

# Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve *in Silico* pK<sub>a</sub> Prediction

Robert Fraczekiewicz,<sup>\*,†</sup> Mario Lobell,<sup>\*,‡</sup> Andreas H. Göller,<sup>‡</sup> Ursula Krenz,<sup>‡</sup> Rolf Schoenneis,<sup>‡</sup> Robert D. Clark,<sup>†</sup> and Alexander Hillisch<sup>‡</sup>

<sup>†</sup>Simulations Plus, Inc. 42505 10th Street West, Lancaster, California 93534, United States

<sup>‡</sup>Global Drug Discovery, Bayer Pharma AG, Wuppertal, Germany

## Supporting Information

**ABSTRACT:** In a unique collaboration between a software company and a pharmaceutical company, we were able to develop a new *in silico* pK<sub>a</sub> prediction tool with outstanding prediction quality. An existing pK<sub>a</sub> prediction method from Simulations Plus based on artificial neural network ensembles (ANNE), microstates analysis, and literature data was retrained with a large homogeneous data set of drug-like molecules from Bayer. The new model was thus built with curated sets of ~14,000 literature pK<sub>a</sub> values (~11,000 compounds, representing literature chemical space) and ~19,500 pK<sub>a</sub> values experimentally determined at Bayer Pharma (~16,000 compounds, representing industry chemical space). Model validation was performed with several test sets consisting of a total of ~31,000 new pK<sub>a</sub> values measured at Bayer. For the largest and most difficult test set with >16,000 pK<sub>a</sub> values that were not used for training, the original model achieved a mean absolute error (MAE) of 0.72, root-mean-square error (RMSE) of 0.94, and squared correlation coefficient (R<sup>2</sup>) of 0.87. The new model achieves significantly improved prediction statistics, with MAE = 0.50, RMSE = 0.67, and R<sup>2</sup> = 0.93. It is commercially available as part of the Simulations Plus ADMET Predictor release 7.0. Good predictions are only of value when delivered effectively to those who can use them. The new pK<sub>a</sub> prediction model has been integrated into Pipeline Pilot and the PharmacophorInformatics (Plx) platform used by scientists at Bayer Pharma. Different output formats allow customized application by medicinal chemists, physical chemists, and computational chemists.



## INTRODUCTION

Protonation and deprotonation influence properties and behavior of chemical compounds in solution in ways that are especially relevant to biochemistry, pharmaceutical science, medicinal chemistry, ecology, and agrochemistry. In particular, 80% of contemporary drugs contain at least one ionizable group.<sup>1</sup> Changes in dominant protonation states can alter pharmacological interactions drastically and thereby influence potency. A drug candidate's predominant charge state at a given pH can be a major determinant of pharmacological activity, aqueous solubility, permeability, plasma protein binding, cardiotoxicity, and metabolism, and the level of general interest in such ionization phenomena is evident from the large number of recent publications on the topic.<sup>2–25</sup>

The importance of pK<sub>a</sub> is also reflected in the commercial availability of automated instruments for high-throughput measurements of ionization constants. Nonetheless, exhaustive pK<sub>a</sub> measurements for all compounds in libraries numbering in the millions is impractical. Such measurements are altogether impossible for compounds that are not yet synthesized, so the use of *in silico* methods for predicting pK<sub>a</sub> from molecular structure so as to provide insights into ionization patterns in virtual libraries is also of interest. Pharmaceutical scientists at Bayer Pharma have long expressed dissatisfaction with available *in silico* methods for predicting ionization constants; none of

the available global models performed well enough on in house compounds with known pK<sub>a</sub> values. The best models had RMSEs around 1 log unit, which is in accordance with reports by other industrial researchers. This was disappointing because Bayer's experimental pK<sub>a</sub> data showed a natural spread of only ±0.1 log units for multiple measurements of the same molecular species. We hypothesized that in spite of their "global" label, models trained on the literature data did not adequately cover the chemical space occupied by the Bayer compounds and that disappointing performance on multiprotic compounds could be addressed by microstate analysis (see Discussion). In addition, a model built using a large amount of Bayer data should benefit from greater experimental consistency than that seen for models built on data extracted from hundreds of papers published over the last decades. Here, we demonstrate that these hypotheses are correct.

In 2011 Bayer and Simulations Plus undertook a collaborative effort to extend the pK<sub>a</sub> model ("S+pKa") in version 6.0 of Simulations Plus' ADMET Predictor program<sup>26</sup> to cover Bayer's chemical space. That earlier version had been trained on approximately 11,000 compounds collected from scientific literature. Bayer provided an additional ~16,000

Received: September 26, 2014

Published: November 25, 2014

compounds with  $pK_a$  values measured internally. The collaboration came to fruition in 2013 when the expanded S+pKa model was delivered to Bayer and was subsequently incorporated into version 7.0 of ADMET Predictor. This article describes how that model was constructed, validated, and applied in industrial settings.

## DATA SETS

One data set—the “Public Set”—was seeded with data from the 2005 version of Biobyte’s Masterfile compilation of measured  $pK_a$  values reported in the scientific literature.<sup>27</sup> Over the years, this subset has been curated at Simulations Plus: duplicates were removed, erroneous chemical structures were corrected, doubtful values were verified with original sources, etc. In addition, Simulations Plus has added a large amount of published  $pK_a$  data not present in the Masterfile. The resulting Public Set consists of 10,810 chemical compounds with a total of 14,176  $pK_a$  values.

A second data set—the “Industrial Set”—consists of 15,980 small molecules from drug discovery programs at Bayer with 19,464 associated  $pK_a$  values experimentally determined at Bayer Pharma. Bayer originally provided Simulations Plus with 25,008  $pK_a$  values and their associated chemical structures, all of which underwent rigorous quality and consistency checks. Many of these molecules had multiple ionizable groups, making the automated predicted vs observed matching procedure built into ADMET Predictor<sup>6</sup> crucial in associating observed  $pK_a$  values with specific macrostate transitions. Those  $pK_a$  values that could not be rationalized by intermediate models were automatically flagged for detailed consideration. Analysis of these outliers by scientists at Simulations Plus and Bayer resulted either in their annotation and use in model training or, if no unambiguous interpretation of the  $pK_a$  transition in the structural context could be made, in their exclusion from further consideration. This rigorous and extensive vetting process was performed to ensure that only high quality data were used to train the model.

