

Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds

Luca Settimo · Krista Bellman · Ronald M. A. Knegtel

Received: 18 June 2013 / Accepted: 14 October 2013 / Published online: 19 November 2013
© Springer Science+Business Media New York 2013

ABSTRACT

Purpose Assessment of the accuracy of experimental and theoretical methods of pKa determination for acids and bases as separate classes.

Methods Four literature pKa datasets were checked for errors and pKa values assigned unambiguously to a single acidic and/or basic ionisation centre. A new chemically diverse and drug-like dataset was compiled from high-throughput UV–vis spectrophotometry pKa data. Measured pKa values were compared with data obtained by alternative methods and predictions by the Epik, Chemaxon and ACD pKa DB software packages.

Results The pKa values of bases were considerably less accurately predicted than those of acids, in particular for structurally complex bases. Several new chemical motifs were identified for which pKa values were particularly poorly predicted. Comparison of pKa values obtained by UV–vis spectrophotometry and different literature sources revealed that low aqueous solubility and chromophore strength can affect the accuracy of experimental pKa determination for certain bases but not acids.

Conclusions The pKa prediction tools Epik, Chemaxon and ACD pKa DB provide significantly less accurate predictions for bases compared to acids. Certain chemical features are underrepresented in currently available pKa data sets and as a result poorly predicted. Acids and bases need to be considered as separate classes during pKa predictor development and validation.

KEY WORDS bases · pKa measurement · pKa prediction · solubility · UV–vis spectrophotometry

ABBREVIATIONS

ADMET	Absorption, distribution, metabolism, excretion and toxicity
DMSO	Dimethyl sulfoxide
hERG	Human ether-a-go-go-related gene
HOMO	Highest occupied molecular orbital
LUMO	Lowest unoccupied molecular orbital
MAD	Median absolute deviation
MW	Molecular weight

INTRODUCTION

The acid–base dissociation constant of a molecule, commonly reported as its negative logarithm or pKa, is an important physical property to consider during the development of new drugs. For example, the pKa affects the affinity of a ligand to its receptor (1–3), pH-dependent aqueous solubility and the choice of suitable excipients and counterions during drug formulation (4).

The pKa of bases is of particular relevance because of their widespread use as solubilising groups. This is reflected in the high incidence of basic functionality in marketed drugs. Approximately 63% of drugs in the World Drug Index (5) are ionisable in the pH 2–12 range and respectively ~43% and ~12% of these contain a single basic or acidic centre (6,7). Ionization can profoundly affect the *in vivo* absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of drug candidates (7–14). Certain toxicities like the inhibition of the human ether-a-go-go-related gene (hERG) potassium ion channel (15–17) and potentially phospholipidosis (18) are sensitive to the pKa of a base. By modulating the overall polarity and electron density of

Electronic supplementary material The online version of this article (doi:10.1007/s11095-013-1232-z) contains supplementary material, which is available to authorized users.

L. Settimo · R. M. A. Knegtel (✉)
Vertex Pharmaceuticals (Europe) Ltd, 86-88 Jubilee Avenue
Milton Park, Abingdon OX14 4RW, UK
e-mail: ronald_knegtel@vrtx.com

K. Bellman
11010 Torreyana Road, San Diego, California 92121, USA

molecules variations in pKa can also significantly affect metabolism (13). The ability to correctly predict basic pKa values is therefore of considerable importance to the successful optimisation of lead molecules to development candidates.

The aim of this study is to assess the ability of the pKa prediction tools Epik (Schrödinger, New York, USA), Chemaxon (Chemaxon, Budapest, Hungary) and ACD pKa DB (ACDLabs, Toronto, Canada) to correctly predict the pKa of basic compounds that are representative of the chemical space explored in drug discovery. This requires careful consideration of issues concerning the quality and scope of published pKa datasets, experimental determination of pKa data in an industrial setting and the validation of pKa prediction tools.

Several pKa predictors have been developed that allow the pKa values of prospective molecules to be estimated prior to their synthesis (19–24). The validation of pKa prediction tools requires the availability of accurate experimental pKa data for compounds relevant to the desired area of application. The chemical content of published pKa datasets is, however, often less structurally complex than the chemical space explored in drug discovery. Many published pKa datasets contain data extracted from secondary sources. Sometimes pKa values were copied incorrectly from original references and information regarding experimental conditions is often lacking. In addition, the pKa values in these datasets are often not assigned to specific ionisation centres in each molecule. Predicted pKa values have therefore often been compared to the nearest experimental value in studies that assessed the performance of pKa predictors. This can result in overestimating the accuracy of the pKa prediction tool. pKa datasets containing only a single acid and/or base per compound are therefore preferred in order to enable the unambiguous assignment of predicted pKa values. Published assessments of the accuracy of pKa prediction algorithms usually only report the correlation coefficient (r^2) over the entire pKa range, without further distinction regarding the quality of predictions for acids and bases as separate classes (21–23). Whole dataset correlation coefficients can be misleading because the majority of pKa values for acids and bases usually lie below and above pKa = 7 (7). Good r^2 values can thus be obtained for an entire dataset, even if in reality a straight line has been fitted between two poorly defined clusters of acidic and basic pKa values. The currently available validation studies thus provide only limited insight into how accurately the pKa of bases can be predicted compared to acids.

In this study, the contents of four literature pKa datasets were verified using the original references and pKa values were unambiguously assigned to a single acidic and/or basic centre in each molecule. In addition, a new pKa dataset was compiled from in-house high-throughput UV–vis spectrophotometry pKa measurements in order to expand

the chemical space covered by the available literature pKa datasets. UV–vis spectrophotometry is frequently used in a high throughput fashion to provide rapid pKa determinations in an industrial setting. Other techniques, such as capillary electrophoresis, are considered to be more accurate (19) but are less practical for the routine pKa determination of large numbers of compounds.

A number of factors can affect the accuracy of pKa determination by means of UV–vis spectrophotometry. A molecule needs to contain a sufficiently strong chromophore in proximity to its ionisation centre(s) so that the protonated and deprotonated species can be distinguished in the UV–vis spectral range (25). Low aqueous solubility presents another challenge for the accurate measurement of pKa values. Although UV–vis spectroscopy has been shown to generally yield accurate pKa data (19,26–29), low solubility remains a potential source of experimental error (30). This is commonly addressed by determining apparent pKa (psKa) values in aqueous mixtures containing a less polar co-solvent (e.g. methanol). The aqueous pKa can then be obtained by extrapolating the psKa values to zero co-solvent content using the Yasuda–Shedlovsky (YS) method (28). Discrimination between acidic and basic centres is achieved using the sign of the slope of the YS plot. Acids have a positive slope in YS plots because their apparent pKa increases with increasing co-solvent content and the opposite trend is observed for bases (31,32). Erroneous inversions of the sign of the slope in YS-plots can occur during automatic curve-fitting when the slope is small (33) and thus yield incorrect pKa assignments. This becomes more likely when the average diameter of the ionised molecule is large or the charge density at its ionisation centre is low (34–36). The pKa dataset compiled from UV–vis spectrophotometry data allows us to investigate the occurrence of such experimental errors and assess the performance of pKa predictors on a new, structurally diverse dataset.

MATERIALS AND METHODS

Preparation of the Compound Datasets

Four datasets were extracted from the literature and comprise the Liao dataset (23) (a subset of pKa values measured with high accuracy as published by Prankerd (37)), the Avdeef or GOLD dataset (38), the Morgenthaler dataset (39) and the Luan dataset (20). The latter consists solely of bases for which pKa data were reported by Lombardo and co-workers (11). The dataset selected from our in-house collection will be referred to as the Vertex dataset.

