JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Benefit of Retraining pK_a Models Studied Using Internally Measured Data

Peter Gedeck,*[,†] Yipin Lu,[‡] Suzanne Skolnik,[§] Stephane Rodde,[∥] Gavin Dollinger,[‡,○] Weiping Jia,[‡] Giuliano Berellini,[§] Riccardo Vianello,[∥] Bernard Faller,[∥] and Franco Lombardo[§,⊥]

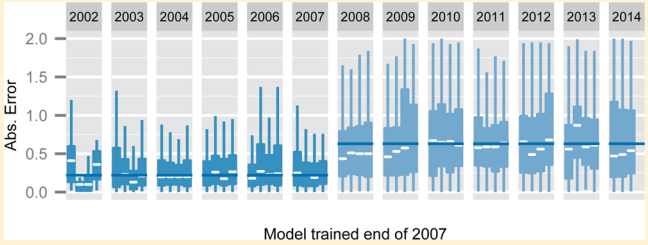
[†]Novartis Institute for Tropical Diseases Pte. Ltd., 10 Biopolis Road, #05-01 Chromos, Singapore 138670, Singapore

[‡]Novartis Institute for Biomedical Research, 5300 Chiron Way, Emeryville, California 94608, United States

[§]Novartis Institute for Biomedical Research, 250 Massachusetts Ave, Cambridge, Massachusetts 02139, United States

[∥]Novartis Institute for Biomedical Research, Postfach, CH-4002 Basel, Switzerland

ABSTRACT: The ionization state of drugs influences many pharmaceutical properties such as their solubility, permeability, and biological activity. It is therefore important to understand the structure property relationship for the acid−base dissociation constant pK_a during the lead optimization process to make better-informed design decisions. Computational approaches, such as implemented in MoKa, can help with this; however, they often predict with too large error especially for proprietary compounds. In this contribution, we look at how retraining helps to greatly improve prediction error. Using a longitudinal study with data measured over 15 years in a drug discovery environment, we assess the impact of model training on prediction accuracy and look at model degradation over time. Using the MoKa software, we will demonstrate that regular retraining is required to address changes in chemical space leading to model degradation over six to nine months.



1. INTRODUCTION

The acid−base dissociation constant pK_a directly influences the ionization state of compounds, and consequently, physicochemical and pharmaceutically relevant properties such as solubility, permeability, and biological activity.[1] The fact that almost 80% of all drugs contain ionizable groups[2] underlines the importance of understanding pK_a values and the structural features that determine them.

A variety of methods were developed over the years to predict pK_a values using computational methods. They range from computational demanding high-level ab initio approaches to fast structure-based QSAR methods. The ab initio approaches have the advantage that they calculate pK_a values from first principle and can therefore be used to calculate pK_a values not only in water, see e.g. Brown and Mora-Diez,[3] but also in nonaqueous solvents; see the work of Ding et al.[4] for an example.

The second approach requires parametrization based on experimental data. For example, Yu et al.[5,6] correlated semiempirically calculated atom properties with pK_a data for a larger set of acids and bases and obtained a number of ionization site-specific models. Similarly, Jelfs et al.[7] used semiempirical descriptors, which they combined with structural descriptors of the ionization sites to correct for specific local environments. Milletti et al.[8−10] derived atomic descriptors based on GRID's molecular interaction fields. For a new structure, these descriptors are assigned by looking up similar structural features in a fragment database and used to derive a count-based fingerprint that describes the local environment of an ionization site. These count-based fingerprints are correlated with experimental pK_a values using partial least-squares regression. The lookup of descriptors in a fragment database makes prediction very fast as it avoids time-consuming calculations. A purely atom-type based approach was proposed by Xing et al.[11]

A number of publications compared commercially available software packages. The performance characteristics vary considerably. In studies that use public domain data, most methods perform well with average prediction errors of 0.5 log units.[12−14] Manchester et al.[15,16] assessed the performance using 211 compounds from Astra-Zeneca's research projects. They report average prediction errors of 1 to 3 log units. As public domain data were used to develop most methods, it is not surprising that we see this discrepancy between public domain and proprietary data. This highlights the need for ensuring that the model is applicable to the chemical structures of interest and if it is not, the model should be trained to learn the new structural features. In a recent study, Milletti et al.[17] used an experimental data set of 9000 compounds from Roche to evaluate the impact of training using the MoKa software. They showed that training can reduce the average prediction error for new chemotypes to what is observed for public domain data. A similar study was published by Fraczkiewicz et

Received: March 30, 2015
Published: June 8, 2015

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

al.[18] for the pK_a prediction method from Simulations Plus. The initial model was trained on about 14 000 literature pK_a values. An additional set of almost 20 000 pK_a values from Bayer was used to create an updated model. Additional pK_a values from Bayer were used for testing. The additional structures from Bayer were in general larger and had a higher logP and more basic centers. Similar to the experience from Milletti et al. retraining with the Bayer data reduced the mean absolute error on the test sets from 0.8 to 0.4−0.5.

A number of studies have looked into the effect of updating models regularly with new data. Using a large data set of human plasma protein binding, Rodgers et al.[19] studied the impact of retraining QSAR models over a 2-year time period. The data set was split into an initial training set, representing the data available at the beginning of the 2 years, and several additional data sets, representing each month during this 2-year period. This procedure allowed the comparison of a static model, which was built using only the initial training set, with models that were built each month, and so made full use of all the available data at the time each model was built. They observed that the prediction quality for all compounds was not reduced with increasing model size; however, predictions for new chemotypes benefited from the updates of the models. In a more recent study, Rodgers et al.[20] extended the range of study to almost three years, again highlighting the importance of model retraining.

Wood et al.[21] studied the effect of monthly changes with an emphasis on different model updating approaches over a period of 15 months. Their results confirmed the benefit of model retraining.

In a comparable study, Gavaghan et al.[22] monitored the performance of a hERG model over 15 months. The predictive performance of the model deteriorated within 4 months of building, which illustrated the necessity of regularly updating global models. Sherer et al.[23] studied the effect of retraining of a passive permeability model. Their observation is in line with the result of Rodgers et al.

In this publication, we will describe the development of a web application for curating our internal pK_a data. This application was used to develop a curated data set of measurements determined over a period of 15 years. This data set of about 11 000 structurally annotated, experimental pK_a values was used to study the effect of retraining of the software program MoKa. First, we assessed training using a standard cross-validation approach. Second, we simulated how quarterly retraining improved the prediction of MoKa on our in-house compounds. The observed improvement will be correlated with similarity of ionization sites from the test set to the training set.

