PCCP

View Article Online / Journal Homepage / Table of Contents for this issue

Dynamic Article Links

Cite this: *Phys. Chem. Chem. Phys.*, 2011, **13**, 11264–11282

www.rsc.org/pccp

PAPER

pK_a Prediction from an *ab initio* bond length: part 2—phenols†

A. P. Harding[ab] and P. L. A. Popelier*[ab]

Received 12th February 2011, Accepted 15th April 2011
DOI: 10.1039/c1cp20379g

The prediction of pK_a continues to attract much attention with ongoing investigations into new ways to predict pK_a accurately, where predicted pK_a values deviate less than 0.50 log units from experiment. We show that a single descriptor, *i.e.* an *ab initio* bond length, can predict pK_a. The emphasis was placed on model simplicity and a demonstration that more accurate predictions emerge from single-bond-length models. A data set of 171 phenols was studied. The carbon-oxygen bond length, connecting the OH to the phenyl ring, consistently provided accurate predictions. The pK_a of *meta*- and *para*-substituted phenols is predicted here by a single-bond-length model within 0.50 log units. However, accurate prediction of the pK_a of *ortho*-substituted phenols necessitated their splitting into groups called *high-correlation subsets* in which the pK_a of the compounds strongly correlated with a single bond-length. The highly compound-specific single-bond-length models produced better predictions than models constructed with more compounds and more bond lengths. Outliers were easily identified using single-bond-length models and in most cases we were able to determine the reason for the outlier discrepancy. Furthermore, the single-bond-length models showed better cross-validation statistics than the PLS models constructed using more than one bond length. For all of the single-bond-length models, RMSEE was less than 0.50. For the majority of the models, RMSEP was less than 0.50. The results support the use of multiple high-correlation subsets and a single bond-length to predict pK_a. Six one-term linear equations are listed as a starting point for the construction of a more comprehensive list covering a larger variety of compound classes.

1. Introduction

One of the most important physiochemical properties of small molecules and macromolecules are the dissociation constants, generally expressed as pK_a. Pharmacokinetic properties can be strongly affected by pK_a, which therefore influences a compound's absorption, distribution, metabolism, excretion and toxicity (ADMET) profile. A drug generally has to pass through at least one biomembrane *via* passive diffusion, or by carrier-mediated uptake, before it can produce any biological effect. Neutral molecules are easily absorbed by phospholipid membranes while the lipid bilayers in the cell walls have very low permeability for ions and most polar molecules. Ionisable groups also affect the ability of molecules to interact with biological targets. Therefore, pK_a can be important in determining the rate and site of metabolism. It is estimated that ninety-five percent of medicinal compounds are ionisable, to some extent at physiological pH,[1] while approximately sixty percent of drug molecules listed in the World Drug Index can be ionised between pH 2 and 12.[2] Improving the ADMET profile of drug candidates is therefore a critical step in lead optimisation. Beyond ADMET profiles, pK_a can be important in drug formulation and chemical synthesis. The benefit of *in silico* pK_a prediction is that physical samples are not needed. Therefore, predictions can be made to aid with decision making in a drug development process before expensive and time consuming synthetic work is undertaken.

pK_a estimation continues to receive much attention. A recent perspective by Lee and Crippen[3] highlighted the importance of the equilibrium constant and the multitude of methods available for predicting pK_a values for both proteins and small molecules. In the context of small molecules the methods generally fall into two main categories: (i) predictive models, using a range of descriptors and learning methods,[4–9] and (ii) *ab initio* quantum chemical methods based on different thermodynamic cycles.[10–16] The first category was reviewed by Lee and Crippen and includes linear free energy relationships and quantitative structure activity/property relationships (QSAR/QSPR). The second category regards continuum solvent pK_a computation, which involves the calculation of free energies in solution. This is usually performed *via* a thermodynamic cycle in which solution-phase reaction free energies are obtained as the sum of the corresponding gas-phase free energy and

[a] *Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street, Manchester, M1 7DN, Great Britain*
[b] *School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain*

† Electronic supplementary information (ESI) available: The identification numbers, experimental pK_a values, SMILES, names, and references for all the phenols used in this work are provided in Table S1. See DOI: 10.1039/c1cp20379g

the solvation free energy. The choice of thermodynamic cycle, level of theory, and solvation model can all affect pK_a calculation.[17] Ho and Coote[17] suggest that a realistic error margin should be in the vicinity of 2 pK_a units, with a partial cancelling of errors. The calculation of pK_a based wholly on first principles can be criticised for being too computationally demanding as it requires thermodynamic analysis and high levels of theory.[18] Zhang *et al.*[19] produced a number of one-term equations to predict pK_a. These equations rely on entropic effects cancelling and use only the enthalpy energy difference between the protonated and deprotonated forms, in conjunction with the COSMO continuum solvation model,[20] to describe the solvation. After an initial investigation using 34 molecules to compare methods (B3LYP, OLYP, HF and PW91) and basis sets (3-21G(d) to 6-311++G(3df,3pd)) for geometry optimisation and single point energy calculations, they concluded that OLYP/3-21G(d) for geometry optimisation, and OLYP/6-311+G(d,p) for the energy calculation, were the best compromise between computational expense and accuracy.[18] Extending the work to a dataset of 370 different organic acids, including carboxylic acids, phosphonic acids, alcohols, thiols, and oximes, they produced linear regression equations for separate classes of compounds with mean absolute deviations of 0.4 pK_a units.[19]

Over recent years there have been a number of publications comparing pK_a prediction methods. Dearden *et al.*[21] compared ten prediction software packages (ADME Boxes,[22] VCCLAB,[23] ADMET Predictor,[24] Pipeline Pilot,[25] SPARC,[26] Marvin,[27] QikProp,[28] ACD/Labs,[29] Pallas,[30] ChemSilicopK_a[31]) using an undivulged test set of 653 molecules and found a package called ADME Boxes to be the most accurate judged by r^2 and the *mean absolute error* (MAE). As Lee and Crippen highlighted,[3] the VCCLAB predictions were actually made by ADME Boxes, since VCCLAB links to ADME Boxes to make the predictions. The differing results for these two packages were attributed to the difference in SMILES handling. Pharma-algorithms, the company responsible for ADME Boxes, has merged with ACD/Labs keeping the ACD/Labs company name. Therefore, VCCLAB now uses ACD/Labs pK_a predictions. The r^2 and MAE range for the ten packages were 0.96 to 0.57, and 0.32 to 1.48, respectively. This comparison was based on a test set provided by ChemSilico. ChemSilico had verified that none of the compounds were part of their training set, which was not the case for the other packages. This may be one of the reasons for ChemSilico performing the worst. Meloun *et al.*[32] used the REGDIA regression diagnostics algorithm, in the package S-Plus, to compare the pK_a predictions of 64 drug molecules from four packages: ACD/Labs, Marvin, Pallas and SPARC. They found that ACD/Labs achieved the best predictive power and the most accurate results. Balogn *et al.*[33] used 248 drugs, agrochemicals and intermediates to compare ACD/Labs, Epik,[34] Marvin, Pallas and VCCLAB. It is clear from their paper that at the time the predictions were made, VCCLAB was still using ADME Boxes predictions. VCCLAB was found to be the most predictive. However, it was suggested that ACD/Labs and Marvin are the most suitable methods for medicinal chemistry as VCCLAB only calculates pK_a for the most acidic and basic groups. The r^2 and MAE ranged from 0.95 to 0.49, and 0.30 to 1.79, respectively. Liao and Nicklaus[35] have compared nine programs to predict pK_a, both commercially available and free. They used 197 pharmaceutical substances with 261 pK_a values and found ADME Boxes, ACD/Labs and SPARC to rank the highest based on r^2 and MAE. The r^2 and MAE for all nine programs ranged from 0.94 to 0.58, and from 0.39 to 1.28, respectively. It is interesting to note that when pK_a was predicted for sites for which the experimental pK_a was determined to be between the medicinally more relevant interval of 5.4 to 9.4 log units, the r^2 ranged from 0.68 to 0.35, and the MAE from 0.45 to 1.04. The relatively poor performance of Jaguar[36] confirms the discussion above on the second category of methods. The Jaguar method uses quantum mechanical calculations to calculate the free energy change in going from the protonated state to the deprotonated state. Jaguar employs empirical correction terms to repair deficiencies in both the *ab initio* calculations and the solvent models, which brings the MAE below 2 pK_a units. This error is suggested as satisfactory for this type of methods.

Quantum Topological Molecular Similarity (QTMS)[37–39] is a new approach to solving QSAR/QSPR problems using properties defined by Quantum Chemical Topology (QCT).[40–43] QCT defines so-called critical points appearing inside a given molecule, where quantum mechanical functions such as the electron density are evaluated. These and other values are QTMS descriptors. In Part 1 of this series of publications we modelled the pK_a of 228 carboxylic acids using the QTMS methodology, in which equilibrium bond lengths are usually added to the descriptor pool.[44] Indeed, as early as 2002, *ab initio* equilibrium bond lengths featured in the rationalisation of antitumor activity of (*E*)-1-phenylbut-1-en-3-ones.[37] Better models were achieved using the descriptors defined by QCT than with bond lengths alone. This has generally been the case in previous QTMS studies that predicted pK_a and other properties.[45–48] Superior models were achieved when the benzoic acids were split into *ortho*- and *meta*-/*para*-substituted groups. However, we believe that if the focus is placed on accuracy rather than globality (which means splitting chemical classes beyond the common aliphatic, *ortho*-, *para*- and *meta*-substituted groups), then strong correlations between *ab initio* bond lengths and pK_a are achievable (*e.g.* pK_a = m*r(C–O) + c). This is the approach and strategy in this paper. We will demonstrate that it is not possible to construct a global model for phenols with very few (*i.e.* one or two bond lengths) descriptors. However, by splitting the phenols into chemically meaningful subsets it is possible to construct simple linear equations using just one bond length. These will be shown to be equal if not better than using several bond lengths in more sophisticated PLS multi-term equations. Quantum mechanical methods are becoming standard in computational drug design[49] and the equations presented here offer a simple and practical way to predict pK_a using information generated from first principles.

Using the accuracy of first-principle methods, Han and co-workers studied the complete series of chlorophenols.[50] Using B3LYP/6-311++G(d,p) for geometry optimisation, in conjunction with a probing molecule to simulate the acid–base interaction, they found that several molecular parameters correlated well with the acidity of the phenols. They found ammonia to be a better probing molecule than water because it is a stronger base and induces larger measurable changes in the molecular properties. The C–O bond length (r(C–O)), O–H bond length (r(O–H)) and O–H···N hydrogen bond length (r(O–H···N)) all correlated well with the experimental pK_a, with correlation coefficients (r^2) ranging from 0.89 to 0.97 for the phenol-ammonia complexes. The complete series of bromophenols and fluorophenols was also investigated using the same methods and similar correlations were noted.[51] The authors of these papers demonstrated that weaker correlations were observed with the molecular properties of monomeric phenols.[52,53] It was suggested that bond lengths were more practical to use because the calculation of vibration frequencies was computationally more demanding.

With the end-user in mind we will demonstrate that the required accuracy in pK_a prediction can be achieved with a relatively low level of theory. This offers the opportunity for pK_a predictions of large data sets within an acceptable time. A comparison with previous work[50,51] will show that the use of the probe molecule is unnecessary. The advantage of single-term linear regression equations over multi-term equations will also be discussed. One advantage is the easier detection of outliers, which means that the experimental data can be challenged. A second advantage is a reduced potential risk of over-fitting. Finally, we will describe a procedure to predict pK_a. While this work (Part 2) is limited to phenols, a subsequent publication (Part 3) will demonstrate the method on benzoic acids and anilines. Because the method is generic it is expected to be applicable to diverse classes of compounds.

2. Methods and computational detail

2.1 Data sets

Table 1 provides the constitution of the data set for the phenols. The experimental pK_a values were taken from a paper[54] by Tehan *et al.*, unless otherwise stated. These authors had previously applied a variety of filters in order to remove non-druglike molecules. Where we have used other sources for experimental pK_a values to correct experimental values from our original data source, expand the data set or test our models, we explicitly highlight these occurrences in the text. The experimental pK_a values and identification numbers used in this work are listed in Table S1 of the Electronic Supplementary Information (ESI). We note that Tehan *et al.* do not provide the likely errors associated with the experimental pK_a values. Concerns about the influence of temperature on pK_a measurement are partially addressed by this group's statement that all experimental data were measured between 10 °C and 30 °C. In the 1960s the IUPAC criteria for classifying its compilation of pK_a values for weak organic acids and bases were: "*very reliable*" (pK_a error $<$ ± 0.005), "*reliable*" (pK_a error $<$ ± 0.005 to ± 0.02), "*approximate*" (pK_a error $<$ ± 0.02 to ± 0.04), and "*uncertain*" (pK_a error $>$ ± 0.04).[35] These error criteria may seem overly restrictive. However, pK_a values are logarithmic representations of the acid dissociation constant K_a and therefore small errors can have profound effects.[55] The original pK_a values were taken from the PHYSPROP database, which provided references to the original measurement. Without checking all the individual references, the errors in the pK_a values are unknown and must influence the validation of our models to some extent.

