consistent Reaction Field model).

The tautomerization facility uses SMARTS patterns to recognize tautomerizable groups within molecules. Multiple tautomerizable functional groups within the same molecule will be recognized and all combinations of tautomers for non-overlapping groups will be considered. The current implementation will also generate some combinations of overlapping tautomers. For molecules with more than one tautomerizable group the population of a particular net tautomeric form is estimated as the product of the populations for the individual tautomerizable functional groups. By default, this facility removes molecular forms with populations lower than P_{min}, however, the tautomeric form with the highest population is always retained to ensure that molecules with many tautomers are not completely eliminated if all of the tautomers have low populations.

The current scheme for tautomerization does not adjust the probabilities of the tautomers based upon what is attached to them. Given this limitation the tautomeric SMARTS patterns have been constructed to exclude their use when strongly perturbing functional groups, such as amino groups, are bonded to the tautomerizable framework. With the current facility for tautomerization, such situations can be handled on a case-by-case basis by adding a new tautomerization pattern that explicitly includes the amino group and thus has appropriate populations. Presently, the tautomer database has many patterns that include such perturbing groups but the coverage is far from being complete. As such, this tautomerization facility in its current form is not a comprehensive tautomerization tool but rather a useful mechanism to introduce more tautomeric variation within Epik. However, its use in combination with tautomerizations achieved via combinations of protonations and deprotonations provides an effective mechanism for sampling tautomeric states. We note that Epik 1.5 is not designed to only produce tautomers but rather to produce collections of protonations states as describe in the following section. Table 1 includes information on the number of tautomer patterns in the database.

Procedure for generating protonation states

Perhaps the most important use for Epik is to process drug-like molecules to produce one or more protonation states that have significant populations in solution for each molecule. As mentioned earlier, once a proton is added or removed or a different tautomeric form generated, the pK_a values for many if not all of the ionizable sites in the molecule change significantly. In addition, different charge states of a molecule may have very different tautomeric equilibria. Multiple protonation changes are needed for some drug molecules given the way their structures are sometimes recorded in databases.

Epik deals with these strong couplings by using an iterative approach, with each iteration involving tautomerizing and then ionizing a collection of protonation states originating from a single input structure as illustrated in Fig. 3. Since such a procedure can generate numerous variations, many of which are often unimportant, culling procedures are necessary to keep processing time down and the output structures focused on the important protonation states. These culling procedures involve eliminating protonation forms that are identical or equivalent to those already in the collection of protonation states and to eliminate improbable states for the molecules involved. The latter involves a scheme for estimating the populations of the various protonation states present. Here are some terms that we will use while describing this procedure:

- ICS is the input collection of structures for the current stage,

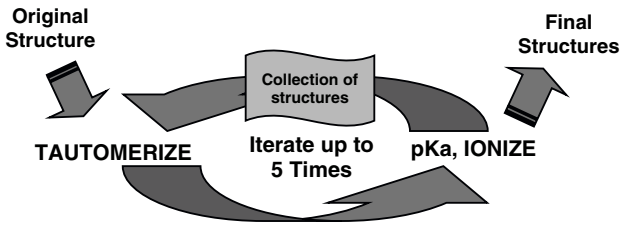


Fig. 3 Epik's iterative process for generating protonation states. The original structure is first tautomerized to produce one or more structures and the output from that stage is then ionized to produce a new collection of protonations states. This process of tautomerization and ionization is iterated up to five times or until no change is introduced by an iteration

- NCS is the new collection of structures being created in the current stage,
- Populations are estimates of the fraction of a set of structures that a particular structure will constitute. For mature collections of structures these populations should resemble but not necessarily coincide with the real populations in solution.

The initial structure is assigned a population of 1.0 and is regarded as the ICS for the first iteration of protonation state adjustment.

During the tautomerization stage each structure k in the ICS is tautomerized separately. The resulting tautomers are added to the NCS if they are not identical or equivalent to one already present there. Each tautomer, t, added to the NCS is initially assigned a weight, T_t given by:

$$T_t = P_k^{ICS} p_t / p_k \tag{9}$$

where P_k is the population of structure k in the ICS and, p_k and p_t are the populations of the structures k and t amongst the various tautomeric forms considered for structure k by the tautomerization facility. When all of the ICS has been processed, the tautomeric weights in the NCS are renormalized to give populations, P_t:

$$P_t = T_t / \sum_{t'} T_{t'} \tag{10}$$

All structures with $P_t < P_{min}$ are eliminated from the NCS to generate the ICS for the ionization stage.

