BEST PRACTICES

Schrödinger Solutions for Small Molecule Protonation State Enumeration and pK_a Prediction

Executive Summary

The pK_a of a drug is a key physicochemical property to consider in the drug discovery process given its importance in determining the ionization state of a molecule at physiological pH. Schrödinger provides several solutions for predicting pK_a values, protonation state distribution and derived properties that can be applied across a range of drug discovery stages, from screening through lead optimization. Here we provide an overview of each technology solution and use case examples of how they can be applied in drug discovery.

Background

Small molecules can undergo ionization in solution where they either lose or gain protons (H^+) at different ionizing sites. The measure of the propensity of a site or molecule to ionize by the association/dissociation of one or more protons is quantified by a pK_a value. If the pK_a value refers to a particular site ionizable site the value is a microscopic pK_a (micro-pK_a), and it is a macroscopic pK_a (macro-pK_a) if the value refers to the entire molecule. The specific arrangement of protons around the ionizing sites constitutes a protonation state, and different protonation states of the same charge level are called tautomers. Each protonation state is in thermodynamic equilibrium with the others and therefore has a free energy associated with its population within this collection of protonation states, which may be derived either from micro-pK_a values through thermodynamic equations or obtained directly by comparing the free energies of the states. In drug design, understanding the different protonation states of a molecule is critical, since they will drive properties including solubility, membrane permeability, and activity.

Challenges of pK_a Prediction

Determining which states predominate at a given pH and by how much is a challenging task both experimentally and computationally because the number of states that are all in thermodynamic equilibrium grows ~2^n with the number, n, of singly protonatable sites. Thus, molecules with many titratable sites can potentially have a large number of different protonation states, all of which need to be enumerated and energetically scored.

Computationally, Schrödinger uses two main approaches to score states: 1) through evaluating thermodynamic equilibrium equations with micro-pK_a values, and 2) directly predicting the states' relative free energies. Predicting pK_a values is an important step to calculating state distributions, which in turn enables prediction of important related quantities that would otherwise be inaccessible.

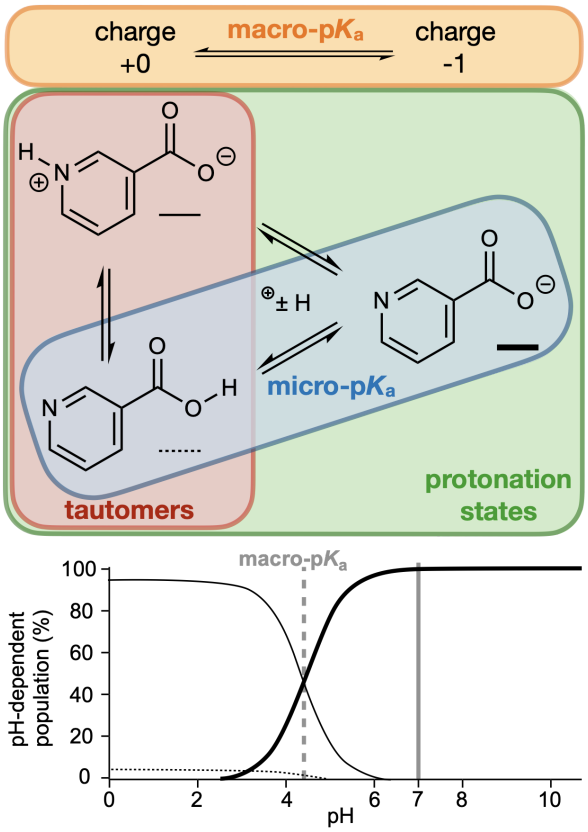

Figure 1. Relationships between macro-pK_a, micro-pK_a, protonation states, and tautomers and the corresponding speciation diagram.