Table 1 A summary of the data sets investigated

Compound class		# of compounds
Phenols		**171**
	Meta/*Para*[a]	55
	Ortho[b]	90
	Ortho, capable of forming Internal Hydrogen Bonds (*ortho*-phenols-IHB)[c]	26

[a] Two iodine containing compounds were removed: 4-iodophenol (compound 35) and 3-iodophenol (compound 50) because the basis set was not readily available for iodine. 4-hydroxyacetophenone (compound 58) was also removed since the CAS number and name provided did not match, therefore causing ambiguity. The name given was hydroxyacetophenone whilst the CAS number relates to 4-hydroxyphenylacetaldehyde. The experimental pK_a quoted is to be 8.05, which is the same as that of 4-hydroxyacetophenone (compound 12). [b] One iodine containing compound was removed, which is 2-iodophenol (compound 134). Secondly, 2-methyl-4-chlorophenol (compound 174) was already in the data set as compound 162 but both compounds appeared with different CAS numbers and different experimental pK_a values. The CAS number of compound 174 was trusted to avoid duplication in the dataset and its structure was corrected to 2-chloro-4-methylphenol. [c] The incorrect name provided for compound 83 was corrected to 3,5-4′-trichloro-2′-nitrosalicylanilide.

2.2 Data generation and analysis

The discussion below provides a general overview of the data generation and analysis. More details about the exact analysis of the data set are given in the results (section 3). An initial guess of the geometry of each compound was provided by MOLDEN.[56] Using the program GAUSSIAN03,[57] geometries were optimised at HF/6-31G(d) level. The bond lengths of interest were then extracted and a Partial Least Squares (PLS)[58] analysis was carried out to fit the bond lengths to the experimental pK_a values. SIMCA-P[59] was used for the majority of the data analysis. Models using all the bond lengths of interest were initially created using the predefined criterion for determining the significant number of Latent Variables (LVs) to appear in the PLS equation. This criterion states that if the value of q^2 of the newly constructed LV is less than 0.097, then no more LVs are computed; the PLS regression is then deemed complete. Separate models were also created for the *ortho*- and *para-*/*meta*-substituted phenols. Variable Importance in the Projection (VIP) plots for the models were subsequently examined. VIP plots provide a condensed summary of the relative importance of each variable to the model, in this case the contribution of specific bond lengths. The bond lengths that contributed the most to the models were then used to construct one-term bond length models for the compound classes and the results analysed. Attempts were then made to separate these models into chemically meaningful groups of compounds where one common bond length showed high correlation with the experimental pK_a values. We refer to these groups of compounds as *high-correlation subsets*. Through the analysis of the single-bond-length equations the influence of conformation was investigated.

Outliers and errors were detected and where possible corrected. Outliers were visually detected from the observed-*versus*-predicted plots and subsequently investigated. If an error was found (*i.e.* wrong experimental pK_a value or structural error) in the original data set then this was corrected. In the Results section the models and associated statistics are provided before and after outlier removal. We investigated a number of techniques to detect outliers, which included the use of the 3σ rule, variations of this rule and other criteria for outlier omission. However, they were found to be unsatisfactory because the high-correlation subsets needed specific (*i.e.* local) treatment and therefore no general rule across all subsets could be established. The application of a rule based on a multiple of σ is more justified for larger and hence more global datasets, while our approach ends up with small and local subsets.

Higher levels of theory were examined and comparisons of the results with and without an ammonia probe were made for one of the high-correlation subsets. The predictions made from the high-correlation subsets were compared to the predictions made from models constructed using all the bond lengths and more diverse training sets. Models were validated using leave-many-out and compared using a variety of statistics discussed in section 2.3 below.

2.3 Statistics

Given that the literature is not always as clear as it could (or should) be about the statistics it uses, we believe it is important to clarify the statistics used in this paper. The construction of the models using SIMCA-P provides three statistics to give an indication of the goodness-of-fit and goodness-of-prediction. The first statistic is the squared correlation coefficient (r^2),

$$r^2 = \frac{\sum_{i=1}^{n} (y_{\text{calc},i} - \bar{y})^2}{\sum_{i=1}^{n} (y_{\text{obs},i} - \bar{y})^2} \qquad (1)$$

where n is the number of observations in the *entire* data set, $y_{\text{calc},i}$ is the pK_a value for molecule i calculated from the regression equation, $y_{\text{obs},i}$ is the corresponding experimental pK_a, and $\bar{y}$ is the mean pK_a value of the entire experimental (*i.e.* observed) data set. The Root Mean Squared Error of Estimation (RMSEE) is calculated as,

$$\text{RMSEE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{\text{obs},i} - y_{\text{calc},i})^2}{n - 1 - \alpha}} \qquad (2)$$

where α is the number of LVs used to construct the PLS model. Another common error measure is the Root Mean Squared Error (RMSE), which is defined as,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{\text{obs},i} - y_{\text{calc},i})^2}{n - 1}} \qquad (3)$$

In an alternative expression sometimes encountered in the literature the denominator in the RMSE equation is sometimes set to n. Because the denominator of the RMSEE is always smaller than that of RMSE and both share the same numerator, RMSEE is always larger than RMSE. Hence, RMSEE penalises the deviation between observed and calculated data more than RMSE, and is therefore a more severe error measure. In turn, the RMSE is more severe than the Mean Absolute Error (MAE), which is given by,

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_{\text{obs},i} - y_{\text{calc},i}|}{n} \qquad (4)$$

In summary, the RMSEE is always larger than the RMSE, which is in turn always larger than the MAE, or RMSEE > RMSE > MAE. This sequence is also the ranking of severity, starting with the most severe error. This discussion should be kept in mind when comparing results to the literature, where RMSE and MAE frequently appear. The MAE is also less sensitive to larger outliers than both the RMSE and RMSEE because the MAE does not square the discrepancies between observed and calculated values. This means that these two measures give a relatively higher weight to large errors compared to the MAE. In this work we assess our models by means of RMSEE, which is the most severe criterion of quality, both in terms of outlier detection and the ranking of error measures given above.

We base our comparisons on RMSEE and use this as a guide to indicate which models should be cross-validated. We only preformed full K-fold Cross-Validation (CV) on the most promising models. In K-fold CV the original set is randomly partitioned in K CV *subsets*. In the current work, K is set to exactly 7 throughout, provided there are more than 7 data points in the total data set. Note that in this work there are datasets with less than 7 points. If the total number of compounds n is divisible by 7 then there are n/7 compounds in each of the 7 CV subsets. If n is not divisible by 7 then the remaining compounds will be evenly distributed over the 7 CV subsets. Note that the number of subsets therefore does not vary. The compounds in the first CV subset are predicted from a model constructed from the remaining 6 CV subsets, all combined in one training set. The compounds in the second CV subset are then predicted from a different model, now constructed from the new remaining 6 CV subsets, excluding the second CV subset. Again these 6 subsets were combined in one (new) training set. This process is repeated for the third and higher CV subsets, until each compound has been excluded exactly once. Each compound will then have been predicted by its corresponding training set. The predicted pK_a value for compound i, denoted by $\hat{y}_{\text{pred},i}$, is obtained from the regression equation constructed from each training set. Now we are in a position to assess the success of these predictions.

The cross-validated r^2, also known as q^2, is calculated as,

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_{\text{obs},i} - \hat{y}_{\text{pred},i})^2}{\sum_{i=1}^{n} (y_{\text{obs},i} - \bar{y})^2} \qquad (5)$$

The generated q^2 is based on 'leave-one-seventh' of the data out rather than 'leave-one-out', which is not recommended because of its known pitfalls.[60,61] However, if the total dataset contains less than seven compounds then the CV is 'leave-one-out'.

One may question if the use of CV is justified against an assessment based on splitting the data set into a training and test set. Hawkins *et al.*[62] recommend that, when the data set is small (< 100 compounds), then CV may be better than splitting the data set into training and test sets. The high-correlation subsets we use to create models are small in the sense of

Hawkins *et al.* who advocate CV when the data set contains less than 100 compounds. CV should involve using a suitable variable selection technique to select the variables important to the training set, each time a CV subset is excluded. This procedure renders a 'true q^2' rather than a 'naïve q^2', where variable selection is not performed each time a CV subset is removed. Below we argue that our assessment procedure is generating a true q^2. Essentially, the main argument is that variable selection does not apply to our way of setting up a model. This is because we only consider either an all-bond-length model or a single-bond-length model. The latter type of model was based on the most important variable in the VIP plot of the all-bond-length model. During the CV of the all-bond-length model the VIP plots for the seven models created in CV were monitored. In the vast majority of cases the most important bond length remained the most important to the models created during CV. Furthermore, SIMCA-P automatically selects the number of LVs to construct the all-bond-length models. Because a new model is constructed for each of the seven training sets, variable selection is performed by default. For these reasons we consider the q^2 value quoted to be the 'true q^2'. By default, SIMCA-P automatically produces a q^2 value when models are constructed.

The Root Mean Square Error of Prediction (RMSEP) is provided by

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_{\text{obs},i} - \hat{y}_{\text{pred},i})^2}{n}} \tag{6}$$

and differs from the RMSEE, looking at the numerator, because $\hat{y}_{\text{pred},i}$ is obtained from the models constructed from the training sets during CV.

The squared correlation coefficient obtained through CV and denoted by r^2_{CV}, which is not be confused with q^2, is calculated as,

$$r^2_{\text{CV}} = \frac{\sum_{i=1}^{n} (\hat{y}_{\text{pred},i} - \bar{y})^2}{\sum_{i=1}^{n} (y_{\text{obs},i} - \bar{y})^2} \tag{7}$$

where the variables have already been explained above.

We also use a further metric, denoted r^2_m, which is calculated as,

$$r^2_m = r^2_{\text{CV}} \times \left(1 - \sqrt{r^2_{\text{CV}} - r^2_{\text{CV},0}}\right) \tag{8}$$

Here, r^2_{CV} and $r^2_{\text{CV},0}$ are the squared correlation coefficient values between the observed (X-variable) and predicted (Y-variable) pK_a values, obtained through CV, with intercept *not* set to zero and set to zero, respectively. Note that $r^2_{\text{CV},0}$ can be negative.

A high r^2_{CV} value does not necessarily indicate that the predicted values are very close to the experimental values. There may be considerable numerical differences between the observed and predicted values in spite of the presence of a good overall correlation. When this is the case there will be substantial differences between r^2_{CV} and $r^2_{\text{CV},0}$, which the r^2_m statistic penalises heavily. Mitra *et al.*[46] have shown that in the case of small data sets, r^2_m calculated from a CV when variable selection is performed at each CV step, reflects the external validation characteristics of the developed model. Based on the reasoning above about 'true q^2' we believe that our quoted r^2_m values are 'true $r^2_{m(\text{leave-one-seventh-out})}$'. Ultimately, we judge the performance of the models based on RMSEP. However, we stress that RMSEE is used to decide which models to perform full CV on.

3. Results

3.1 Phenols

Fig. 1 shows the common skeleton and bonds screened to predict the pK_a of the phenol compounds (Table 1). In previous QTMS studies, conformation has not been taken into account with the knowledge that the substitution effects have a greater influence on the models. We show that conformation is important in some cases (discussed later). We initially investigated the use of the 8 phenol bond lengths of the common skeleton in order to model all the phenol compounds together and the common groupings of *ortho*, *meta*/*para* and *ortho*-phenols that we deemed capable of forming internal hydrogen bonds.[54] During the course of this work, suspected errors in the initial dataset were corrected and the importance of conformation was examined.

Inspection of the VIP plot modelling all the phenol compounds showed that r(C–O) and r(O–H) contributed most to the model. Therefore, these bonds were monitored to see if they could model the 171 phenol compounds individually (Table 2), in line with our motivation discussed in the Introduction. The r^2 decreased and the RMSEE increased when these two bond lengths were individually used. The reduction in the quality of the model was less for r(C–O) on its own than for r(O–H) on its own. We determined what influence splitting the data set into *meta*-/*para*- and *ortho*-substituted phenols (common for compounds of this type) had on the quality of the models. The RMSEE for the model constructed using all the bond lengths for *meta*-/*para*-substituted phenols decreased by approximately 50% but the RMSEE for the *ortho*-substituted phenols increased. The reduction in the quality of the models of the *meta*-/*para*-substituted phenols when using r(C–O) on its own or r(O–H) on its own is also small compared to the all-bond-length model. This suggests that the *meta*-/*para*-substituted phenols can be modelled using just one bond length. The RMSEE for the *ortho*-substituted model increased when using r(C–O) on its own or r(O–H) on its own. This is not surprising since different *ortho*-substituents can affect the pK_a of compounds because of their close proximity to the acidic hydrogen.