Each structure, k, in the ICS is ionized separately using the following procedure:

1. Estimate the pK_a values of all of the ionizable groups
2. Accumulate structures into the NCS by examining each ionizable group in turn:
   I. If $pK_a < tpH + pHtol$, and the group is:
      i. in a basic form add that form to the NCS
      ii. in an acidic form remove the proton and add it to NCS
   II. If $pK_a > tpH - pHtol$, and the group is:
      i. in an acidic form add that form to the NCS
      ii. in a basic form add a proton and add the resulting structure to the NCS

In the preceding if an equivalent structure is already present in the NCS a redundant copy is not added.

Ionization weights, $u_{i,l}$, are calculated for each ionizable site, i, molecule, l, within the NCS using:

$$u_{i,l} = 1/\left(10^D + 1\right) \tag{11}$$

and ionization weight for molecule l as a whole is given by:

$$U_l = \prod_i u_{i,l} \tag{12}$$

This formula inherently makes the approximation that the ionization equilibrium for each ionizable site is independent.

The net weighting V_l for each molecule within the NCS is given by:

$$V_l = P_k^{ICS} U_l \tag{13}$$

where $P_k^{ICS}$is the population of the state k within the ICS from which state l was generated. These weights are normalized within the NCS to give populations:

$$P_l = W_l / \sum_{l'} W_{l'} \tag{14}$$

and structures with $P_l < P_{min}$ are eliminated from the NCS.

At this stage the NCS is usually used as the ICS for another cycle of tautomerization and ionization. However, if this is the fifth such iteration or if the NCS is the same as that produced by the ionization stage in the last iteration, the NCS becomes the output from structural adjustment.

The population for each output protonation state is converted into an energy penalty using:

$$E_l = -RT \ln(P_l) \tag{15}$$

and recorded along with the output structures. These penalties may prove useful in subsequent processing using other applications.

By default, to prevent excessive processing times and to limit the number of output protonation states, the number of structures that survive each culling is limited to 16. The user can specify a different number. However, if it is less than 3, the internal cullings retain up to three structures to permit the evolution of the collection of structures through less probable states. Sometimes the less probable states must be visited in order to reach most of the high probability states.

Despite the limitations and approximations used in the protonation state generating process, it produces the most important protonation states for the vast majority of the structures for which Epik has been parameterized to give accurate pK_a values.

Procedure for pK_a scanning

Neither of the approaches for estimating pK_a values for the current structure or generating the most populated protonation states is suitable for comparing with experimental pK_a values, either for evaluation or fitting parameters. The main reason for this is that the experimental pK_a values

may not all come from an ionization immediately accessible from a single structure. For example, if a structure has three experimental pK_a values at least one must involve ionizing any input molecule more than once. Epik's pK_a scanning feature is suitable for this type of comparison. In this approach, each input structure is processed in three stages:

1. Structure adjustment as described in the previous section is used to produce what Epik regards as the most probable form for the molecule at pH 7.0. This structure will be referred to as the pH-adjusted structure.
2. The pK_a of the most acidic proton is noted. This proton is removed and the pK_a values are recalculated. This process is repeated until no more acidic hydrogen atoms remain or the pK_a for removing the next proton rises above pH 17.0.
3. The pH-adjusted structure is reinstated. The pK_a of the most basic atom is noted. A proton is attached to this atom and the pK_a values are recalculated. This process is repeated until no basic atoms remain or the pK_a for adding the next proton drops below –3.0.

For each input structure, the pH-adjusted structure is saved in the output structure file. The pK_a values for removing or adding protons are recorded in the .log file in ascending order along with the type of ionization relative to the pH-adjusted structure (acid or base), the heavy atom bonded to the proton and the type of HT pattern used. As mentioned earlier, Epik estimates pK_a values and not effective pKa values which are what is measured in titration experiments. The practical effect of this is that when two or more sites have pK_a values within 1 pK_a unit of each other the predictions from this pK_a scanning approach should still roughly coincide with the effective ones but with a greater uncertainty.