All  $pK_a$  measurements were performed at Bayer Pharma’s Wuppertal and Berlin research sites, most in an automated medium throughput format with the Sirius T3 system from Sirius Analytical. The Sirius T3 can measure one sample in around 4 min, using 5  $\mu$ L of 10 mM DMSO stock solution and collecting up to 50 data points from pH 2.0 to pH 12.0 by UV detection. Some of the  $pK_a$  data were measured with the older Sirius system SGA Profiler previously in use. Warfarin (mean  $pK_a = 4.12 \pm 0.06$  from 73 measurements) and diclofenac (mean  $pK_a = 4.96 \pm 0.05$  from 84 measurements) were used as internal reference standards. Generated results were visually checked for consistency and plausibility before entry into the Bayer database. In our experience, most erroneous experimental  $pK_a$  values in the Bayer database originate from misassignment of structures rather than from errors in measurement per se. Such mistakes can be made at the point of registration after synthesis or because of subsequent degradation. Tautomerism is also a source of structural ambiguity.

The combined Public and Industrial Sets were split into the Training Pool and Internal Test subsets with the aid of Kohonen Mapping<sup>28</sup> as implemented in the ADMET Modeler module of ADMET Predictor. This split resulted in a Training Pool of 25,509  $pK_a$  values: 10,594 from the Public Set and 14,915 from the Industrial Set. The Internal Test pool consisted of 3582 + 4549 = 8131  $pK_a$  values, 3582 from the Public Set and 4549 from the Industrial Set. The Training pool

was used to build 10 ANNE submodels using ADMET Modeler ANN engine<sup>26</sup> augmented with specialized in-house software. The Internal Test pool was used to select the 10 ANNEs that appear in the final model but were not otherwise used for model training.

Three additional sets of small molecules from drug discovery programs at Bayer that were not used for training were used for external validation. Test Set 1 consisted of compounds with relatively high structural similarity to the Industrial Set. Structural similarity was judged using the Tanimoto similarity score based on “166 MACCS” fingerprints.<sup>29</sup> Almost all (98%) of the 4730 compounds (with 5644  $pK_a$  values) in Test Set 1 have a Tanimoto similarity score  $\geq 0.80$ .

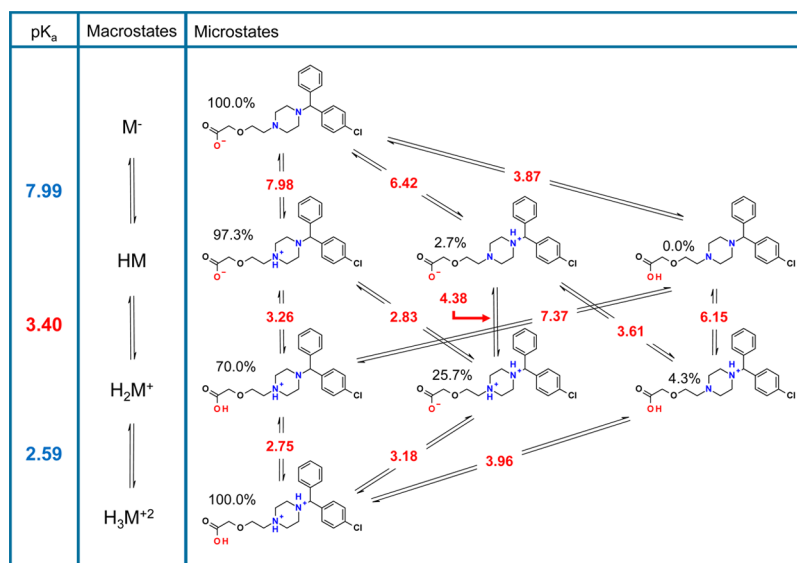
The 9168 experimental  $pK_a$  values in Test Set 2 represent the strongest predominantly acidic (provided  $pK_a \leq 9$ ) or basic (provided  $pK_a \geq 5$ ) transition in a molecule. All of the acid and base  $pK_a$  assignments in Test Set 2 were verified by visual inspection. The rationale behind the selection process for Test Set 2 was that such  $pK_a$  values determine the predominant charge state in the physiologically most relevant pH range between 5 and 9. These compounds are significantly less similar to the Industrial Set; only 60% of the 8931 compounds have a Tanimoto similarity score  $\geq 0.80$ . Test Set 3 contains 12,951 compounds (with 16,404  $pK_a$  values) which are most dissimilar to the Industrial Set with only 45% having a Tanimoto similarity score  $\geq 0.80$ . The  $pK_a$  measurements for these compounds were all performed after the measurements for those in the Industrial Set used to construct the model. Test Set 3 is regarded as the most challenging of the three external test sets.

## METHODS

**Implementation of Predictive Model S+pKa.** The S+pKa prediction models described here are generated by an exhaustive microstates analysis, the general principles and implementation of which have been described elsewhere.<sup>5,6</sup> Briefly, the fully protonated and fully deprotonated forms of the molecule of interest are generated, along with all the intermediate protonation states. The dissociation constants between microstates differing in protonation at one site are estimated by QSPR models, and the system of equations produced is solved to yield the observable macroscopic  $pK_a$  values as well as the relative populations of microstates that have the same overall charge, i.e., those which make up a single macrostate. Figure S1 of the Supporting Information illustrates the overall mechanics of predicting ionization microconstants for specific protonated atoms in specific microstates with the aid of QSPR models for the various ionizable centers.

