Predicting pK_a Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods

Peter Hunt[1], Layla Hosseini-Gerami[2], Tomas Chrien[1], Jeffrey Plante[3], David J. Ponting[3], Matthew Segall[1]

1 Optibrium Ltd, F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, UK
Tel. +44 (0) 1223 815 900

2 Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK Tel. +44 (0) 1223 336 300

3 Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK Tel. +44 (0)113 394 6020

Abstract

The acid dissociation constant (pK_a) has an important influence on molecular properties crucial to compound development in synthesis, formulation and optimisation of absorption, distribution, metabolism and excretion properties. We will present a method that combines quantum mechanical calculations, at a semi-empirical level of theory, with machine learning to accurately predict pK_a for a diverse range of mono- and polyprotic compounds. The resulting model has been tested on two external data sets, one specifically used to test pK_a prediction methods (SAMPL6) and the second covering known drugs containing basic functionalities. Both sets were predicted with excellent accuracy (root-mean-square errors of 0.7 – 1.0 log units), comparable to other methodologies using much higher level of theory and computational cost.

Introduction

The dissociation of a proton from a heteroatom has a significant influence 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.

It is possible to model the free energy of the proton dissociation and gain very high accuracies using a quantum mechanical (QM) method based on density functional theory (DFT);[1,2] however these free energy relationships tend to be modelled on specific functional groups (carboxylic acids, anilines, phenols, etc.) and are applied to small data sets due to the computational cost of these high-level methods. Varekova *et al.* generated quantitative structure-activity relationship (QSAR) models on a set of phenols with partial charge descriptors for the Hydrogen, Oxygen and directly bonded Carbon atoms calculated using QM methods.[3] Their semi-empirical method using the Austin-Method 1 (AM1) Hamiltonian performed poorly, achieving a squared correlation coefficient (r^2) of only 0.45 and a root-mean-square error (RMSE) of 1.64; however the much more computationally expensive second-order Møller–Plesset (MP2) method with a 6-31G* basis set achieved an excellent r^2 of 0.97 and an RMSE of 0.40, albeit on a single class of ionizable sites. Harding and Popelier also investigated *ab initio* calculated descriptors for the prediction of pK_a for 171 phenols, this time considering bond length descriptors, with a similar r^2 of 0.92 and RMSE of 0.67.[4]

QSAR models have been built using other semi-empirical QM descriptors. Tehan *et al.* published two papers detailing a correlation with Fukui frontier molecular orbital (FMO) descriptors.[5,6] The authors looked at several different classes of protonation site, including phenols, carboxylic acids, amines and heterocycles, and used the AM1 method for all calculations. The r^2 coefficients for individual classes of site varied from 0.55 (heterocycles) to 0.94 (amines), with predictions improving with the addition of partial charge descriptors.

Several commercial vendors have developed pK_a models (ACDLabs,[7] SimulationsPlus,[8] Schrodinger,[9,10] ChemAxon,[11] Molecular Discovery[12]), which rely on linear free energy relationships (LFER) or QSAR methods and are derived from large databases of pK_a values (>10,000 data points in most cases). Liao and Nicklaus[13] compared nine different programs applied to the predicted pK_a values of 197 pharmaceutical compounds and found a range of performances (r^2 values from 0.58 to 0.94 and RMSEs from 1.8 to 0.65), where surprisingly the worst performing was a DFT-based method. However, it is very difficult to ascertain whether the more successful methods simply had more of the test compounds within their own training sets. Other reviews of pK_a tools have been published, at least in part, based on proprietary compounds Balogh *et al.* published results on their in-house dataset of 95 compounds, [14] Manchester *et al.* evaluated the predictive performance on 211 drug-like compounds,[15] Settimo *et al.* from Vertex used datasets of in-house and external sources to show how varied the performance can be between models.[16] The success of each method depended on the test dataset used and on the ability of each model to recognise potential ionisable sites. Hence, the performance varied depending on the ionisable site considered (bases tended to be less well predicted compared to acids).

Our aim was to create a model using a semi-empirical QM approach combined with machine learning. We wished to retain the physically relevant contributions to the electrostatics, provided by consideration of the whole molecular environment, while balancing this with acceptable computational cost and speed of calculation. The model we describe herein employs a single model (not specific to any ionisable group), derived from a comparatively small, but carefully curated data set of aqueous pK_a values for mono- and diprotic compounds, where the atomic assignment of pK_a values to specific heteroatoms or groups was unambiguous. This is essential for a correct characterisation of the ionised site to reduce the potential for noise in the data, where the association of a pK_a with an exact site can be lost in databases containing very large numbers of polyprotic molecules.

The next section describes the methods and data used to generate a model to predict pK_a. Following this, the Results and Discussion section provides a detailed analysis of the performance of the model on independent validation and test sets, as well as additional published benchmarking sets. Finally, the Conclusions section summarises the outcome of this study.

Methods

Data Set

A data set was collated from public data sources[17,18,19] and comprises aqueous pK_a determinations at temperatures close to 25 degrees. Each data point was carefully inspected to ensure that the assignment of the pK_a value to a specific site or group was clear and unambiguous.

The total data set contained 2,435 pK_a values, which were split into 1,722 values in the training set, 359 in the validation set and 354 in the test set. These corresponded to 2,243 unique compounds of which 1,570 were in the training set, 336 in the validation set and 337 in the test set (a split equivalent to 70%:15%:15% of the compounds created by random selection). The purpose of the validation and test sets are to provide two independent assessments of any model generated, the validation set allows the selection of the best performing methodology and the test set is a check to ensure that this selection has not been influenced by random chance. Valid models should perform equally well on both sets whilst a much poorer performance on the test set would indicate a biased model or simply a good validation result by random chance.

The distributions of pK_a values in the training, validation and test sets are shown in Figure 1. Distributions of simple compound properties for these sets are shown in Figure 2 and one can see that the molecules included are distributed more towards the smaller, less functionalised end of the property ranges. However, this is in keeping with the dataset construction strategy to include only molecules with unambiguous assignment of pK_a

values to atomic sites. Despite this, there are representative ‘drug like’ molecules included in the training, validation and test sets. Other property distributions are illustrated in Figure S1 in the supplementary information which cover the lipophilicity, three dimensionality, and aromaticity.