2. METHOD

2.1. Data Set. There are three experimental methods for the determination of ionization constants (pK_a) available at Novartis:

- High-throughput determination of UV-metric ionization constants
- Potentiometric determination of ionization constants
- Capillary electrophoresis determination of ionization constants

UV-metric ionization constants were determined on the commercial Spectral Gradient Analyzer (SGA) or T3 instrument (Sirius Analytical Ltd.,[24] sirius-analytical.com) as described by Allen et al.[25] Test compounds were diluted to 0.04 mM in a cosolvent mixture and titrated three times in 20−40% wt methanol. The titrations were performed at 25 °C and 0.15 M ionic strength, from pH 2 to 12 or 12 to 2 (with delta pH of 0.2) depending on the acidic or basic nature of the test compound. A linear buffer was added to allow fast pH stabilization after each titrant addition. Wavelengths from 230 to 450 nm were typically monitored for UV absorbance change due to the ionization state of the compound. Target factor analysis (TFA) was used to calculate apparent pK_as from the multiwavelength absorption data at a given percent of cosolvent (psK_a), followed by Yasuda−Shedlovsky extrapolation to 0% methanol to provide the aqueous pK_a. Acid/base assignment was performed based on the slope of extrapolation. Experimental variability was determined from 183 duplicate measurements from different days and experimentalists, with a standard deviation of 0.17 (0.35 log units).

Potentiometric ionization constants were determined on the commercial GlpKa or T3 instruments (Sirius Analytical) as described by Takács-Novák et al.[26] Briefly, 0.3 to 1 mM test solutions were titrated from pH 2 to 12 or 12 to 2, depending on the acidic or basic nature of the test compound. Titrations were conducted at 25 °C and 0.15 M ionic strength. Aqueous titrations were performed in triplicate in 0.15 M KCl, while sparingly soluble test compounds were titrated in 10−60% wt methanol, 1,4-dioxane, or dimethyl sulfoxide cosolvent. A minimum of three titrations in varying amounts of cosolvent were performed for Yasuda−Shedlovsky extrapolation to the aqueous pK_a. For each titration, initial estimates of pK_a values were obtained from Bjerrum difference plots (number of bound protons versus pH) and further refined by a weighted nonlinear least-squares procedure (see Avdeef[27,28]) available in the instrument software. Experimental variability was determined from 389 duplicate measurements from different days and experimentalists, with a standard deviation of 0.28 (0.4 log units).

Capillary electrophoresis ionization constants were determined on the commercial pK_a PRO instrument (Advanced Analytical Technologies, Inc.) as described by Shalaeva et al.[29] This system is equipped with 96 uncoated fused-silica capillaries (75 μm i.d., 200 μm o.d) and a UV absorbance detector (214 nm). Test compounds were diluted to 50−150 μg/mL in aqueous buffer premade by Advanced Analytical Technologies, Inc. The solvent consisted of 0.1−0.2% (v/v) DMSO (neutral marker) and 1−5 mM HCl or NaOH to enhance the solubility of basic or acidic compounds, respectively. In the case of some low solubility compounds, samples could be further diluted to approximately 5−10 μg/mL to avoid precipitation. The experiment was performed at 20 °C and 50 mM ionic strength from pH 1.75 to 11.2 (a total of 24 aqueous pH buffers; purchased from Advanced Analytical Technologies, Inc.) The sets of 24 aqueous pH buffers were prepared from phosphoric acid, formic acid, sodium acetate, sodium phosphate, or boric acid with the addition of sodium chloride or/and sodium hydroxide in various proportions to a level ionic strength of 50 mM. The program pK_a Estimator (Advanced Analytical Technologies, Inc.) was used to calculate the pK_a values of compounds from the electrophoresis data. Experimental variability for duplicate measurements is about 0.2 pK_a units.

For the computational estimation of pK_a values, we use MoKa 2.5.4[30] from Molecular Discovery in this study. This software implements the method developed by Milletti et al.[10]

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

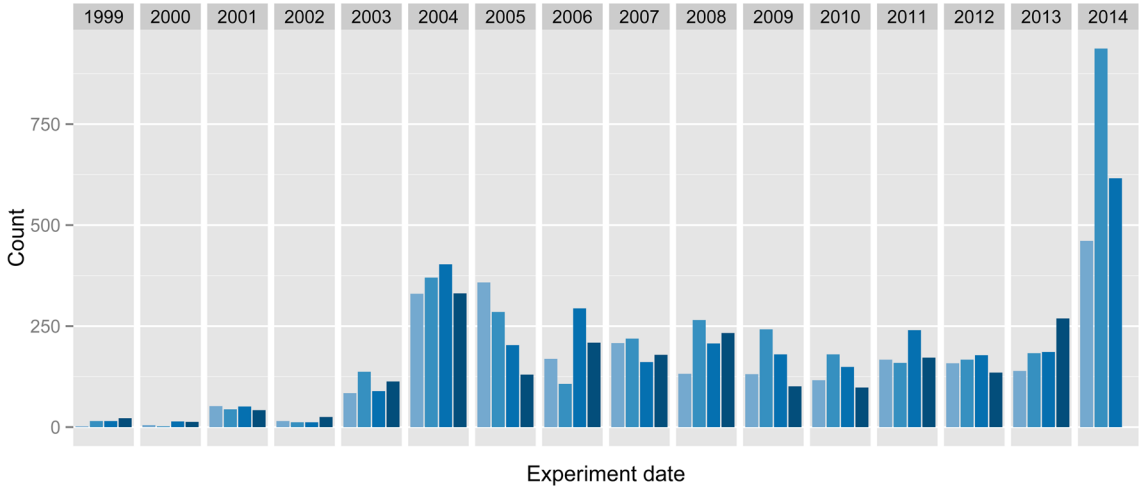


Figure 1. Number of assigned pK_a values over time grouped by year (panel) and quarter (color). Before the curation effort started, we can assign about 150−200 pK_a values per quarter. Since then, this number goes up highlighting the strong focus of the curators on more recent data. High-throughput determination of ionization constants was introduced during 2003 leading to an increase of experimental pK_a values. The additional increase in 2014 is explained by the regular use of the curation web application as part of the experimental data analysis and publishing workflow.

MoKa 2.5.4 comes with command line tools, which allow prediction of pK_a values (moka), retraining of the pK_a model (kibitzer), and prediction of tautomers (tauthor) from the command line instead of the GUI application.

2.2. Data Curation. Over the last 15 years, we measured pK_a values for over 58 000 compounds giving us a data set of almost 100 000 experimental pK_a values. Today, this number grows on average by 600 pK_a values a month. In order to organize this data set so that it can be used to train a pK_a model, it is necessary to assign the individual experimental pK_a values to ionization sites of the chemical structure. We developed a web application to facilitate this curation process and store the decisions of the experimentalists. The data are stored in a relational database.