The chemical structures of the compounds in the Avdeef (38) and Luan (20) datasets were obtained by searching for their names and smiles strings in the PUBCHEM database using a Python script (40). The smiles strings for the compounds in the

Morgenthaler dataset (39) were obtained with the help of optical structure recognition software (41). All structures and associated pKa data from the five datasets were visually inspected to assure the correct assignment of pKa values.

For simplicity, only amines (primary, secondary and tertiary) and 4-aminopyridines were considered as bases and carboxylic acids and N-aryl secondary amides as acids. The inclusion of other acids such as sulfonamides was considered, however, only three and four compounds from the Liao and Avdeef datasets contained a sulfonamide as the single acidic centre. The pKa values of 75 acidic sulfonamides from the Vertex collection were measured (data not shown) but published pKa data obtained by other methods were only available for four of these compounds. The lack of alternative experimental data for comparison purposes did not allow for an in depth comparison and therefore sulfonamides were not included in the analyses. The pKa predictions for this set of Vertex sulfonamides did, however, correlate well with our measured values with r^2 values of 0.71 and 0.63 for Chemaxon and Epik respectively.

The names, smiles strings and experimental pKa values of the compounds in the final processed Liao, Avdeef, Morgenthaler and Luan pKa datasets are available as Supplementary Material in Tables SIV–SVII.

In Silico Predictions

Epik 2.1209 is a pKa prediction tool available from Schrödinger Inc. This method uses Hammett-Taft equations with extensions, such as mesomer standardisation, charge cancellation and charge spreading approaches (42). It can also be used to generate and predict different protonation states and tautomers. pKa values were calculated using the command-line script called “epik” using the “–scan” option.

Chemaxon is the pKa prediction tool (43) available in Marvin (version 5.3.2). This software application predicts the microspecies distribution by calculating the sum of increments from partial charge, polarisability and structure specific features of the molecule. Macro ionisation constants associated with the ionisable functional groups of the molecule are calculated and displayed in the graphical user-interface.

The ACD/pKa DB software (ACDLabs pKa DB, version 5.13, ACDLabs, Toronto, 2001) uses Hammett-Taft equations derived from a large library of experimental data to predict pKa.

Given that pKa predictors are prone to random errors, sometimes >4 pKa units in cases where an ionization center is not parameterized, error estimates can be skewed by large outliers. Therefore the Median Absolute Deviation (MAD) was used alongside correlation coefficients as a robust measure of prediction error with reduced sensitivity to random outliers.

All cLogPs were calculated with Marvin, whereas polar surface areas (PSA) were calculated according to the method

described by Ertl and co-workers (44). Chemical space was analysed by performing a principal component analysis on 2D MACCS fingerprints calculated with MOE (version 2010.10, Chemical Computing Group, Montreal, Canada) using a script provided by their customer support. MOE was also used to compute van der Waals volumes and the distance in Å between the ionisation centre and the centroid of the closest aromatic ring. For 4-aminopyridines the distance was measured from the pyridine nitrogen atom to the centroid of the aromatic ring whereas for N-aryl secondary amides, the distance was measured from the amide nitrogen atom to the centroid of the aromatic ring.

The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) energies (in eV) were calculated using the semi-empirical quantum mechanic Modified Neglect of Differential Overlap method (MNDO), as implemented in MOE. These calculations were performed on the ionised and the neutral form of each molecule. HOMO and LUMO energies could not be calculated for amphotericin B (compound 13 in the Vertex dataset), due to the size and complexity of this molecule.

Experimental pKa Determinations

All pKa measurements were performed on the GLpKa instrument and data analysed using the RefinementPro software, both from Sirius Analytical Instruments (East Sussex, UK). A hybrid pH-metric/UV method (Fast-D-PAS) was used to monitor changes in UV absorbance of analytes during titration runs that spanned a range between 2 and 12 pH units. Standardisation of the pH electrode, measurements of UV lamp output intensity, and calibrations were all performed according to the manufacturer's guidelines. Samples were dissolved to 10 mM in dimethyl sulfoxide prior to running experiments. Then, 250 mL of a proprietary buffer solution from Sirius Analytical Instruments was manually added to 50 mL of 10 mM of each test compound resulting in a compound concentration of 1.67 mM prior to the start of titrations. An individual assay consisted of three titrations per vial. Due to concerns regarding compound solubility, each of the three titrations was performed in the presence of methanol co-solvent, at 50%, 40%, and 30% v/v methanol. The temperature was held constant at 25°C. Each titration curve was first corrected for background absorbance of the buffer solution. psKa values were obtained at 50%, 40%, and 30% methanol concentrations. Aqueous pKa values were determined by extrapolating the psKa values to zero methanol content using the YS procedure (28). The slope of the YS curve was used to assign pKa values to the acids and bases of zwitterions. The purity of compounds showing large deviations from their predicted pKa values was confirmed to be >90% by HPLC or LC/MS.

Solubility Measurements

Solubility for a subset of 156 randomly chosen compounds from the Vertex dataset was measured using a high throughput shake flask assay at pH 7.4. These compounds were supplied as samples containing typically 2% of dimethyl sulfoxide (DMSO) which may result in some overestimation of their aqueous solubility. The assay determines the equilibrium solubility of compounds in isotonic neutral buffer solution (Gibco Dulbecco Phosphate Buffered Saline solution). The method relies on segregating a precipitating solid from the buffer by centrifugation and determining the compound concentration in the supernatant by liquid chromatography using UV–vis spectroscopy for detection.

RESULTS AND DISCUSSION

Accuracy, Physicochemical Property Distributions and Chemical Diversity of the pKa Datasets

The assignment of pKa values to the correct ionisation centre is critical to performing accurate pKa predictions but can be challenging for compounds containing more than one ionisation centre. In order to avoid misassignment of pKa values, only compounds containing a single basic and/or acidic centre were selected from the pKa datasets.

A number of errors were identified in the published pKa datasets. For example, the pKa values for the amines of terbutaline and labetalol from the Luan dataset differ from those in the original reference cited in their work (11). Comparing pKa values with those reported for the same or highly similar molecules in other sources identified additional errors. For example, pyridoxamine (compound 176 in the Liao dataset depicted in Fig. 1) contains a primary amine, a pyridine and a phenol. In the pKa prediction study carried out by Liao and Nicklaus, the experimental pKa values of the amine and phenol were assigned to be 8.11 and 10.34 (23). The experimental pKa values for the closely related compound pyridoxine (Fig. 1) are 8.87 and 4.84 for the phenol and pyridine (38), suggesting that the pKa of the amine

and phenol in pyridoxamine should be 10.34 and 8.11. Errors of this type were corrected in all datasets used in this study. It was not always reported in the original references whether co-solvents and YS extrapolation were used. Some of the pKa values in the Morgenthaler set were measured in 10% methanol but without YS extrapolation, which could introduce additional errors.

The contents of the final five processed datasets used for pKa predictions are listed in Table I. The Morgenthaler and Luan datasets contained only bases, whereas the remaining datasets contained compounds with only one basic and/or one acidic centre after processing. Out of the 477 compounds comprising the Vertex dataset, 106 (~22%) were commercially available. The measured pKa values for these compounds are provided as Supplementary Material (Table SI).