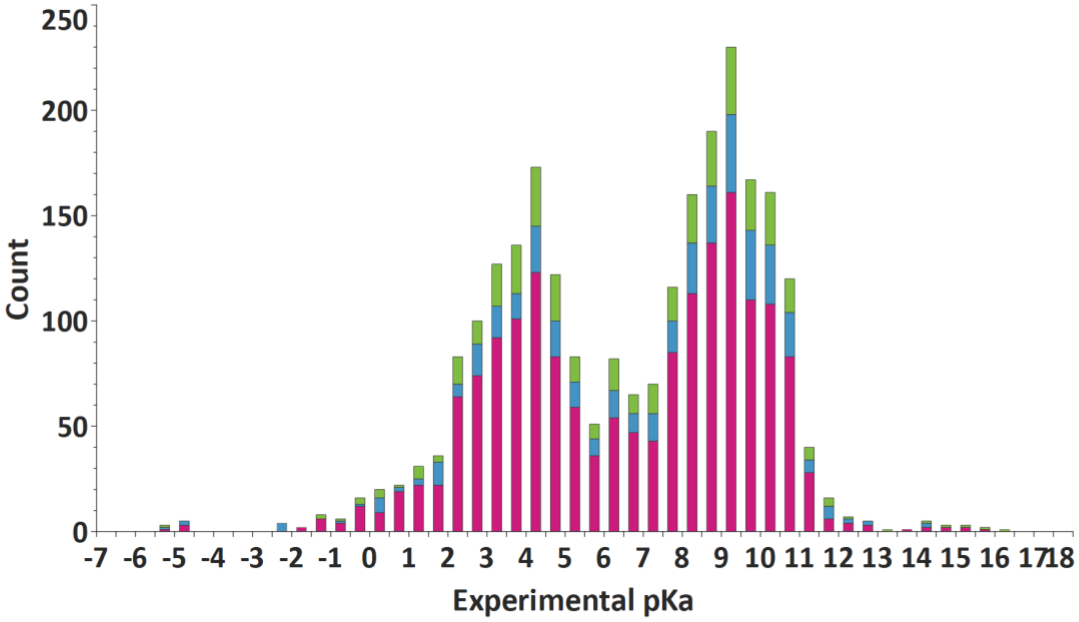
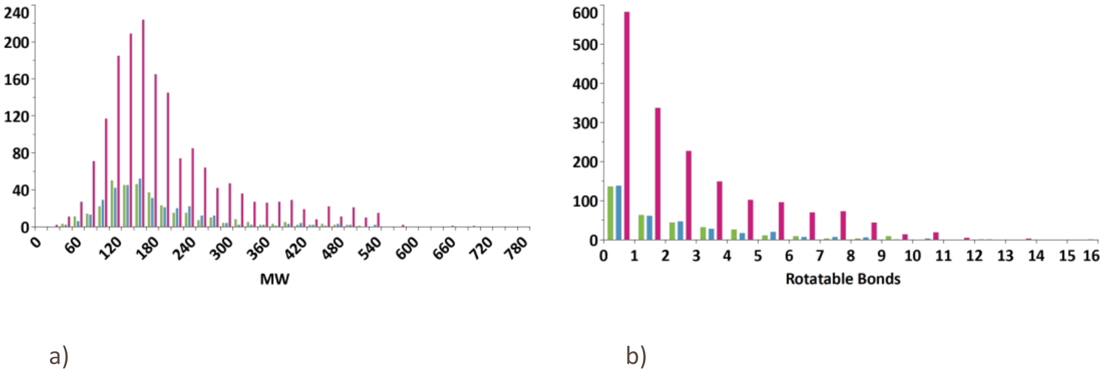
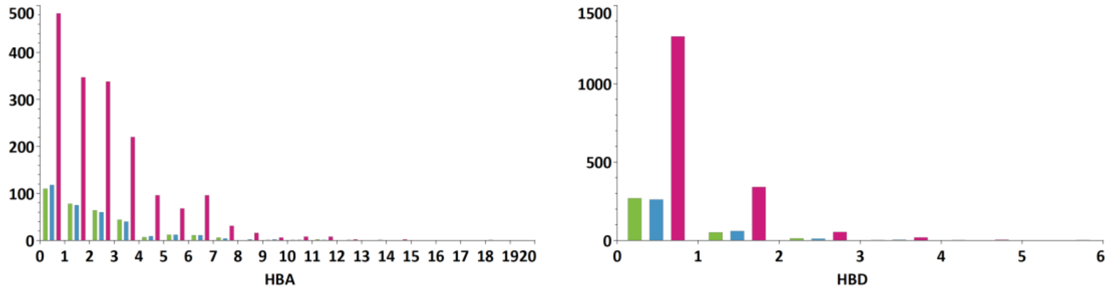


Figure 1. Plot of the distribution of the experimental pK$_a$ values across the data set. The histogram bars are coloured according to the proportion of that bar in the training (pink), validation (blue) and test (green) sets. The distributions show the relative paucity of pK$_a$ values around the 5.5 to 7.5 region as well as those at the extremities of the range.





c) d)

Figure 2. Plots of the distributions of simple properties across the training (pink bars), validation (blue bars) and test (green bars) sets. a) Molecular weight, b) Rotatable bonds, c) Hydrogen bond acceptors, d) Hydrogen bond donors.

Figure 3 shows the distribution of the training, validation and test sets relative to the 'chemical space' of launched small molecule drugs. This shows good coverage of the diversity of drug-like compounds, with the exception of steroids (the cluster in the top right) and large macrocyclic and linear peptides (the bottom right region). Further analysis of the coverage of the data set relative to launched small molecule drugs, in the space of the descriptors used in the modelling process, is provided in Figure S2 in the supplementary information.