The database contains several tables of which three are relevant for the curation process. The main table contains information about the chemical structure (identifier, SMILES). These data are added on initial loading or update of experimental pK_a data from our data warehouse. We realized early on that the registered structure does not always represent the dominant tautomer. The web application therefore also contains functionality that allows replacing the chemical structure with a tautomeric form (e.g., 2-hydroxy-pyridine with 2-pyridone). The replacement makes use of the tautomer enumeration functionality in MoKa (tauthor).

After the loading of a new structure or replacement of an existing one, we calculate its pK_a values using the default MoKa and store the predictions in a calculation table that is linked to the structure table with a one-to-many relationship. If available, this table not only contains the calculated pK_a values (and prediction error information) of a structure but also includes information about the acidic or basic character of the ionization center and its structural type. Structural types are assigned using a set of substructure queries defined using SMARTS patterns or mol-files.

The experimental data are stored in a third table that is again linked using a one-to-many relationship with the structure table. For an individual compound, there can be not only multiple ionization sites but also multiple test occasions for different sample batches or experimental methods. While for most compounds we have only a single experiment that may have detected zero or more pK_a values, there are several cases where we have multiple experiments, each with zero or more pK_a values. In order to simplify the user interface and the training process, we decided to allow only one experimental pK_a value to be assigned to a calculated pK_a value. Each association is linked with the user that made the assignment and a level of trust (certain, likely, and potential; see below for details).

Additional tables support the curation effort by giving the curators specific views of the data. One table with precalculated ionization site environment fingerprints (see below) allows fast searches for similar ionization sites. Other tables store information about research projects and if ionization sites are assigned, represented, or novel with respect to their environment.

The web application was developed using Python 2.7[31] and the Django framework.[32] For manipulation of chemical data, we use RDkit.[33] Data are updated from the Novartis data warehouse every 6 hours to enable experimentalists curating their latest measurements in a timely fashion.

2.3. Assignment of Experimental pK_a Values to Ionization Sites. The assignment of pK_a values to ionization sites is made difficult by the fact that the experimentally observed pK_a values (macroconstants) are the net effect of the interaction of the actual ionization microequilibra (microconstants).[18,34] As the MoKa software is not differentiating between macro- and microconstants in the training, we need to consider this problem in our curation effort by not blindly assigning experimental pK_a values to ionization sites. Instead we restrict the assignment to cases where we can unambiguously make it and where the macroconstant is a good approximation of the microconstant.

We automatically create an association between experiment and calculation in the database for cases where experimentally only one pK_a value was observed and only one pK_a value was calculated for the range 1 to 13. The assigned trust is determined by the difference in experimental and predicted pK_a values. For differences less than 0.5 the trust is *certain*; if the difference is below 1.0, we assign a trust of *likely*; and for values up to 2.0 the trust is *potential*. For cases with differences above 2 log units, we do not assign the association automatically. The

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

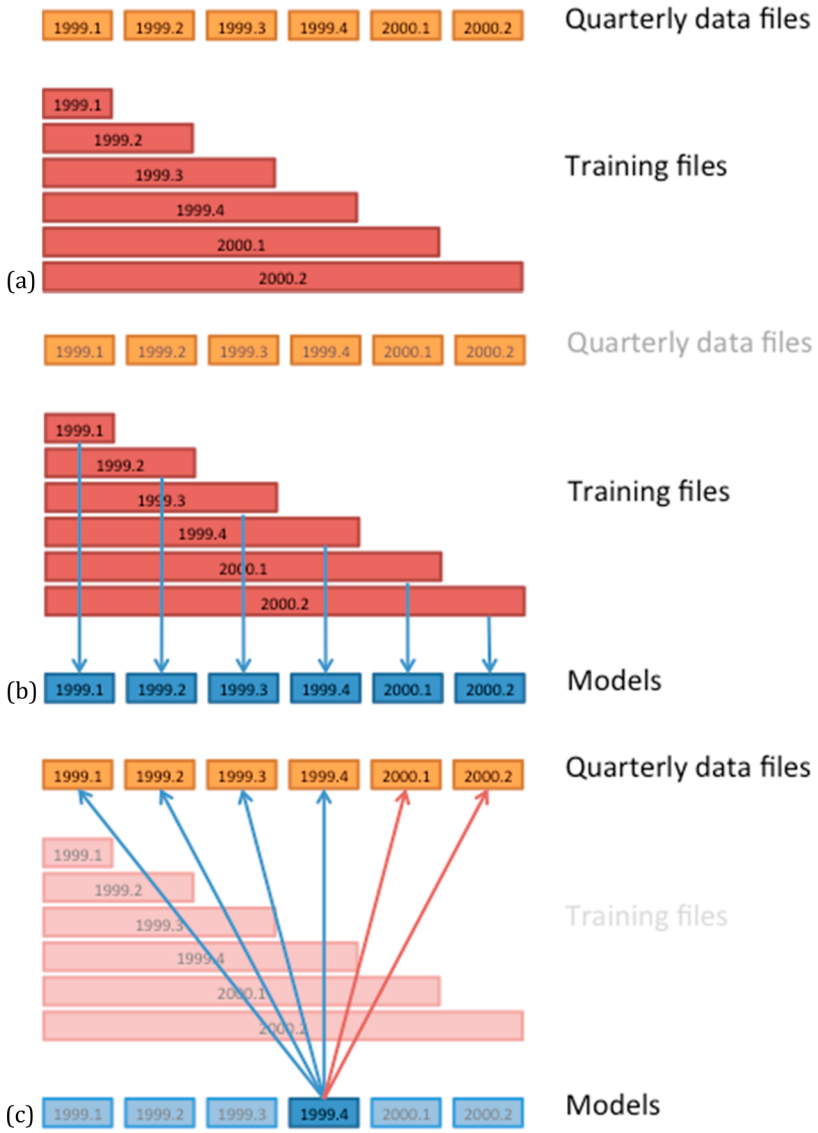


Figure 2. Design of retrospective study. The data set is split into quarterly data files that are assembled into cumulative training files. These training files are used to simulate the creation of quarterly models. Applying them to the quarterly data files allows studying the development of model performance over time. (a) Split of data set into quarterly data files and assembly of cumulative training files. (b) Training of quarterly models using cumulative training files. (c) Prediction of quarterly data files using individual model. Predictions on the training data are indicated with blue arrows, predictions on future data in red.

data curators can acknowledge automatic assignments and convert them to manual. Polyprotic molecules are not assigned automatically.