Figure 2 shows the distribution of cLogP and molecular weight (MW) for the datasets used in this study. Drug-like molecules are often larger and more complex than the general organic chemicals often used in the parameterisation and validation of pKa predictors. Therefore the percentage of compounds with drug-like characteristics and the degree of chemical diversity were determined for each dataset.

According to Ghose and co-workers (45), drug-like compounds are characterised by having $160 < MW < 480$ and $-0.4 < cLogP < 5.6$. Applying these ranges to the five datasets demonstrated that ~84%, ~46% ~65%, ~60%, ~96%, of the compounds in the Vertex, Liao, Avdeef, Morgenthaler and Luan datasets met these criteria. A more detailed breakdown of the datasets in terms of how they conform to the Ghose criteria for drug-likeness is provided in the Supplementary Material as Table SII. The dataset containing the least drug-like molecules is the Liao dataset as it contains a significant proportion of compounds with $MW < 160$ and/or $cLogP < -0.4$. Figure 3 compares the chemical space occupied by the five datasets used in this study. The Liao and Avdeef datasets were similar in terms of the chemical space covered, whereas the remaining datasets showed a much greater diversity. The Vertex and Morgenthaler

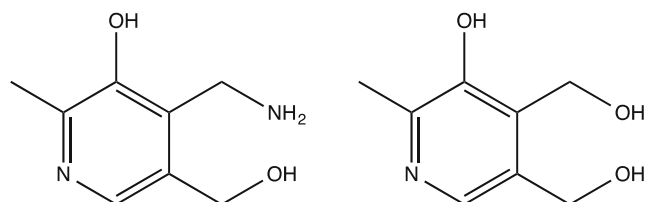


Fig. 1 Chemical structures of pyridoxamine and pyridoxine.

Table I Composition of the Datasets. Twenty (a), Nine (b) and Thirteen (c) Compounds in the Vertex, Liao and Avdeef Datasets Contained One Basic and One Acidic Centre Within the Same Molecule

Dataset	Number of unique compounds	Number of acidic centres predicted	Number of basic centres predicted
Vertex	477 ^a	167	330
Liao	105 ^b	43	71
Avdeef	122 ^c	49	86
Morgenthaler	174	0	174
Luan	67	0	67

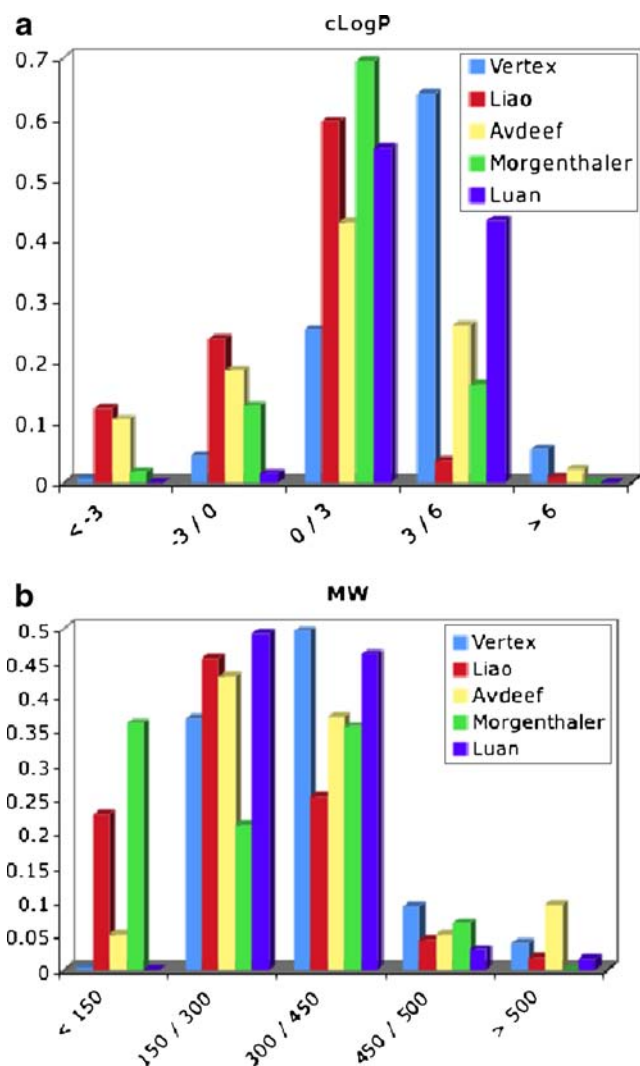


Fig. 2 Distribution for the cLogP (a) and molecular weight (MW) (b) for the five chemical datasets.

datasets showed the highest chemical diversity and are therefore the most relevant datasets for validating and expanding the scope of pKa prediction methods.

Figure 4 shows the distribution of pKa values for acids and bases in the Vertex dataset. Acidic pKa values generally cluster in the 2–6 pKa range while the N-aryl secondary acidic amides provide a second cluster of pKa values around pKa = 10. Bases show a larger variation of pKa values in comparison to acids with values ranging from less than 6 to ~11.

Sources of Error in UV–vis Spectrophotometric pKa Determination

The Vertex dataset is the only dataset for which all pKa values were obtained using the same methodology: UV–vis spectrophotometry using water/methanol mixtures with the

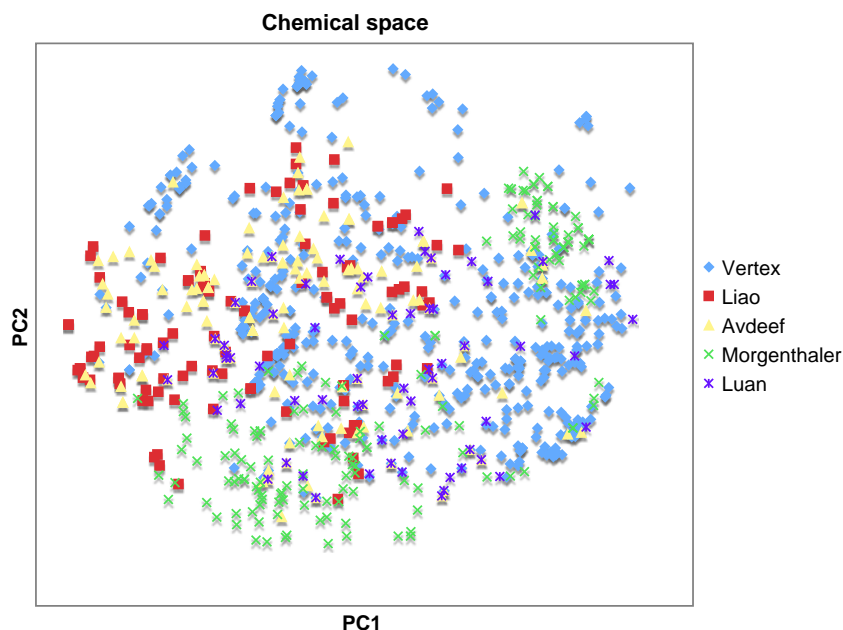
final aqueous pKa values determined by YS extrapolation. This provided the opportunity to assess potential sources of error unique to this methodology in more detail. One acid was identified for which erroneous YS curve extrapolation yielded a negative slope. Such errors can occur when the slope is small and might result in its pKa value being incorrectly assigned to a base. This error was identified and corrected by comparing the YS plots of structurally related compounds (see [Supplementary Material](#)).