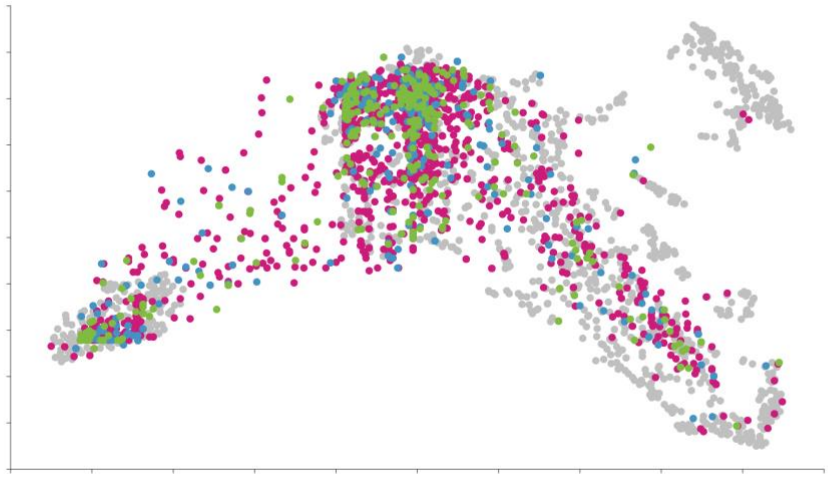


Figure 3. A chemical space representation of approximately 1300 launched small molecule drugs (grey points) with the compounds in the pK_a data set overlaid. The training (pink), validation (blue) and test (green) sets are shown. This illustrates that the data set used herein covers a large majority of the small molecule drug space. In this plot the proximity of two points represents the structural similarity between the corresponding compounds defined using a Tanimoto index based on a 2-dimensional path-based fingerprint. The distribution of points is generated using the t-distributed stochastic neighbour embedding algorithm.[20]

The full data set, including compound structures, set membership, experimental and predicted data are included in the supplementary information.

Descriptor Calculation

A single 3-dimensional conformation was generated for each compound using OMEGA.[21] If stereochemistry was defined in the SMILES then this was maintained in the 3D structure, but if no stereochemistry was defined then a single stereochemical configuration was taken from OMEGA. We have found during the course of this research that intramolecular H-bonding groups can promote H-transfers during the optimisation steps (especially with the ionised sites) so these groups are, where possible, rotated away from the site of ionisation prior to optimisation. Any proton transfer would make the QM descriptors for that system inconsistent with the descriptors generated for systems where this transfer does not take place, hence for consistency we alter the OMEGA torsions for some substructures prior to optimisation. A list of the SMARTs patterns we use to identify these potential problematic sites and torsion modifications applied can be found in TABLE T1 in the supplementary information.

The parent and subsequent structures were optimised using the AM1 Hamiltonian in MOPAC[22] without solvent corrections, using the following keywords "AM1 EF PRECISE MMOK VECTORS T=3600 CYCLES=1000 EIG-EXOUT GEO-OK CHARGE=*" where the charge keyword value '*' was set appropriately for the compound charge state.

Hydrogen atoms were then added or removed from the parent structure to create the appropriate conjugate acid or conjugate base and the Hydrogen atom positions were adjusted as required using OMEGA. This newly ionised structure was again optimised through MOPAC and the descriptor calculation repeated. The descriptors used capture atomic and bond properties, describing the Hydrogen, heteroatom (X) and all of the heavy atom neighbours (R) of the heteroatom in the conjugate acid and base forms for the ionisable site, as illustrated in Figure 4. Nucleophilic and electrophilic delocalisabilities (matching those obtained when the "SUPER" keyword[22] is used in recent MOPAC versions), bond lengths and atom charge descriptors were calculated for the heteroatom and nearest heavy-atom neighbours. Each of the descriptors calculated for the individual heavy atoms, R, were averaged across the n heavy atoms to produce a single value for that descriptor. The HOMO and LUMO energies and heats of formation of the conjugate acid and base were also used as descriptors.

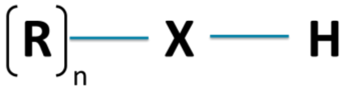


Figure 4. The atoms considered in the parameterisation of the model where 'n' refers to all the directly bonded heavy atoms of the heteroatom X.

This process was repeated for each site considered as likely to be ionised in a reasonable pK_a range. Currently alkyl or unsubstituted amide and sulphonamide NH groups and Carbons are excluded; however, these could be introduced by the inclusion of accurate pK_a determinations to the training set. For polyprotic compounds, each site was protonated or deprotonated individually and the most basic and/or most acidic site was identified. A second round of calculations was then performed where this primary site was ionised and the other sites were examined to determine the second-most basic/acidic site; if the molecule was zwitterionic, the process was repeated with the other acidic/basic site charged as this enables a better prediction for the most acidic/basic sites. Once these two most acidic/basic sites were determined, a final round of calculations were performed for any other isonisable site, with these two sites in their appropriate charge states. These heuristics provide a way of capturing the effect of ionisation of the most acidic and basic sites in a molecule on the estimation of the acidity/basicity of subsequent sites.

We are not attempting to explicitly calculate specific micro pK_a states of a molecule, with appropriate Boltzmann-weighted contributions for each state, as this extends the calculation time and provided little gain in accuracy. Our aim, with these heuristics, is to produce a set of predictions from a simple, single QSAR model, that are more realistic evaluations of the charge state of a molecule in water and, as such, the ease or otherwise of introducing a second or subsequent charged site into a system that already bears a charge (or charges) should be considered and predicted appropriately.

Model Generation

Models were generated and tested using the data set and descriptors described above, employing four methods implemented in the Auto-Modeller™ module of the StarDrop™ software[23]: a linear method, partial least square (PLS);[24] radial basis functions (RBF);[25,26] random forests (RF);[27] and Gaussian processes (GP) with conjugate gradient optimisation of hyperparameters.[28]

The best model identified in the study, by the use of an independent validation set, is based on the RBF method. RBFs have been praised for their simplicity, robustness and ease of implementation in multivariate scattered data approximation. Such techniques have been applied with success in problems ranging from training neural networks to image compression and broadening to other disciplines such as engineering.[25,26] A radial basis function is a non-linear transfer function which aims to approximate a real valued function $y(\underline{x})$ by $\Psi(\underline{x})$, given the set of sample values $Y = \{y_1, \dots, y_N\}$ at the points $X = \{\underline{x}_1, \dots, \underline{x}_N\}$. To achieve this, we choose $\Psi(\underline{x})$ to be of the form

$$\Psi(\underline{x}) = \sum_{i=1}^{N} a_i \phi(\|\underline{x} - \underline{x}_i\|),$$

where a_i is a real valued weight, ϕ is a basic function and $\|\cdot\|$ denotes the Euclidian distance metric. Vectors $\underline{x}_i$ represent *N* points where the radial basis functions are centred and represent the descriptor vectors for the *N* compounds in the training set. A variety of basis function types may be used, including linear, cubic, multi-quadratic and Gaussian. Based on previous experiments, the model herein uses a linear function.