Overview of Schrödinger Solutions

Epik Classic

Epik Classic, previously known simply as Epik[1], is an expert system for rapidly and accurately predicting the micro-pK_a values and the most populated protonation states for a ligand at a given pH. The underlying pK_a prediction technology is the empirical Hammett-Taft linear free energy relationship (LFER), which identifies an ionizing group, takes its root pK_a value, perturbs it by the bonded chemical fragments, and applies charge spreading to arrive at its effective micro-pK_a value. Epik Classic then uses the predicted pK_a values to enumerate a ligand's protonation states, rank them by energy, and then return the most populated states. Because Epik Classic uses SMARTS patterns-based rules, it is fast enough for high-throughput, although at the expense of being unaware of both conformational and stereochemical effects.

Epik 7

Epik 7[2] is a complete redesign of Epik that leverages Schrödinger's powerful machine learning (ML) technology for more accurate results across broader chemical space. Ionizing groups are initially identified by SMARTS patterns and are then used to enumerate the protonation states for a range of ionizations. The micro-pK_a values of each site in each state are predicted with 3-layer atomic graph convolutional neural networks (GCNNs) extending out radially six bonds from the ionizing atom. The predicted pK_a values for the states are then used to predict the relative energies of the states to both allow determination of the most populated states at a pH and calculation of macro-pK_a values. The topological nature of the ML approach means that Epik 7, like previous versions, is rapid but agnostic to 3D geometry and stereochemistry.

Jaguar pK_a

Jaguar pK_a takes a third, more physics-based approach to predicting micro-pK_a values for a ligand. This workflow calculates the pKa values at the user-defined ionizing sites in a query ligand by first generating the conjugate pair, on which are then executed conformational searches to locate the lowest energy structures,[3] followed by density functional theory (DFT) based geometry optimizations and single-point energy evaluations. These resulting conformationally-averaged, "raw" micro-pK_a values are then corrected using empirically-parametrized relationships to give accurate predictions. Jaguar pK_a performs best on non-tautomerizable structures. Being physics-based, it does take into account geometric and stereochemical effects, but at the expense of speed.

Macro-pK_a

Macro-pK_a follows the same philosophy as Jaguar pK_a by combining physics-based DFT calculations with empirical corrections, but extends its applicability to enable calculation of tautomerizable ligands. Macro-pK_a automatically identifies ionizing sites, enumerates the protonation states, and calculates the micro-pK_a values following a similar workflow to Jaguar pK_a, but with an enhanced scheme for generating empirical corrections. Finally, the calculated micro-pK_a values are used to rank the protonation states by energy, return the most populated states for a user-supplied pH, and determine the macro-pK_a values for the ligand. The exhaustiveness of this approach comes at a larger time and resource cost than Jaguar pK_a.

Use Cases

Here we outline several use cases for pK_a prediction in the drug discovery workflow.
Note: Each use case example outlines below could be approached with any of the listed solutions within that section. The dataset presented highlights the applicability of just one of the possible solutions.

I. Querying microscopic pK_a values

Applicable Solutions

- Epik Classic
- Epik 7
- Jaguar pK_a

When investigating the binding modes of a ligand, the micro-pK_a value of an ionizing site is an indicator of the propensity for it to become ionized at a given pH. The ionization state of the ligand directly influences how it interacts with another molecule such as a protein, e.g., whether or not it can participate in a salt bridge.