The pKa values for 50 out of 106 commercially available compounds within the Vertex dataset could be compared with independently determined pKa values and are provided in Table SIII in [Supplementary Material](#). Acids showed excellent agreement with the literature data but bases compared poorly with an r^2 of 0.52 and a slightly higher median absolute deviation (MAD) in comparison to bases (Table II). The same analysis was performed for the four literature datasets and yielded r^2 values greater than 0.9 for both acids and bases in all datasets (data not shown). These results suggest that high-throughput UV–vis spectrophotometry yielded erroneous pKa data for at least some bases. Careful inspection revealed that the low r^2 value obtained for the bases in our dataset was due to five outliers that differed by more than 0.8 pKa units from their literature values. When the correlation coefficient was recalculated with these compounds omitted the r^2 for bases improved to 0.83 and MAD to 0.15 (Table II). In order to better understand the underlying causes of these errors, the influence of solubility, chromophore strength and distance to the ionisation centre on experimental error were further investigated.

Solubility

The aqueous solubility was measured for 46 out of the 50 compounds from the Vertex dataset with published pKa values. The relation between solubility and the difference between pKa values obtained through UV–vis spectrophotometry and published data is shown in Fig. 5a. Bases show significantly larger experimental discrepancies than acids and tend to be less soluble. The distribution of aqueous solubility for all 156 compounds with solubility data in the Vertex dataset is depicted in Fig. 6. The majority of acids are highly soluble (>200 μ M) whereas the solubility of bases is generally lower and more variable, in agreement with previous analyses of drug-like compounds (12). At the pH used to determine the solubility of these compounds (pH=7.4), carboxylic acids are fully ionised whereas a significant portion of bases from the Vertex dataset are not (Fig. 4) which may explain their reduced solubility (12). In addition, amines are often used in medicinal chemistry to solubilise poorly soluble compounds. Measuring the pKa of bases with low aqueous solubility using UV–vis spectrophotometry may therefore benefit from repeating the experiment

Fig. 3 Chemical space of the five datasets used in this study. Principal components (PC) were derived from principal component analysis on MACCS fingerprints.



using higher co-solvent concentrations or a less polar co-solvent. Among the five outliers, fluspirilene and tamoxifen were found to have an aqueous solubility $<10 \mu\text{M}$.

Chromophore Strength and Distance to the Ionisation Centre

A key requirement for UV-vis spectrophotometric detection in pKa measurements is that a significant change occurs in the UV-vis spectrum of a compound upon ionisation. This requirement is thought to be met if a molecule contains a strong chromophore in proximity to its ionisation centre (25).

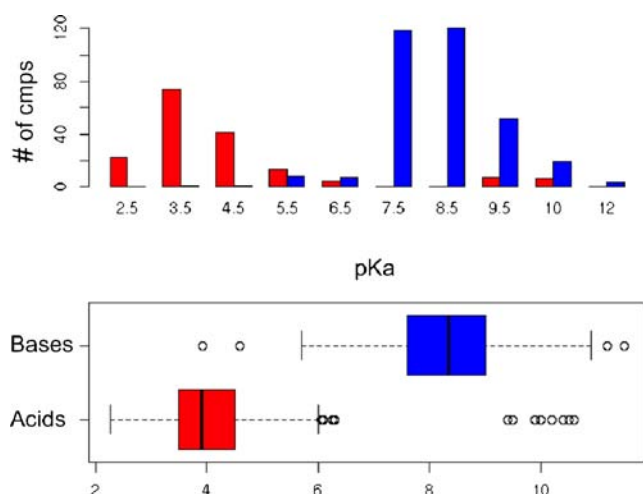


Fig. 4 Distribution (histogram and box plot) of experimental pKa values in the Vertex dataset. Acids and bases are shown in red and blue, respectively. The median, the first and third quartiles are shown. Empty circles represent outliers.

The fact that carboxylic acids and N-aryl amides absorb in the UV-vis range and their chromophores thus coincide with their ionisation centres is likely to aid the accuracy with which their pKa values can be determined (Table II). The 4-aminopyridines in the Vertex dataset present a special case among the basic compounds as their ionisation centre coincides with their chromophore and for these compounds a significant smaller average prediction error (0.35 ± 0.27) was obtained in comparison with other bases (0.54 ± 0.52).

A smaller difference between the HOMO and LUMO energy levels of a molecule, correlates with increased conjugation and molar absorptivity (i.e. chromophore strength) (46). This is an approximation since other factors such as steric effects, alkyl substitutions, geometrical strain, solvent polarity and interactions with other chromophores in the same molecule can also influence the maximum absorptivity in the UV-visible spectrum (46,47). Unfortunately, experimental absorptivity data at λ_{max} were unavailable for the Vertex dataset so the calculated HOMO-LUMO gap was used as a substitute descriptor.

Figure 5b shows the differences between pKa values obtained with UV-vis spectrophotometry and literature data plotted against the HOMO-LUMO energy difference calculated for the ionised molecules. The remaining three outliers (ranitidine, amitriptyline and cyclobenzaprine) had HOMO-LUMO energy differences $>5 \text{ eV}$. The pKa value of ranitidine was 9.6 as determined by UV-vis spectrophotometry, compared to published values in the 8.31–8.47 range (11,38,48). Ionisation of the amino group of ranitidine does not significantly alter its UV-vis spectrum and this appears to be reflected in its large HOMO-LUMO gap of $\sim 7 \text{ eV}$. The HOMO-LUMO energy difference of a base can be easily calculated and a value $>5 \text{ eV}$

Table II Squared Correlation Coefficients and Median Absolute Deviation (MAD) Values Obtained by Comparing Experimental pKa Values Measured Using UV–vis Spectrophotometry and Literature Data. Total Number of Acidic Centres = 20; Total Number of Basic Centres = 35; Total Number of Zwitterions = 5

	r^2 /MAD (all ionisation centers)	r^2 /MAD (acids)	r^2 /MAD (bases)
Vertex vs published experimental pKa (50 compounds) ^a	0.96/0.16	0.97/0.13	0.52/0.17 (0.83/0.15) ^a

(a): r^2 and MAD recalculated after the removal of five outliers

could warrant a critical assessment of its pKa value when determined by UV–vis spectrophotometry.

The distance between the ionic centre and the centroid of the closest aromatic ring in the molecule can be considered as a descriptor for the proximity of the chromophore to the ionisation centre in a molecule. This is again an approximation since amides, esters, double or triple bonds and sulphur atoms are also known chromophores (46,47). Figure 5c does not show any correlation between this descriptor and the variation in experimental pKa values. These results suggest that UV–vis spectrophotometry can still be used to accurately determine the pKa of a molecule even if the distance between chromophore and the ionisation centre is relatively long.

Results of pKa Predictions

Tables III, IV and V list the r^2 and MAD values for the prediction of the pKa values of all ionisation centres by Chemaxon, Epik and ACD pKa DB. Correlation coefficients and MAD values are also provided for the acids and bases in each dataset as separate classes. In terms of overall prediction quality the Vertex dataset gave comparable results to four literature data sets. Comparing the r^2 values obtained for acids and bases with the r^2 for both classes combined provided insight into the true accuracy of the pKa prediction methods. This is illustrated in Fig. 7, where the Epik pKa predictions for the acids and bases in the Vertex dataset are shown separately. The squared correlation coefficients and MAD values equal 0.35/0.60 for bases and 0.83/0.40 for acids, compared to an r^2 and MAD of 0.84/0.55 for the entire dataset. The correlation found for the entire dataset seems artificially high and appears to be achieved by fitting a line through the clusters of acids and bases below and above pKa = 7. Predictions for the bases and acids from the Vertex dataset as separate classes follows the same trend as found for the literature datasets where basic pKa values are generally less accurately predicted compared to acidic pKa values.