The inputs for the QSPR models were atomic descriptors whose values depend on each ionizable atom’s type and molecular environment, including the protonation status of other ionizable atoms. Figure S2 of the Supporting Information shows some of the atomic descriptors used, with cetirizine (an antiallergy drug known in the United States under the name Zyrtec) as an illustrative example.<sup>26</sup>

In our implementation, the QSPR models are artificial neural network ensembles (ANNEs)<sup>28</sup> trained using the ADMET Modeler model-building module of the ADMET Predictor augmented with specialized in-house software at Simulations Plus, Inc. A set of 10 unique ANNE covers the following major types of ionizable atoms contained in the training set of chemical compounds:



**Figure 1.** Detailed model output of the v7.0 S+pKa predicted macroscopic (left column) and microscopic (right column) dissociation constants for the respective macrostates and microstates of cetirizine, which is triprotic. “M” represents the base structure. Black percentage numbers shown next to microstates illustrate relative contribution (probability) of this microstate to the corresponding macrostate. Red numbers along the reaction arrows illustrate pK<sub>a</sub> microconstants.

AP predicted pK <sub>a</sub>	AP predominant charge state at pH 7.40	AP CSno	AP highest basic pK <sub>a</sub> (strongest base)	AP most basic atom number (group)	AP lowest acidic pK <sub>a</sub> (strongest acid)	AP most acidic atom number (group)	AP 2nd highest basic pK <sub>a</sub> (strongest base)	AP 2nd most basic atom number (group)
	zwitterionic [79.5%]	2	7.99	6(>N-) [97.3%]	3.4	15(-OH) [74.2%]	2.59	3(>N-) [70.0%]

**Figure 2.** Medicinal chemist use case illustrated for the drug cetirizine. Predictions were generated with version 7.0 of the S+pKa model.

1. Hydroxy acids
2. Acidic amides
3. Acids of aromatic NH
4. Thioacids
5. Carboacids
6. Amines
7. Bases of aromatic N
8. N-oxides
9. Thiones
10. Carbobases (protonable C in certain  $\pi$ -excessive rings)<sup>30–32</sup>

The 10 ANNEs were trained in parallel against the Training subset of 25,509 pK<sub>a</sub> values described in the section above. Once all microconstants were predicted, the known mathematical relationships between them<sup>18</sup> were used to calculate pK<sub>a</sub> macroconstants, microstate percentages, proton dissociation probabilities, etc. A plot of predicted macroconstants and microconstants for cetirizine is shown in Figure 1.

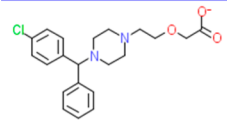
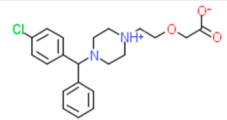
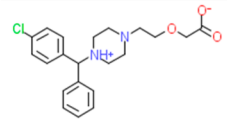
Experimental determination of microstates populations requires more specialized experimental techniques than those that are routinely employed.<sup>4,10,13,17–19</sup> Such work is being done nonetheless, particularly by Hungarian chemists, and it has been successful in certain relatively simple cases. One recent report involved the aqueous ionization of cetirizine.<sup>9</sup> Figure S3 of the Supporting Information shows the

experimentally determined ionization constants and microstate distributions for cetirizine, which are in good agreement with those generated using S+pKa 7.0 (Figure 1).

**pK<sub>a</sub> Predictions.** In order to gauge the impact of including the Industrial Set on predictive performance, two series of predictive pK<sub>a</sub> calculations were performed on each of the external test sets: one with an older version of the S+pKa model (from ADMET Predictor 6.0) trained with Public Set data only and another with S+pKa from ADMET Predictor 7.0, which was trained with the combined Public and Industrial Sets. In the following sections, the two series are labeled as “v6.0” and “v7.0”, respectively.

Predicted and experimental pK<sub>a</sub> were paired using the built-in matching function of ADMET Predictor.<sup>26</sup> Matched data sets were exported as tab delimited text files and further analyzed in Microsoft Excel to plot graphs and calculate prediction statistics. Bayer Pharma’s proprietary Pharmacophore Informatics software Pix was used to calculate error distributions for the three Bayer external test sets.

**Implementation and Deployment of Final pK<sub>a</sub> Prediction Tool at Bayer.** Simulations Plus had delivered the final pK<sub>a</sub> prediction tool as a LINUX executable to Bayer. The LINUX executable was then programmed into a Pipeline Pilot node that supports four different groups of users:

Structure	AP overall charge state	AP microstate prevalence [%] at pH=7.40	AP macroscopic pKa
	-1	20.46	
	0	77.39	7.99
	0	2.14	7.99

Microstate prevalence threshold:

50% } 10% } 1%

**Figure 3.** Computational chemist use case illustrated for the drug cetirizine. Predictions were generated with version 7.0 of the S+pKa model.

1. *Medicinal Chemists.* Plx contains chemical spreadsheet functionality frequently used by Bayer's medicinal chemists, who typically analyze  $pK_a$  values in the context of chemical structures and determine the predominant charge state at pH 7.4 (e.g., for SAR analysis). For this application, a simplified table view on macrostate  $pK_a$  values resulting from microstate equilibria is provided as output. An illustration for the drug cetirizine is given in Figure 2. The predominant charge state at pH 7.4 is reported together with its percentage prevalence. For easier table sorting and grouping, the predominant charge state is also given as a number (1, negatively charged; 2, zwitterionic; 3, neutral; and 4, positively charged). Not more than two most acidic and basic  $pK_a$  values in the range 0 to 14 are reported. Assignment of a macrostate to specific groups/atoms can be risky, but medicinal chemists want to have an idea of which atoms have the strongest influence on the  $pK_a$  of interest. This challenge was addressed by computing the *relative degree of influence* ( $RDI(j)$ ) for each ionizable group  $j$  in a given macrostate

$$RDI(j) = 100 \times \sum_{i=1}^M MC_i \times dp(j)_i \quad (1)$$

where  $M$  is the number of microstates in a given macrostate,  $MC_i$  is the  $i$ -th microstate contribution (expressed as a fraction), and  $dp(j)_i$  stands for the dissociation probability of the  $j$ -th proton in the  $i$ -th microstate. For example, using experimental data for the  $H_2M^+$  to  $HM$  transition ( $pK_a = 3.01$ ) from Figure S3 of the Supporting Information, one can calculate  $RDI(-OH) = 100 \times (0.648 \times 0.9999 + 0.012 \times 0.998) = 66\%$ , which identifies the  $-OH$  group as the ionizable group dominating this transition and is well matched by the value of 74.2% in Figure 2. The RDI values for dominant groups are shown in brackets. The same principle of dominance governs the "acidic" and "basic" labeling. In rare cases, a dominant atom may not be found. For example, each ionization transition in mellitic acid is characterized by an equal  $RDI = 1/6 = 16.7\%$  for each carboxylate, so no single group is dominant.

