Prediction of Ionization Constants for Complex Multicenter Electrolytes Utilizing Proprietary 'In House' Data

Pranas Japertas[a,b], Rytis Kubilius[a], Dainius Simelevicius[a,c]

[a] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania; [b] Faculty of Chemistry, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania; [c] Faculty of Mathematics and Informatics, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania

Pharma Algorithms

www.pharma-algorithms.com info@pharma-algorithms.com

Introduction

Ionization is one of the key parameters that affects absolute majority of the physicochemical properties and biological activities that are of interest to the developers of any new marketed chemicals, regardless of their intended use. Therefore, estimation of pK_a values has always been the field where prediction accuracy received special attention from the industry. In this work we present the methodology of pK_a prediction developed by Pharma Algorithms which utilizes a novel similarity based routine, allowing estimation of reliability for each prediction (evaluation of the Model Applicability Domain) and providing a possibility to expand the Applicability Domain of the model with the help of any user-defined proprietary 'in house' databases of experimental pKa values.

Calculation of Baseline pK_a values

Microscopic pK_a constant, or *microconstant*, is a value that describes dissociation of a certain ionogenic group of the molecule. As it can be seen in the below scheme (see Fig. 1), representing a complete ionization profile of the cysteine molecule, there are different microconstants of the same group for each protonation state. This is due to the obvious fact that the presence of other charges in the molecule strongly influences the dissociation. Therefore the calculation of the microconstant for any ionogenic group in a particular protonation state starts with the estimation of micro pK_a value for that group in a hypothetical state of an uncharged molecule ("fundamental microconstant"). These values are then corrected in order to account for the actual surroundings of the reaction center including charge influences of any neighboring ionization centers.

In total, algorithm utilizes a data set of >18,000 compounds with experimental pKa measurements, a database of 4,600 ionization centers, a set of ca. 500 various interaction constants and four interaction calculation methods for different types of interactions, producing a full range of microconstants from which pK_a *macroconstants* are obtained.

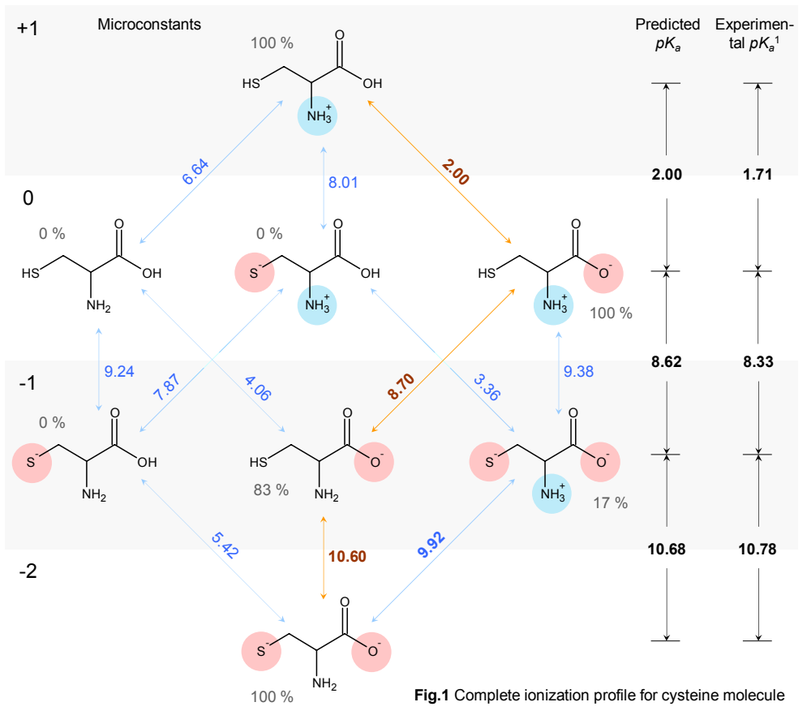


Fig.1 Complete ionization profile for cysteine molecule

When the difference of microconstants in a particular protonation state is big enough (> 2), practically there is only one group dissociating during that stage, and it is evident that the experimentally measured pK_a value would belong to this group. In that case, the predicted pK_a macroconstant and the microconstant of the group coincide. This is the case observed in the first stage of cysteine ionization above (see Fig. 1).