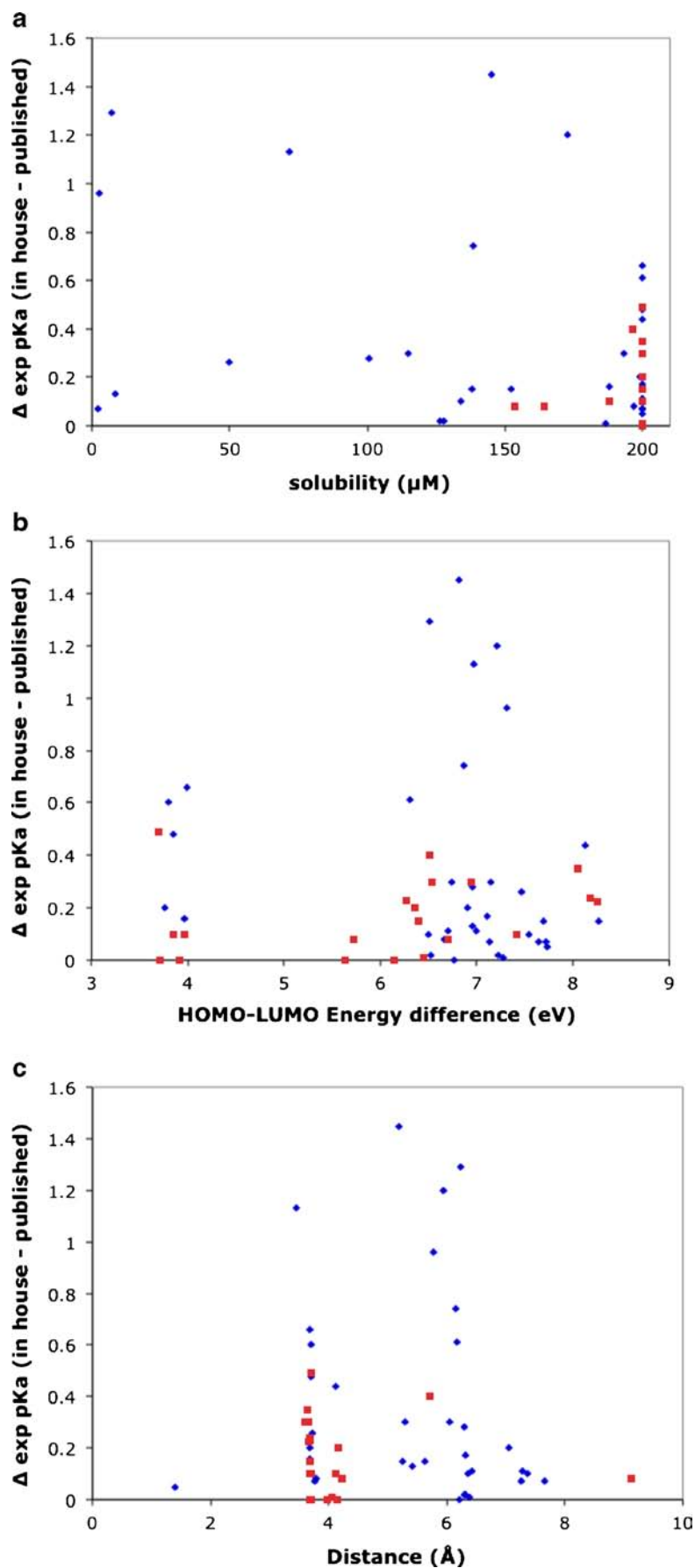
It is difficult to directly compare our pKa predictions with those previously reported in the literature due to the use of older versions of the pKa prediction software and our selections from and corrections to the original literature datasets. Nevertheless, the r^2 values 0.97 for Chemaxon, 0.93 for Epik and 0.96 for ACD pKa DB obtained for the entire Liao data set in this study (Tables III, IV and V) compare favourably to the equivalent r^2 values of 0.763, 0.802 and 0.908 reported previously (23). Our prediction results for the entire Avdeef dataset ($r^2=0.94$, 0.87 and 0.93 using Chemaxon, Epik and ACD pKa DB, respectively) show an improvement over the equivalent r^2 values of 0.892 and 0.485, 0.923 reported by Balogh *et al.* (22). The r^2 values obtained by predicting the entire Vertex pKa dataset (0.87, 0.84 and 0.81 for Chemaxon, Epik and ACD pKa DB, respectively) compare well with published comparisons and the range of r^2 values found for the literature datasets used in the current study (0.25–0.97, 0.45–0.93 and 0.69–0.96 and for Chemaxon, Epik and ACD pKa DB, respectively).

As shown in Tables III, IV and V, it is difficult to judge which is the best performing pKa predictor because their performance is highly dependent on the dataset being predicted. For example, the pKa values of bases in the Morgenthaler dataset were predicted well by Epik ($r^2=0.81$) but poorly by Chemaxon ($r^2=0.25$). In contrast, the pKa values of acids in the Avdeef dataset were predicted well by Chemaxon ($r^2=0.70$) but poorly by Epik ($r^2=0.25$). Epik tends to yield the highest MAD values while Chemaxon and ACD pKa DB yield comparable, lower prediction errors.

In general the pKa values of acids are predicted with considerably higher r^2 ($r^2>0.7$) and lower MAD (<0.4) values than those of bases (see Tables III, IV and V). As reported by Gleeson (12) and confirmed by the distribution of pKa values from the Vertex dataset as shown in Fig. 4, pKa values of basic compounds are generally distributed over a wider range of pKa in comparison to acids. This could explain why it is more challenging to predict the pKa of bases correctly.

Many prediction errors were caused by the presence of specific chemical features, which appear to be well parameterised for one pKa predictor but poorly for the other and *vice versa*. The more structurally diverse Morgenthaler and

Fig. 5 The relationship between $\Delta_{\text{exp pKa}}$ (the difference between pKa values measured with UV-vis spectrophotometry and independently determined values from the literature) with the aqueous solubility (**a**), the HOMO-LUMO energy difference (**b**) and the distance from the ionisation centre to the nearest aromatic ring (**c**), for a subset of compounds from the Vertex dataset. Acids and bases are coloured in red and blue, respectively. Compounds with solubility $>200 \mu\text{M}$ are shown in the plot as if they had a solubility of $200 \mu\text{M}$.



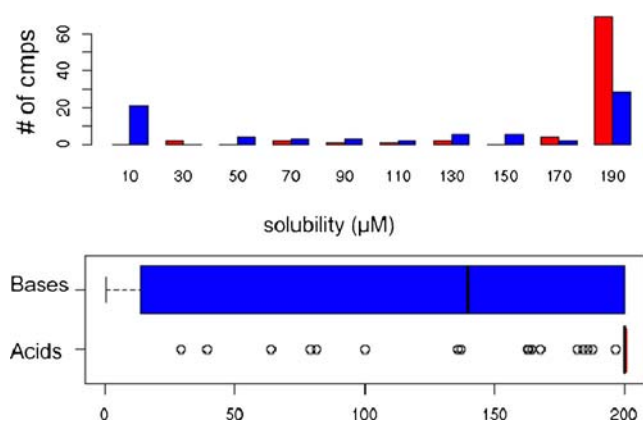


Fig. 6 Distribution of the solubility (histogram and box plot) measured for a subset of 156 compounds from the Vertex dataset. Compounds with solubility $>200 \mu\text{M}$ are shown as if they had a solubility of $200 \mu\text{M}$. Acids and bases are coloured in red and blue, respectively. The median, the first and third quartiles are shown. Empty circles represent outliers. Most of the acids have solubility $>200 \mu\text{M}$.