2. *Computational Chemists.* Different microspecies can be expected to interact differently with a binding site, and one particularly challenging task for this group of users is the preparation of large libraries for high-throughput docking applications (e.g., virtual screening). Such preparation requires

rapid automatic adjustment of structures' protonation states to reflect their dominant charge state(s) at a given pH. An illustration for the drug cetirizine is given in Figure 3. The pH (default 7.4) can be set as an input parameter to reflect assay conditions (some proteins operate *in vivo* in an acidic or basic environment). The user can also set the microstate prevalence threshold (default 1%). In virtual screening, it makes sense to also consider minor protonation states if they occur with an appreciable prevalence because it might very well be a minor protonation state that interacts with high affinity with the target protein. Considering minor protonation states as well will increase the number of structures generated for docking, however. In certain cases, one might therefore opt for a larger microstate prevalence threshold, which can be increased up to 50%, at which point only the most dominant protonation state will be generated for each molecule. The software's fast calculation speed (>100,000 compounds/h [CPU: Intel Xeon L5420, 2.5 GHz]) allows overnight processing of large libraries with several million structures.

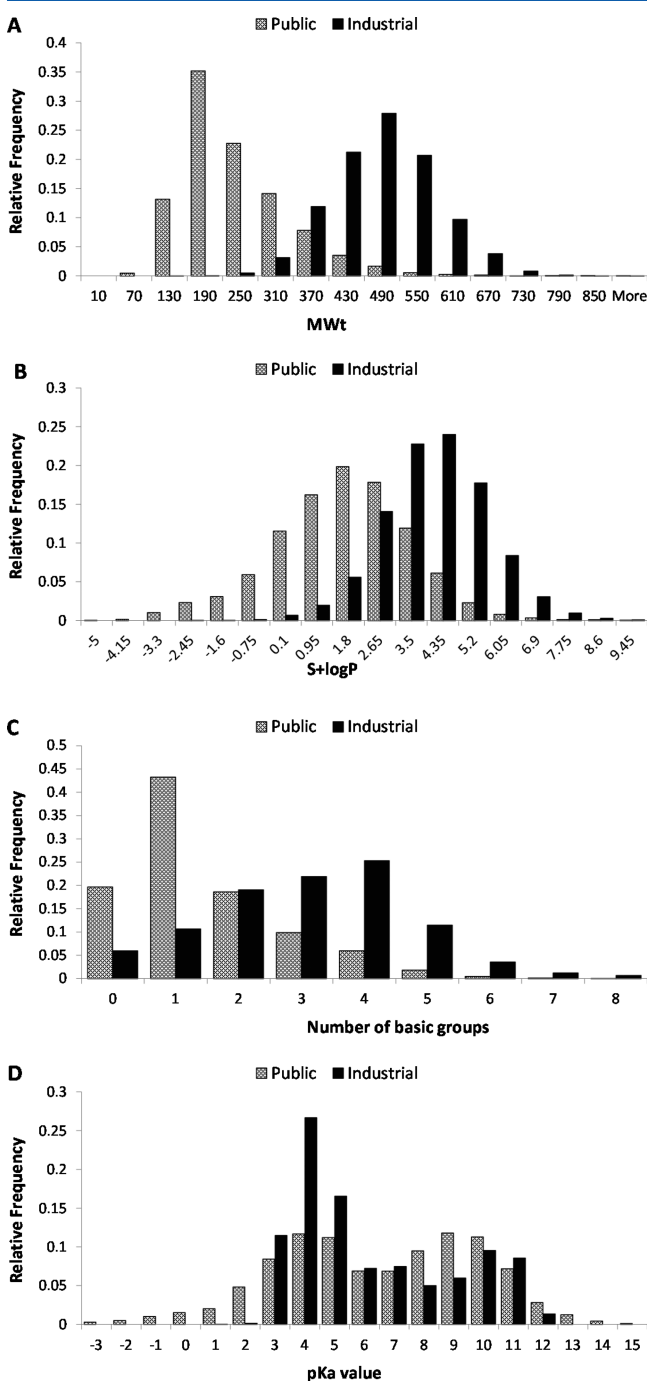
3. *Physical and Analytical Chemists.* This workflow calls for a table of all microstates including their calculated microstate prevalence percentage up to a given microstate prevalence threshold, which is provided as input. The default value for this threshold is 0%, so information on all possible microstates is provided analogous to the cetirizine example in Figure 1. Raising the threshold to a higher value will filter out microstates with a prevalence below this threshold, which is a useful way to filter out microstates with a very low prevalence and thereby simplify the resulting table.

4. *Cheminformaticians.* These users develop predictive models, e.g., for *in silico* ADMET analysis. The pH can be set as an input parameter to reflect assay conditions (e.g., human plasma protein binding at pH 7.4, buffer solubility at pH 6.5 [pH of the lower small intestine], logD at pH 2.3). The predominant charge state with its calculated prevalence percentage is delivered as output and can be used as descriptors for the development of predictive models.