For manual assignments, the team of curators agreed on using only two categories for trust. *Certain* should be used for cases were the curator has high confidence, and *likely*, for cases with low confidence and the wish for a second opinion.

The web application comes with different interactive representations of the data to facilitate the curation effort and to concentrate on experimental results that are of most value to the model retraining. The default view is showing compounds in reverse experimental order, so that experimentalists can quickly start curating their latest measurements. One view allows analyzing structures with exactly one experimental and one calculated pK_a value that could not be automatically assigned (i.e., where the difference is greater 2 log units). Another analysis allows focusing on cases where MoKa gives predictions with low confidence. These may be cases where the structure needs to be replaced with a tautomer and the application uses the tautomer enumeration from MoKa to suggest alternative structures to the curator.

2.4. Curated Data Set. At the end of the third quarter of 2014, the data set contained 58 400 structures with 93 300 experimental and 136 100 calculated pK_a values. Initially, we could assign about 6000 pK_a values automatically; over time, we assigned another 5000 pK_a values manually. Figure 1 shows the development of associated pK_a values over time. Automatic assignment allows growing the data set by 150−200 pK_a associations each quarter. However, with our manual curation effort, we now get about 800 pK_a associations in the more recent quarters. As will become evident later on, this manual effort will have a strong impact on improving the quality of the pK_a predictions for the current chemistry at Novartis.

2.5. Retraining. In version 2.5.4, MoKa introduced a command line version of the retraining module kibitzer. Our web application is able to create the input file for the retraining

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

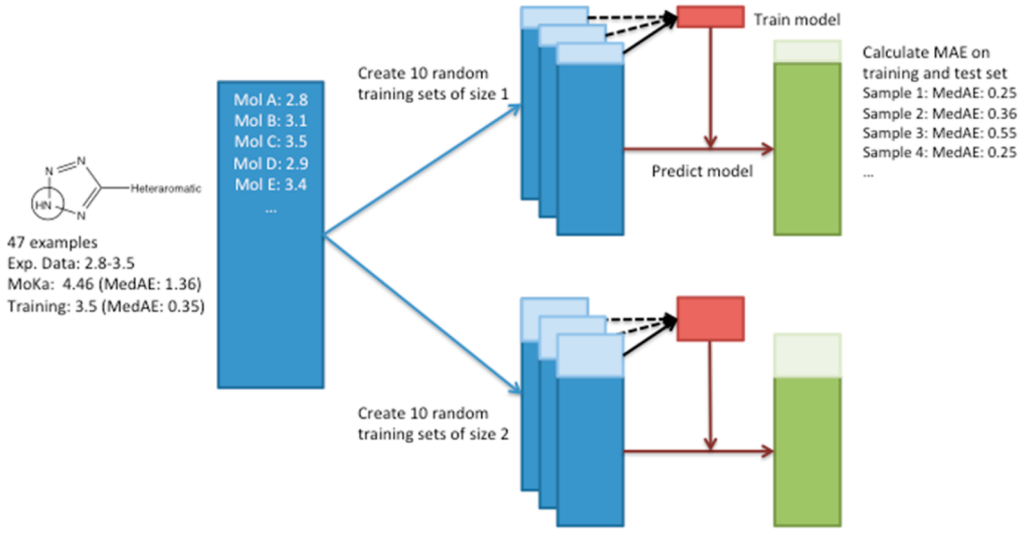


Figure 3. Process used to analyze the effect of training. The ionization site similarity was used to identify sets of compounds with identical environments up to seven bonds away from ionization site (e.g., tetrazole shown here). As a reference, the MedAE of all the examples in each set was determined for the untrained MoKa and for MoKa trained with all examples in the set (here 1.36 and 0.35, respectively). Next, a training set of size one was created from the set by random sampling and a model trained using this one example. The model obtained this way was applied to the remaining ionization sites in the set and the MedAE determined. This was repeated ten times for a training set size of one and after that for increasing training set sizes up to 15 examples.

module based on the current associations of experimental pK_a values to ionization sites. We include all manually and automatically assigned pK_a values in the training process.

2.6. Cross-Validation Study. The recommended approach in QSAR studies is to split the data set into a training and an *independent* test set, create a model using the training set, and finally analyze its performance on the test set.[35] It is recommended to repeat this analysis several times to estimate the variation of the statistics and to avoid any potential bias in a specific training set/test set split. The actual model training step in the MoKa software uses a cross-validation approach to determine the optimal number of components in a partial least-squares (PLS) model.[8]

In the study, we randomly split the data set into training and test set varying the size of the training set from 10% to 90% of the full data set in 10% increments.

For the analysis, we determined the absolute prediction error $|pK_{a,pred} - pK_{a,exp}|$ for individual ionization sites and looked at their distribution summarized by first quartile, median, and third quartile. These statistics are determined for the untrained MoKa and the trained models.

2.7. Retrospective Analysis. The availability of data over a time span of 15 years allows analyzing the effect model retraining has over time. This time period is long enough to allow a realistic assessment of long-term model behavior.

On the basis of the experimental date, we split the data set by quarter into 64 data files ranging from the first quarter in 1999 (Q_1999.1) to the third quarter in 2014 (Q_2014.3). These quarterly data files were then combined into cumulative training files. The training file for quarter 2010.2 (T_2010.2) contains the data from all quarterly files below and including quarter 2010.2. The data set assembly process is schematically shown in Figure 2a.

Kibitzer is used with each training file to create quarterly models (e.g., T_2010.2 is used to create model M_2012.2; see Figure 2b). For the studied time period, this gives us 64 individual models. Applying these models to the quarterly data files, gives us information about the performance on training data (e.g., applying M_2012.2 on Q_2012.2 and older) and improvement in prediction of new structures (e.g., applying M_2012.2 on Q_2012.3 and newer).

2.8. Ionization Site Similarity. To compare different ionization sites, we characterized them using rooted fingerprints as introduced by Vulpetti et al.[36] as a description of local environment of fluorine atoms. A generic implementation is available in RDKit. We used paths up to a length of seven bonds to create count-based fingerprints for the local environment of each ionization site. The Dice similarity was used to compare two ionization sites.

$$\mathrm{sim}(\mathrm{FP}_i, \mathrm{FP}_j) = \frac{2\sum_b \min(\mathrm{FP}_{ib}, \mathrm{FP}_{jb})}{\sum_b \mathrm{FP}_{ib} + \sum_b \mathrm{FP}_{jb}}$$

where FP_{ib} is the count for bit b in FP_i; FP_{jb} is the count for bit b in FP_j, and $\min(\mathrm{FP}_{ib},\mathrm{FP}_{jb})$ is the lower count for bit b in the two fingerprints.