Vertex datasets showed the poorest predictions for their basic pKa values. This suggests the presence of unique structural features in these bases, which were not present in the datasets used to calibrate pKa prediction methods. In the next section structural motifs for which one of the two pKa predictors performed poorly are exemplified.

Analysis of Errors in pKa Prediction

Cases where Epik or Chemaxon failed to correctly predict pKa values were identified as compounds that were predicted well by one algorithm (absolute error <0.5 pKa units) but poorly by the other (absolute error >1.5 pKa units). Table VI lists cases where Epik but not Chemaxon gave large prediction errors. Epik failed to predict the pKa correctly of several amines with an electron-withdrawing group two carbon atoms removed from the amine. Some carboxylic acids (such as 4-oxo-1,4-dihydropyridine-3-carboxylic acid) also yielded high prediction errors (in some cases >3 units of pKa). The pKa of several N-aryl secondary amides was also poorly predicted by Epik by ~ 2 pKa units. Table VII illustrates cases where only Chemaxon yielded large pKa prediction errors. Chemaxon incorrectly predicted the pKa values of various tertiary amines with low pKa <7 . Some of these amines from the Morgenthaler dataset were predicted with an error >4 units of pKa. Table VIII provides examples where ACD pKa DB performed better than both Epik and Chemaxon.

Even though Epik and Chemaxon were validated with large, albeit undisclosed chemical datasets (42,43), the poor predictions for the ionisation centres shown in Tables VI, VII and VIII suggest that some chemical features were still underrepresented. These examples could be used to improve the scope of both pKa predictors.

Table III Squared Correlation Coefficients and Median Absolute Deviation (MAD) Values for pKa Predictions Using Chemaxon

Dataset	r^2/MAD (all predictions)	r^2/MAD (acids)	r^2/MAD (bases)
Vertex	0.87/0.51	0.89/0.37	0.39/0.64
Liao	0.97/0.33	0.76/0.41	0.75/0.31
Avdeef	0.94/0.29	0.7/0.33	0.55/0.29
Morgenthaler	0.25/0.52	NA	0.25/0.52
Luan	0.57/0.33	NA	0.57/0.33

NA: not applicable because no acidic centres were present in the set

CONCLUSIONS

The development and validation of pKa prediction tools commonly relies on the assumption that the pKa values of acids and bases can be measured and predicted with equal accuracy. Predictions for these two types of ionisation centres have therefore rarely been compared in any detail. Validation studies often use historic literature pKa datasets containing compounds with limited structural diversity and multiple, sometimes erroneous pKa values that are not unambiguously assigned to their ionisation centres. All these factors can result in overestimation of the accuracy of pKa prediction tools.

The validation and improvement of pKa prediction tools requires the availability of accurate and unambiguously assigned pKa data in order to gain a realistic appreciation of their true accuracy. Analysis of four literature pKa datasets identified several errors and a lack of chemical diversity and structural complexity in certain datasets. In order to expand the chemical space available for the development and validation of pKa prediction tools a new, more chemically diverse and drug-like dataset obtained by UV-vis spectrophotometry has been made available.

Table IV Squared Correlation Coefficients and Median Absolute Deviation (MAD) Values for pKa Predictions Using Epik

Dataset	r^2/MAD (all predictions)	r^2/MAD (acids)	r^2/MAD (bases)
Vertex	0.84/0.55	0.83/0.40	0.35/0.60
Liao	0.93/0.43	0.73/0.28	0.49/0.58
Avdeef	0.87/0.43	0.25/0.37	0.44/0.45
Morgenthaler	0.81/0.83	NA	0.81/0.83
Luan	0.45/0.45	NA	0.45/0.45

NA: not applicable because no acidic centres were present in the set

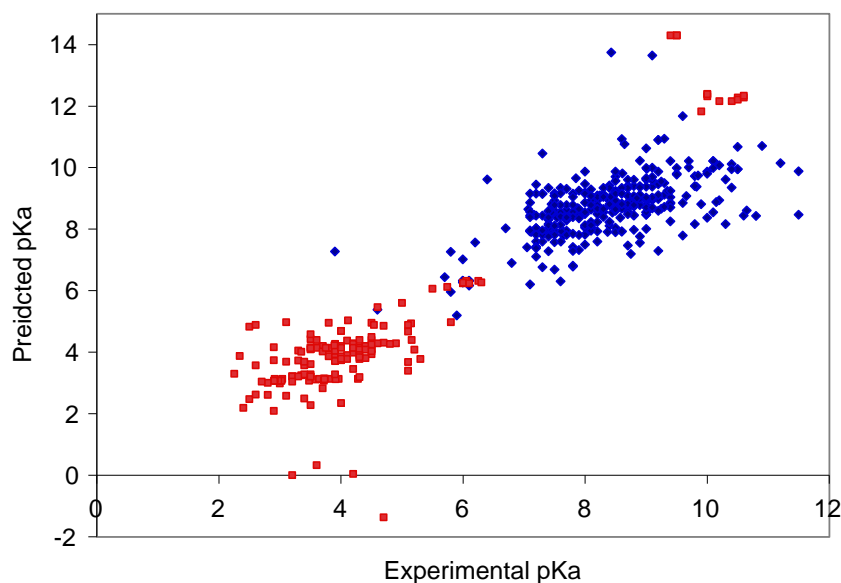
Table V Squared Correlation Coefficients and Median Absolute Deviation (MAD) Values for pKa Predictions Using ACD pKa DB

Dataset	r^2 /MAD (all predictions)	r^2 /MAD (acids)	r^2 /MAD (bases)
Vertex	0.81/0.52	0.82/0.43	0.37/0.61
Liao	0.96/0.21	0.76/0.11	0.69/0.31
Avdeef	0.93/0.25	0.31/0.19	0.74/0.3
Morgenthaler	0.91/0.47	NA	0.91/0.47
Luan	0.69/0.27	NA	0.69/0.27

NA: not applicable because no acidic centres were present in the set

Analysis of this dataset and literature data showed that the experimental determination of the pKa of basic compounds is more prone to errors because of the generally lower aqueous solubility of bases compared to acids. Increased co-solvent concentrations or the use of alternative, less polar co-solvents could render pKa measurements for such bases more accurate. The determination of basic pKa values by UV–vis spectrophotometry can yield additional errors due to the presence of weak chromophores. Bases are more sensitive to the presence of a weak chromophore since, unlike acids, their chromophores usually do not coincide with the ionisation centre. Bases with calculated HOMO-LUMO energy differences greater than 5 eV may possess weak chromophores and alternative experimental methods of pKa determination may have to be considered in such cases. Despite the increased risk of errors that often accompanies high-throughput methodologies, pKa values obtained by high-throughput UV–vis spectrophotometry compared well with pKa values obtained by manually performed measurements using different experimental techniques.

Fig. 7 pKa prediction plot for the Vertex dataset using Epik. Acids and bases are coloured in red and blue, respectively. The r^2 for the whole dataset is 0.84 whereas the r^2 is 0.83 for acids and 0.35 for bases.