Training and performance

This section will describe both of Epik's training processes and performance on a non-training set. Training consists of adding new types of parameters corresponding to new classes of functional groups and less often adjusting the existing parameters to better fit the experimental data. Epik's goals are somewhat atypical for pK_a software in that one of its main goals is the robust automated production of the appropriate protonation states when applied to large collections of medicinal chemicals—compounds with a complex and varied protonation behavior. This means that Epik must not only deal well with standard functional groups like amines and carboxylic acids but it must also perform well on more complex chemical species involving heteroaromatic rings including fused heterocycles along with strongly perturbing substituents. At the same time we strive to use a fairly limited number of parameters to achieve broad coverage in order to develop an approach that should handle molecules outside of its training set well.

Initially Epik's parameter set was created using parameters from the literature [2, 18] which usually come from studies targeted at a particular type of functional group. As a result, Epik should perform well on common, simple functionalities (e.g. carboxylic acids). The real challenge beyond this has been to detect when an ionizable group is present that is either not identified at all by the current coverage or is matched with an existing set of parameters yet is distinct enough that it should be described by new parameters.

Epik development has utilized three data sets:

I. Systematically constructed monofunctional molecules with pK_a values from Jaguar's pK_amodule[8] or from selected experiments. Currently there are a number of subsets to this collection comprising in total 500–600 systems. This data set is used to detect cases where Epik may need new HT parameters. The protonation states in the database are to the best of our knowledge appropriate so Epik is run in pKa prediction mode without adjusting the protonations state of the molecules. If Epik's predictions differ significantly from the values in the database that functional group is flagged for more attention. For such cases if only Jaguar pK_a values are present, experimental data for related compounds is sought to determine if Epik's predictions need improvement and if so that data is used for fitting the parameters. Only in rare instances, where no experimental data could be found and Jaguar's pKa estimates are expected to be very reliable were Jaguar's values fit to.

II. 3295 molecules with high quality experimental pK_a data from the literature [19–22] which comprises a fairly random subset of the data available. There are a total of 4,114 pK_a values because a significant fraction of these molecules report pK_a values for more than one ionization. Of these 57 have been tagged as unreliable because of signs that the molecules may be undergoing chemical reactions outside protonation state changes. The remaining 4,057 pK_a values are used to measure how well the parameter set is doing, to identify gaps and for limited refitting of parameters. Jaguar is sometimes used to help understand the experimental pK_a values.

III. 123 drug-like molecules from the Drugbank database [23] that possess pK_a data were used to detect problems that need addressing, however they were not used in fitting parameters for predicting pK_a values.

Epik is run in scan mode, involving the prediction of both the protonation states and pK_a values for data sets II and III.

Data sets I and II will continue to grow and we expect to add more data sets in the future. No molecules have been excluded from data sets II or III because of the type or classes of protonation behaviors they exhibit. One additional class of testing is the internal application of Epik to large databases of drug-like compounds by people who have detailed knowledge of the protonation behavior of pharmaceutical agents. This process identifies relevant problematic cases that warrant further attention and constitutes an important part of Epik's on-going improvement process.

Figure 4 presents the distribution of the errors (experiment minus predicted) for data set II. This distribution is fairly closely reproduced by a combination of two Gaussians with σ values of 0.65 and 1.9. The 0.65 Gaussian accounts for most of the main peak however the wider Gaussian is needed to account for the broad shoulders on this peak. The need for two Gaussians can be rationalized as follows. Epik uses pattern matching to select which parameters are to be used to predict pK_a values. If a suitable pattern is present the predictions should be accurate. However, if a suitable pattern is not present the error in the predictions will be significantly larger. Predictions within roughly 1 pK_a unit are accurate enough to be useful. For instance, when generating protonation states with the default value for *pHtol* 2 units, the important states should be present in the output if the error is 1 pK_a unit. The main

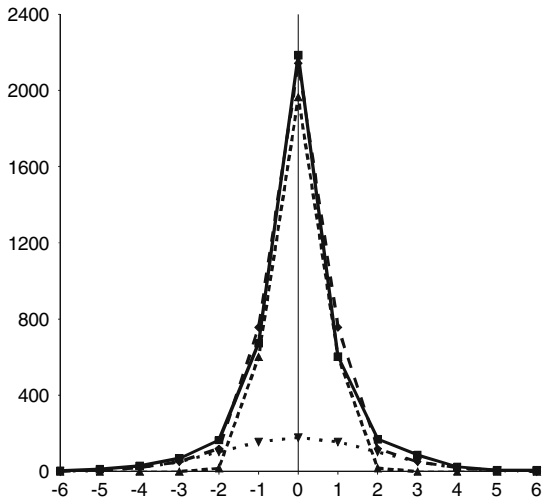


Fig. 4 Distribution of differences between experimental pK_a values and those estimated by Epik. The squares and the solid line combination correspond to the observed differences placed in bins 1 pK_a unit wide. The diamonds and long dashed line correspond to the sum of two Gaussians with σ constants of 0.65 and 1.9 units represented by the upward pointing triangle and medium dash line, and the downward pointing triangle and short dash, respectively

challenge in improving Epik has been in identifying new, distinct ionizable functionalities and then fitting the corresponding parameters so that the number of pK_a values that fall under the wider Gaussian distribution is reduced.

