Predicting $pK_a$ Using a Combination of Quantum and Machine Learning Methods

Peter Hunt, Layla Hosseini-Gerami, Tomas Chrien, Matthew Segall

Optibrium Limited, Cambridge, UK. info@optibrium.com

Introduction
The dissociation of a proton from a heteroatom has a significant impact on the charge distribution and interactions of a molecule. These influence many important molecular properties, including binding to target and off-target proteins, absorption, distribution, metabolism and excretion (ADME) and pharmacokinetic (PK) properties such as solubility, tissue or cellular distribution and permeability. Therefore, the ability to predict the propensity of a molecule to lose or gain a proton in water is crucial for the development of new chemical entities with desirable PK, ADME and binding properties.

Method
Quantum-mechanical descriptors for polarizability, bond length and charge were calculated for the (de)protonated heteroatom (X), the bound hydrogen (H) and the adjacent heavy atoms (R) (Figure 1), for both the conjugate acid and base forms, using the semi-empirical AM1 method.

A dataset of 2473 carefully curated $pK_a$ values from ChEMBL and other sources, representing 1968 unique compounds that are a mix of mono and diprotic species, was used to train and test the model. This was split into a training, validation and test sets of 1722, 377 and 374 $pK_a$ values respectively. The Auto-Modeller™ module in StarDrop [1] was used to apply a variety of machine learning methods to build models. The Radial Basis Function method produced the most predictive model.

Results
Table 1 shows the coefficient of determination ($R^2$) and root-mean-square error (RMSE) on the independent validation and test sets. These correlations are further illustrated in Figure 2.

Table 2 shows a comparison of these results with seven previously-published methods.

Conclusion
The model described herein predicts the $pK_a$ for a large range of mono- and di-protic compounds with high degree of accuracy (< 1 log unit RMSE). The model also performs excellently on the external SAMPL6 test set, specifically created to benchmark $pK_a$ prediction methods. The high level of performance on this data set is only bettered by much more computationally expensive and time consuming methods relying on ab initio density functional methods and conformational sampling.

References