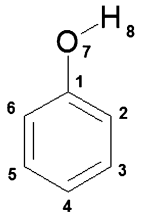


Fig. 1 The eight bond lengths used to predict the pK_a of the phenol compounds. The main text refers to bond lengths 1–7 and 7–8 as r(C–O) and r(O–H), respectively. Reference to other bonds makes use of this numbering scheme to distinguish between the C–C bonds, *e.g.* $r(C_1–C_2)$.

Table 2 The results of the phenol compounds modelled with the bond lengths calculated at the HF/6-31G(d) level of theory

Subsets	# LV	# Bonds	# Compounds	r^2	q^2	RMSEE
All	4	All	171	0.92	0.88	0.67
All	1	*r*(C–O)	171	0.86	0.85	0.88
All	1	*r*(O–H)	171	0.52	0.51	1.62
Meta/*Para*	2	All	55	0.91	0.87	0.34
Meta/*Para*	1	*r*(C–O)	55	0.87	0.85	0.41
Meta/*Para*	1	*r*(O–H)	55	0.84	0.83	0.45
Ortho	4	All	116	0.92	0.86	0.72
Ortho	1	*r*(C–O)	116	0.85	0.85	0.99
Ortho	1	*r*(O–H)	116	0.47	0.46	1.84
Ortho without *Ortho*-IHB	5	All	90	0.94	0.88	0.65
Ortho without *Ortho*-IHB	1	*r*(C–O)	90	0.88	0.87	0.94
Ortho without *Ortho*-IHB	1	*r*(O–H)	90	0.59	0.58	1.72

These effects include steric hindrance to protonation or deprotonation and internal hydrogen bonding.

To investigate the large deterioration of the *ortho*-substituted phenol model, when using all the bond lengths compared to using just *r*(C–O) or *r*(O–H) on their own, we inspected the predicted-*versus*-observed pK_a plots. Fig. 2a shows such a plot for all 171 phenols using a regression model using only *r*(C–O). Inspection of Fig. 2a suggests subsets of phenols that have a higher r^2 value than the full set of 171 phenols.

It was rewarding to find that such *high-correlation subsets*, identified by eye, later turned out to be meaningful chemical subsets. For example, in Fig. 2b it is clear that *o*-halogen phenols (shown in dark blue) and *o*-nitro (shown in light blue) phenols are separate high-correlation subsets. This was also seen for other *o*-phenols depending on the *o*-substituent. It appeared that *meta*-/*para*-phenols were a high-correlation subset irrespective of the different substituents. A number of compounds appeared to be outliers from the high-correlation subset to which they would have been expected to belong to. An example of this is shown in Fig. 2b for 4,6-dinitro-*o*-cresol (compound 135). This compound appeared to belong to the *o*-halogen high-correlation subset (dark blue in Fig. 2b), which was inconsistent with the compound's structure. Inspection of the optimised structures showed that this was caused by the anti conformation being used instead of the syn conformation that was used for the other *o*-nitrophenols. When this compound was optimised in the syn conformer, it correctly moved into the *o*-nitrophenol high-correlation subset (light blue in Fig. 2b). This example is representative for a number of other compounds that appeared to belong to high-correlation subsets different to the chemically meaningful subsets we had identified. All the *ortho*-phenols were subsequently optimised in the syn and anti form and the energies were used as a guide to decide which high-correlation subset they belonged to. Because of symmetry, conformation plays no role in di-*ortho*-substituted phenols with identical substituents. However, for the asymmetrical di-*ortho*-substituted and the mono-*ortho*-substituted compounds, the orientation of the acidic hydrogen can have a large influence on bond lengths. The results of the detailed modelling of the *o*-phenols and identification of high-correlation subsets are reported in the following section.

3.2 *Ortho*-phenols

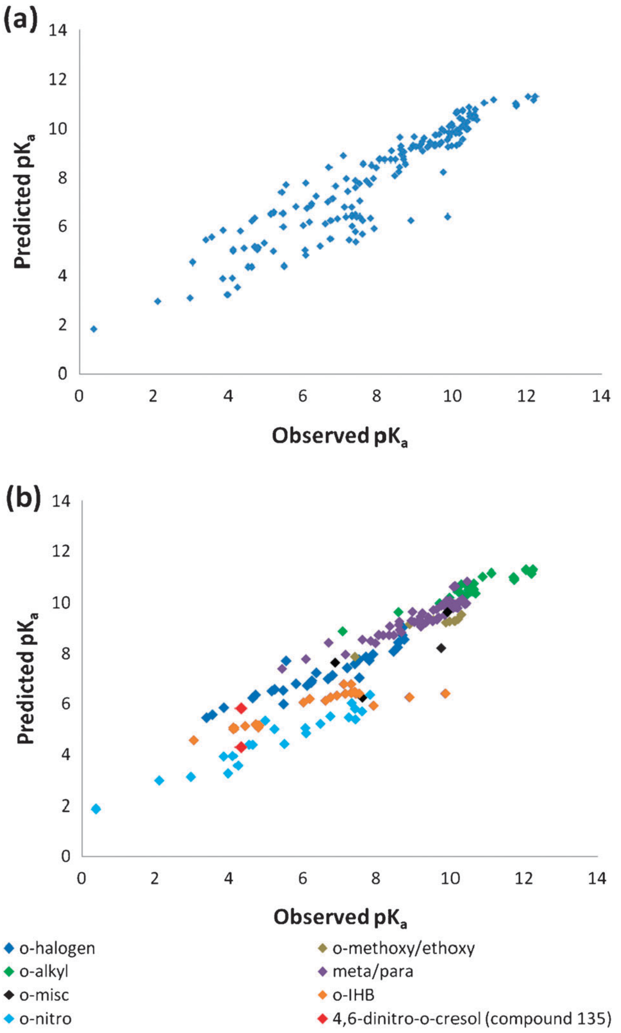


Fig. 2 (a) Plot of predicted *vs.* observed pK_a for the phenols using *r*(C–O). (b) Plot of the predicted *vs.* observed pK_a for the phenols using *r*(C–O) separated by colour into chemically meaningful high-correlation subsets. The different pK_a values of 4,6-dinitro-*o*-cresol calculated from *r*(C–O) for the syn and anti conformer is highlighted in red as an example.

Inspection of the structures of the compounds belonging to high-correlation subsets and their energies in the data set's *o*-phenols, led to rules to assign the compounds to specific *o*-phenol high-correlation subsets. These rules were confirmed by the detailed investigation of the high-correlation subsets. In section 3.2.1 to 3.2.6 we discuss the results that allowed us to state the rules here. In the case of the *o*-phenols it was fortuitous that the energies could be used as a guide, without exception. In all the high-correlation subsets we determined that the lowest energy conformation of all compounds was the same (*i.e.* either syn or anti) without exception. For example, for all the *o*-nitrophenols the syn conformation was the lowest energy and for all the *o*-alkylphenols the anti conformation was the lowest energy. These rules are encapsulated in the flow chart below showing which high-correlation subset a phenol of interest should be predicted from.

It should be noted that certain phenols can belong to more than one high-correlation subset. For example, 2-nitro-6-chlorophenol can be predicted by the *o*-nitro or the *o*-halogen models, depending on the direction of the acidic hydrogen. We will show that the better prediction is made by the *o*-nitro model because nitro substituents decrease the pK_a more than chlorine substituents as the former are more electron-withdrawing. We will also show that for *meta-*/*para*-substituted phenols the influence of conformation on the quality of the models is minimal. We screened the compounds in search of high-correlation subsets from different classes of compounds in the ortho subset, previously (see RMSEE values larger than 0.50 in Table 2) shown to produce poor correlations when modelled together. We compare all-bond-length models

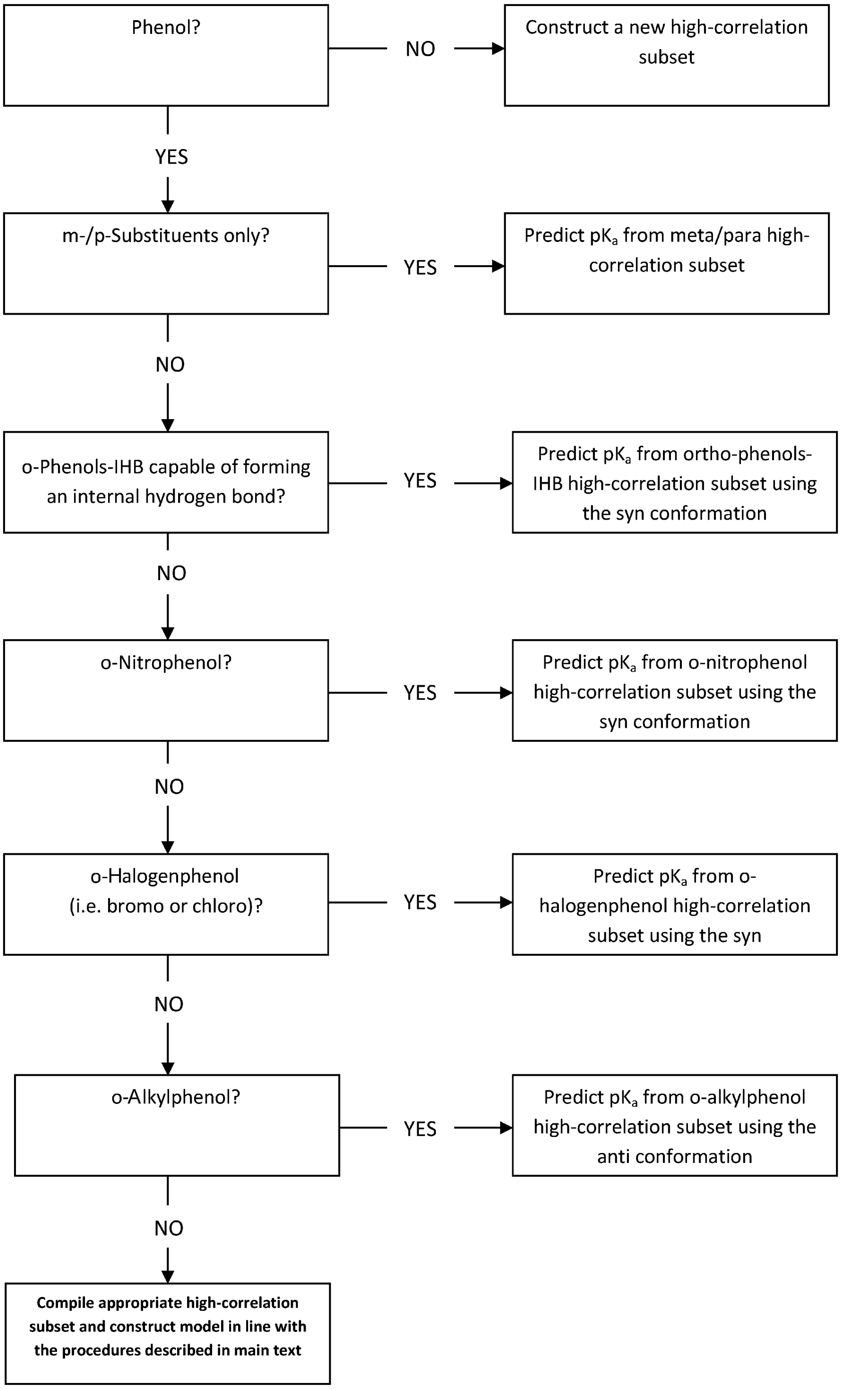

Phenol?
NO
Construct a new high-correlation subset
YES
m-/p-Substituents only?
YES
Predict pK$_a$ from meta/para high-correlation subset
NO
o-Phenols-IHB capable of forming an internal hydrogen bond?
YES
Predict pK$_a$ from ortho-phenols-IHB high-correlation subset using the syn conformation
NO
o-Nitrophenol?
YES
Predict pK$_a$ from o-nitrophenol high-correlation subset using the syn conformation
NO
o-Halogenphenol (i.e. bromo or chloro)?
YES
Predict pK$_a$ from o-halogenphenol high-correlation subset using the syn
NO
o-Alkylphenol?
YES
Predict pK$_a$ from o-alkylphenol high-correlation subset using the anti conformation
NO
Compile appropriate high-correlation subset and construct model in line with the procedures described in main text

Table 3 The statistical details of the models created for the *o*-nitrophenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	*Syn*	All	4	23	0.97	0.37	0.91	0.65	0.88	0.87	0.79
HF	*Syn*	*r*(C–O)	1	23	0.91	0.58	0.91	0.58	0.91	0.90	0.82
HF	*Syn*	*r*(O–H)	1	23	0.85	0.77	0.84	0.78	0.83	0.80	0.70
HF	*Syn*	All	2	22 (−87)	0.97	0.38	0.93	0.51	0.94	0.94	0.93
HF[a]	***Syn***	***r*(C–O)**	**1**	**22 (−87)**	**0.94**	**0.48**	**0.94**	**0.50**	**0.93**	**0.93**	**0.87**
HF	*Syn*	*r*(O–H)	1	22 (−87)	0.88	0.71	0.87	0.73	0.85	0.84	0.75
HF	*Anti*	All	2	23	0.98	0.33	0.88	0.81	0.82	0.80	0.70
HF	*Anti*	*r*(C–O)	1	23	0.79	0.90	0.78	0.91	0.77	0.71	0.58
B3LYP	*Syn*	All	2	23	0.94	0.50	0.90	0.61	0.89	0.89	0.82
B3LYP	*Syn*	*r*(C–O)	1	23	0.91	0.58	0.91	0.59	0.90	0.89	0.81

[a] The bold line in this and subsequent tables marks the best single-bond-length model as judged by the lowest RMSEP value, where appropriate, as discussed in the main text.

to single-bond-length models using CV, as discussed in section 2.3.