## RESULTS

**Comparison of Chemical Subspaces.** Access to the Industrial Set (see Data Sets section) provides a unique opportunity to compare the chemical space covered by it with the chemical space covered by the Public Set. The results were

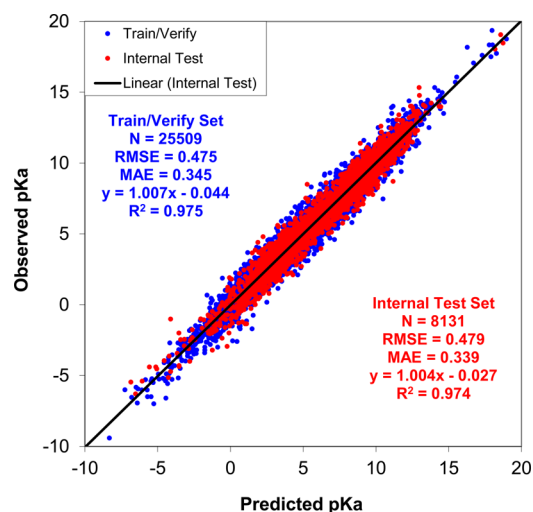
quite compelling; Industrial and Public Sets differ substantially (Figure 4). For example, Industrial compounds are generally significantly heavier, more lipophilic, and possess more basic ionizable groups. The distribution of  $pK_a$  values in the Public Set has a balanced bimodal distribution with peaks at  $\sim 4$  and  $\sim 9$ , whereas the Industrial Set  $pK_a$  values are significantly skewed toward 4. Taken together with the differential distribution of the number of basic groups, this clustering of  $pK_a$  values around 4 reflects the prevalence of weak bases (aromatic nitrogen) rather than carboxylic acids in this set.



**Figure 4.** Comparative distribution of (A) molecular weight in daltons, (B) lipophilicity as predicted S+logP, (C) number of basic functional groups, and (D)  $pK_a$  values for the Public and Industrial Sets described in text.

A similar observation was made by Mannhold et al. when comparing the public literature's chemical space with that of Pfizer and Nycomed drug-like research compounds.<sup>33</sup> They cited this discrepancy as the main factor responsible for unsatisfactory performance of logP predictors trained on public data. By analogy, the significant difference between Public and Industrial chemical spaces is most likely responsible for the shortcomings of previous commercial  $pK_a$  predictors.

**Validation of S+pKa Model.** Figure 5 shows the performance of version 7.0 of the S+pKa model on the



**Figure 5.** Performance graph for the S+pKa model trained on the combined Public and Industrial data. Only a portion of this set (25,509  $pK_a$  values, blue points) was used for the actual S+pKa model building; the remaining 8131  $pK_a$  values (red points) form the internal test set which helped in selection of the final 10 ANNs out of hundreds of prototypes. Predictive statistics: MAE = mean absolute error, RMSE = root-mean-square error, and  $R^2$  = determination coefficient. Linear equations,  $y = ax + b$ , illustrate best fit lines to the respective subsets of points.

Training Pool and Internal Test Set. The root-mean-square errors (RMSEs) and mean absolute errors (MAEs) are very comparable, which confirms that overtraining has been avoided.

The predictive accuracy does not degrade at all in the so-called “physiological range” of 6–9 log units, where the statistics for 5015 training compounds were RMSE = 0.45 and MAE = 0.34. For 1576 test compounds, the corresponding statistics were RMSE = 0.45 and MAE = 0.33. All these numbers are commensurate with the overall statistics.

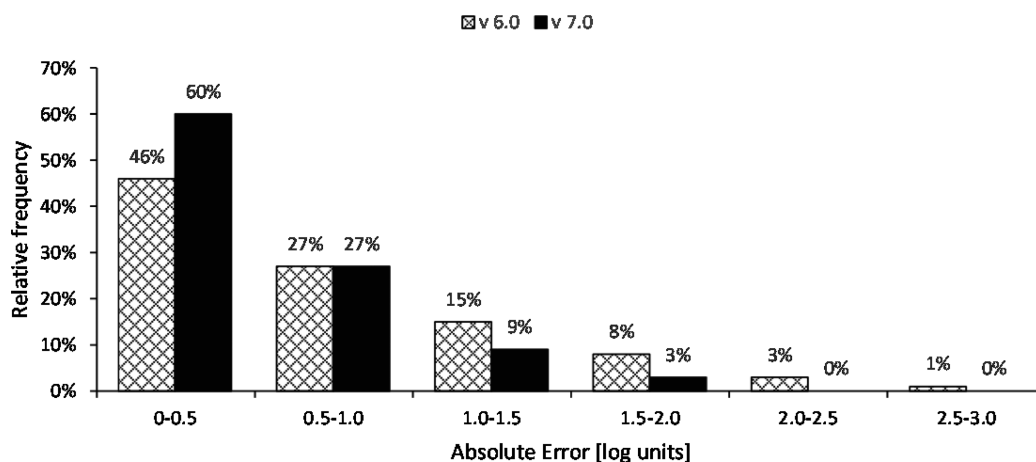
S+pKa v7.0 was also rigorously tested at Bayer to compare its performance to that of its predecessor, S+pKa v6.0, using the three external Test Sets described in the Data Sets section. Note that none of these sets were used in model training. Table 1 presents the results in terms of three key performance statistics.

By all measures, and across all three external test sets, the improvements in the accuracy of  $pK_a$  prediction were dramatic. MAEs, for example, are reduced by 33–50%. Moreover, the new model is robust; MAEs fall near 0.5 log units regardless of the degree of Tanimoto similarity to the Industrial Set used to train the model. Particularly compelling are results for the third most stringent test set of compounds whose  $pK_a$  values were measured *after* the Bayer data sourcing of the Industrial Set. Figure 6 shows the detailed distribution of predictive absolute errors in this case.

**Table 1. Performance Statistics of Two Versions of the S+pK<sub>a</sub> model: One Trained on Public Set Only (marked “v 6.0”) and the Other on the Combined Public and Industrial Sets (marked “v 7.0”) <sup>a</sup>**

test set	number of compounds	number of pK <sub>a</sub> values	average closest Tanimoto similarity to the Industrial Set	fraction of Tanimoto similars (score ≥0.80)	MAE		RMSE		R <sup>2</sup>	
					v 6.0	v 7.0	v 6.0	v 7.0	v 6.0	v 7.0
1	4730	5644	0.88	98%	0.82	0.41	1.03	0.58	0.85	0.95
2	8931	9168	0.82	60%	0.79	0.52	1.04	0.71	0.76	0.89
3	12,951	16,404	0.79	45%	0.72	0.50	0.94	0.67	0.87	0.93

<sup>a</sup>External Test Sets 1, 2, and 3 have been described in the Data Sets section. Predictive statistics: MAE = mean absolute error, RMSE = root mean square error, and R<sup>2</sup> = determination coefficient.