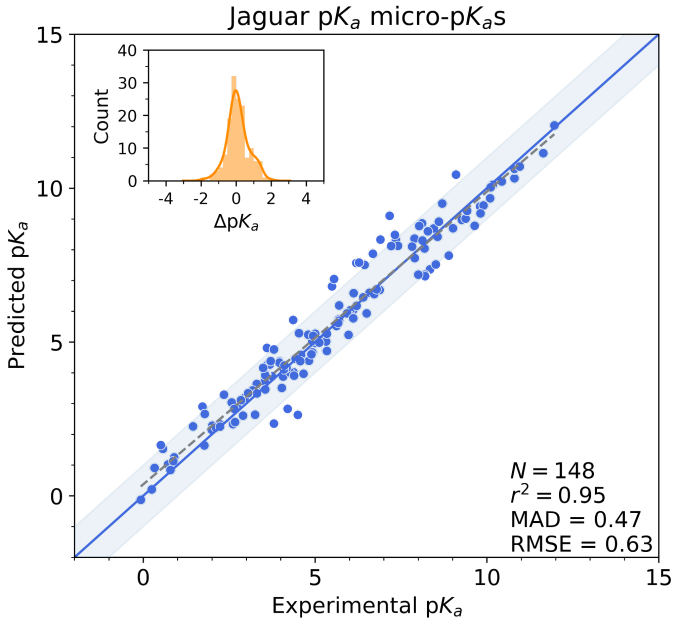

Figure 2. Jaguar pK_a micro-pK_a predictions for a dataset of small molecules.

II. Querying apparent or macroscopic pK_a values

Applicable Solutions

- Epik 7
- Macro-pK_a

For monoprotic or polyprotic compounds with a single dominant tautomer at each charge level, micro-pK_as may very closely match the apparent or macro-pK_a value that is most commonly obtained through titration experiments. However, for compounds or ionization states with multiple competitive tautomers, the micro-pK_a value of a single tautomer may not fully reproduce the experimentally observed macroscopic value. To obtain this apparent value, all states' must first be enumerated and evaluated so that all their micro-pK_a values are considered in the macro-pK_a calculation.

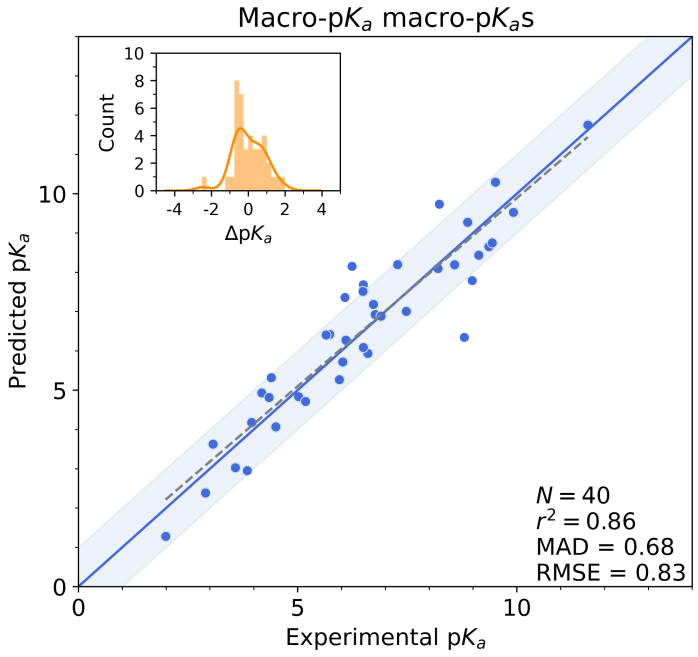

Figure 3. Macro-pK_a macro-pKa predictions for a dataset of tautomeric molecules.

III. Ligand preparation and high-throughput screening

Applicable Solutions

- Epik Classic
- Epik 7

Physics-based simulations typically require specification of all atoms in the simulation system, including all hydrogen atoms. Thus, structure-based simulations including Glide docking, molecular dynamics, and free energy perturbation with FEP+ should be performed using an ensemble of the highly-populated protonation states of a ligand. Therefore, a crucial first step in any structure-based screen of a small molecule ligand library is to prepare the ligands by obtaining the most populated protonated states. Epik Classic and Epik 7 are integrated with our automated ligand preparation workflow, LigPrep, to allow preparation of large ligand libraries for high-throughput screening. Additionally, both Epik Classic and Epik 7 and their LigPrep implementations allow for the generation and scoring of additional states that may potentially bind to metal ions in the pocket.