3.2.1 ***o*-Nitrophenols.** The results from CV are reported in Table 3 to allow comparisons between models. Inspection of the VIP plot for the all-bond model using all the *o*-nitrophenols revealed that *r*(C–O) was the most important descriptor followed by *r*(O–H). For this reason we created separate models for each of these two bond lengths. Looking at r^2 it is surprising that this value remains high for either of the single-bond-length models compared to the all-bond-length model. Inspection of the plot showing observed-*versus*-predicted pK_a values for the single-bond-length models caused suspicion about the experimental pK_a of 2,3-dinitrophenol (compound 87, experimental pK_a given as 4.96). Another source[46] quoted the experimental pK_a of this compound to be 5.24. This increase in pK_a moves it towards a value of approximately 6 log units predicted by our different models. Removal of compound 87 from the fitting procedure improved the model statistics. When this compound was removed during CV of the all-bond-length model, the resulting model used only two LVs compared to the three LVs making up the models with it included. This suggests that the program SIMCA-P had added a LV to fit compound 87. This was not the case for the single-bond-length models as the fitting was minimal here. During CV the VIP plots of the models were inspected when each CV group was removed in turn. *r*(C–O) followed by *r*(O–H) were the most important bonds to all the models in CV. The *r*(C–O) model, when compound 87 was removed, produced the lowest RMSEP (0.50) and a high r^2_m (0.87). This was pleasing considering only one bond length is used.

Table 3 also provides the statistics relating to the models built using anti conformations. The all-bond-length model has the highest r^2 value in conjunction with the lowest RMSEE. However, the model is shown to be weaker when CV is performed compared to that constructed using the lower energy syn conformations. The *r*(C–O) model using the anti conformations is also poorer than when the syn conformers are used. The VIP plot for the all-bond-length anti model showed *r*(C–O) to be the most important. However, *r*(C_5–C_6) was the next most important and *r*(O–H) was ranked sixth.

Table 3 also shows that the higher level of theory, B3LYP/6-311+G(2d,p), produced very similar statistics to the more economical level HF/6-31G(d) suggesting the latter is sufficient. This conclusion is supported by a study[63] of 2,4-dinitrophenol where the HF/6-31G(d) level of theory performed very well for predicting the geometrical parameters. The HF method failed to reproduce the vibration frequency of the O–H bond stretch. However, this is of no import as we only use the bond lengths.

3.2.2 ***o*-Halogen phenols.** The statistics relating to the *o*-halogen phenols are given in Table 4. The *o*-halogen phenols were initially modelled as syn conformations. The models built from all the compounds were inspected. The VIP plot for the all-bond-length model showed that *r*(C–O) was the most important followed by *r*(O–H). The *r*(C–O) model gave better statistics than the all-bond-length model. Inspection of the observed-*versus*-predicted plot for this model revealed six suspicious data points. The structures of the compounds that represented these points are shown in Fig. 3. We will discuss each outlier in turn starting with 2,4,6-tribromophenol. A different experimental pK_a of 6.1 for 2,4,6-tribromophenol (compound 120) was found[46] instead of the value of 6.8 given in the source[54] we used for the experimental pK_a values. The value of 6.1 was much closer to the value predicted from our correlation and was hence adopted.

Next we explain why 6-chloro-2-nitrophenol (compound 141) and 6-chloro-2,4-dinitrophenol (compound 160) should be only be predicted from the *o*-nitro high-correlation subset model. We note that these compounds have both *o*-nitro and *o*-halogen substituents and so the pK_a could be predicted by either the *o*-nitro or *o*-halogen high-correlation subsets. To obtain a reasonable prediction from the *o*-halogen high-correlation subset (for 6-chloro-2-nitrophenol and 6-chloro-2,4-dinitrophenol) we used the conformation in which the acidic hydrogen points towards the halogen, which we note is *not* the lowest energy, but is consistent with the conformations used for the other *o*-halogen phenols. 6-chloro-2,4-dinitrophenol (compound 160) was predicted reasonably well by the *r*(C–O) model (experimental pK_a of 1.6 compared to a predicted pK_a of 2.1). This discrepancy of 0.5 log units decreased to 0.43 when predicted from the *o*-nitro high-correlation subset. A more dramatic improvement occurred for 6-chloro-2-nitrophenol (compound 141), which had an error of 1 pK_a unit by the *r*(C–O) model of *o*-halogen phenols, but an error only 0.55 log units when predicted from the *o*-nitro high-correlation subset. When predicting

Table 4 The statistical details of the model created using *o*-halogen phenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	*Syn*	All	1	32	0.75	0.89	0.71	1.05	0.64	0.55	0.45
HF	*Syn*	*r*(C–O)	1	32	0.88	0.61	0.87	0.62	0.87	0.86	0.79
HF	*Syn*	*r*(O–H)	1	32	0.63	1.08	0.63	—[a]	—	—	—
HF	*Syn*	All	2	26	0.95	0.40	0.91	0.44	0.93	0.91	0.81
HF	***Syn***	***r*(C–O)**	**1**	**26**	**0.97**	**0.27**	**0.97**	**0.29**	**0.97**	**0.97**	**0.94**
HF	*Syn*	*r*(O–H)	1	26	0.60	1.06	0.58	—	—	—	—
HF	*Anti*	All	2	32	0.83	0.74	0.70	—	—	—	—
HF	*Anti*	rC–O)	1	32	0.83	0.72	0.81	—	—	—	—
HF	*Anti*	*r*(O–H)	1	32	0.61	1.12	0.53	—	—	—	—
HF	*Anti*	All	2	26	0.94	0.40	0.91	0.46	0.92	0.91	0.81
HF	*Anti*	*r*(C–O)	1	26	0.96	0.35	0.96	0.35	0.95	0.95	0.91
HF	*Anti*	*r*(O–H)	1	26	0.78	0.79	0.76	—	—	—	—
B3LYP	*Syn*	All	1	32	0.71	0.97	0.65	—	—	—	—
B3LYP	*Syn*	*r*(C–O)	1	32	0.87	0.64	0.86	—	—	—	—
B3LYP	*Syn*	*r*(O–H)	1	32	0.52	1.24	0.44	—	—	—	—
B3LYP	*Syn*	All	2	26	0.93	0.46	0.86	0.67	0.83	0.81	0.70
B3LYP	*Syn*	*r*(C–O)	1	26	0.96	0.32	0.96	0.33	0.96	0.96	0.92
B3LYP	*Syn*	*r*(O–H)	1	26	0.72	0.89	0.70	—	—	—	—

[a] A dash in this Table 5–7 indicates that various cross-validation statistics were not collected as justified in the main text.

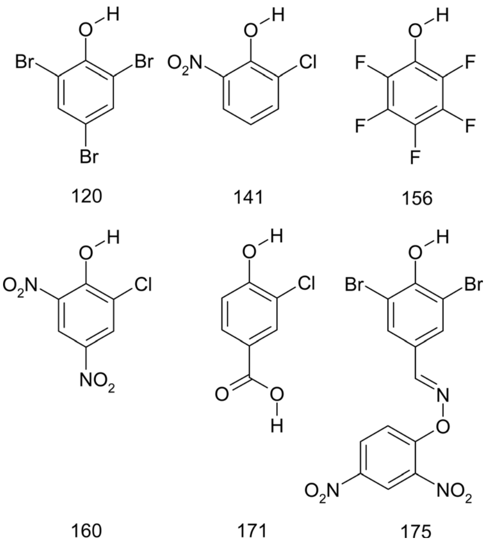


Fig. 3 Structures (and their compound numbers) that belong to the data points that seemed to be outliers in the *o*-halogenphenol models.

by the *o*-nitro high-correlation subset, we used the opposite conformations, where the acid hydrogen points towards the nitro groups, which were the lowest energies. From these two compounds we conclude that phenols with a nitro and a halogen substituent in both ortho positions should be predicted from the *o*-nitro high-correlation subset and not the *o*-halogen high-correlation subset.

Pentafluorophenol (compound 156) needs a separate discussion. This compound showed the largest discrepancy between observed and predicted pK_a. 2-fluorophenol (compound 129) was the only other compound in our data set that had an *o*-fluoro substituent and appeared to belong to the *o*-halogen high-correlation subset. It was clear that *o*-chlorophenols and *o*-bromophenols formed a single high-correlation subset. However, because we only had two *o*-fluorophenols in the dataset it was impossible to prove if this class of compounds needed to be modelled separately or if the experimental pK_a of pentafluorophenol should be challenged. The experimental value we adopted from the work of Tehan *et al.* was verified against an alternative literature source,[51] where the same pK_a value of 5.53 was used to produce excellent correlations. This check confirmed that the experimental value used is accurate. We therefore calculated the bond lengths of a further three *o*-fluorophenols, for which we had experimental pK_a values, to verify that they produce a separate high-correlation subset. The *r*(C–O) of 2,4-difluorophenol, 2,6-difluorophenol and 2,3,5,6-tetrafluorophenol were calculated and the correlation between *r*(C–O) of the five *o*-fluorophenols and pK_a was checked. An r^2 of 0.91 and RMSEP of 0.40 suggested that *o*-fluorophenols indeed produce their own high-correlation subset and cannot be included with the other *o*-halogen compounds. To confirm that this was not a fortuitous result based on the HF/6-31G(d) level of theory, we compared our result to that obtained by Han and Toa[51] using B3LYP/6-311++G(d,p) and an ammonia probe. Using their *r*(C–O) equation to predict the pK_a for the same five *o*-fluorophenols, we obtained an r^2 and RMSEP of 0.90 and 0.41, respectively. After this confirmation we removed 2-fluorophenol and pentafluorophenol from subsequent analysis of the *o*-halogen high-correlation subset because it was clear they produced a separate *o*-fluorophenol high-correlation subset.

A plausible explanation for the exclusion of 3-chloro-4-hydroxybenzoic acid (compound 171) stem from the fact that it has micro pK_a values. In the modelling, we only considered the neutral form of this compound, that is, the COOH and OH groups both retained their proton. Since the carboxyl group is more acidic than the OH group, the former's proton will dissociate first, leaving behind an anionic phenol. Future work can tackle this situation probably at the cost of introducing diffuse basis functions, which are known to be necessary for anions. Secondly, bromofenoxim (compound 175) was also excluded due to a large discrepancy between model and experiment but the reason is not clear.

Now we focus on the influence of the omission of outliers. Table 4 shows how the models improved when the six compounds were removed, resulting in good CV results. Table 4 shows that the single-bond-length models benefit approximately equally from this omission compared to the all-bond-length models. For example, upon omission of six outliers the RMSEE for the *r*(C–O) model roughly halves, from 0.61 to 0.27. Equally, the RMSEE for all-bond-length model also halves from 0.89 to 0.40. A similar trend is observed for RMSEP. The most dramatic improvement due to the omission of outliers is seen in the r_m^2 statistic. For the all-bond-length models with outliers included, a r_m^2 value of 0.45 suggests that poor predictions are made in CV, while reasonable predictions are made for the single-bond-length model, suggested by an r_m^2 of 0.79. After outlier omission, the r_m^2 value for the all-bond-length model improves to 0.81, suggesting a large improvement in prediction. However, the *r*(C–O) model without outliers is superior based on an r_m^2 of 0.94. It is interesting to note that the *r*(C–O) models are always superior to the all-bond-length models in the original model fit and in CV. The *r*(O–H) models were not cross-validated as they were inferior to the other models based on the original fitting statistics. This is why the corresponding CV statistics are not listed in Table 4.