On the other hand, when there are two or more ionogenic groups with pK_a microconstants of comparable magnitude, they dissociate simultaneously. The dissociation runs at different extent, unless the microconstants are equal.

If we assume that there are *n* ionogenic groups with equal pK_a microconstants, then in the first stage of ionization the molecule has *n* equivalent ways of losing a proton but only one site to which the proton can be restored (there is still only one atom ionized). Therefore, the measured dissociation constant would become *n* times higher than the microconstant of a single group. Consequently, in this case:

$$pK_a(predicted) = pK_a(microscopic) - \log n \qquad (1)$$

In the second stage of ionization the correction (or "*statistical factor*") becomes log[(n – 1)/2], then log[(n – 2)/3], and so on. In the intermediate case, when the pK_a microconstants are not equal, the correction is less than the calculated statistical factor. The correction quickly comes down with increasing difference of microconstants.

'Trainable pK_a' model and the reliability of predictions

Every model, no matter what data, descriptors or modeling techniques were used building it, has a certain applicability domain, beyond which the quality of predictions becomes highly questionable. The fact that the literature based training sets rarely cover the specific part of the chemical space occupied by the compounds that a certain company is working with, makes this issue especially relevant in the application of third-party methods, trained on such public data sets, to the proprietary 'in house' compounds in the industry.

Therefore Pharma Algorithms has developed the 'Trainable pK_a' model utilizing a novel similarity based methodology that has already been proved efficient in the prediction of various other properties. It allows correcting the baseline pK_a values, simulated in the manner described previously, according to the experimental pK_a measurement results for the most similar compounds present in the user defined Self-training Library. Since pK_a is a property associated with a certain ionization site in the molecule, the approach to similarity determination in this case is different from the whole-molecule properties, e.g. *LogP*. The influence of atoms and fragments on the similarity of two molecules diminishes with increasing distance from the reaction center. In other words – the closest environment of the ionization center makes the greatest impact, with the possession of the same ionizable group, of course, being the main prerequisite for the molecular similarity. The below example comparing the 3 most similar compounds retrieved for different ionization centers of the Cetirizine molecule illustrates the actual performance of this similarity concept in practice (see Fig. 2).

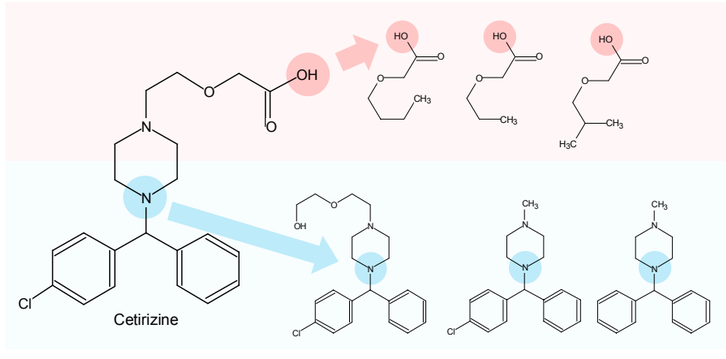


Fig.2 Examples of the most similar molecules retrieved for different ionization centers of Cetirizine

The possibility to add new data to the model without retraining it, enables the user to cover parts of his interest in the chemical space not initially included in the training set with great ease. In addition, this methodology allows quantitative assessment of the prediction reliability with the help of estimated Reliability Index (RI). This index, that is provided for every prediction, can have values in the range [0 to 1] and its estimation takes into account the following aspects:

- similarity of the tested compound to the training set
- consistence of experimental values for similar compounds

In the following example, a 'Trainable pK_a' model was prepared using the Self-training Library containing 6283 acidic and 8335 basic ionization centers with experimental pK_a measurements. This model was used to predict pK_a values of the first ionization stage for the separate dataset containing exactly the same number of ionizable groups (6283 acidic and 8335 basic). The bar-charts in the next column (see Fig. 3) present the RMSE values for acid and base pK_a predictions, obtained when the results are analyzed with regard to the Reliability Indices of predictions as well as the number of resulting predictions of different reliability for the considered test set.

