Comparsion of pK_a predictors on in-house dataset

György T. Balogh, Ákos Tarcsay

Gedeon Richter Plc., Lead Discovery Unit, Budapest, 10. Pf. 27. H-1475, Hungary.

Introduction

The success of the virtual screening is fundamentally influenced by the optimal tautomer form and ionization state that underlines the importance of accurate pK_a prediction for even large libraries. Moreover, ADME and PK optimization supported by in silico methods also requires information on the ionization state, therefore, fast and accurate prediction of pK_a values is in high demand. Testing commercially available in silico tools using unbiased compound set provides information about overall performance and uncovers typical errors that might need further attention. Our test set involved ~ 200 experimental pK_a values of ~100 compounds that were predicted by different pK_a predictors including Marvin, ACD, Epik, Pallas and Pharma Algorithm (ADME Box). Statistical analysis and specific examples will be discussed evaluating the predictive power of different commercially available pK_a predictors.

Materials and Methods

Potenciometric pK_a measurement

GlpKa automated pK_a analyzer (Sirius Analytical Instrument Ltd., Forest Row, UK) fitted with combination Ag/AgCl pH electrode was used for determination of dissociation constants of Gedeon Richter Plc.'s in-house compounds.

Potentiometric titration in aqueous medium. In general, 10.00 mL of a 1 mM aqueous solution of sample was pre-acidified to pH 2.0 with 0.5 M HCl, and then titrated with 0.5 M KOH to an appropriately high pH, usually 12. Titrations were carried at constant ionic strength (I=0.15 M KCl) and temperature (t=25.0±0.5 °C), and under nitrogen atmosphere. A minimum of three parallel measurements were carried out and the pK_a values of samples were calculated by RefinementPro software (Sirius Analytical Instrument Ltd., Forest Row, UK).

Titration in MeOH-water mixture. The co-solvent dissociation constants (p$_s K_a$ values) of the compounds with relatively low aqueous solubility were also determined in various MeOH-water mixtures between 15 and 65 wt%. Each sample was measured at least in four different MeOH-water mixtures. To obtain aqueous pK_a value from p$_s K_a$ data, the Yashuda-Shedlovsky extrapolation method (p$_s K_a$ + log [H2O] = (a/e) + b) was used.

UV-pH titration. The same titration protocol was performed as it was described at potentiometric titration in aqueous medium. Although, required sample concentration was two order of magnitude lower than in potentiometric titration. Multi-wavelength UV spectrophotometric titrations were performed with the Sirius D-PAS^{TM} ultra-violet spectrometer (Sirius Analytical Instrument Ltd., Forest Row, UK) attachment for the GlpKa. The D-PAS was fitted with bifurcated fibre-optic probe with path length of 1 cm (Hellma, UK). Spectrophotometry can applied for log *K* measurement provided that the compound has a chromophore in proximity to the ionization center, and the absorbance changes sufficiently as a function of pH.

All of the above mentioned experimental protocol was published by Balogh et al.[1].

Softwares used for pK_a predictions

Five pK_a predictor tools were evaluated: ACD[2], Epik (Schrödinger)[3], Marvin (ChemAxon)[4], Pallas[5] and PharmaAlgorithm (ADME Suit)[6]. Technical details as well as version numbers are summarized in Table 1.

Table 1. Computational details.

Software	Method	Version	Options
ACD	Hammet-Taft	v12.0	Apparent
Epik	Hammet-Taft extended with mezomer standardization, charge cancellation and charge spreading approaches.	v20211	(i) Ligprep neutralization, 3D conformer generation without tautomerization (ii) epik -scan -ph 7.0 -imae <input.mae> -omae <output.mae> -lowest_pka -10.0 –highest_pka 20.0
Marvin	Partial charge increments, polarizability increments and structure specific increments based site-specific regression	5.3.2	excalc pka -M true (take major tautomeric form) -a 4 -b 4
Pallas	Similarity based	3.5.1.4	pK_a by pKalc
Pharma Algorithms	Adapted after Hammett and Taft takes into account all necessary electronic, steric and other effects	ADME Box 5.0	Trainable pK_a (read-only) database

Data set

100 drug-like in-house compounds with 194 ionizing groups were selected from our compound collection. The distribution of measured pK_a values and molecular weight is shown in Figure 1. and Figure 2, the mean Tanimoto distance in test collection was 0.78, using ChemAxon fingerprints. Molecules were stored in 2D sdf file format.