High-correlation subsets using the anti conformations were also investigated with and without the identified outliers. The outliers were still suspicious data points in the inspected high-correlation subsets. These were removed and models with all bond lengths and *r*(C–O) were cross-validated to compare to the models constructed using the syn conformations. Using just the *r*(C–O) provided a better model than using all bond lengths, as with the syn conformer models. The models were not as good as the models where the syn conformers were used. Using *r*(O–H) once again provided a poor correlation. Inspection of the observed-*versus*-predictive pK_a plot from *r*(O–H) revealed high-correlation subsets different to those seen when all the phenols were modelled with *r*(C–O). The structures of the *o*-halogenphenols producing these separate high-correlation subsets were inspected and showed that di-*o*-bromophenols, di-*o*-chlorophenols and mono-orthophenols (*i.e.* those substituted with a chlorine or bromine at the ortho position) belong to their own subset. This is not surprising, as *r*(O–H) is affected by the substituent that it points towards, resulting in separate models for the di-*o*-phenols and a single high-correlation subset for the mono-*o*-phenols as the acid hydrogen points towards a hydrogen in each case. This observation is confirmed by 2-chloro-6-methylphenol (compound 90) not belonging to any high-correlation subset as the methyl group has a different influence to that of a hydrogen. These results confirm the success of *r*(C–O) and the syn conformations. The statistics of the models constructed with and without the six outliers, using B3LYP/6-311+G(2d,p) geometries and the syn conformation, offer no improvement to those created using HF/6-31G(d).

3.2.3 *o*-Alkylphenols.

The anti conformation is the lowest energy for the *o*-alkylphenols, which is opposite to the syn conformation being favoured by the *o*-nitro and *o*-halogenphenols. However, for symmetrical 2,6-substituted phenols syn and anti cannot be assigned, but the position of the phenolic hydrogen can be in the aromatic plane or out-of-plane. In the former case the hydrogen points towards an alkyl group, which is the conformer with a lower energy than that of the latter case. For asymmetrical 2,6-substituted phenols, *e.g.* 2-(1,1-dimethyl)-4,6-dimethylphenol (compound 164), the conformation with the hydrogen pointing towards the methyl group has the lowest energy. Four compounds that also had *o*-nitro and *o*-halogen substituents were initially included in the modelling (Table 5) as the conformation where the phenolic hydrogen pointed towards the alkyl substituent, which we note is not the lowest energy. These compounds fitted into the alkyl high-correlation subset relatively well. However, they are modelled better in the lower energy conformation by the *o*-nitro and *o*-halogen high-correlation subset. Therefore the compounds were excluded from CV in agreement with two rules listed in section 3.2. The models using all the bond lengths are comparable to the model using just *r*(C–O) (Table 5). The correlation obtained using *r*(O–H) was inferior to that obtained using *r*(C–O). There was no notable improvement when using B3LYP/6-311+G(2d,p) generated bond lengths compared to those calculated with HF/6-31G(d).

3.2.4 *o*-Phenols capable of forming internal hydrogen bonds.

Twenty-three out of the 26 compounds in this high-correlation subset had the same extended common skeleton (*i.e.* the phenol common skeleton extended by the phenylamide part) with different substituents around both aromatic rings (Fig. 4).

Table 5 he statistical details of the model created using *o*-alkylphenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r_{CV}^2	$r_{CV,0}^2$	r_m^2
HF	*anti*	All	3	29	0.96	0.44	0.94	—	—	—	—
HF	*anti*	*r*(C–O)	1	29	0.93	0.58	0.30	—	—	—	—
HF	*anti*	*r*(O–H)	1	29	0.79	1.00	0.78	—	—	—	—
HF	*anti*	All	2	25	0.94	0.28	0.83	0.42	0.79	0.76	0.64
HF	***anti***	***r*(C–O)**	**1**	**25**	**0.91**	**0.34**	**0.90**	**0.37**	**0.89**	**0.87**	**0.78**
HF	*anti*	*r*(O–H)	1	25	0.36	0.90	0.36	0.91	0.30	−1.06	−0.05
B3LYP	*anti*	All	3	29	0.95	0.52	0.91	—	—	—	—
B3LYP	*anti*	*r*(C–O)	1	29	0.92	0.62	0.92	—	—	—	—
B3LYP	*anti*	*r*(O–H)	1	29	0.83	0.91	0.83	—	—	—	—
B3LYP	*anti*	All	2	25	0.92	0.33	0.76	—	—	—	—
B3LYP	*anti*	*r*(C–O)	1	25	0.92	0.32	0.91	—	—	—	—
B3LYP	*anti*	*r*(O–H)	1	25	0.37	0.90	0.35	—	—	—	—

Two different internal hydrogen bonds can be formed, *i.e.* O···H–N and O–H···O. Because the latter structure corresponds to the lowest energy this was the only conformation considered. The three compounds that did not have this extended common skeleton are shown in Fig. 5.

The observed-*versus*-predicted plots and the statistics (Table 6) for the models containing all 26 compounds were inspected. The three compounds, 2-hydroxybenzamide (compounds 59), methyl salicylate (compound 61) and 2-vanillin (compound 62), which did not have the same common skeleton as the majority of the compounds in this high-correlation subset, appeared to be outliers. Methyl salicylate was identified as an outlier by Tehan *et al.*[54] One would not be surprised if this compound did not fit the models because of its similarity to compounds 59 and 62. The reason for these outliers could be the lack of structural similarity between these three compounds and the rest of the *o*-phenols capable of forming an internal hydrogen bond. We suggest that these compounds belong to their own high-correlation subset needed to predict pK_a using just r(C–O). This was confirmed by an r^2 value of 0.95 for the correlation of pK_a and r(C–O), although more data points for these types of compounds are needed to confirm this. The pK_a of the remaining 23 compounds were modelled using all the bond lengths (Table 6). Once again, the r(C–O) was most important in the VIP plot, however, it was followed by r(C$_2$-C$_3$) and not r(O–H). The r(C–O) model gave the best statistics for the original model and CV statistics compared to the all-bond-length model and the other single-bond-length models.

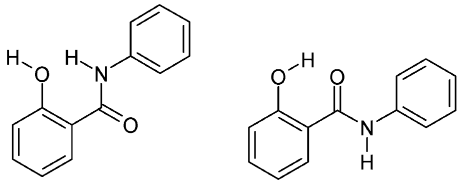
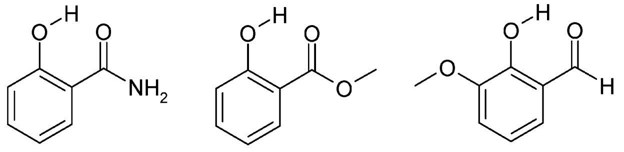


Fig. 4 The common skeletons and different hydrogen bonds than can be formed by the *o*-phenols capable of forming internal hydrogen bonds.



59 61 62

Fig. 5 The three compounds in the class of phenols capable of forming internal hydrogen bonds that did not have the same extended common skeleton as the rest of the compounds.

3.2.5 *o*-Methoxy/ethoxyphenols.

This high-correlation subset consisted of only eight compounds. The pK_a range was small (7.4 for vanillin (compound 124) to 10.28 for 4-methyl-2-methoxyphenol (compound 104)). Removing vanillin, which had a much lower pK_a value than the rest, resulted in a range of only 0.74 log units. The syn conformation is the lowest energy in all cases. According to the statistics, the models deteriorate when vanillin is included (Table 7). To increase the size of the dataset we sourced 21 compounds from Ragnar *et al.*[64] Five of these compounds were already present in our dataset. A comparison of the given pK_a values in that publication and in our dataset showed they were in good agreement, the largest difference being 0.05 pK_a units. We used the 16 remaining compounds as a test set for the syn models. It was pleasing to note that including vanillin gave lower values for RMSEP in all cases and that the r(C–O) bond length model gave the lowest RMSEP (Table 8). The models created without vanillin had rather poor CV statistics (*i.e.* q^2) because of the small pK_a range. However, these models actually produced reasonable predictions for the test set. We added the 16 compounds to the Tehan compounds and created new models containing more compounds to increase the domain of applicability of the model (Table 7).

3.2.6 Miscellaneous *o*-phenols.

2-cyanophenol, 2-phenylphenol, 2-amino-4-nitrophenol and 2-aminophenol are the only representatives of these classes of *o*-phenol compounds. It is expected that these would produce separate high-correlation subsets but as there are few examples this was not investigated.

3.3 *Meta*- and *para*-phenols

The *meta*/*para* phenol models were already of high quality using just r(C–O) with an r^2 value of 0.87 and an RMSEE of 0.41, without taking into account conformation (Table 2). The r(C–O) and r(O–H) were the most important to the all-bond-length model according to the VIP plot. We investigated conformation to see if it was important as in the case of the *o*-phenols. Different conformations are only possible for the asymmetrical *meta*- and *meta*-/*para*-phenols. Different conformations based on the orientation of the phenolic hydrogen were optimised and an r(C–O) model was created using all the

Table 6 The statistical details of the model created using *o*-phenols capable of forming internal hydrogen bonds

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	O–H–O	All	2	26	0.83	0.71	0.80	—	—	—	—
HF	O–H–O	r(C–O)	1	26	0.77	0.81	0.76	—	—	—	—
HF	O–H–O	r(O–H)	1	26	0.30	1.41	0.27	—	—	—	—
HF	O–H–O	r(C$_2$-C$_3$)	1	26	0.72	0.90	0.69	—	—	—	—
HF	O–H–O	All	1	23	0.88	0.50	0.83	0.79	0.74	0.74	0.71
HF	**O–H–O**	**r(C–O)**	**1**	**23**	**0.95**	**0.32**	**0.95**	**0.33**	**0.94**	**0.94**	**0.92**
HF	O–H–O	r(O–H)	1	23	0.35	1.16	0.23	1.31	0.15	−1.50	−0.04
HF	O–H–O	r(C$_2$-C$_3$)	1	23	0.85	0.56	0.82	0.62	0.80	0.78	0.69

Table 7 The statistical details of the model created using *o*-methoxy/ethoxyphenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{\mathrm{CV},0}$	r^2_m
HF	*syn*	All	1	8	0.76	0.52	0.25	0.75	0.42	0.25	0.24
HF	*syn*	*r*(C–O)	1	8	0.87	0.37	0.85	0.63	0.83	0.78	0.64
HF	*syn*	*r*(O–H)	1	8	0.82	0.44	0.82	0.44	0.77	0.67	0.52
HF	*syn*	All	1	7 (−124)	0.76	0.26	0.14	—	—	—	—
HF	*syn*	*r*(C–O)	1	7 (−124)	0.47	0.39	0.17	—	—	—	—
HF	*syn*	*r*(O–H)	1	7 (−124)	0.15	0.49	−0.10	—	—	—	—
HF	*anti*	All	1	8	0.82	0.44	0.29	—	—	—	—
HF	*anti*	*r*(C–O)	1	8	0.88	0.36	0.86	—	—	—	—
HF	*anti*	*r*(O–H)	1	8	0.91	0.31	0.90	—	—	—	—
HF	*anti*	All	1	7 (−124)	0.82	0.23	0.44	—	—	—	—
HF	*anti*	*r*(C–O)	1	7 (−124)	0.51	0.37	0.09	—	—	—	—
HF	*anti*	*r*(O–H)	1	7 (−124)	0.64	0.32	0.28	—	—	—	—
HF	*syn*	All	1	24	0.84	0.39	0.79	0.82	0.39	0.19	0.21
HF	***syn***	***r*(C–O)**	**1**	**24**	**0.91**	**0.29**	**0.89**	**0.53**	**0.69**	**0.57**	**0.45**
HF	*syn*	*r*(O–H)	1	24	0.85	0.37	0.83	0.58	0.61	0.43	0.35

Table 8 The results of testing 16 methoxyphenols in the methoxy/ethoxyphenol models

					Model statistics			16 Compound test set
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP
HF	*syn*	All	1	8	0.76	0.52	0.25	0.43
HF	*syn*	*r*(C–O)	1	8	0.87	0.37	0.85	0.30
HF	*syn*	*r*(O–H)	1	8	0.82	0.44	0.82	0.41
HF	*syn*	All	1	7 (−124)	0.76	0.26	0.14	0.53
HF	*syn*	*r*(C–O)	1	7 (−124)	0.47	0.39	0.17	0.77
HF	*syn*	*r*(O–H)	1	7 (−124)	0.15	0.49	−0.10	0.42

conformations and all the compounds. The differences between the predicted pK_a values for the same compounds in the different conformations were calculated for the 55 *m*/*p*-substituted phenols. The average difference was found to be less than 0.1 log unit. For this reason we decided that conformational differences would not be considered in the subsequent investigations for the *meta*- and *para*-phenols. After modelling the *para*-phenols and *meta*-phenols separately and finding little improvement to the models we investigated high-correlation subsets. The dataset contained 6 nitrophenols, including 3-trifluoromethyl-4-nitrophenol and 3-nitro-4-cresol, 14 halogen phenols, including 3-trifluoromethylphenol, 4-trifluoromethylphenol, 4-chloro-3,5-dimethylphenol, 3-methyl-4-chlorophenol, 15 alkylphenols, 5 methoxy/ethoxyphenols, 2 hydroxybenzaldehydes, 2 hydroxyacetophenones, and 11 compounds we classed as miscellaneous, which included compounds such as *m*/*p*-cyanophenol, *m*/*p*-phenylphenol and *m*/*p*-aminophenol. We investigated the nitro, halogen and alkylphenols to see if treating these classes of compounds separately produced high-correlation subsets.