2.9. Effect of Training. Another question with respect to training is how many examples are required to lead to an improved prediction? In order to address this important question, we analyzed the data according to the scheme shown in Figure 3. The rooted fingerprints are used to identify subsets of ionization sites with identical structural environment. We found 97 ionization sites in our data set with ten or more assigned experimental pK_a values. For each of these sets of compounds, we determined a baseline effect by estimating the default MoKa performance as well as the performance after training with all examples. To study the effect of training, we created random training sets with different size from one up to 15 examples. For each training set, we create a new MoKa model and applied it to the corresponding test set to determine MedAE of prediction. To get better statistics, we created ten random training sets for each size. We then looked at how many examples are required to improve the model prediction.

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

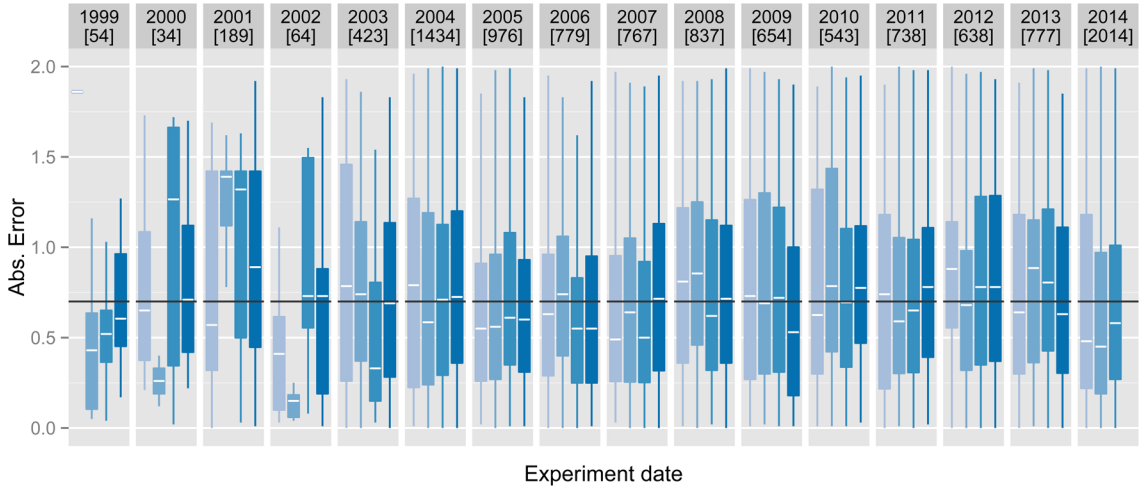


Figure 4. Distribution of absolute error over time for MoKa 2.5.4 predictions. The distributions of errors are aggregated by quarter and shown as box plots. The box plots are grouped by year (panel) and quarter (color) with the counts of data points given in square brackets. The horizontal black line indicates the overall median error of 0.7 log units. The number of examples per quarter up to 2003 is too small to give good estimations of the distribution.

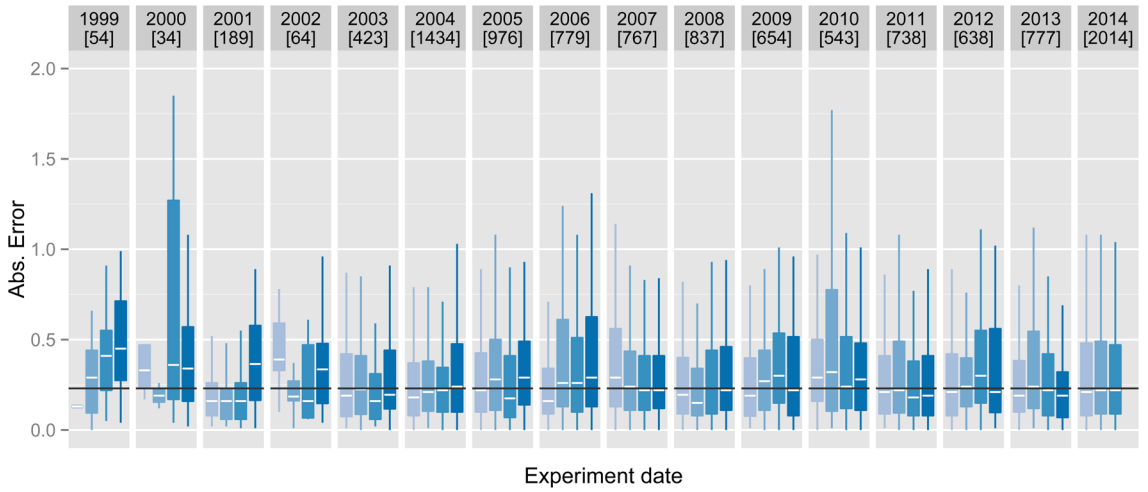


Figure 5. Distribution of absolute error of retrained MoKa when applied on the training set. MoKa was retrained using more than 11 000 internal data points. Compared to the untrained MoKa program (see Figure 4), the median absolute Error is reduced from 0.7 to 0.23. It is important to note that this improvement greatly overestimates performance on new data. Results are grouped by year (panel) and quarter (color).

3. RESULTS AND DISCUSSION

3.1. Baseline Performance. In order to assess the impact of training, it is important to understand the prediction performance of the untrained software. The distribution of absolute residuals between experimental and predicted pK_a values is shown in Figure 4. The median absolute error is 0.7 log units. The distribution of this error is relatively stable when we analyze our data set over time. For individual quarters the median absolute error varies from 0.6 to 0.8. The overall distribution varies similarly over time as can be seen in the box plots in Figure 4.

As a second baseline number, we looked at a model trained using the full data set. The distribution of errors for applying this model on the same training set is shown in Figure 5. The training reduces the median absolute error from 0.7 to 0.23 log units. This improvement is of course greatly overestimating the impact of training, however understanding it will help to assess the effect of training on new data points in the cross-validation and retrospective studies. Figure 6 compares the two baseline distributions.

3.2. Cross-validation Study. The results of the cross-validation study are summarized in Figure 7. The estimates of the different statistics, first, second, and third quartile, based on the 200 repeats are very tight; the interquartile range for each of them is below 0.03 log units. For the standard MoKa program, the 25% quartile of the absolute errors is 0.3, the median is 0.69, and the 75% quartile is 1.23. If we train a model using the full data set and apply it on the same data, we get 0.1 (25% quartile), 0.23 (median), and 0.52 (75% quartile). The predictive performance on the test set depends greatly on the split. If only 10% of the data set is used for training, the median absolute error is 0.45; this goes down to 0.33 with a 50% training set and gets closer to the performance on the training set with 0.31 for a 90% training set. The third quartile shows an even larger decrease. It reduces from 0.91 for a 10% training set to 0.65 for a 90% training set.