In fitting the RBF to the data, the conventional method is to require that $\Psi(\underline{x})$ pass through all the training data points, which gives the following linear system of equations:

$$y_j = \sum_{i=1}^{N} a_i \phi(\|\underline{x}_j - \underline{x}_i\|), \qquad j = 1,\dots,N.$$

This linear system can be solved for the weights $a_i, (i = 1 \dots N)$. Then the fitted function $\Psi(\underline{x})$ can be used for predicting new data points.

This RBF method can be considered as a generalisation of a *k*-nearest-neighbour approach to use all N compounds in the training set, where the contribution of each training point is weighted by its Euclidean distance to a test point. We have found it to be particularly effective where the relevant descriptor space is densely populated by training points. However, the RBF method does not extrapolate well beyond these regions, so it is important to include an assessment of the domain of applicability of the model. Furthermore, because this function is guaranteed to pass through the training points, it is essential to validate the resulting model with an independent test set to ensure that it generalises to new compounds and is not over trained.

Results and Discussion

The results for the four methods applied to the independent validation and test sets are summarised in *Table 1*, which shows the coefficient of determination (R^2), RMSE, Mean Absolute Deviation (MAD), the maximum error, and the total number of pK_a predictions which exceeded cutoffs at either 1 or 2, 3, 4 or 5 log unit deviations for each model.

The predictions from the different models, in particularly the RBF, GP and RF models were strongly correlated and there were no consistent patterns in outliers or mispredictions from any of the models (see Figure S3 in the supplementary information). Therefore, the best model was determined to be that derived using the RBF methodology and the results on the independent validation and test sets are shown in Figure 5. The remaining analyses, described below, relate to this model. Please note that in cases where multiple potential sites of ionisation are present in the same functionality and a clear assignment cannot be made, e.g. a piperazine, both were included in the validation and test sets and hence the associated error represents an average between these to give a fair assessment. However, including only the closest agreement between the model predictions and the experimental value only reduced the RMSE values by around 0.02 in the test set and 0.07 in the validation set.

Table 1. The performance of the models for the independent validation and test sets. The coefficient of determination (R^2) root-mean-square error (RMSE), mean absolute deviation (MAD), maximum deviation and the distribution of erroneous predictions above 1,2,3,4, or 5 log units are shown for each set. GP is Gaussian processes, RBF is radial basis function, RF is random forests, PLS is partial least squares

Method	Validation Set					Test Set				
	R^2	RMSE	MAD	Max Error	Error distribution	R^2	RMSE	MAD	Max Error	Error distribution
RBF	0.90	1.05	0.65	5.3	80,27,11,4,1	0.92	0.91	0.59	4.2	67,20,6,1,0
GP	0.86	1.28	0.84	7.4	111,40,15,6,2	0.85	1.21	0.80	6.2	96,38,12,8,4
RF	0.87	1.25	0.82	5.2	95,33,17,7,2	0.88	1.12	0.74	5.6	95,35,13,4,1
PLS	0.66	2.0	1.43	9.9	183,92,39,20,11	0.69	1.81	1.32	8.2	183,80,36,21,5

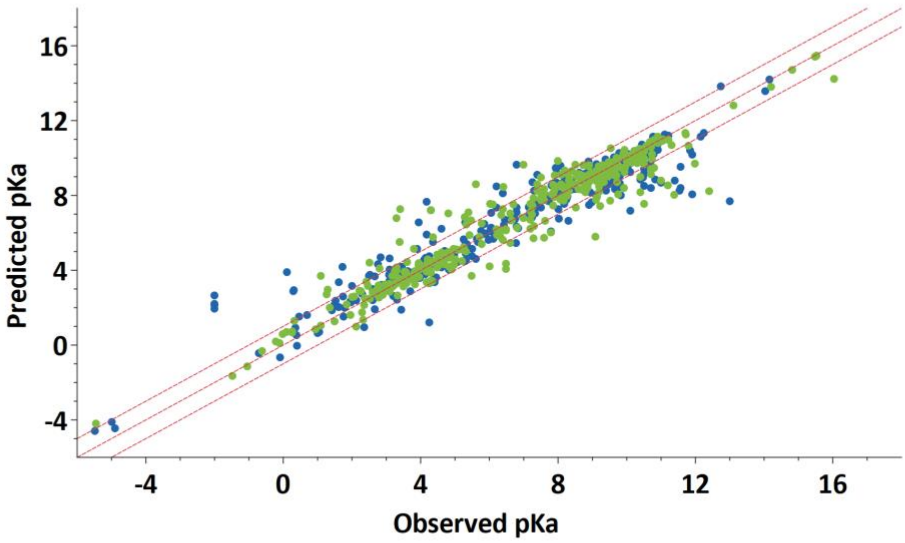


Figure 5. A plot of predicted versus observed pK_a values for the validation (blue points) and test (green points) sets using the RBF model. The identity line and lines deviating by ±1 log unit are shown as red dotted lines.

An alternative view of the model performance is to analyse the error distribution, as RMSE values can be influenced by a small number of large outliers. For the validation and test sets 34% and 32% of the pK_a values are predicted within 0.2 log units; this rises to 68% for the validation and 69% for the test set within 0.6 log units and 79% and 83% within 1 log unit respectively (shown in Figure 6) 90% of the validation set has a deviation below 1.6 whilst for the test set this deviation is only 1.46 log units.