3.3.1 *m*-/*p*-Nitrophenols. The *r*(O–H) model produced the highest correlation and the lowest RMSEE (Table 9). As there were only 6 compounds we tested the model using 5 compounds for which pK_a values could be found in the literature. These were 3-methyl-4-nitrophenol, 3,5-dimethyl-4-nitrophenol and 3-chloro-4-nitrophenol, 3-fluoro-4-nitrophenol and 3,5-difluoro-4-nitrophenol.[65–67] The RMSEP of prediction for these compounds is shown in Table 10. The results are above the 0.5 pK_a unit threshold that we aim for but it must be considered that there are no halogen-substituted compounds in the training set.

3.3.2 *m*-/*p*-Halogen phenols. The all-bond-length model produced the best statistics, however, the *r*(C–O) model was very similar in terms of RMSEE (Table 11). Three compounds were tested in the models with the RMSEP shown in Table 12. The test compounds were 3-chloro-4-nitrophenol, 3,5-difluoro-4-nitrophenol and 3,5-difluoro-4-nitrophenol. We acknowledge that three compounds is very small for a test set but this consists of approximately twenty percent of the training set. Their predictions are extrapolations because no compounds in the training set are stronger acids. The predictions are poor, suggesting that the nitro group has the greatest effect and they should be predicted by the nitro model.

3.3.3 *m*-/*p*-Alkylphenols. Modelling of the alkylphenols was attempted but as Table 13 shows, this proved unsuccessful because of the small pK_a range (0.53 pK_a units) of this class. For these compounds the best prediction would come from using the mean pK_a value of this high-correlation subset (10.2) knowing that the error is ~0.25 pK_a units.

3.3.4 Comparison of the models created for the high-correlation subsets of phenols to those constructed using different subsets of all the phenols. Table 14 provides the statistics for different subsets of *o*-phenols to compare to the predictions from the high-correlation subset models constructed separately for *o*-nitro, *o*-halogen, *o*-alkyl, *o*-methoxy/ethoxy and the *o*-phenols capable of forming internal hydrogen bonds. We performed this analysis to prove that the predictions from the single-bond-length high-correlation subset models were better than those made by models constructed using all the *o*-phenols and all bond lengths. The eight outliers that were identified

Table 9 The statistical details of the model created using *m*-/*p*-nitrophenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{\mathrm{CV},0}$	r^2_m
HF	N/A	All	2	6	0.98	0.25	0.91	0.52	0.83	0.83	0.83
HF	N/A	*r*(C–O)	1	6	0.90	0.45	0.89	0.47	0.83	0.81	0.69
HF	N/A	*r*(O–H)	1	6	0.98	0.21	0.97	0.27	0.95	0.95	0.94

Table 10 The statistics relating to the 5-compound test set

					Model statistics			5-Compound test set
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP
HF	N/A	All	2	6	0.98	0.25	0.91	0.62
HF	N/A	*r*(C–O)	1	6	0.90	0.45	0.89	0.64
HF	N/A	*r*(O–H)	1	6	0.98	0.21	0.97	0.62

Table 11 The statistical details of the model created using *m*-/*p*-halogen phenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{\mathrm{CV},0}$	r^2_m
HF	N/A	All	2	14	0.93	0.17	0.78	0.32	0.74	0.73	0.66
HF	**N/A**	***r*(C–O)**	**1**	**14**	**0.86**	**0.23**	**0.81**	**0.30**	**0.76**	**0.75**	**0.69**
HF	N/A	*r*(O–H)	1	14	0.79	0.29	0.70	0.38	0.65	0.63	0.55

Table 12 Results of the 3-compound test set

					Model statistics			3-Compound test set
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP
HF	N/A	All	2	14	0.93	0.17	0.78	1.05
HF	N/A	*r*(C–O)	1	14	0.86	0.23	0.81	1.06
HF	N/A	*r*(O–H)	1	14	0.79	0.29	0.70	1.46

Table 13 The statistical details of the model created using *m*-/*p*-alkylphenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{\mathrm{CV},0}$	r^2_m
HF	N/A	All	2	15	0.38	0.13	0.13	0.18	0.02	−1.94	−0.01
HF	**N/A**	**C–O**	**1**	**15**	**0.19**	**0.15**	**0.15**	**0.15**	**0.09**	**−3.04**	**−0.07**
HF	N/A	O–H	1	15	0.02	0.16	−0.04	0.17	0.04	−19.52	−0.15

from the high-correlation subsets have been removed to give a fair comparison. The lowest energy conformation was used for all the compounds. The models created for all the *o*-phenols with the eight outliers removed (116 compounds − 8 outliers = 108 compounds) have lower RMSEEs than the models created with the outliers included (Table 2). Removal of the miscellaneous compounds has only a small effect on the statistics. However, the models improve slightly when the *o*-phenols capable of forming internal hydrogen bonds are removed. In all cases the internal statistics and CV statistics are the best for the models created using all the bond lengths compared to those created using *r*(C–O) and *r*(O–H). The CV statistics confirm that the models created using *r*(C–O) are better than those that created using *r*(O–H).

Table 15 provides the statistics for different subsets of *m*/*p*-phenols to compare to the predictions from the high-correlation subsets. The internal model statistics are the same as those given in Table 2 since no outliers were identified. Here we also provide the CV statistics for these models. The all-bond-length model has the best statistics followed by the *r*(C–O) and the *r*(O–H) models, respectively. The CV statistics confirm that models of high quality and the RMSEP is below 0.5 pK_a units for all the models. An improvement in the models is noticeable when only the *m*-/*p*-nitro, halogen and alkylphenols investigated as high-correlation subsets are used to construct models. The all-bond-length, *r*(C–O) and *r*(O–H) models display virtually the same statistics.

Table 16 provides the statistics for the models created using all the phenols without the eight *o*-phenol outliers identified

Table 14 The statistics relating to the models constructed for subsets of *o*-phenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest energy	All	4	108 (- outliers identified in previous sections)	0.93	0.67	0.90	0.81	0.90	0.89	0.83
HF	Lowest energy	*r*(C–O)	1	108	0.88	0.89	0.88	0.89	0.88	0.86	0.77
HF	Lowest energy	*r*(O–H)	1	108	0.53	1.76	0.52	1.76	0.52	0.14	0.20
HF	Lowest energy	All	4	104 (ibid but without misc compounds)	0.95	0.60	0.92	0.72	0.92	0.92	0.87
HF	Lowest energy	*r*(C–O)	1	104	0.89	0.89	0.89	0.88	0.88	0.87	0.78
HF	Lowest energy	*r*(O–H)	1	104	0.54	1.76	0.54	1.76	0.53	0.18	0.22
HF	Lowest energy	All	5	81 (ibid with *o*-phenols IHB)	0.96	0.57	0.93	0.67	0.94	0.93	0.88
HF	Lowest energy	*r*(C–O)	1	81	0.89	0.92	0.89	0.91	0.89	0.87	0.79
HF	Lowest energy	*r*(O–H)	1	81	0.62	1.68	0.62	1.68	0.61	0.41	0.34

Table 15 The statistics relating to the models constructed for subsets of *m*-/*p*-phenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	N/A	All	2	55	0.91	0.34	0.87	0.37	0.89	0.88	0.81
HF	**N/A**	***r*(C–O)**	**1**	**55**	**0.87**	**0.41**	**0.85**	**0.43**	**0.85**	**0.83**	**0.72**
HF	N/A	*r*(O–H)	1	55	0.84	0.45	0.83	0.48	0.82	0.78	0.66
HF	N/A	All	2	35 (nitro, halogen, alkyl)	0.96	0.26	0.94	0.30	0.94	0.94	0.88
HF	N/A	*r*(C–O)	1	35	0.95	0.30	0.94	0.31	0.94	0.94	0.89
HF	N/A	*r*(O–H)	1	35	0.95	0.29	0.95	0.30	0.95	0.94	0.90

from the high-correlation subsets. Small improvements in the internal statistics are observed compared to the models with the eight outliers included in Table 2. The all-bond-length models has the best CV statistics followed by *r*(C–O) and *r*(O–H), respectively.

We now compare the predictions made from the high-correlation subsets to those made by the models constructed from combinations of compounds from the high-correlation subsets and the models constructed with *all* the phenol compounds (171 − 8(outliers) = 163). To compare the predictions from the high-correlation-subsets to those obtained from the different subset models of the *o*-phenols, *m*-/*p*-phenols and all the phenols shown in Table 14–16, respectively, we use the RMSEP. From the CV of the different models we calculated the RMSEP for only the compounds that belonged to high-correlation subsets. For the *o*-phenols this involved five high-correlation subsets (*i.e.* *o*-nitro, *o*-halogen, *o*-alkyl, *o*-phenols capable of forming IHB and *o*-methoxy/ethoxy) and three different models (*i.e.* all-bond-length, *r*(C–O) and *r*(O–H)). We then calculated the mean RMSEP from all-bond-length, *r*(C–O) and *r*(O–H) models constructed using all the compounds that belonged to high-correlation subsets. This provided three average RMSEP values. This was repeated for only the compounds that formed high-correlation subsets from the models constructed from the *o*-phenols without the *o*-phenols capable of forming internal hydrogen bonds and miscellaneous compounds, all the *o*-phenols without the miscellaneous compounds, all the *o*-phenols (Table 14) and all the phenols (Table 16). The average RMSEP values obtained for the compounds belonging to high-correlation subsets are given in Table 17. Note that the models constructed without the *o*-phenols capable of forming internal hydrogen-bonds were not used to calculate an RMSEP for this high-correlation subset. Therefore the value is based on the remaining four high-correlation subset compounds. One concludes that the RMSEP from the high-correlation subsets is lower in all cases and the *r*(C–O) models provide the lowest RMSEP compared to the all-bond-length models and the *r*(O–H) model. In other words, better predictions are made by the high-correlation subsets and a single *ab initio* bond length is able to predict pK_a of *o*-phenols.

Table 18 compares the RMSEP values, obtained for the *m*-/*p*-phenols identified as forming high-correlation subsets, to the RMSEP values, obtained from models constructed from more phenols. The improvement in the RMSEP from using the predictions made by the high-correlation subsets is much less than that observed for the *o*-phenols. However, it is interesting to note that both the *r*(C–O) and *r*(O–H) models have lower RMSEP than the models constructed from all the bond lengths. These results suggest that either *r*(C–O) or *r*(O–H) can be used to predict the pK_a for *m*-/*p*-phenols.