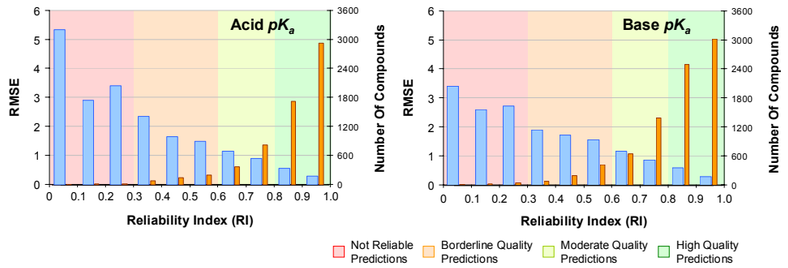


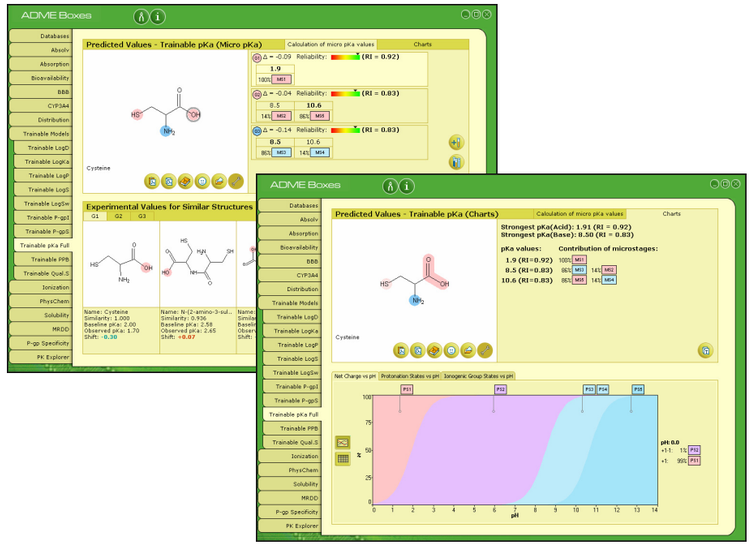
Fig.3 RMSE values for the predictions of different reliability and the number of compounds with corresponding predictions in the test set

Observed correlation between RI and RMSE values in this case clearly show that Reliability Index serves as a proper evaluation of whether a compound the model is trying to make prediction for is in the chemical space of the model. Lower values suggest compound being further from the model space and prediction less reliable, on the other hand high RI values indicate, that one can be quite confident about the quality of the prediction.

As it can be seen from the investigated example, the methodology behind the 'Trainable pK_a' model is efficient in evaluation of the Model Applicability Domain. Among other benefits, Reliability Index values can be also useful in measurement prioritization and experimental planning. In general, given the ease and throughput with which pK_a values are experimentally measured in today's industry, the ability to utilize this wealth of available 'in-house' information provides considerable advantages over standard ionization prediction tools.

Ionization prediction software

All parts of the 'Trainable pK_a' algorithm have been developed with the intention of their implementation as a predictive module inside ADME Boxes software. This allows the access to the whole predictive power of this complex calculation routine via the straightforward graphical user interface of this program illustrated below. A simplified version of the module without the ability to add user-defined data is also made available for the users dealing with compounds falling well within the Applicability Domain of the model based solely on the Pharma Algorithms pK_a training set.



References

1. The Merck Index. An Encyclopedia of Chemicals, Drugs and Biologicals. Thirteenth Edition. Eds., O'Neil, M.J., Smith, A., Heckelman, P.E. Merck & Co Inc, Whitehouse Station, NJ. 2001.