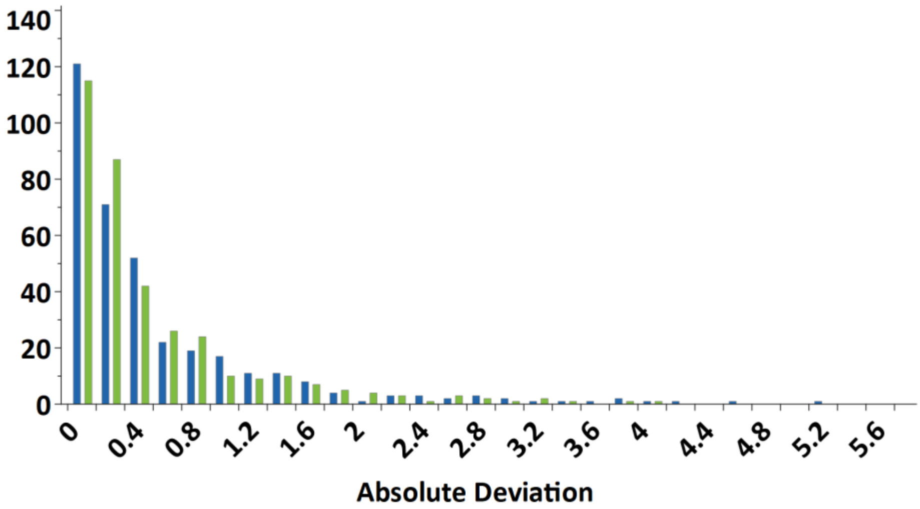


Figure 6. A plot of absolute deviations for the validation (blue bars) and test (green bars) sets using the RBF model. The largest deviations were found to occur for some of the extreme measured pK_a values

The distribution of deviations is very similar for those experimental pK_a values that are either above 7, or 7 and below, as shown in Figure 7, although it is interesting to see the higher proportion of small deviations in the predictions of the test set relative to the validation set for those in the $pK_a \leq 7$ group .

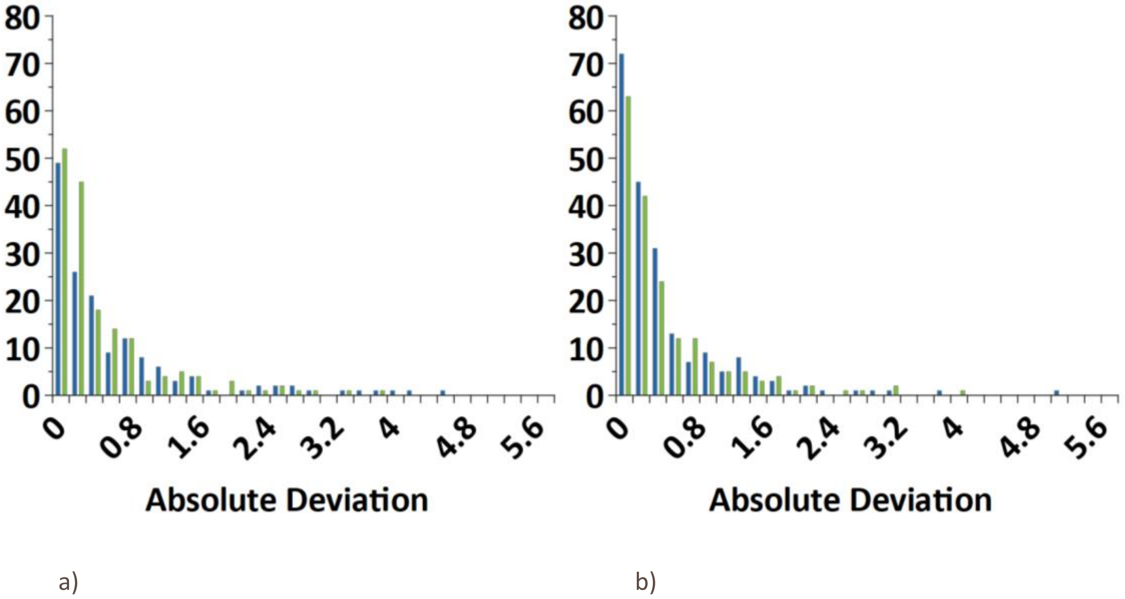


Figure 7. A plot of absolute deviations for the validation (blue bars) and test (green bars) sets using the RBF model for experimental pK_a values that are either above 7, or 7 and below; (a) is $pK_a \leq 7$, b) is $pK_a > 7$)

A more detailed analysis of the distributions of the absolute deviations in prediction relative to the experimental pK_a is shown in Figure 8. This shows that the extreme pK_a values are amongst the poorest predictions, but there are a small number of outlier predictions across the pK_a range. The median of the absolute deviations dip in the

pK_a ranges 8-10 and 3.5-5, which coincide with the maxima in the distribution of the experimental pK_a values in the data set, as shown in Figure 1. This is as expected because of the better representation of these pK_a values in the training set and is desirable because it suggests that the model will be more accurate for the most relevant ranges of pK_a.

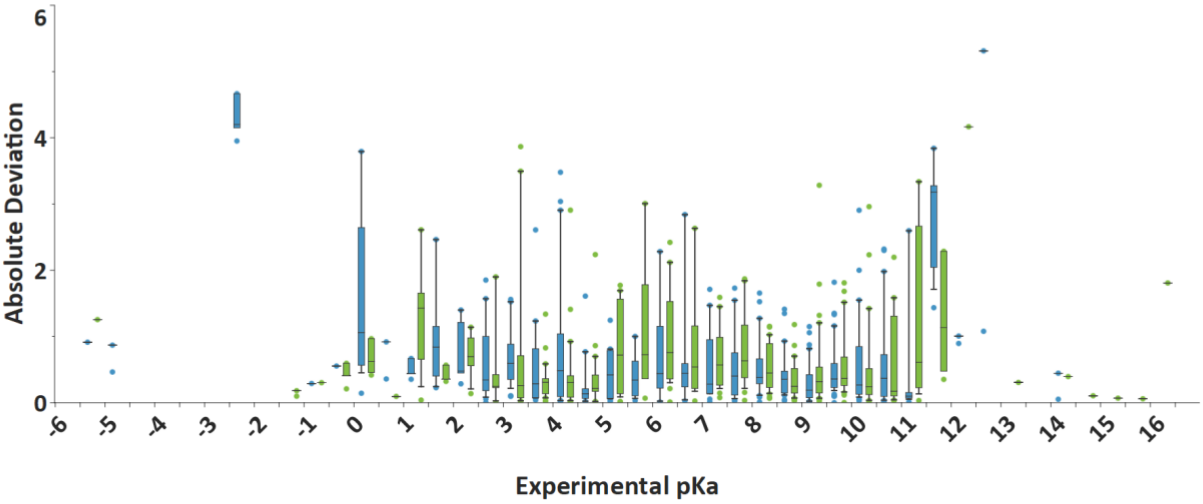
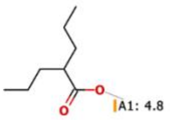


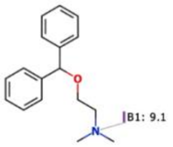
Figure 1. A box plot showing distributions of absolute deviations of the RBF model predictions as they are distributed across the range of experimental pK_a values. The boxes are coloured by the data set, with the validation set in blue and the test set in green. Attention should be drawn between the shape of this distribution and the frequency distribution of experimental pK_a values shown in Figure 1.