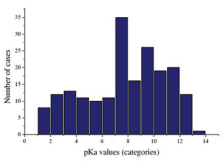


Figure 1. pK_a distribution.

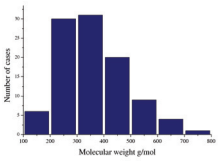


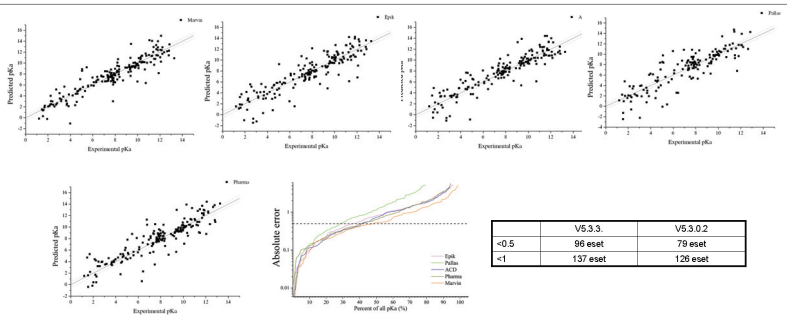
Figure 2. Molecular weight distribution.

Results

Performance of the prediction was evaluated in terms of the number of non-calculated (no predicted pK_a or 0< predicted pK_a <15) cases, correlation between predicted and experimental data, Fischer-F value (F), standard error of estimate (SEE) and mean absolute error (MAE), results are summarized in Table 2. Marvin was found to predict the highest amount of the pK_a values, while Pallas could not calculate the pK_a in 40 cases. Considering the r^2, SEE and MAE no significant difference can be observed. Marvin and ACD slightly outperformed Epik and Pharma, while Pallas gave somewhat less accurate prediction (see Figure 3.).

Table 2. Statistical parameters of pK_a predictors made by the investigated tools

	Epik	Pallas	ACD	Pharma	Marvin V5.3.3.	Marvin v5.3.0.2
r^2	0.7871	0.7551	0.8241	0.7975	0.847	0.824
F	680.0687	468.6440	848.0970	712.9395	1050.185	878.505
SEE	1.5893	1.5091	1.3251	1.3608	1.223	1.312
MAE	1.1109	1.2541	0.9951	1.0086	0.8723	0.9580
Noncalculated	8	40	9	11	2	4



	V5.3.3.	V5.3.0.2
<0.5	96 eset	79 eset
<1	137 eset	126 eset

Figure 3. Evaluation of pK_a prediction capabilities made by the investigated tools

Impact of multiple ionization constants on the reliability of the prediction was also investigated. The MAE in function of the number of ionizable groups is shown in Figure 4. No correlation between the number of ionizable groups and the MAE was observed except in case of Pallas, where the MAE increased with the increasing number of ionizable groups. Next, the effect of pK_a range on the MAE was analyzed (see Figure 5.). It can be concluded, that the used tools had their highest accuracy in the interval of pK_a 7-10.

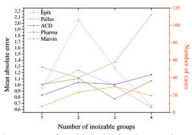


Figure 4. Impact of multiple ionization on the performance.

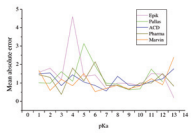


Figure 5. Mean absolute error dependence on the pK_a.

Finally, we investigated the cases, where all predictors made mistakes resulting error above 0.5 pK_a unit. These moieties are shown in Figure 6. The non-predicted cases are shown in Figure 7.