Standard deviations are often used to describe the overall magnitude of the error between predicted and observed values and we will also follow this practice. However, since the distribution of the error is not well described by a single Gaussian we will also report the average and median of the magnitude of the error. These are less sensitive to outliers in non-Gaussian distributions. The errors for data sets II and III are presented in Table 2. Data set I is not listed because it is only used to identify functionalities that are missing, which are typically fixed accurately and immediately.

For both data sets most of the experimental pK_a values are reproduced within 1 pK_a unit. In addition, while the errors for data set III are larger than data set II, a data set used in fitting, the degradation in performance is not dramatic, illustrating that the strategy of using a limited set of parameters to achieve broad coverage seems workable. The supplemental material lists pK_a values for data set III from both experiment and Epik.

Epik requires roughly 15 s on a 2 GHz PIV class processor to read and process the parameter files prior to processing molecules. Epik can process many ligands in a single run. After initialization, Epik takes roughly 0.25 s on average per drug-like molecule to estimate the pK_a values for the structure as given. To generate one or more protonation states for each input structure it takes roughly 3 s per ligand on average. This process takes longer because often multiple states are processed, and tautomerization and the iterative approach are used. pK_a scanning involves first generating the most important protonation state (usually at pH 7) followed by sequential removal and addition of protons so it requires more time than protonation state generation, roughly 7 s per ligand on average. Epik is fast enough to be applied to large databases of drug-like molecules. There is a script, para_epik that makes it easy to distribute jobs across a number of processors so that the wait time for results is reduced.

Summary

Epik is a program for predicting pK_a values and generating protonation states rapidly enough to process large databases of drug-like molecules. Epik can be used by LigPrep to adjust protonation states as part of an effective 3D ligand structure preparation process and is used to generate candidate ligand protonation states in a protein preparation workflow, PrepWizard, in Maestro. There is room for further improvement through speed enhancements, more

Table 2 Data set size and performance

Data set	Number of molecules	Number of pK_a values	Number predicted	Standard deviation	Average of abs error	Median of abs error
II	3,295	4,057	4,056	1.27	0.78	0.44
III	123	123	123	1.37	0.92	0.64

patterns for ionizable groups and new technologies for predicting pK_a values. However, in its current form, Epik 1.5 is sufficiently accurate and reliable, in part due to the new charge cancellation and charge spreading technology, to be a useful tool for handling protonation state behavior in drug-like compounds.

Appendix: Prioritizing SMARTS patterns for ABGs

All SMARTS patterns for ABGs are assigned numeric priorites. The pattern with the highest priority that matches a particular ABG is selected and the parameters (e.g. pKa^0 and ρ) associated with that pattern are used in the HT calculations. These priorities are assigned in one of three ways:

1. The ABG was manually assigned a negative priority and thus are matched only if a more specific pattern is not found. This was only done for very general patterns (e.g. primary, secondary or tertiary amines) which would match many functionalities, most of which are better described by more specific patterns (e.g. amides and anilines). Roughly 5% of the patterns in the database are assigned negative priorities.
2. The priority for the ABG was calculated from the SMARTS pattern.
3. In a couple of cases the priority was calculated as described in the last item (2) except that a manually assigned shift was added to distinguish very closely related patterns.

The procedure for calculating the priority from the SMARTS pattern will be outlined in detail below.

All SMARTS patterns for ABGs are recorded in the acidic form and begin with the acidic hydrogen followed immediately by the atom to with the acidic hydrogen is bound. We will refer that atom as the first heavy atom.