The error distributions were also plotted on the chemical space illustrated in Figure 3 and a related space derived from the property matrix used to generate the model, however no notable biases in terms of chemical structure or atomic property could be identified. This is illustrated in Figure S4 in the supplementary information.

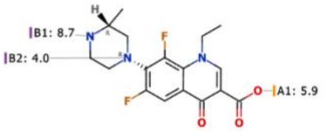
The speed of the predictions is of interest in the application of this methodology to drug design. The submission of the validation and test sets to the model, without prior calculation of the descriptors, resulted in average calculation times per molecule of 17 and 20 seconds respectively on a single threaded CPU (Intel® i7-8550U @ 1.8GHz). Parallelisation of this methodology is eminently feasible and would enable even large collections of molecules to be predicted in a reasonable timeframe. A few examples of pharmaceutical or agrochemical marketed compounds from the validation set are shown below in Figure 9 along with the approximate time taken to predict all the various potential ionisable sites within each molecule, starting from the SMILES string.



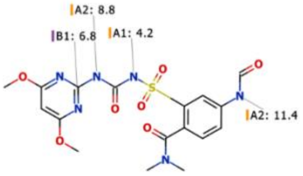
VALPROIC ACID – expt. pKa = 4.6
Time taken 2.5 seconds



DIPHENHYDRAMINE – expt. pKa = 9.1
Time taken 17 seconds



LOMEFLOXACIN – expt. pKas = 5.82, & 8.87
Time taken 52 seconds



FORAMSULFURON – expt. pKa = 4.6
Time taken 238 seconds

Figure 9. A selection of drugs and agrochemicals and the times taken to predict all the perceived ionisable sites using the RBF model within the StarDrop™ application[23]

The RBF method does not naturally produce a measure of descriptor importance. However, examination of the GP model indicates that the average of the R-X bond lengths, the X atom charge in the conjugate base and acid forms, the X-H bond length in the conjugate acid, and the nucleophilic delocalisabilities of the X atom in the conjugate acid and the average value for this descriptor for the R atoms in the conjugate base form, are important in that model (see Figure 2 for the relationship of the R, X atoms at the ionisable site). The PLS model also shows similar descriptors with high absolute coefficients indicating a consistent pattern between models.

Comparison on Benchmarking Data Sets

This model was further validated by testing against two published data sets. The SAMPL6 data set comprises 24 kinase inhibitor-like molecules with 31 experimental pK$_a$ values and was specifically devised to test pK$_a$ prediction methods.[29] The second was a set of 48 amine-containing drug molecules (53 pK$_a$ values) devised by Jensen *et al*.[30], which had been used to evaluate semi-empirical pK$_a$ prediction methodologies, similar to that used in this paper.

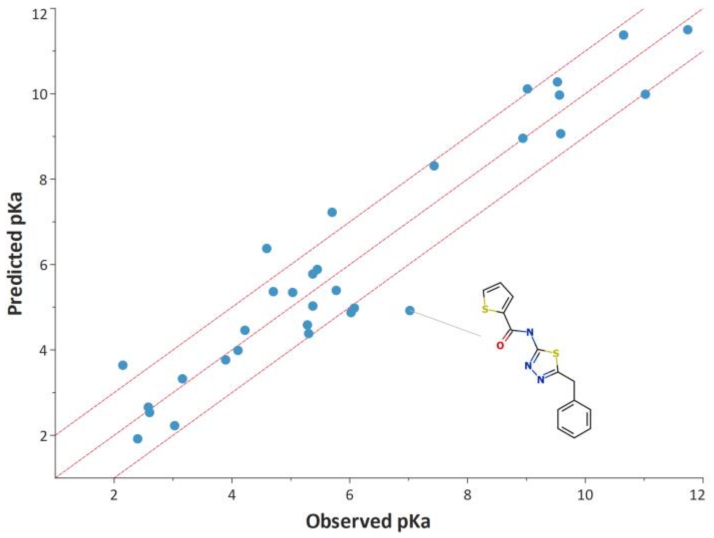


Figure 10. Plot of predicted versus observed pK$_a$ values for the SAMPL6 data set using the RBF model. The identity line and lines deviating by ±1 log unit are shown as red dotted lines, the compound with the largest misprediction is shown.

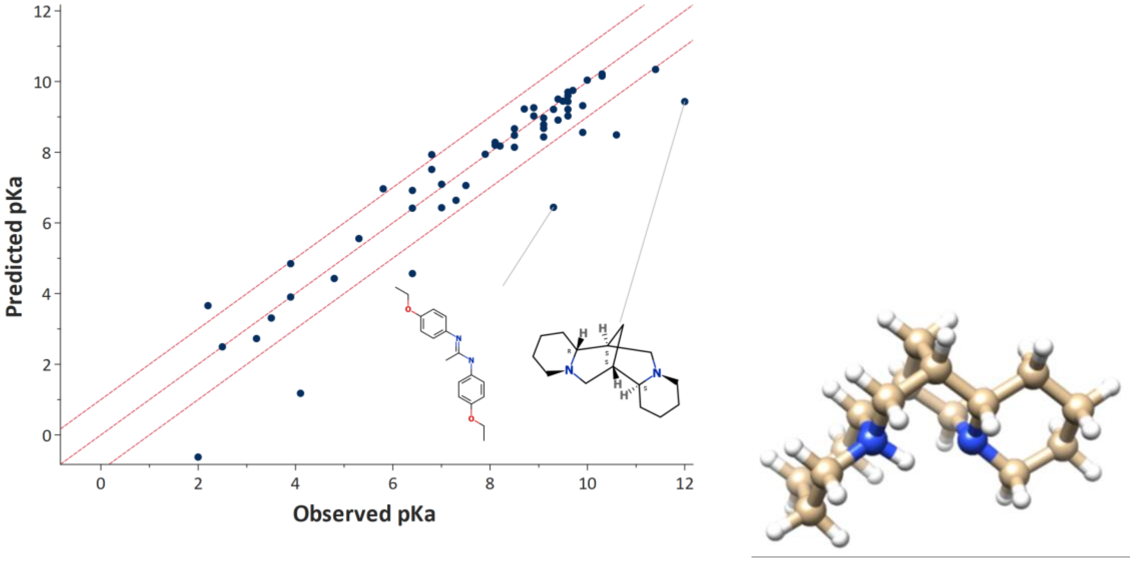
The RMSE on the whole SAMPL6 set was 0.85 and the predicted versus observed graph is shown in Figure 10. This compares very favourably with the performances of the other methods published in the same journal issue as the challenge data set in reference 28, shown in *Table 2*, where the authors noted that some outliers were removed due to sites not being thought of as ionisable or extra modifications were required such as conformational sampling. Further submissions to the SAMPL6 challenge are summarised on GitHub[31,32] and examination of the RMSE values quoted for the 38 submissions listed shows that our model is 4th in the list, which we believe to be notable given the poorer performances from many more complex methodologies. None of the compounds from this data set were included in the training or validation sets for the models described herein, so this represents a fair comparison.

