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.

Figure 1. Atoms for which descriptors are generated in the QM calculations

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.

Validation R^2	Validation RMSE	Test R^2	Test RMSE
0.90	1.06	0.92	0.91

Table 1. Results for the independent validation and test sets.

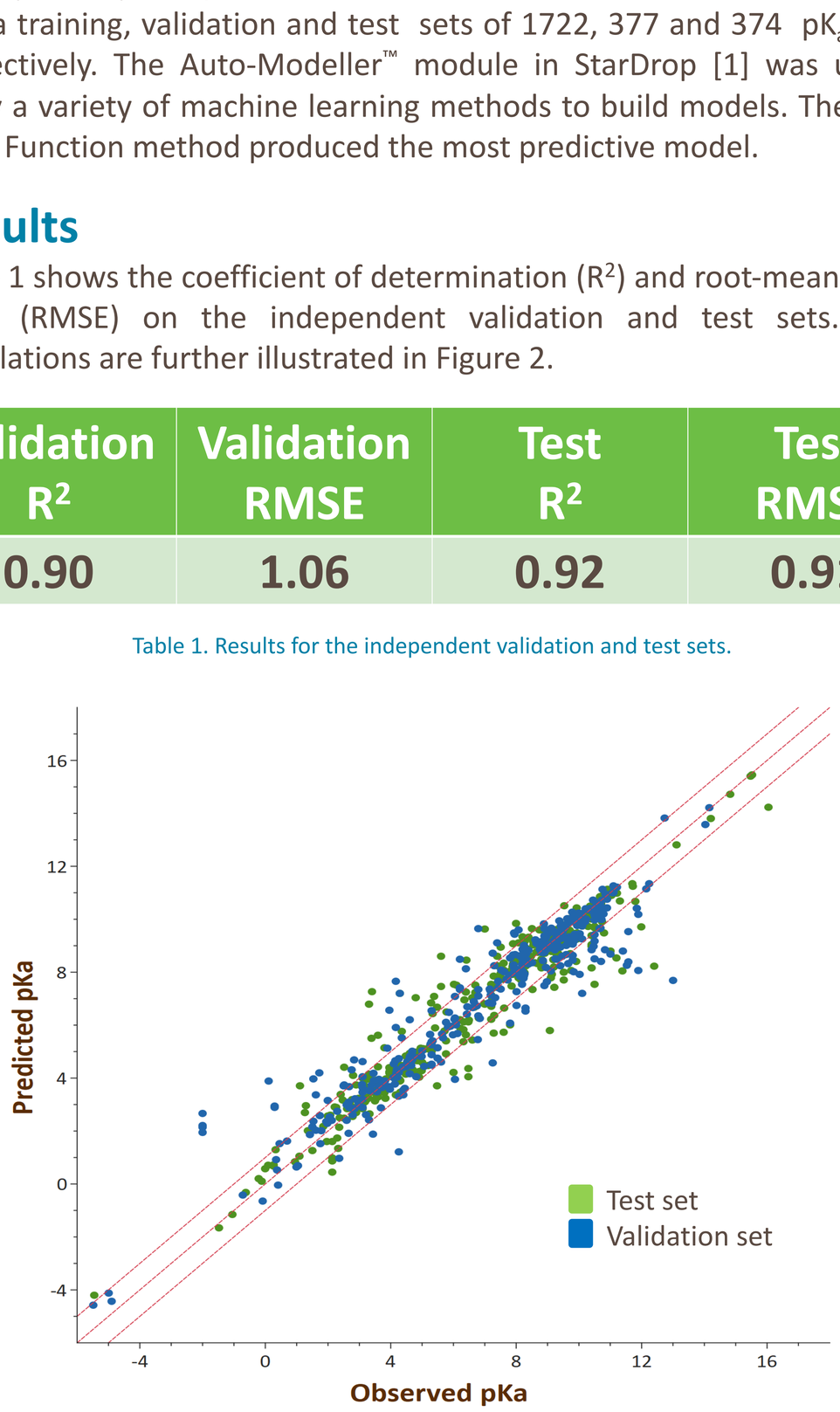

Figure 2. Predicted versus actual for the validation and test sets. The identity line and ± 1 log unit are shown as dotted red lines.