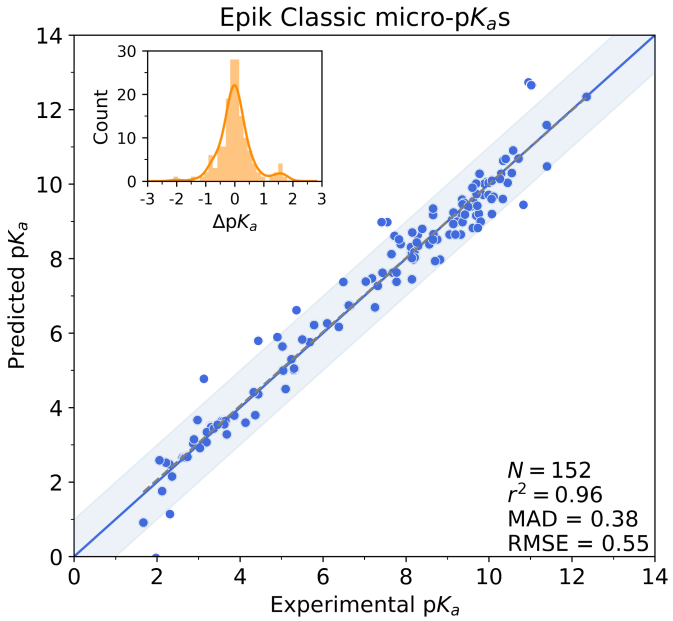

Figure 4. Epik Classic micro-pK_a predictions for a dataset of 152 drug molecules

IV. Hit-to-lead optimization

Applicable Solutions

- Epik Classic
- Epik 7

Once hits are identified, a series of analogs are synthesized to explore the relevant chemical space in greater detail to arrive at improved behavior. It is important to be able to screen potential candidates rapidly and accurately to assess which to optimize further. The < 0.5 log unit accuracy and sub-second calculation speed of Epik Classic and Epik 7 make them excellent tools for rapid idea generation and testing. In addition to pK_a value and protonation state distribution prediction, they have been implemented in other ADMET or property predictors, such as for membrane permeability and solvation energy.

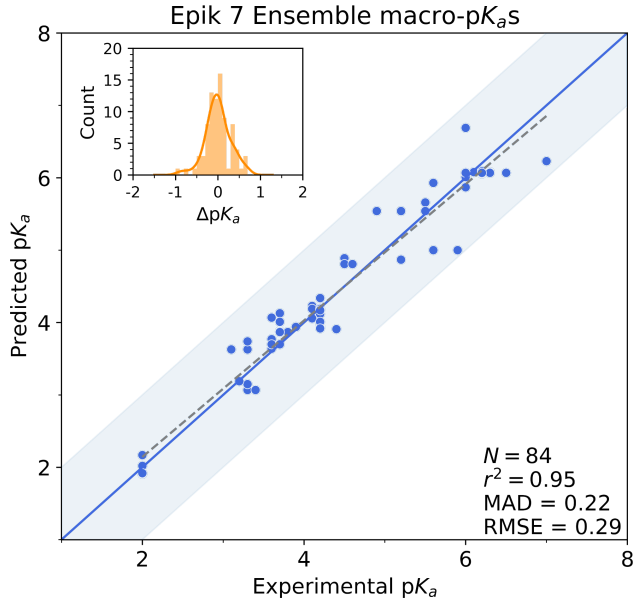

Figure 5. Epik 7 macro-pK_a predictions for a dataset of congeneric tricyclic thrombin inhibitors.

V. Early-stage lead optimization

Applicable Solutions

- Epik 7
- Jaguar pK_a
- Macro-pK_a

Optimizing the many physical characteristics required can be laborious and costly, from ideation, through synthesis and assay. In this environment, where high quality property predictions are required and time permits, Schrödinger's physics-based predictors, Jaguar pK_a and Macro-pK_a, take into account more molecular characteristics, including conformational and stereochemical effects to improve pKa prediction accuracy.

Additionally, Macro-pK_a and Epik 7 both offer detailed speciation reports for a queried ligand. These are especially helpful for understanding the distribution of tautomeric states across the pH spectrum.