**Figure 6.** Distribution of absolute errors of prediction (in log units) determined on Test Set 3. The graph compares two versions of the S+pK<sub>a</sub> model: one trained on Public Set only (marked “v 6.0”, cross-hatched bars) and the other on the combined Public and Industrial Sets (marked “v 7.0”, solid bars).

Comparing S+pK<sub>a</sub> v 7.0 to S+pK<sub>a</sub> v 6.0, the percentage of small errors (under 0.5 log units) increased from 46% to 60%. Unlike the latter, the former has no deviations above 2 log units for this test set.

To put the performance of the earlier model in proper perspective, a comparison with another commercial pK<sub>a</sub> predictor, the desktop version of the ACD/pK<sub>a</sub> DB (v 12.0), was made.<sup>34</sup> Due to the limited throughput of this desktop version, a representative subset of 1000 compounds was chosen from Test Set 3 and processed with the ACD/Labs software. Nineteen of these could not be processed because of missing specific functional groups in the software’s database, so the final test set was comprised of 981 compounds. ACD/pK<sub>a</sub> DB v 12.0 (MAE = 0.77) and S+pK<sub>a</sub> v 6.0 (MAE = 0.73) showed comparable predictive accuracy on this subset, whereas S+pK<sub>a</sub> v 7.0 was superior in performance (MAE = 0.51), with an accuracy comparable to that for the whole of Test Set 3 (MAE = 0.50).

## DISCUSSION

The S+pK<sub>a</sub> model we describe is based on empirically derived equations known as quantitative structure–property relationships (QSPR). The advantage of QSPR models is the fact that predictions can be made for every submitted structure regardless of its origin. In contrast, many other computer programs for predicting pK<sub>a</sub> rely on an older perturbation-based approach.<sup>6</sup> Such programs carry vast databases of experimental pK<sub>a</sub> data. If a submitted chemical structure is found in the underlying database, then the program simply looks up and returns the stored experimental ionization constants for this compound (known as “pK<sub>a</sub><sup>0</sup>” values). No

predictive work is done in this case. If the submitted structure is not found, then such a program first finds similar structures and combines their experimental pK<sub>a</sub> to obtain an appropriate “pK<sub>a</sub><sup>0</sup>” as baseline. Next, the predictive work is done by adding empirical correction factors (perturbations) to the obtained “pK<sub>a</sub><sup>0</sup>” values. For this reason, the predictive performance of perturbative models tends to be uneven. They usually perform well against data sets drawn from the open literature,<sup>7</sup> which is also the source of their internal databases of “pK<sub>a</sub><sup>0</sup>” values, but they may perform poorly when truly external test data are used,<sup>2,3,8,12,15,24,25</sup> for example, in-house compounds from AstraZeneca.<sup>8</sup> On average, the RMSEs of the external predictions were 1 log unit or greater. The report by Milletti et al., where the predictive pK<sub>a</sub> model MoKa was successfully expanded to and validated on the Roche chemical space, is an exception.<sup>11</sup> The drop of RMSE from 1.09 to 0.49 for the Roche test set was the result of retraining MoKa with 6226 additional pK<sub>a</sub> values from the Roche in-house library. However, the MoKa model, retrained with Roche data, is not publically available at this time.

Many pharmaceuticals and agrochemicals of current interest are multiprotic and have complex dissociation patterns. Many significant issues arise from this fact, as has been discussed elsewhere.<sup>5,6,14</sup> Knowledge of all the microconstants for a multiprotic compound provides a great deal of useful information about the protonation microstates, such as their relative contributions to the corresponding macrostates as well as, for a given microstate, the relative probability of dissociation for each of its attached protons.

Ambiguities in the assignment of specific functional groups to pK<sub>a</sub> transitions make direct modeling of macroconstants

from chemical structure difficult at best. Models constructed in that way are generally limited in scope, for example, to a narrow range of chemical classes. Our desire to have a robust globally predictive model of protic ionization led us to rely on rigorous microstates analysis (see Supporting Information for further details). In contrast to macroconstants, microconstants are properties of individual groups and depend directly on the structure of the microstate involved. Therefore, microconstants are natural candidates for use in QSPR modeling. Unfortunately, microconstants are functions of the molecular environment. For example, the basicity of the distal nitrogen in cetirizine strongly depends on the protonation states of the other two groups. Molecular descriptors must be sophisticated enough to take this fact into account.

Taken together, these considerations lead us to believe that the predictive power of the S+pKa model stems from a combination of several factors, such as its application of rigorous microstates analysis, its use of high-performance ANNE modeling tools, the large amounts of high quality data from Bayer upon which it is based, and careful curation of data taken from the open literature.

Our results show that the regions of chemical space occupied by compounds in the public domain differ substantially from that occupied by compounds currently of interest to the pharmaceutical industry. Years of predictive modeling experience have taught us that the coverage of chemical space is one of the most important features of any predictive model. It is no surprise then that the S+pKa predictions have improved almost 2-fold from v 6.0 (trained exclusively with data in the public domain only) to v 7.0, which was trained with data from both the public and industry domains.

The value of the S+pKa model described here goes beyond the raw pK<sub>a</sub> values and performance statistics it provides. It also offers a detailed insight into multiprotic ionization, accounting for distributions across all microstates. Information on the prevalence of minor microstates, for example, can be critical to understanding the binding of ligands to their biological targets and to reaction kinetics. The different microstates contributing to any given macrostate are simply tautomers of each other, which means that their distribution can ultimately be used to estimate the relative prevalence of various tautomeric forms. Work on exploiting this fact is ongoing at Simulations Plus.

The uniqueness of the collaboration between Simulations Plus and Bayer is grounded in the fact that a huge portion of the available data on pK<sub>a</sub> values at the pharma company were made available to the software partner company. In order to collaboratively curate the data sets, detailed structural information had to be exchanged. Moreover, in contrast to similar undertakings by others, the resulting product is commercially available.