Table 2. Summary of the performance of published methods on the SAMPL6 benchmarking data set. The method, root-mean-square error (RMSE) and comments made by the authors in the corresponding references are shown.

Author	Method	RMSE	Comments
Bannan *et al.*[33]	Gaussian process model	2.2	reducing to 1.7 by removing an outlier SM06 – an amide anion
Pracht *et al.*[2]	LFER with conf. sampling and DFT	0.68	
Prasad *et al.*[34]	Hybrid QM/MM with explicit solvent	2.4	“protocol needs work”
Selwa *et al.*[35]	*ab initio* QM free energies	1.95	
Tielker *et al.*[36]	EC-RISM	1.7	reducing to 1.5 with improved electrostatics and 1.1 with conf. sampling
Zeng *et al.*[37]	M06-2X DFT with SMD solvation model	1.4	falling to 0.73 with linear correction

The results obtained for the set published by Jensen *et al.*[30] are complicated by the presence of some of the compounds from this set in the training set of our model. However, the RMSE for the prediction of all of the pK_a values within the set is 0.98, which rises only slightly to 1.05 for the 45 pK_a values not included in our training set. These results are equivalent to those obtained by Jensen *et al.* with solvent corrected values using COSMO and either the PM3 or AM1 semi-empirical methods, but *only* when the zwitterion Cefadroxil is excluded from their analysis.

A graph of predicted versus observed pK_a is shown for the Jensen *et al.*[30] data set in Figure 11(a), in which some of the outlier compounds are highlighted. These compounds indicate where our methodology might be improved either by additional data, or by the higher computational cost incurred by conformational sampling. The amidine Phenacaine highlights the relative paucity of acyclic amidines in our training set, whilst the Sparteine structure highlights the stabilisation of a protonated structure through internal Hydrogen bonding. Our procedure essentially describes the equilibrium between the Hydrogen on the ionisable compound and water, whereas the environment of the Hydrogen in Sparteine is influenced by the other Nitrogen within the Sparteine molecule (as shown in Figure 11(b)). Hence the specific Nitrogen-Hydrogen environment evaluated in our calculations gives rise to a less basic prediction than observed experimentally, as the internal H-bond is not captured by molecules in our training set and we only consider one conformation of the training and test molecules. The other outliers at the bottom left are the second pK_a (i.e. the di-protonated pK_a) values quoted for Hydroquinine and Procaine respectively; the monoprotonated values are predicted within 1 log unit error.



a) b)

Figure 11. (a) Plot of predicted versus observed pK$_a$ for the data set published by Jensen et al[30] using the RBF model. The identity line and lines deviating by ±1 log unit are shown as red dotted lines. Two main outliers, Sparteine and Phenacaine are highlighted the other main outliers are secondary rather than primary pk$_a$ values. (b) a 3-dimensional conformation of Sparteine illustrating the interaction between the proton and two Nitrogens that is not accurately captured by the model, giving rise to an inaccurate prediction.

Conclusion

We have described a method combining semi-empirical QM and machine learning methodologies for the prediction of pK$_a$ for both mono- and poly-protic species. This gives rise to a single model that can be applied reliably across acidic and basic functionalities. The model performs as well as more computationally intensive methods on two published test sets, which were devised to specifically benchmark pK$_a$ prediction methods and cover important areas of drug-like molecules. Our methodology considers the effects of the whole molecule on the immediate vicinity of the ionisable site and therefore incorporates more information about the molecular environment than is considered in simple isolated fragment-based QSAR methodologies. The speed of our method is still more than acceptable for the calculation of many hundreds of molecules encountered in a drug design or agrochemical optimisation process.

Supporting Information Available: A word file containing additional figures relating to the property and error distributions within the data set and a table of torsional modifications prior to initial geometry optimisation. A zip file of csv format files which contain the compound SMILES strings, experimental pK$_a$ data and predicted pK$_a$ data from the RBF model for the training, validation, and test sets used in the model building.

Conflict of interest statement

The authors Peter Hunt, Layla Hosseini-Gerami, Tomas Chrien, and Matthew Segall are all current or former employees of the company Optibrium Ltd in whose StarDrop software the model described herein has been implemented.

Funding sources

No funding sources outside of employment by Optibrium Limited or Lhasa Limited were used in this work.

References

1. Zhang, S.; Baker, J.; Pulay, P. A Reliable and Efficient First Principles-Based Method for Predicting pK_a Values. 1. Methodology, *J. Phys. Chem. A* **2010,** *114*, pp. 425-431.

2. Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pK_a values in the context of the SAMPL6 challenge *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1139–1149.

3. Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.; Skřehota, O.; Kudera, M.; Sehnal, D.; Bouchal, T.; Abagyan, R.; Huber, H.; Koča, J. Predicting pK_a Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes *J. Chem. Inf. Model.* **2011,** *51*, 1795-1806.

4. Harding, A.; Popelier, P. pK_a Prediction from an ab initio bond length: part 2—phenols *Phys. Chem. Chem. Phys.* **2011,** *13*, 11264.

5. Tehan, B.; Lloyd, E.; Wong, M.; Pitt, W.; Montana, J.; Manallack, D.; Gancia, E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids *Quantitative Structure-Activity Relationships*, **2002,** *21*, 457-472.