External Validation

The model was applied to the SAMPL6 data set, previously used to test pK_a prediction methods [2]. This comprises a collection of 24 kinase inhibitor-like compounds with 31 experimental pK_a values. The results are illustrated in Figure 3 with the main outliers marked (the quinazolinone outlier is the third pK_a value on this compound).

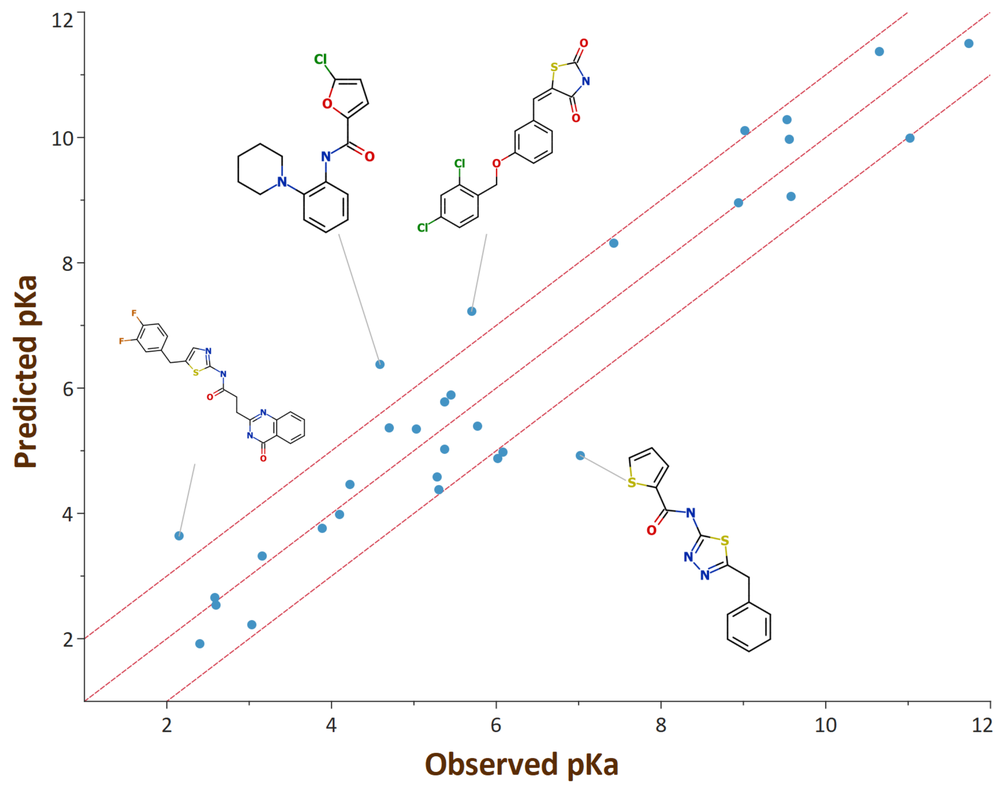

Figure 3. Predicted versus actual for the SAMPL6 set pK_a values. The identity line and ± 1 log unit are shown as dotted red lines.

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

Method	RMSE	Comments	Authors
This work	**0.85**		
Gaussian process model	2.2	reduces to 1.7 by removing an outlier SM06 – amide anion	Bannan et al,
LFER with conf. sampling and DFT	0.68	Very expensive *ab initio* QM method	Pracht et al
Hybrid QM/MM with explicit solvent	2.4	"protocol needs work"	Prasad et al
ab initio QM free energies	1.95		Selwa et al
EC-RISM	1.7	reduces to 1.5 with improved electrostatics and 1.1 with conformational sampling	Tielker et al
M06-2X DFT with SMD solvation model	1.4	falling to 0.73 with linear correction to DFT	Zeng et al

Table 2. Comparison with published pK_a prediction methods on the SAMPL6 external data set.

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

[1] StarDrop https://www.optibrium.com/stardrop/

[2] M. Isik *et al*. pK_a measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. J. Comp.-Aid. Mol. Des., (2018), **32**, 1117-38