Another beneficial aspect of the collaboration between Simulations Plus and Bayer is the unique insight offered to model builders; it provided a chance to learn exactly how and by whom the S+pKa model would be used. The wealth of information produced by the model enables its output to be tailored to the needs of the various groups of scientists who will use it in pharmaceutical R&D: medicinal, computational, physical and analytical chemists, and cheminformaticians.

Bayer Pharma has been very active over the past 10 years in developing a portfolio of *in silico* ADMET prediction tools<sup>35–41</sup> and has firmly integrated them into its drug discovery process. We believe that being able to make good predictions of pH-dependent ionization and charge states from a compound's

chemical structure will enable us to make further progress with the *in silico* prediction of PhysChem and ADMET properties.

Experimental scientists at Bayer not only measure pK<sub>a</sub> values but also annotate reported results. For this purpose, the new pK<sub>a</sub> prediction tool is regularly used to help them classify transitions as predominantly acidic or basic and add comments to differentiate between multiple acidic or basic transitions for single molecules. This offers the opportunity to give immediate feedback if novel ionizable groups are encountered that are not yet recognized or not well predicted by the new pK<sub>a</sub> prediction tool. The new pK<sub>a</sub> model has also been implemented at Bayer CropScience and was perceived as a significant improvement over previously used pK<sub>a</sub> prediction tools.

## CONCLUSION

It is amazing what a combination of good methodology and good data can achieve. In spite of recent warnings of the demise of QSAR,<sup>42–44</sup> our work clearly shows that putting these two “goods” together can produce a very useful predictive model of a critical physicochemical property. Generic aspects of good QSAR methodology have been discussed at length.<sup>45–49</sup> The mentioned negative perception of QSAR stems from the neglect of either or both of these key ingredients.

Here, “good methodology” refers to the need to build a model on top of a physicochemical scheme that accurately represents the phenomenon of interest. In particular, any global model of macroscopic pK<sub>a</sub> must take microstates into account if it is to be robust and widely applicable.

“Good data” refers not just to the experimental quality and rigorous curation but also to the coverage of chemical space of interest to the end user. Including small molecules from drug discovery programs at Bayer in model training greatly enhanced predictive performance on pharmaceutically relevant compounds. This is not a surprising observation, but it is one that bears repeating. If other companies follow Bayer's example and open at least some of their internal data to external collaborators, QSPR and QSPR predictions are likely to improve, which will benefit their own researchers as well as others.

## ASSOCIATED CONTENT

### Supporting Information

Fundamental modeling paradigm, examples of atomic descriptors, concise explanation of the science behind multiprotic ionization, and model comparisons on the publicly available industrial data sets. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Authors

\*E-mail: [robert@simulations-plus.com](mailto:robert@simulations-plus.com) (R.F.).

\*E-mail: [mario.lobell@bayer.com](mailto:mario.lobell@bayer.com) (M.L.).

### Notes

The authors declare no competing financial interest.

## REFERENCES

- (1) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51*, 817–834.
- (2) Balogh, G. T.; Gyarmati, B.; Nagy, B.; Molnar, L.; Keseru, G. M. Comparative evaluation of *in silico* pK<sub>a</sub> prediction tools on the gold standard dataset. *QSAR Comb. Sci.* **2009**, *28*, 1148–1155.