6. Tehan, B.; Lloyd, E.; Wong, M.; Pitt, W.; Manallack, D.; Gancia, E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds *Quantitative Structure-Activity Relationships*, **2002,** *21*, 473-485.

7. ACD/pKa :: Predict accurate acid/base dissociation constants from structure :: ACD/Labs Percepta Predictors, *Acdlabs.com*, 2017. [Online]. Available: http://www.acdlabs.com/products/percepta/predictors/pka/ Last accessed: 10th Oct. 2019.

8. Fraczkiewicz, R.; Lobell, M.; Göller, A.; Krenz, U.; Schoenneis, R.; Clark, R.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico pK_a Prediction *J. Chem. Inf. and Model.* **2015,** *55*, 389-397. [Online]. Available: https://www.simulations-plus.com/software/admetpredictor Last accessed: 10th Oct. 2019.

9. Shelley, J.; Cholleti, A.; Frye, L.; Greenwood, J.; Timlin, M.; Uchimaya, M. Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules *J. Comp.-Aid. Mol. Des.* **2007,** *21*, 681-691.

10. Klicić, J.; Friesner, R.; Liu, S.; Guida, W. Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods *J. Phys. Chem. A* **2002,** *106*, 1327-1335.

11. pKalc. [Online]. Available: http://www.compudrug.com/pkalc. Last accessed 10th Oct. 2019.

12. Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields *J. Chem. Inf. Model.* **2007,** *47*, 2172-2181. MoKa from Molecular Discovery https://www.moldiscovery.com/software/moka/ last accessed 10th Oct. 2019.

13. Liao, C.; Nicklaus, M. Comparison of Nine Programs Predicting pK_a Values of Pharmaceutical Substances *J. Chem. Inf. Model.* **2009,** *49*, 2801-2812.

14. Balogh, G. T.; Tarcsay, A.; Keseru, G. M. Comparative evaluation of pK_a prediction tools on a drug discovery dataset, *J. Pharm. Biomed. Anal.* **2012,** *67-68,* 63-70.

15. Manchester, J.; Walkup, G.; Rivin, O.; You, Z. Evaluation of pK_a Estimation Methods on 211 Druglike Compounds *J. Chem. Inf. Model.* **2010,** *50*, 565-571.

16. Settimo, L.; Bellman, K.; Knegtel, R. Comparison of the Accuracy of Experimental and Predicted pK_a Values of Basic and Acidic Compounds *Pharm. Res.* **2013,** *31*, 1082-1095.

17. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** *40*, D1100-7.

18. Perrin, D. D. Dissociation constants of organic bases in aqueous solution London : Butterworths, 1965 and 1972

19. Kortüm, G.; Vogel, W.; Andrussow, K. Dissociation constants of organic acids in aqueous solution London : Butterworths, 1961.

20. Van der Maaten, L.; Hinton, G. Visualizing High-Dimensional Data Using t-SNE *Journal of Machine Learning Research* **2008***,* 9, 2579-2605.

21. Hawkins, P.C.D.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database, *J. Chem. Inf. Model*. **2010**, *50*, 572-584. https://www.eyesopen.com/omega last accessed 10th Oct. 2019.

22. Stewart, J. J. P. MOPAC: A General Molecular Orbital Package, *Quant. Chem. Prog. Exch*., **1990,** *10*, 86. SUPER keyword http://openmopac.net/manual/super.html last accessed 26th Mar. 2020.

23. StarDrop https://www.optibrium.com/stardrop last accessed 10th Oct. 2019.

24. Wold, S.; Sjöström, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. Encyclopedia of Computational Chemistry; Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P.; Schaefer, H. F. III; Schreiner, P. R. Eds.; Wiley: Chichester, U. K., **1998,** *3*, 2006-2022.

25. Radial Basis Functions Buhmann, M. D., Cambridge University Press, 2003 ISBN: 9780511543241, https://doi.org/10.1017/CBO9780511543241

26. Fast radial basis functions for engineering applications Biancolini, M. E., Springer International, 2018 ISBN: 9783319750095, https://www.worldcat.org/title/fast-radial-basis-functions-for-engineering-applications/oclc/1017892596?referer=di&ht=edition

27. Breiman, L. Random Forests, *Machine Learning*, **2001,** *45*, 5-32.

28. Obrezanova, O.; Csanyi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modelling of ADME Properties *J. Chem. Inf. Model*. **2007,** *47*, 1847-1857.

29. Isik, M.; Levorse, D.; Rustenburg, A. S.; Ndukwe, I. E.; Wang, H.; Wang, X.; Reibarkh, M.; Martin, G. E.; Makarov, A. A.; Mobley, D. L.; Rhodes, T.; Chodera, J. D. pK_a measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1117-1138.

30. Jensen, J. H.; Swain, C. J.; Olsen, L. Prediction of pK_a values for druglike molecules using semiempirical quantum chemical methods *J. Phys. Chem. A* **2017,** *121*, 699-707.

31. https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/pKa/analysis/analysis_of_typeIII_predictions/analysis_outputs_closest/statistics.pdf, last accessed 22nd Jan. 2020

32. https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/pKa/analysis/analysis_of_typeIII_predictions/analysis_outputs_hungarian/statistics.pdf, last accessed 22nd Jan. 2020

33. Bannan, C. C.; Mobley, D. L.; Skillman, A. G. SAMPL6 challenge results from pK_a predictions based on a general Gaussian process model *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1165–1177.

34. Prasad, S.; Huang, J.; Zeng, Q.; Brooks, B. R. An explicit-solvent hybrid QM and MM approach for predicting pK_a of small molecules in SAMPL6 challenge *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1191–1201.

35. Selwa, E.; Kenney, I. M.; Beckstein, O.; Iorga, B. I. SAMPL6: calculation of macroscopic pK_a values from *ab initio* quantum mechanical free energies *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1203–1216.

36. Tielker, N.; Eberlein, L.; Güssregen, S.; Kast, S. M. The SAMPL6 challenge on predicting aqueous pK_a values from ECRISM theory *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1151–1163.

37. Zeng, Q.; Jones, M. R.; Brooks, B. R. Absolute and relative pK_a predictions via a DFT approach applied to the SAMPL6 blind challenge *J. Comp.-Aid. Mol. Des*. **2018,** *32*, 1179–1189