3.3. Retrospective Analysis. In the retrospective analysis, we can assess how training impacts predictions as they happen

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

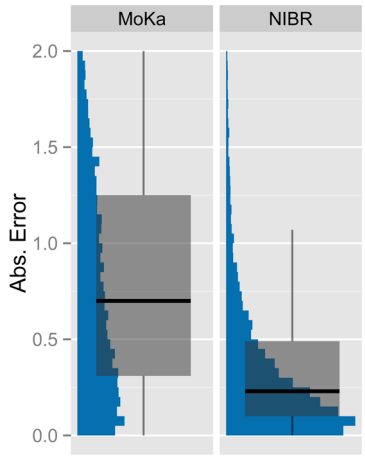


Figure 6. Comparison of absolute error distribution for the untrained and retrained MoKa prediction of pK_a values. The graphs summarize the distribution as histograms and box plots.

in a real-world context. As an example, Figure 8 shows the impact of training MoKa with all experimental data points up to the end of 2007. As expected, applying the trained model on the training data reduces the median absolute error to 0.21. For *novel* data, the median absolute error is reduced; however less drastically, to 0.56. This is better than the untrained method, however shows only a small improvement. If we look at the quarters in 2008 following the training, we can see performance improvements below average; the median absolute error is below 0.5.

A similar behavior can be observed for all models. In Figure 9, the change in median absolute error of prediction is shown for the different models from the retrospective analysis. The large fluctuations in the median are mainly due to different structural composition of the quarterly tests sets. However, despite these variations, a general increase post-training can be seen for all models. This becomes even clearer after smoothing the results over all curves. Over a period of one year, the model degrades quickly having a median absolute error of about 0.37 in the first quarter past training and approaching 0.6 in the third quarter. From then on the increase is slowly approaching the default MoKa performance over a period of several years.

3.4. Influence of Ionization Site Similarity on Model Performance. We've seen from the cross-validation study and even more from the retrospective analysis that the predictive performance of a retrained MoKa model depends greatly on the composition of the test set in relation to the training set. Over a short time scale, it is likely that related structures are submitted for measurement. To analyze if this is the case, we determined the maximum similarity of ionization sites of compounds not involved in training to our training set. Figure 10 shows the distribution of these similarities as a function of time after model training. The distributions are grouped by quarters and summarized as box plots. It is clear from this graph, that the similarity is higher the closer in time the compounds were screened to the training time point.

In Figure 11, we group the absolute prediction errors of the test sets by the similarity to the training set. For ionization sites that are represented in the training set (marked as Identical), the distribution is identical to what is observed for predicting the training set. For ionization sites that have a similarity between 0.6 and 1.0, the prediction is improved; however for similarities below 0.6, the distributions of absolute prediction errors is basically identical to the performance of the standard MoKa software.

These results show that good performance of retrained MoKa software on new data, depends greatly on the similarity of the ionization site to the training set. This also explains the results of the cross-validation study. There we observed that predictive performance improved with increasing size of the training set. From the analysis of similarity study we now understand that this improvement is due to increasing probability of finding a highly similar ionization site for test set compounds in the training set. To avoid this sampling bias

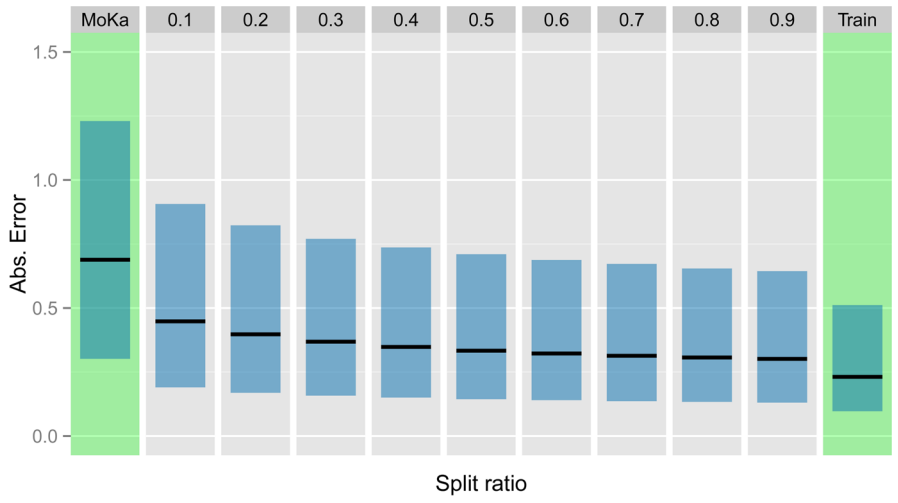


Figure 7. Results of cross-validation study. The total curated data set was split randomly into training and test sets using different split ratios. A split ratio of 0.1 corresponds to a training set of 10% of data set. The training set was used to create a new model and the model applied to training and test set. The absolute prediction errors on the test set were determined and its distribution summarized using first quartile, median, and third quartile. This process was repeated 200 times for each split ratio, and the statistics were averaged by split ratio and reported in the graph as box plots (bottom of blue bar, black line, top of bar). The variation of these statistics is below 0.03 in all cases. In addition, we determined the absolute errors for predictions with the untrained MoKa program and summarized in the same way. The graph shows from left to right: performance of standard MoKa, performance on independent test set with split ratio increasing from 0.1 to 0.9, and last the performance of the models on the training set.