The SMARTS pattern is translated into a list of atoms and a list of bonds. The type of bond is noted or inferred from the SMARTS pattern consistent with the SMARTS standard. Each atom is classified SP3-like unless it meets one of two conditions:

1. If any of the bonds involving this atom are double, triple or aromatic
2. If it is a O, S or N and bonded to an aromatic atom

The priority, P, of a SMARTS pattern for an ABG is calculated using the equation:

$$P = 2^* a_2 + \sum_{i>2} p_i^* a_i \tag{16}$$

where: a_i is the weighting for atom i in the SMARTS pattern and p_i is an attenuation factor that depends on the shortest topological path from atom 2, the first heavy atom, to atom i. All atoms in the SMARTS pattern are included in the sum except the acidic hydrogen atom (atom 1). The a_i values were determined by trial and error and are given in Table 3. The p_i values were calculated using the equation:

$$p_i = \prod_j s_j \tag{17}$$

where s_j is a attenuation factor corresponding to a portion of the shortest path from atom 2 to atom i. Each non-aromatic bond has a separate propagation factor while each set of consecutive aromatic bonds gets a single factor. Aromatic bonds are treated differently because the influence of atoms in aromatic systems does not monotonically

Table 3 Atom weighting factors used in prioritizing SMARTS matches for ABGs

Element or SMARTS symbol	SP3-like	Bound to aromatic	a_i
H			0.1
C	Y		0.5
	N		0.1
N	Y		1.0
	N		1.1
O	Y		2.0
	N	Y	4.0
		N	2.2
S	Y		1.3
	N	Y	3.7
		N	1.5
F			2.0
Cl			1.6
Br			1.55
I			1.4
A,a			0.07
*			0.04
All others			0.5

Table 4 Attenuation factors, s_j, for different non-aromatic bond types

SMARTS Bond type	s_j
–, ∼, @ (single and wild-card)	0.5
= (double)	0.7
# (triple)	0.85

Table 5 Attenuation factors, s_j, as a function of aromatic path length

# of aromatic bonds	s_j
1	1.0
2	0.7
3	0.85
4	0.5
5	0.6
>5	$0.7^{n-4}/2$

decrease with the number of bonds. The attenuation factors for non-aromatic bonds are given in Table 4 while those for aromatic bonds are given in Table 5.

References

1. Epik 1.5 (2007) Schrödinger, LLC, New York, NY
2. Perrin DD, Dempsy B, Sergeant EP (1981) pKa prediction for organic acids and bases. Chapman and Hall, London
3. LigPrep 2.1 (2007) Schrödinger, LLC, New York, NY
4. Resonance structures and mesomers are synonyms as discussed on the webpage: http://en.wikipedia.org/wiki/Resonance_structures. In the context of Epik we prefer to use mesomer because of its connection with the terminology "mesomeric effect" which often arises in discussions of pK_a values
5. ACD: http://www.acdlabs.com/products/phys_chem_lab/
6. ChemAxon: http://www.chemaxon.com/product/pka.html
7. Sparc software: http://www.epa.gov/ATHENS/publications/reports/EPA_600_R_03_033.pdf
8. Klicic JJ, Friesner RA, Liu S-Y, Guida WCJ (2002) Phys Chem A 106, 1327 http://www.schrodinger.com
9. http://www.chemsilico.com/CS_products/products.html
10. SMARTS, SMiles ARbitrary Target Specification, is a registered trademark of Daylight Chemical Information Systems
11. Jaffé HH (1953) Chem Rev 53:191
12. Clark J, Perrin DD (1964) Quart Rev 18:295
13. Longuet-Higgins HC (1950) J Chem Phys 18:265. ibid. 275. ibid. 283
14. Perrin DD (1965) J Am Chem Soc 5590
15. Maestro 8.0 (2007) Schrödinger, LLC, New York, NY
16. Hägele G, Holzgrabe U (1999) In: Holzgrabe U, Wawer I, Diehl B (eds), NMR spectroscopy in drug development and analysis. Wiley-VCH, Weinheim Germany, pp 61–76
17. Jaguar 7.0 (2007) Schrödinger, LLC, New York, NY
18. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR, hyrdophobic, electronic and steric constants. American Chemical Society, Washington, DC
19. Serjeant EP, Dempsey B (1979) Ionization constants of organic acids in aqueous solution. Pergamon Press, Oxford England
20. Perrin DD (1965) Dissociation constants of organic bases in aqueous solution. Butterworths, London
21. Perrin DD (1972) Dissociation constants of organic bases in aqueous solution: supplement 1972. Butterworths, London
22. CrossFire Beilstein, version 7.0; MDL Information Systems GmbH, Frankfurt am Main, Germany (http://www.mdl.com)
23. DrugBank http://redpoll.pharmacy.ualberta.ca/drugbank/