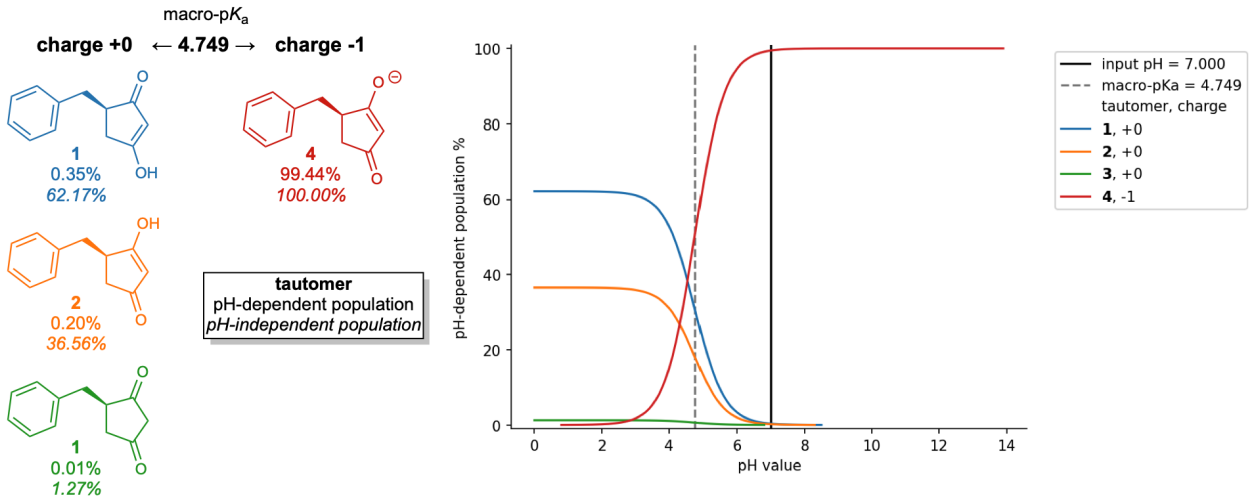

Figure 6. A Macro-pK_a report detailing the macro-pK_a value and the distribution of protonation states across a pH range.

Feature Comparison Table

Feature\Program	Epik Classic	Epik 7	Jaguar pK_a	Macro-pK_a
Technology	Hammett-Taft LFERs	Atomic GCNNs	DFT + Empirical Fitting	DFT + Empirical Fitting
Year Introduced	2007	2022	2005	2023
Ionization Enumeration Distance	±5	±2[a]	±1	±2[a]
Average Single Ligand Run Time	0.2 s	0.4 s	1 h	1 d
Practical Maximum Ligand Size	150 atoms	200 atoms	100 atoms[b]	100 atoms[b]
Predicts micro-pK_a values	Yes	Yes	Yes	Yes
Predicts macro-pK_a values	No	Yes	No	Yes
Enumerates States	Yes	Yes	No	Yes
Predicts Populations	Yes	Yes	No	Yes
Accurate pK_a values	Yes	Yes	Yes	Yes
Remote Intramolecular Interactions	No	No	Yes	Yes
Conformational Effects	No	No	Yes	Yes
Stereochemistry	No	No	Yes	Yes
Pseudochirality	No	No	No	Yes
DMSO Solvent	Yes	No	Yes	No
Metal Binding States	Yes	Yes	No	No
Trainable	No	Yes[c]	Yes	Yes[c]
Panel	Yes	Yes	Yes	Yes
Speciation Report	No	Yes	No	Yes
LigPrep Integration	Yes	Yes	No	No

[a] Easily adjustable; [b] Strongly influenced by the number of conformers (and tautomers in Macro-pK_a); [c] Only by internal experts at this time.

Table 1. Comparison of Features of the Small Molecule Protonation State Enumeration and pK_a Prediction Technologies

References

(1) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pK_a Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided Mol. Des.* **2007**, *21* (12), 681–691.

(2) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Elk, J. C.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. Epik: pK_a and Protonation State Prediction through Machine Learning. *J. Chem. Theory Comput.* **2023**, *19* (8), 2380–2388.

(3) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based pK_a Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* **2016**, *12* (12), 6001–6019.