Our results demonstrate the need for significant improvement in the pKa prediction of basic compounds. For the most chemically diverse datasets, prediction of the pKa values of bases by the Epik, Chemaxon and ACDLabs pKa predictors was found to be significantly less accurate than for acids or both classes combined. The prediction of the pKa of acids generally yielded higher r^2 values than those for bases. Similarly, the MAD values for the more chemically diverse bases from the Vertex and Morgenthaler data sets was found to reach values in the 0.5–0.8 range while acids generally had MAD values in the 0.3–0.4 range. This result confirms the necessity to consider both classes separately during the development and validation of pKa prediction tools, which is currently not standard practice.

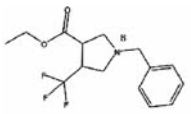
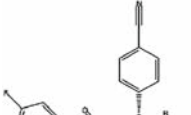
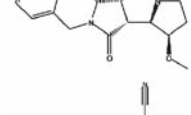
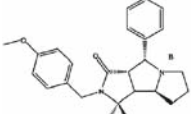
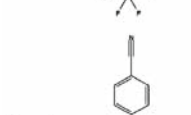

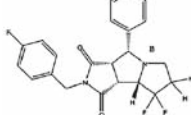
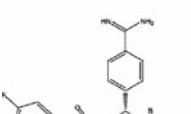
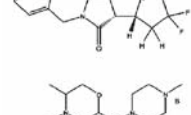
The discrepancy in accuracy between basic and acidic pKa predictions may be partially due to the larger pKa range exhibited by bases compared to acids. A scarcity of datasets containing more diverse and structurally complex compounds is another limiting factor in the development of more accurate pKa predictors. The Vertex and Morgenthaler datasets, which contained the most structurally diverse compounds, yielded the worst predictions for the pKa of basic compounds. These results suggest that these two datasets contain structural motifs that were poorly represented in the datasets used to parameterise the pKa prediction algorithms used in this study. A number of new chemical substructures were identified in the five datasets for which the pKa was predicted with errors greater than 1.5 pKa units by either of the Chemaxon and Epik predictors. The curated literature data and the Vertex pKa dataset together form a new collection of unambiguously assigned pKa data covering a more diverse range of chemistries of use to the development of improved pKa prediction tools.

Table VI Cases Where Epik Gives A Substantial pKa Prediction Error in Comparison to Chemaxon (Absolute Errors > 1.5 and < 0.5 units Respectively)

Structure	Dataset	Name of the compound in the dataset	Experimental pKa	CHEMAXONEPIK pKa	Epik pKa
	Vertex	86	9	8.96	10.63
	Vertex/Avdeef	81/ tetracaine	8.6/8.49	8.42	10.93
	Luan/Liao/Avdeef	lidocaine	7.94/7.928/7.95	7.75	9.51
	Luan	fentanyl	8.43	8.77	9.94
	Avdeef/Liao	procaine	9.04/9.04	8.96	10.63
	Avdeef	ranitidine	8.31	8.09	11.68
	Morgenthaler	fig13b2	5.4	5.27	7.04
	Morgenthaler	fig15-2	10.2	10.31	8.44
	Morgenthaler	50	6.5	6.43	8.36
	Morgenthaler	fig17-4	8.1	7.97	10.16
	Morgenthaler	fig11a5	8.4	8.42	10.7
	Avdeef	Aspartic acid	9.67	9.61	8.15
	Avdeef	Nalidixic acid	6.01	5.95	-1.12
	Vertex	34	4.7	4.93	-1.36
	Vertex	41	10.4	10.6	12.17
	Vertex	77	10.6	10.68	12.28
	Vertex	78	10.6	10.7	12.34
	Vertex	52	10.5	10.68	12.28
	Vertex	75	10.5	10.6	12.22
	Vertex	45	10.2	10.6	12.17

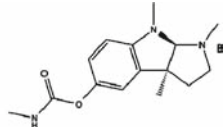
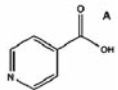
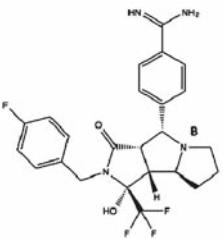
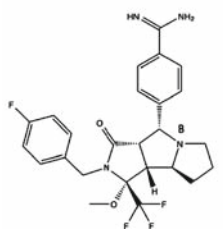
For each compound the ionisation centre predicted is indicated with A for the acid or B for the base. The names of the original set and of the compound are the same as reported in [Supplementary Material](#). If the compound is present in more than one set, the names and the experimental pKa values of each compound are separated by a forward slash

Table VII Cases Where Chemaxon Gave A Substantial pKa Prediction Error in Comparison to Epik (Absolute Errors > 1.5 and < 0.5 units Respectively)

Structure	Dataset	Name of the compound in the dataset	Experimental pKa	CHEMAXON pKa	EPIK pKa
	Morgenthaler	fig18 -6	5.4	7.92	5.44
	Morgenthaler	fig2SI -6	4.2	9.06	4.31
	Morgenthaler	fig3SI -6	5.6	8.66	5.91
	Morgenthaler	fig2SI -13	2	6.04	1.84
	Morgenthaler	32	2	6.32	2.42
	Morgenthaler	40	2	6.31	2.29
	Avdeef	ofloxacin	8.31	6.2	7.93
	Vertex	2	8.2	6.06	7.93
	Vertex	87	9.3	7.05	9.5

For each compound the basic centre predicted is indicated with the letter B. The names of the original dataset and of the compound are the same as reported in [Supplementary Material](#)

Table VIII Cases Where Chemaxon and Epik gave A Substantial pKa Prediction Error (Absolute Errors > 1.5) in Comparison to ACD (Absolute Error < 0.5 Units)

Structure	Dataset	Name of the compound in the dataset	Exp. pKa	ACD pKa	Chemaxon pKa	Epik pKa
	Liao	161	8.17	8.44	6.59	6.34
	Liao	149	1.77	1.94	3.73	3.37
	Morgenthaler	23	5.5	5.75	8.82	7.35
	Morgenthaler	24	5.2	5.6	9.36	7.18

For each compound the ionisation centre predicted is indicated with A for the acid or B for the base. The names of the original dataset and of the compound are the same as reported in [Supplementary Material](#)