- (3) Balogh, G. T.; Tarcsey, A.; Keseru, G. M. Comparative evaluation of  $pK_a$  prediction tools on a drug discovery dataset. *J. Pharm. Biomed. Anal.* **2012**, *67–68*, 63–70.
- (4) Borkovec, M.; Brynda, M.; Koper, G. J. M.; Spiess, B. Resolution of microscopic protonation mechanisms in polyprotic molecules. *Chimia* **2002**, *56*, 695–701.
- (5) Fraczkiewicz, R. *In silico* Prediction of Ionization. In *Comprehensive Medicinal Chemistry II*; Testa, B., van de Waterbeemd, H., Eds.; Elsevier: Oxford, U.K., 2006; Vol. 5, pp 603–626.
- (6) Fraczkiewicz, R. *In silico* Prediction of Ionization. In *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering [Online]*, Reedijk, J., Ed.; Elsevier: The Netherlands, 2013, Vol. 5, Chapter 5.25. <http://www.sciencedirect.com/science/article/pii/B978012409547202610X> (accessed November 4, 2014).
- (7) Liao, C.; Nicklaus, M. C. Comparison of nine programs predicting  $pK_a$  values of pharmaceutical substances. *J. Chem. Inf. Model.* **2009**, *49*, 2801–2812.
- (8) Manchester, J.; Walkup, G.; Rivin, O.; You, Z. Evaluation of  $pK_a$  estimation methods on 211 druglike compounds. *J. Chem. Inf. Model.* **2010**, *50*, 565–571.
- (9) Marosi, A.; Kovacs, Z.; Beni, S.; Kokosi, J.; Noszal, B. Triprotic acid-base microequilibria and pharmacokinetic sequelae of cetirizine. *Eur. J. Pharm. Sci.* **2009**, *37*, 321–328.
- (10) Mernissi-Arifi, K.; Schmitt, L.; Schlewer, G.; Spiess, B. Complete resolution of the microscopic protonation equilibria of D-myo-inositol 1,2,6-tris(phosphate) and related compounds by  $^{31}\text{P}$  NMR and potentiometry. *Anal. Chem.* **1995**, *67*, 2567–2574.
- (11) Milletti, F.; Storch, L.; Goracci, L.; Bendels, S.; Wagner, B.; Kansy, M.; Cruciani, G. Extending  $pK_a$  prediction accuracy: High-throughput  $pK_a$  measurements to understand  $pK_a$  modulation of new chemical series. *Eur. J. Med. Chem.* **2010**, *45*, 4270–4279.
- (12) Milletti, F.; Storch, L.; Sforza, G.; Cruciani, G. New and original  $pK_a$  prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.
- (13) Peinhardt, G.; Wiese, M. Microionization constants: Novel approach for the determination of the zwitterionic equilibrium of hydroxyphenylalkylamines by photometric titration. *Int. J. Pharm.* **2001**, *215*, 83–89.
- (14) Rupp, M.; Koerner, R.; Tetko, I. V. Predicting the  $pK_a$  of small molecules. *Comb. Chem. High Throughput Screening* **2011**, *14*, 307–327.
- (15) Settimo, L.; Bellman, K.; Knegtel, R. A. Comparison of the accuracy of experimental and predicted  $pK_a$  values of basic and acidic compounds. *Pharm. Res.* **2014**, *31*, 1082–1095.
- (16) Shields, G. C.; Seybold, P. G. *Computational Approaches for the Prediction of  $pK_a$  Values*; CRC Press: Boca Raton, FL, 2014; p 155.
- (17) Szakacs, Z.; Kraszni, M.; Noszal, B. Determination of microscopic acid-base parameters from NMR-pH titrations. *Anal. Bioanal. Chem.* **2004**, *378*, 1428–1448.
- (18) Szakacs, Z.; Noszal, B. Protonation microequilibrium treatment of polybasic compounds with any possible symmetry. *J. Math. Chem.* **1999**, *26*, 139–155.
- (19) Tam, K. Y. Multiwavelength spectrophotometric determination of acid dissociation constants. Part VI. Deconvolution of binary mixtures of ionizable compounds. *Anal. Lett.* **2000**, *33*, 145–161.
- (20) Tam, K. Y. Multiwavelength spectrophotometric resolution of the micro-equilibria of a triprotic amphoteric drug: Methacycline. *Mikrochim. Acta* **2001**, *136*, 91–97.
- (21) Tam, K. Y.; Quere, L. Multiwavelength spectrophotometric resolution of the micro-equilibria of cetirizine. *Anal. Sci.* **2001**, *17*, 1203–1208.
- (22) Tam, K. Y.; Takacs-Novak, K. Multiwavelength spectrophotometric determination of acid dissociation constants: Part II. First derivative vs. target factor analysis. *Pharm. Res.* **1999**, *16*, 374–381.
- (23) Tam, K. Y.; Takacs-Novak, K. Multiwavelength spectrophotometric determination of acid dissociation constants: A validation study. *Anal. Chim. Acta* **2001**, *434*, 157–167.
- (24) Wan, H.; Ulander, J. High-throughput  $pK_a$  screening and prediction amenable for ADME profiling. *Exp. Opin. Drug Metab. Toxicol.* **2006**, *2*, 139–155.
- (25) Wang, J.; Skolnik, S. Recent advances in physicochemical and ADMET profiling in drug discovery. *Chem. Biodiversity* **2009**, *6*, 1887–1899.
- (26) ADMET Predictor, version 7.0; Simulations Plus, Inc.: Lancaster, CA, 2014.
- (27) BioByte Masterfile, version 2008; BioByte Corp.: Claremont, CA, 2008.
- (28) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999; p 380.
- (29) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 1273–1280.
- (30) Abraham, R. J.; Bullock, E.; Mitra, S. S. Physical properties of alkyl pyrroles and their salts. *Can. J. Chem.* **1959**, *37*, 1859–1869.
- (31) Chiang, Y.; Whipple, E. B. The protonation of pyrroles. *J. Am. Chem. Soc.* **1963**, *85*, 2763–2767.
- (32) Hinman, R. L.; Whipple, E. B. The protonation of indoles: Position of protonation. *J. Am. Chem. Soc.* **1962**, *84*, 2534–2539.
- (33) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2008**, *98*, 861–893.
- (34) ACD/Percepta, version 12.0; ACD/Labs, Inc.: Toronto, Canada, 2012.
- (35) Wunberg, T.; Hendrix, M.; Hillisch, A.; Lobell, M.; Meier, H.; Schmeck, C.; Wild, H.; Hinzen, B. Improving the hit-to-lead process: Data-driven assessment of drug-like and lead-like screening hits. *Drug Discovery Today* **2006**, *11*, 175–180.
- (36) Lobell, M.; Hendrix, M.; Hinzen, B.; Keldenich, J.; Meier, H.; Schmeck, C.; Schohe-Loop, R.; Wunberg, T.; Hillisch, A. *In silico* ADMET traffic lights as a tool for the prioritization of HTS hits. *Chem. Med. Chem.* **2006**, *1*, 1229–1236.
- (37) Göller, A. H.; Hennemann, M.; Keldenich, J.; Clark, T. *In silico* prediction of buffer solubility based on quantum-mechanical and HQSAR and topology-based descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 648–658.
- (38) Hennemann, M.; Friedl, A.; Lobell, M.; Keldenich, J.; Hillisch, A.; Clark, T.; Göller, A. H. CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *Chem. Med. Chem.* **2009**, *4*, 657–669.
- (39) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark data set for *in silico* prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.
- (40) Nisius, B.; Göller, A. H. Similarity-based classifier using topomers to provide a knowledge base for hERG channel inhibition. *J. Chem. Inf. Model.* **2009**, *49*, 247–256.
- (41) Nisius, B.; Göller, A. H.; Bajorath, J. Combining cluster analysis, feature selection and multiple support vector machine models for the identification of human ether-a-go-go related gene channel blocking compounds. *Chem. Biol. Drug Des.* **2009**, *73*, 17–25.
- (42) Doweiko, A. Is QSAR relevant to drug discovery? *IDrugs* **2008**, *11*, 894–899.
- (43) Doweiko, A. QSAR: Dead or alive? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81–89.
- (44) Johnson, S. R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2007**, *48*, 25–26.
- (45) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Env. Res.* **2009**, *20*, 241–266.
- (46) Golbraikh, A.; Tropsha, A. Beware of  $q^2$ ! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
- (47) Scior, T.; Medina-Franco, J. L.; Do, Q. T.; Martinez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to recognize and work around pitfalls in QSAR studies: A critical review. *Curr. Med. Chem.* **2009**, *16*, 4297–4313.



(48) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476–488.

(49) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.