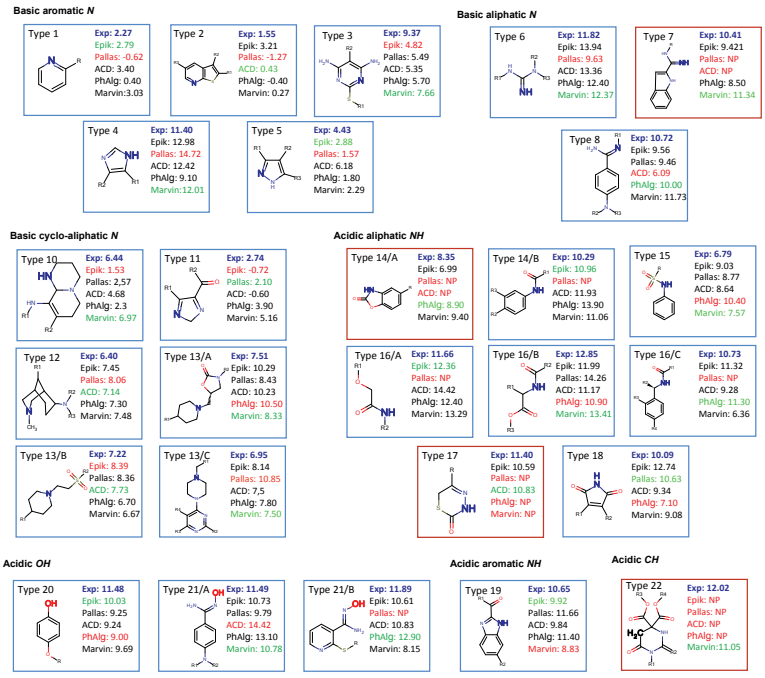


Figure 6. Common outliers identified by all the five pK_a prediction tools. (NP is non-predicted, red is the worst and green is the best predictor)

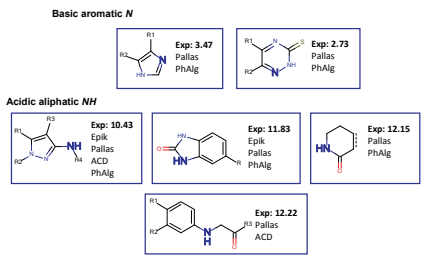


Figure 7. Further non-predicted functions by at least two pK_a prediction tools

Conclusion

The highest performance was observed in case of Marvin, while ACD, PharmaAlgoritm and Epik had moderate accuracy and Pallas proved somewhat less accurate results. It is noteworthy to state that among miscalculated cases weak acidic NH and oxime moiety were marked to be common types.

References

[1] Gy. T. Balogh, B. Gyarmati, B. Nagy, L. Molnár, Gy. M. Keserű Comparative Evaluation of in Silico Prediction Tools ont he Gold Standard Dataset. QSAR Comb. Sci. 2009, 28(10), 1148-1155.
[2] ACD/pKa DB, Advanced Chemistry Development, Inc., Toronto, Ontarion, Canada; http://acdlabs.com/products/phys_chem_lab/pka
[3] J.C. Shelley, A. Cholletti, L.L. Frye, J.R. Greenwood, M.R. Timlin, M. Uchimaya, J. Comput. Aided Mol. Des. 2007, 21, 681-691. http://www.schrodinger.com/
[4] J. Szegezdi, F. Csizmadia A Method for Calculating the pK Values of Small and Large Molecules, 233rd ACS National Meeting, CINF41 2007, Chicago, USA. http://chemaxon.com/product/pka.html
[5] F. Csizmadia, J. Szegezdi, F. Darvas, in QSAR and Modeling (Ed.: C.G. Wermuth), ESCOM, Leiden 1993, pp. 507-510; http://www.compudrug.com/
[6] ACD/pKa DB, Advanced Chemistry Development, Inc., Toronto, Ontarion, Canada; http://acdlabs.com/products/pc_admet/