Table 16 The statistics relating to the models constructed using all the phenols

					Model statistics		CV statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest energy	ALL 8	4	163	0.93	0.63	0.91	0.68	0.92	0.91	0.84
HF	Lowest energy	C–O	1	163	0.89	0.80	0.88	0.80	0.88	0.87	0.78
HF	Lowest energy	O–H	1	163	0.57	1.55	0.56	1.55	0.56	0.24	0.25

Table 17 The average RMSEP for the *o*-phenols predicted from the relevant models

# Bonds	Compounds used to build models All phenols	All *o*-phenols	All *o*-phenols without miscellaneous *o*-phenols	All Ortho without miscellaneous *o*-phenols and *o*-phenols capable of forming IHB	High-correlation subsets
All	0.69	0.77	0.72	0.73	0.58
r(C–O)	0.83	0.81	0.81	0.84	0.42
r(O–H)	1.67	1.68	1.70	1.66	0.85

Table 18 The average RMSEP for the *m*/*p*-phenols predicted from the relevant models

# Bonds	Compounds used to build models All phenols	All *m*-/*p*-phenols	Nitro, halogen and alkyl phenols	High-correlation subsets
All	0.41	0.36	0.32	0.34
r(C–O)	0.67	0.39	0.34	0.31
r(O–H)	1.21	0.43	0.30	0.27

3.4 Comparison of the correlation obtained with and without an ammonia probe

As mentioned in the Introduction, the complete series of chlorophenols, bromophenols and fluorophenols has previously been investigated separately for correlations of molecular properties with pK_a by Tao and co-workers.[50–53] Having demonstrated that strong correlations between pK_a and one bond length can be achieved for halogen phenols we queried whether better results could be achieved using an ammonia probe and a higher level of theory. For all the monomeric *o*-halogen phenols in the dataset, the syn conformation was the lowest in energy, apart from the two 2-halogen-6-nitrophenols. The authors stated that when a probe molecule is introduced, the anti conformer (where the hydroxyl hydrogen points away from the closest halogen) is the one with the lowest energy.[51,53] This can only be the case for mono-*ortho*-substituted halogen phenols. The ammonia, as a probe molecule, is positioned with its lone pair pointing directly towards the hydroxyl hydrogen of the halogen phenols, conserving the C_s symmetry if other *meta*-/*para*-substituents are ignored (Fig. 6).

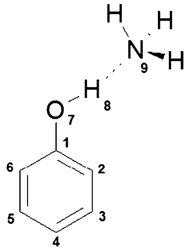


Fig. 6 The general structure and number scheme for the phenol-ammonia complex.

Full geometry optimisations were performed at the HF/6-31G(d) and B3LYP/6-311++G(d,p) level of theory on the data set of *o*-halogen phenols (30 compounds), without the two 2-halogen-6-nitrophenols, in the presence of the ammonia probe. For the asymmetric halogen phenols geometry optimisations were performed on both the syn and anti conformers. Contrary to the calculated energies of the monomeric halogen phenols, where the syn conformation was consistently lower in the energy, the same was *not* found for the halogen phenol-ammonia complexes. For the asymmetric halogen phenols, at both levels of theory, we generally found the syn conformation to have the lowest energy. However, in some cases the lowest energy conformer was not consistent at both levels of theory for the same compound. These findings were inconsistent with the work of Han *et al.*[50] For example, we find the syn conformer of 2-chlorophenol to have the lowest energy using both levels of theory and including the basis-set superposition error (BSSE) in the HF calculation. For this reason we constructed models based on the three possible combinations: each compound being in its lowest energy conformer, each in its syn conformer and each in its anti conformer (Table 19) (symmetrically substituted compounds such as 2,6-dichlorophenol can of course not be assigned anti or syn but this fact did not exclude them from the dataset). The four compounds previously identified as outliers were still outliers in the models even after the introduction of the probing ammonia. This can be seen by an improvement in all the models statistics when the outliers are removed.

These results indicate that there is no need to use the more expensive B3LYP/6-311++G(d,p) level of theory as the models generated using HF/6-31G(d) are of equal and sometimes of superior quality. The lowest RMSEEs are produced by the all-bond-length models followed by the *r*(C–O), *r*(O–H) and *r*(O–H···N) models, respectively. This is confirmed by the VIP plot for the all-bond-length models ranking the importance of these bond lengths to the models in the same order. In most cases the difference between the statistics for the all-bond-length models and the *r*(C–O) models is small, suggesting that the single *r*(C–O) models are suitable for predicting pK_a of halogen phenols. Considering the single-bond-length models, the syn conformation generally produces the lowest RMSEEs. This is because the influence of the *o*-halogen substitution is constant for each complex considered. For the lowest-energy and the anti-conformation models, the influence is not constant and therefore corrupts the correlations. For example, the anti conformation models created using *r*(O–H···N) have the highest RMSEE. This is caused by 2-chloro-6-methylphenol (compound 90). The presence of the methyl group causes the O–H···N bond length to be much longer than it is for

Table 19 The statistics relating to the models constructed using an ammonia probe

			HF				B3LYP			
Conformation	Bonds	# Compounds	# LV	r^2	q^2	RMSEE	# LV	r^2	q^2	RMSEE
Lowest energy										
	All	30	4	0.97	0.92	0.31	2	0.94	0.88	0.42
	C–O	30	1	0.90	0.90	0.51	1	0.91	0.90	0.50
	O–H	30	1	0.81	0.78	0.71	1	0.73	0.70	0.84
	O–H···N	30	1	0.54	0.44	1.11	1	0.56	0.49	1.09
	All	26	3	0.99	0.98	0.18	2	0.98	0.96	0.26
	C–O	26	1	0.97	0.97	0.27	1	0.95	0.95	0.36
	O–H	26	1	0.94	0.93	0.42	1	0.88	0.87	0.59
	O–H···N	26	1	0.72	0.70	0.89	1	0.77	0.76	0.81
Syn										
	All	30	4	0.97	0.92	0.30	2	0.95	0.89	0.37
	C–O	30	1	0.90	0.90	0.51	1	0.91	0.91	0.49
	O–H	30	1	0.82	0.79	0.69	1	0.75	0.72	0.82
	O–H···N	30	1	0.59	0.48	1.04	1	0.57	0.51	1.08
	All	26	2	0.98	0.96	0.25	2	0.99	0.97	0.20
	C–O	26	1	0.98	0.98	0.25	1	0.96	0.96	0.34
	O–H	26	1	0.93	0.92	0.44	1	0.90	0.89	0.54
	O–H···N	26	1	0.87	0.86	0.61	1	0.78	0.78	0.78
Anti										
	All	30	1	0.87	0.84	0.58	2	0.93	0.87	0.43
	C–O	30	1	0.91	0.91	0.49	1	0.91	0.91	0.49
	O–H	30	1	0.85	0.82	0.62	1	0.85	0.82	0.65
	O–H···N	30	1	0.46	0.38	1.21	1	0.61	0.55	1.02
	All	26	3	0.98	0.95	0.24	2	0.97	0.95	0.28
	C–O	26	1	0.97	0.97	0.29	1	0.95	0.94	0.39
	O–H	26	1	0.90	0.84	0.54	1	0.93	0.90	0.46
	O–H···N	26	1	0.50	0.30	1.18	1	0.70	0.60	0.91

the other compounds, presumably because of steric hindrance and repulsion between the hydrogens. The slope of the regression between pK_a and r(O–H···N) is positive. Therefore, a longer bond length causes a higher pK_a prediction. This means 2-chloro-6-methylphenol has a predicted pK_a that is much higher than the experimental pK_a value. Han and co-workers[50,51] found that separate correlations were required for pK_a with r(O–H···N) for di-ortho halogen phenols because of steric interference. By using the syn conformation every compound is exposed to steric interference from the *o*-halogen substitution, although it appears that separate correlations may still be needed for r(O–H···N) and possibly r(O–H). Pentabromophenol (compound 142) corrupted the correlations for all the r(O–H···N) models. Han and Tao[51] excluded this compound from their equations on the basis that the full geometry optimisation of its complex with ammonia had not converged (note that our geometry optimisation of this complex did converge though). Removing this compound from the r(O–H···N) model with the syn conformations, which is the best r(O–H···N) correlation, did not improve it enough to be better than the r(C–O) model.

This investigation into using an ammonia probe demonstrates that single bond-lengths can be used to predict the pK_a of *o*-halogen phenols. The results obtained from HF/6-31G(d) and B3LYP/6-311++G(d,p) are comparable. The use of r(C–O) with the syn conformation produced the best statistics for the single-bond-length models and has the advantage of avoiding erratic predictions caused by non-halogen *ortho*-substitutions for di-orthophenols and the need for separate correlations for di-*ortho*-halogenated phenols. However, comparing the r(C–O) model to that obtained using the monomeric phenols where an r^2, q^2 and RMSEE of 0.97, 0.97, and 0.27, were obtained respectively, the use of an ammonia probe is unnecessary considering the increase in time taken to perform the geometry optimisation. Large improvements are seen in the models using the r(O–H) bond length obtained from the *o*-halogen phenol-ammonia complex compared to those models obtained from the monomeric halogen phenols. However, the improvements are not strong enough to make the use of a probe the preferred option over the monomeric r(C–O) model.

4. Discussion

The phenols data set has been deconstructed by identifying high-correlation subsets where one *ab initio* bond length, calculated at the HF/6-31G(d) level of theory, is able to predict pK_a. Lower RMSEEs were found when they were modelled as *o*-phenols and *m*-/*p*-phenols separately. The *o*-phenols had an RMSEE greater than 0.5 pK_a units when they were modelled together using all the bond lengths. However, this highlighted that r(C–O) and r(O–H) were most important to the model. When r(C–O) and r(O–H) were used alone to predict pK_a, the models drastically reduced in quality. Subsequent analysis of the observed-*versus*-predicted plots for these two bond lengths revealed the possibility of improving the predictions by further deconstructing the data set into high-correlation subsets. It was pleasing to note that the high-correlation subsets, identified by eye, were chemically meaningful. These high-correlation subsets, which included *o*-nitrophenols, *o*-halogen-phenols, *o*-alkylphenols, *o*-phenols capably of forming IHBs and *o*-methoxy/ethoxyphenols, were fully analysed by comparing all-bond-length models to single-bond-length models. All-bond-length models differ from single-bond-length models in their capacity to highlight outliers. Outliers are readily exposed in single-bond-length models, where they cannot benefit from the fitting flexibility offered by all-bond-length models. In other words, the simplicity of the single-bond-length models calls for the obligatory investigation of a number of suspicious compounds. The majority of these compounds could be explained by wrong conformations, erroneous experimental pK_a values and structural differences with the rest of the compounds in the high-correlation subset. In most cases, r(C–O) models were the best, compared to all-bond-length models and r(O–H) models.

The *m*-/*p*-phenol models for r(C–O) and r(O–H), constructed using all the compounds, were comparable to the all-bond-length model, which was not the case for the *o*-phenols. However, because improved models were found by separating the *o*-phenols into high-correlation subsets, the same separation was carried out for the *m*-/*p*-phenols. Small improvements were noted but these were not comparable to the improvements seen for the high-correlation subsets of *o*-phenols. For the phenols, r(C–O) consistently provided the best models.

Through analysis of the phenol data set, we proposed rules to decide in which conformation the phenols need to be optimised in order to make the best possible prediction. Secondly, these rules also determined which high-correlation subset scores the best prediction. In the cases of the phenols, these rules were decided based on the energy of each compound. We note that no compound violated these rules.

In this study our emphasis has been on accuracy rather than globality. The results demonstrate that one bond length from the phenols can be used to predict pK_a. In order for this to work however, high-correlation subsets needed to be identified and treated separately. Generally, properties of *ortho*-substituted compounds are notoriously more difficult to predict than *m*-/*p*-substituted compounds. This has been demonstrated by the vast improvements in the statistics of the ortho models when the high-correlation subsets were identified. While high-correlation subsets were identified for the *m*-/*p*-phenols, the improvements in modelling these separately were minor. We have shown for phenols that the use of an ammonia probe of higher level of theory offers no advantage. It is important to postulate how large a high-correlation subset needs to be in order for the correlation to be meaningful. In principle two compounds are sufficient to obtain an equation but of course its reliability would be questionable. One can think of the proposed method in a "dynamic" way. A very small number of compounds can split off as a high-correlation subset from a larger dataset. As new experimental data becomes available this high-correlation subset can grow to a more comfortable

Table 20 The high-correlation subsets with associated equations to predict pK_a (bond lengths in Bohr), the number of compounds used to construct the equation and relevant statistics

Compound class	High-correlation subset	Equation	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
Phenols	*o*-Nitro phenols	pK_a = 137.575 r(C–O) − 337.575	22	0.94	0.48	0.94	0.50	0.93	0.93	0.87
	o-Halogen phenols	pK_a = 147.411 r(C–O) − 365.540	26	0.97	0.27	0.97	0.29	0.97	0.97	0.94
	o-Alkyl phenols	pK_a = 162.106 r(C–O) − 405.207	25	0.91	0.34	0.90	0.37	0.89	0.87	0.78
	o-Phenols-IHB	pK_a = 160.912 r(C–O) − 397.578	23	0.95	0.32	0.95	0.33	0.94	0.94	0.92
	o-Methoxy/ethoxyphenols	pK_a = 128.767 r(C–O) − 318.646	24	0.91	0.29	0.89	0.53	0.69	0.57	0.45
	m-/*p*-Phenols	pK_a = 122.985 r(C–O) − 304.553	55	0.87	0.41	0.85	0.43	0.85	0.83	0.72

Table 21 Details of methods from the literature (including ACD and SPARC) used to predict pK_a for phenols with their associated statistics. The values in parentheses represent test set statistics and LR = linear regression, MLR = multi-linear regression and QTMS = Quantum Topological Molecular Similarity