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

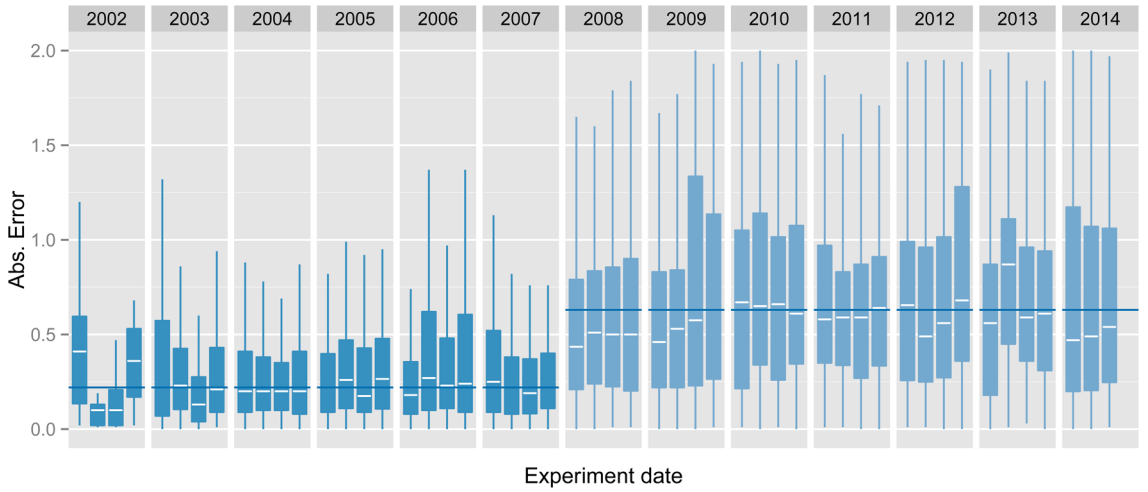


Figure 8. Distribution of absolute errors for a model trained with data up to fourth quarter 2007 (M_2007.4). The dark blue box plots give the distributions of the error on data points that were used in training; the median absolute error is 0.21 (blue horizontal line). The light blue box plots show the distribution for new data; overall the median absolute error for these *future* structures is 0.56 (blue horizontal line).

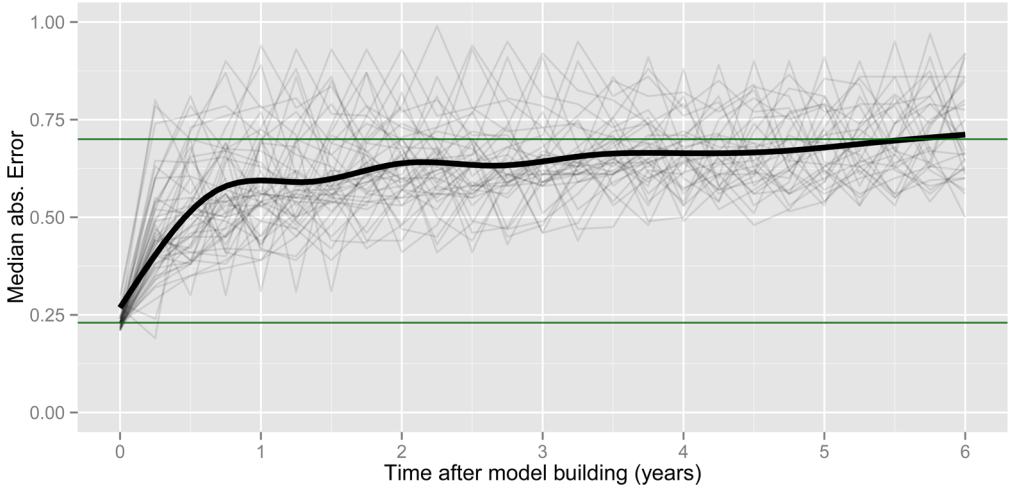


Figure 9. Degradation of model performance by quarter post model building. The graph shows the median absolute error of pK_a predictions for individual quarters post model building. The light gray lines show the results for all models built from 2002 onward to highlight the observed variation. The thick black line is a locally weighted regression (LOESS) line[38] based on all medians. The horizontal green lines give the benchmark performance of the standard MoKa software and the models applied to the training data. The graph clearly shows that the benefit of training is largest during the first year, rapidly increasing to around 0.6 and then slowly increasing over the years to the untrained model.

and to get a more realistic estimation of the generalization capability of a model, a leave-cluster-out cross-validation study would be required.[37]

3.5. Effect of Training on Prediction. Using subsets of compounds with identical ionization sites, we can estimate the effect of training on prediction performance using the approach outlines in Figure 3. We identified 97 different ionization sites that are represented in ten or more compounds. The behavior of the individual sets differs from case to case. Two typical examples are shown in Figure 12.

For about 42% of these ionization sites, the mean prediction error of MoKa is already below 0.4 pK_a units; for 28% it is below 0.3 pK_a units. Adding these compound examples to the training set will in general therefore only have little influence on the models. For the remaining analysis, we use a mean error of 0.3 pK_a units to say that training successfully improved prediction.

The cutoff value of 0.3 pK_a units gives us 72% of the ionization sites that MoKa does not predict well. After training with all example compounds, about 48% will have an average prediction error below the cutoff value, while 24% of the ionization sites still have a large average prediction error. By creating training sets of different size, we can see that for 26% already one example is sufficient to improve the prediction. Table 1 lists the results for this and an additional cutoff value of 0.4 pK_a units. For this value, already 42% of the ionization sites are predicted well and an additional 46% benefit from training.

4. CONCLUSIONS

At Novartis, we determined pK_a values for nearly 60 000 compounds giving almost 100 000 pK_a values. Using an internally developed web application, we were able to associate roughly 10% of the experimental values with specific ionization sites. The manual curation effort leads to assignments of almost 50% of new experimental pK_a values. This curated data set can be used for a number of studies; some of them are described in this contribution.

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

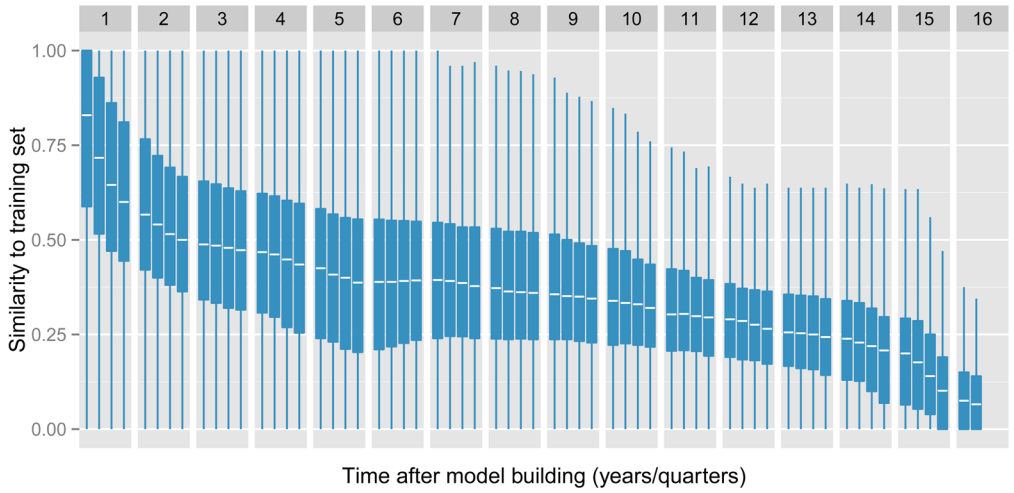


Figure 10. Ionization site similarity as a function of time difference in quarters to model training. In the four quarters immediately following model training, the ionization sites are more similar to the training set than for compounds that are synthesized further in the future.