REFERENCES

- Abraham MH, Duce PP, Prior DV, Barratt DJ, Morris JJ, Taylor PJ. Hydrogen bonding. Part 9. Solute proton donor and proton acceptor scales for use in drug design. *J Chem Soc Perkin Trans.* 1989;2:1355–75.
- Agouridas V, Laios I, Cleeren A, Kizilian E, Magnier E, Blazejewski JC, *et al.* Loss of antagonistic activity of tamoxifen by replacement of one N-methyl of its side chain by fluorinated residues. *Bioorg Med Chem.* 2006;14(22):7531–8.
- Mitra R, Shyam R, Mitra I, Miteva MA, Alexov E. Calculating the protonation states of proteins and small molecules: Implications to ligand-receptor interactions. *Curr Comput-Aided Drug Des.* 2008;169–179
- Stahl PH. *The Practice of medicinal chemistry.* London: Academic; 2003.
- WDI. The world Drug Index available from www.derwent.com (Derwent, London, UK).
- Tam K, Comer J. *Pharmacokinetic optimization in drug research: Biological, physicochemical, and computational strategies.* Weinheim: Wiley-VCH; 2001.
- Manallack DT. The pK(a) distribution of drugs: application to drug discovery. *Perspect Medicin Chem.* 2008;1:25–38.
- Mitani GM, Steinberg I, Lien EJ, Harrison EC, Elkayam U. The pharmacokinetics of antiarrhythmic agents in pregnancy and lactation. *Clin Pharmacokinet.* 1987;12(4):253–91.
- Xie X, Steiner SH, Bickel MH. Kinetics of distribution and adipose tissue storage as a function of lipophilicity and chemical structure. II. Benzodiazepines. *Drug Metab Dispos.* 1991;19(1):15–9.
- Deak K, Takacs-Novak K, Kapas M, Vastag M, Tihanyi K, Noszal B. Physico-chemical characterization of a novel group of dopamine D(3)/D(2) receptor ligands, potential atypical antipsychotic agents. *J Pharm Biomed Anal.* 2008;48(3):678–84.
- Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J Med Chem.* 2004;47(5):1242–50.
- Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. *J Med Chem.* 2008;51(4):817–34.
- Hansch C. Quantitative relationships between lipophilic character and drug metabolism. *Drug Metab Rev.* 1972;1(1):1–13.
- Hansch C, Steward AR, Iwasa J. The use of substituent constants in the correlation of demethylation rates. *J Med Chem.* 1965;8(6):868–70.
- Fermini B, Fossa AA. The impact of drug-induced QT interval prolongation on drug discovery and development. *Nat Rev Drug Discov.* 2003;2(6):439–47.
- Alberati D, Hainzl D, Jolidon S, Krafft EA, Kurt A, Maier A, *et al.* Discovery of 4-substituted-8-(2-hydroxy-2-phenyl-cyclohexyl)-2,8-diaza-spiro[4.5]decan-1-one as a novel class of highly selective GlyT1 inhibitors with improved metabolic stability. *Bioorg Med Chem Lett.* 2006;16(16):4311–5.
- Jamieson C, Moir EM, Rankovic Z, Wishart G. Medicinal chemistry of hERG optimizations: highlights and hang-ups. *J Med Chem.* 2006;49(17):5029–46.

18. Ploemen JP, Kelder J, Hafmans T, van de Sandt H, van Burgsteden JA, van Salemink PJ, *et al.* Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential: a case study with structurally related piperazines. *Exp Toxicol Pathol.* 2004;55(5):347–55.
19. Lee AC, Crippen GM. Predicting pKa. *J Chem Inf Model.* 2009;49(9):2013–33.
20. Luan F, Ma W, Zhang H, Zhang X, Liu M, Hu Z, *et al.* Prediction of pK(a) for neutral and basic drugs based on radial basis function Neural networks and the heuristic method. *Pharm Res.* 2005;22(9):1454–60.
21. Dearden JC, Cronin MTD, Lappin DC. A comparison of commercially available software for the prediction of pKa. *J Pharm Pharmacol.* 2007;59 (Suppl. 1):A–7
22. Balogh GT, Gyarmati B, Nagy B, Molnar L, Keseru GM. Comparative evaluation of in Silico pKa prediction tools on the gold standard dataset. *QSAR Comb Sci.* 2009;28(10):1148–55.
23. Liao C, Nicklaus MC. Comparison of nine programs predicting pK(a) values of pharmaceutical substances. *J Chem Inf Model.* 2009;49(12):2801–12.
24. Manchester J, Walkup G, Rivin O, You Z. Evaluation of pKa estimation methods on 211 druglike compounds. *J Chem Inf Model.* 2010;50(4):565–71.
25. Allen RI, Box KJ, Comer JE, Peake C, Tam KY. Multiwavelength spectrophotometric determination of acid dissociation constants of ionizable drugs. *J Pharm Biomed Anal.* 1998;17(4–5):699–712.
26. Avdeef A, Box KJ, Comer JE, Gilges M, Hadley M, Hibbert C, *et al.* PH-metric log P 11. pKa determination of water-insoluble drugs in organic solvent-water mixtures. *J Pharm Biomed Anal.* 1999;20(4):631–41.
27. Takács-Novák K, Box KJ, Avdeef A. Potentiometric pKa determination of water-insoluble compounds: validation study in methanol/water mixtures. *Intern J Pharm.* 1997;151(2):235–48.
28. Volgyi G, Ruiz R, Box K, Comer J, Bosch E, Takacs-Novak K. Potentiometric and spectrophotometric pKa determination of water-insoluble compounds: validation study in a new cosolvent system. *Anal Chim Acta.* 2007;583(2):418–28.
29. Albert A, Serjeant E. The determination of ionization constants. 3rd ed. London: Chapman and Hall; 1984.
30. Ruiz R, Rafols C, Roses M, Bosch E. A potentially simpler approach to measure aqueous pKa of insoluble basic drugs containing amino groups. *J Pharm Sci.* 2003;92(7):1473–81.
31. Mandic Z, Gabelica V. Ionization, lipophilicity and solubility properties of repaglinide. *J Pharm Biomed Anal.* 2006;41(3): 866–71.
32. Gobry V, Bouchard G, Carrupt PA, Testa B, Girault HH. Physicochemical characterization of sildenafil: ionization, lipophilicity behavior, and ionic-partition diagram studied by two-phase titration and electrochemistry. 2000;83(7):1465–74.
33. Box K, Ruiz R, Cimpan G, Allen R, Mole J, Comer J. A mixed solvent system for use with the ProfilerSGA for rapid measurement of ionisation constants, *LogP 2004 The 3rd lipophilicity symposium*, Zurich, Switzerland, March 2004.
34. Fuoss RM. Properties of electrolytic solutions. III. The dissociation constant. *J Am Chem Soc.* 1933;55(3):1019–28.
35. Ramsey J, Colichman E. Dissociation constants of some substituted phenyltrimethylammonium perchlorates in ethylene chloride; Effect of ion asymmetry. *J Am Chem Soc.* 1947;69(12):3041–5.
36. Shedlovsky T. The behaviour of carboxylic acids in mixed solvents. New York: Pergamon Press; 1962.
37. Prankerd RJ. Profiles of drug substances, excipients, and related methodology. San Diego: Elsevier Academic Press; 2007.
38. Avdeef A. Absorption and drug development: solubility, permeability, and charge state. New York: Wiley-IEEE; 2003.
39. Morgenthaler M, Schweizer E, Hoffmann-Roder A, Benini F, Martin RE, Jaeschke G, *et al.* Predicting and tuning physicochemical properties in lead optimization: amine basicities. *ChemMedChem.* 2007;2(8):1100–15.
40. Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. Washington: American Chemical Society; 2008.
41. Filippov I. <http://cactus.nci.nih.gov/osra/>, NCI/CADD Group, 2007.
42. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J Comput Aided Mol Des.* 2007;21(12):681–91.
43. Szegezdi J, Csizmadia F. A Method for Calculating the pK Values of Small and Large Molecules, *233rd ACS National Meeting, CINF41*, http://www.chemaxon.com/conf/Calculating_pKa_values_of_small_and_large_molecules.pdf, Chicago, USA. 2007.
44. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem.* 2000;43(20):3714–7.
45. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J Comb Chem.* 1999;1(1):55–68.
46. Kalsi PS. Spectroscopy of organic compounds. 6th ed. New Delhi: New Age International Pvt Ltd; 2007.
47. Turro NJ, Ramamurthy V, Scaiano JC. Principles of molecular photochemistry. An introduction, University Science Books, USA, 2009
48. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data. *J Med Chem.* 2002;45(13):2867–76.