Ref.	Phenols	# Compounds	# Descriptors	Method	Level of Theory	r^2	q^2	RMSE/RMSEP
54[a]	All	175	1	LR	AM1	0.81	0.81	—[d]
	All	175	4	MLR	AM1	0.93	0.93	—
	Meta/*Para*	58	1	LR	AM1	0.92	0.91	0.32
	Ortho-phenols-IHB	26	1	LR	AM1	0.83	0.81	0.69
	Ortho	91	1	LR	AM1	0.92	0.91	0.75
68[b]	All	62 (18)	6	QTMS	B3LYP/6-31+G(d,p)	0.83 (0.91)	0.79	(0.58)
69[c]	*Meta*/*Para*	79	1	LR	AM1	0.87	0.87	0.42
	Meta/*Para*	79	1	*r*-LR	AM1	0.87	0.87	0.42
	Meta/*Para*	79	—	ACD	—	0.93	0.93	0.31
	Meta/*Para*	79	—	SPARC	—	0.95	0.95	0.25
	Ortho-phenols-IHB	29	1	LR	AM1	0.76	0.75	0.90
	Ortho-phenols-IHB	29	1	*r*-LR	AM1	0.77	0.77	0.87
	Ortho-phenols-IHB	29	—	ACD	—	0.84	−0.10	1.88
	Ortho-phenols-IHB	29	—	SPARC	—	0.86	0.62	1.11
	Ortho	116	1	LR	AM1	0.90	0.90	0.78
	Ortho	116	1	*r*-LR	AM1	0.90	0.90	0.78
	Ortho	116	—	ACD	—	0.95	0.95	0.55
	Ortho	116	—	SPARC	—	0.96	0.96	0.52

[a] The majority of experimental pK_a values used in this work is taken from this publication. [b] The QTMS method was used to predict pK_a values in this reference. [c] A publication expanding the data set used in Tehan *et al.*[54] and this work. The original models were used to predict the pK_a values of the added compounds. Subsequently the models were recalibrated with the addition of the new compounds and comparisons with the original models, ACD and SPARC were made. Hence, in the method column "r-LR" stands for recalibrated-Linear regression. [d] A dash in the table means the information is not available or relevant.

number (*i.e.* > 20) (as typically appearing in this work). There is a tension between the reliability of a model and the size of a high- correlation subset giving rise to the model. The desire to construct models with a larger number of compounds may lead to grouping several high-correlation subsets into one more "global" model. Although the increased number will benefit reliability, the corresponding accuracy may deteriorate because of the loss of locality.

In Table 20 we present the single bond-length equations for the phenol high-correlation subsets that can be applied to the prediction of pK_a for suitable compounds. Table 21 provides literature references for studies of phenols along with details of the methods and statistics. We note the ambiguity in the way errors are reported in the literature when simple stated as "rms". For example, it is not clear if the errors corresponding to SPARC and ACD derive from CV, from an external test set or even the predicted compounds being in the training set. This is why the utmost right column is headed by "RMSE/RMSEP". The large error values for *ortho*-substituted phenols are striking, especially when an intramolecular hydrogen bond is involved. Secondly, increasing the data set size we used and recalibrating the model made little or no difference to the errors. This suggests that levels of theory beyond AM1, the use of bond lengths as descriptors, and the introduction of high-correlation subsets all improve prediction.

5. Conclusions

We have investigated the prediction of pK_a for phenols. We succeeded in constructing models able to predict pK_a within 0.5 pK_a units using a single bond length from a monomer geometry optimised by an affordable and sufficiently reliable *ab initio* method, which was determined to be HF/6-31G(d). We achieved this by grouping molecules into high-correlation subsets, which were visually identified from observed-*versus*-predicted plots. It was pleasing to note that the structures in each subset contained a common substitution pattern, *e.g.* an OH group adjacent to a NO_2 group. Improvements in model statistics are small for high-correlation subsets of *meta-*/*para*-substituted phenols compared to a single model containing all these compounds. However, for *ortho*-substituted phenols, the statistics of high-correlation subsets improve much compared to a single model for all ortho phenols.

In the majority of cases, the models constructed from a single bond length were superior or, at the very least, similar to the models constructed using all the bond lengths. In each all-bond-length model, the most important bonds (*i.e.* those with the highest VIP value) were associated with the OH functional group where the deprotonation occurs. The use of an ammonia probe or a higher level of theory for the *o*-halogen phenols provided no advantage over the use of single bond lengths generated for the monomer at HF/6-31G(d). We have listed six single-bond-length equations from which the pK_a of relevant compounds can be predicted. In a subsequent publication (Part 3) we extend this list of equations to include benzoic acids and anilines.

Acknowledgements

The work was funded by the EPSRC with sponsorship from GlaxoSmithKline (GSK), Stevenage, Great Britain. Thanks to Dr M Kranz at GSK for helpful discussions.

References

1 J. I. Wells, *Pharmaceutical Preformulation*, Ellis Horwood Ltd, 1998, p. 25.
2 J. Comer and K. Tam, in *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*, ed. B. Testa, H. van de Waterbeemd, G. Folkers and R. Guy, Wiley-VCH, 2001, pp. 275–304.
3 A. C. Lee and G. M. Crippen, *J. Chem. Inf. Model.*, 2009, **49**, 2013.
4 S. Jelfs, P. Ertl and P. Selzer, *J. Chem. Inf. Model.*, 2007, **47**, 450.
5 L. Xing and R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 796.
6 L. Xing, R. C. Glen and R. D. Clark, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 870.
7 A. C. Lee, J. Y. Yu and G. M. Crippen, *J. Chem. Inf. Model.*, 2008, **48**, 2042.
8 F. Milletti, L. Storchi, L. Goracci, S. Bendels, B. Wagner, M. Kansy and G. Cruciani, *Eur. J. Med. Chem.*, 2010, **45**, 4270.
9 J. Zhang, T. Kleinoder and J. Gasteiger, *J. Chem. Inf. Model.*, 2006, **46**, 2256.
10 C. O. da Silva, E. C. da Silva and M. A. C. Nascimento, *J. Phys. Chem. A*, 2000, **104**, 2402.
11 C. O. da Silva, J. B. da Silva and M. A. C. Nascimento, *J. Phys. Chem. A*, 1999, **103**, 11194.
12 M. D. Liptak and G. C. Shields, *J. Am. Chem. Soc.*, 2001, **123**, 7314.
13 M. D. Liptak and G. C. Shields, *Int. J. Quantum Chem.*, 2001, **85**, 727.
14 M. Namazian, M. Zakery, M. R. Noorbala and M. L. Coote, *Chem. Phys. Lett.*, 2008, **451**, 163.
15 H. Lu, X. Chen and C. G. Zhan, *J. Phys. Chem. B*, 2007, **111**, 10599.
16 F. Eckert, M. Diedenhofen and A. Klamt, *Mol. Phys.*, 2010, **108**, 229.
17 J. Ho and M. L. Coote, *Theor. Chem. Acc.*, 2009, **125**, 3.
18 S. Zhang, J. Baker and P. Pulay, *J. Phys. Chem. A*, 2010, **114**, 425.
19 S. Zhang, J. Baker and P. Pulay, *J. Phys. Chem. A*, 2010, **114**, 432.
20 A. Klamt and G. J. Schuurmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799.
21 J. C. Dearden, M. T. D. Cronin and D. C. A. Lappin, *J. Pharm. Pharmacol.*, 2007, **59**, A7.
22 ADME, ADME Boxes ⟨http://pharma-algorithms.com/webboxes/⟩.
23 VCCLAB ⟨http://www.vcclab.org/⟩.
24 ADMET Predictor ⟨http://www.simulations-plus.com/⟩.
25 Pipeline Pilot ⟨http://accelrys.com/⟩.
26 SPARC ⟨http://sparc.chem.uga.edu/sparc/⟩.
27 Marvin ⟨http://www.chemaxon.com/⟩.
28 QikProp ⟨http://www.schrodinger.com/⟩.
29 ACD/Labs ⟨http://www.acdlabs.com/home/⟩.
30 PALLAS ⟨http://www.compudrug.com/⟩.
31 CSpK_a ⟨http://www.chemsilico.com/⟩.
32 M. Meloun and S. Bordovska, *Anal. Bioanal. Chem.*, 2007, **389**, 1267.
33 G. T. Balogh, B. Gyarmati, B. Nagy, L. Molnar and G. M. Keseru, *QSAR Comb. Sci.*, 2009, **28**, 1148.
34 Epik ⟨http://www.schrodinger.com/⟩.
35 C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2009, **49**, 2801.
36 Jaguar ⟨http://www.schrodinger.com/⟩.
37 S. E. O'Brien and P. L. A. Popelier, *J. Chem. Soc., Perkin Trans. 2*, 2002, 478.
38 P. L. A. Popelier and P. J. Smith, *Eur. J. Med. Chem.*, 2006, **41**, 862.
39 P. L. A. Popelier, U. A. Chaudry and P. J. Smith, *J. Chem. Soc., Perkin Trans. 2*, 2002, 1231.
40 R. F. W. Bader and P. L. A. Popelier, *Int. J. Quantum Chem.*, 1993, **45**, 189.
41 P. L. A. Popelier, *Atoms in Molecules: An Introduction*, Pearson Education, 2000.
42 R. F. W. Bader, *Atoms in Molecules. A Quantum Theory*, Oxford Univ. Press, 1990.
43 P. L. A. Popelier, in *Quantum Chemical Topology: on Bonds and Potentials*, Heidelberg, Germany, 2005.
44 A. P. Harding, D. C. Wedge and P. L. A. Popelier, *J. Chem. Inf. Model.*, 2009, **49**, 1914–1924.

45 G. I. Hawe, I. Alkorta and P. L. A. Popelier, *J. Chem. Inf. Model.*, 2010, **50**, 87.
46 I. Mitra, P. P. Roy, S. Kar, R. Ojha and K. Roy, *J. Chemom.*, 2010, **24**, 22.
47 U. A. Chaudry and P. L. A. Popelier, *J. Org. Chem.*, 2004, **69**, 233.
48 S. Kar, A. P. Harding, K. Roy and P. L. A. Popelier, *SAR QSAR Environ. Res.*, 2010, **21**, 149.
49 T. Zhou, D. Huang and A. Caflisch, *Curr. Top. Med. Chem.*, 2010, **10**, 33.
50 J. Han, R. L. Deming and F. M. Tao, *J. Phys. Chem. A*, 2005, **109**, 1159.
51 J. Han and F. M. Tao, *J. Phys. Chem. A*, 2006, **110**, 257.
52 J. Han, R. L. Deming and F. M. Tao, *J. Phys. Chem. A*, 2004, **108**, 7736.
53 J. Han, I. Lee and F. M. Tao, *J. Phys. Chem. A*, 2005, **109**, 5186.
54 B. G. Tehan, E. J. Lloyd, M. G. Wong, W. R. Pitt, J. G. Montana, D. T. Manallack and E. Gancia, *Quant. Struct.–Act. Relat.*, 2002, **21**, 457.
55 D. D. Prankerd, in *Profiles of Drug Substances, Excipients and Related Methodology*, Elsevier Academic Press, 2007.
56 G. Schaftenaar and J. H. Noordik, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 123.
57 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople, in *GAUSSIAN 03*, Wallingford CT, 2004.
58 S. Wold, M. Sjostrom and L. Eriksson, in *Partial Least Squares Projections to Latent Structures (PLS) in Chemistry*, ed. P. v. R. Schleyer, Wiley, Chichester, GB, 1998.
59 UMETRICS, in SIMCA-P 10.0, www.umetrics.com, Umeå, Sweden, 2002.
60 D. J. Livingstone, *Data Analysis for Chemists*, Oxford University Press, 1995.
61 A. Golbraikh and A. Tropsha, *J. Mol. Graphics Modell.*, 2002, **20**, 269.
62 D. M. Hawkins, S. C. Basak and D. Mills, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 579.
63 V. Chis, *Chem. Phys.*, 2004, **300**, 1.
64 M. Ragnar, C. T. Lindgren and N. O. Nilverbrant, *J. Wood Chem. Technol.*, 2000, **20**, 277.
65 E. Chapman, M. C. Bryan and C. H. Wong, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 910.
66 B. Schafer and W. Engwald, *Fresenius' J. Anal. Chem.*, 1995, **352**, 535.
67 W. R. Vaughan, *J. Org. Chem.*, 1956, **21**, 1201.
68 K. Roy and P. L. A. Popelier, *J. Phys. Org. Chem.*, 2009, **22**, 186.
69 H. Yu, R. Kuhne, R.-U. Ebert and G. J. Schuurmann, *J. Chem. Inf. Model.*, 2010, **50**, 1949.