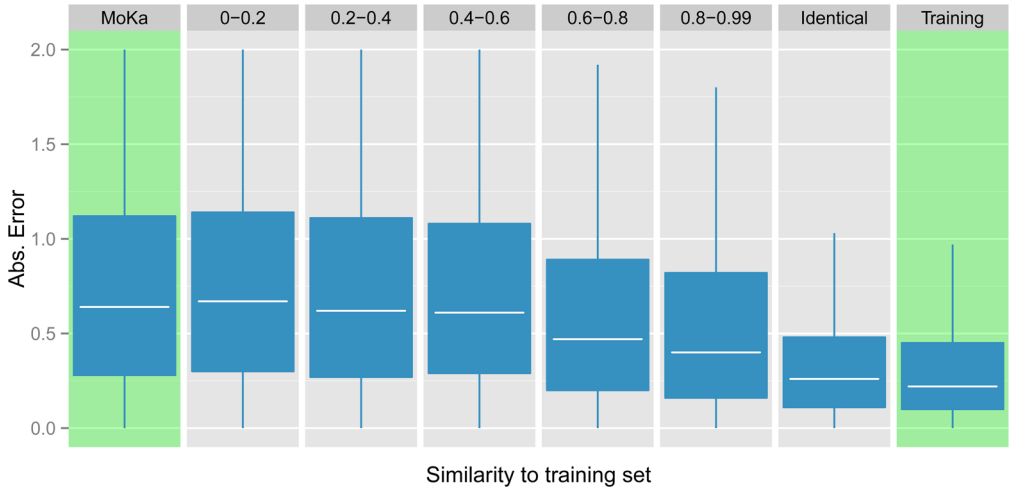


Figure 11. Effect of ionization site similarity to training set on prediction error. The distribution of absolute error is shown as a function of similarity to the training set (gray panels). The data are grouped into bins of increasing similarity. The panel labeled "Identical" summarizes the results for test set ionization sites that have the exact same environment as representatives from the training set. The green panels show the distribution for the untrained MoKa and for prediction on the training set. The graph shows that predictions for ionization sites that are similar to the training set are better than for novel ionization sites. For low similarity, the distribution of errors approaches the standard MoKa performance.

On our data set, we determined a mean absolute error of 0.7 for the prediction quality of the untrained MoKa 2.5.4 software. This is comparable to results previously obtained by Manchester et al.[15] when they applied it to 211 drug-like compounds.

MoKa offers the possibility to improve predictions by training the pK_a model using additional data. We assessed the training using two different validation strategies: cross-validation and a retrospective study.

In the cross-validation study, we split the data set randomly into training and test sets of different sizes. The cross-validation step was repeated 200 times for each split to achieve reliable statistics for the mean absolute error. The trained pK_a models clearly show improved prediction performance compared to the default MoKa model. However, we also see that the predictions get better with increasing size of the test set.

The retrospective analysis was designed to allow studying the effect of training on future predictions with a focus on model performance over time. We split the data set by date and built models for subsequent quarters from 1999 up to now. Each model represents a snapshot in time, and it can be applied to compounds that were screened at a later date. Like the cross-validation study, the retrospective analysis showed that training would improve predictions. However, it also revealed that model improvements could only be seen for six to nine months and approaches the standard MoKa performance in the long term. By looking at ionization site similarity, we could demonstrate that the similarity of ionization sites is higher for compounds screened within a quarter or two. Longer term, the chemical structures submitted to the assays change more drastically.

On the basis of these results, it is clear that regular retraining is crucial to see a benefit from the internal experimental data. We therefore decided on a quarterly release cycle. This is a compromise; ideally, it would be best to continuously retrain the model once new data become available. This will however lead to fluctuations in the predicted values. While these are in general very small for established chemistry and only large for novel ionization sites, too frequent changes may irritate users of the model. The quarterly released models are trained using all

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

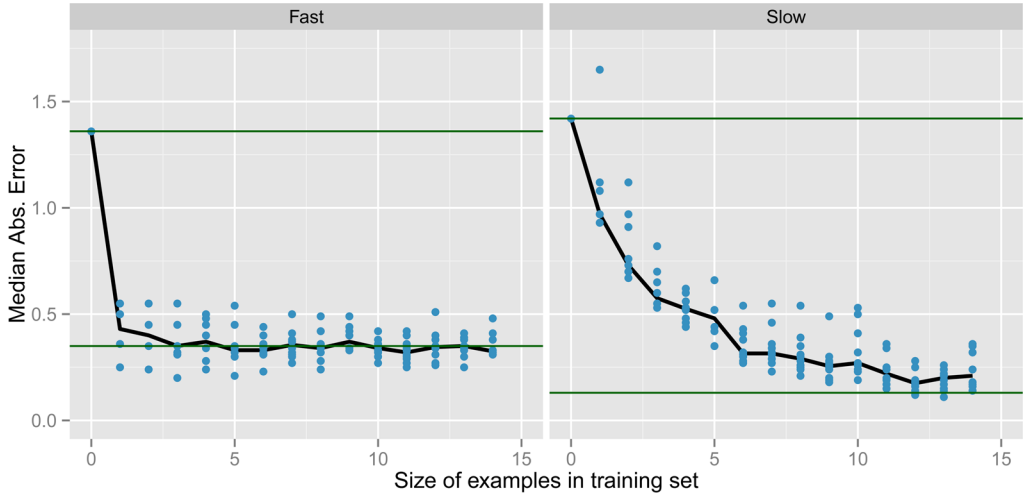


Figure 12. Two examples for the analysis process described in Figure 3. Each panel gives the results for a set of compounds with identical ionization sites. The top green lines indicate the MedAE of the untrained MoKa version; the bottom green line the performance of a model trained with all examples for the ionization site. The blue points give the MedAE on the randomly created test sets and the black line connects the overall MedAE for each of the different training set sizes. The example in the left panel (Fast) shows that already including a single assignment in the training leads to a performance that cannot be improved further by adding more examples. In contrast, the example on the right (Slow) shows a much slower performance improvement. Here at least five examples need to be included to get MedAE below 0.5.

Table 1. Analysis of Impact of Training on Identical Ionization Sites[a]

cut-off value:	0.3 pK_a units	0.4 pK_a units
MoKa predicts below cutoff	28%	42%
training fails to improve below cutoff	24%	12%
training improves test set performance below cutoff	48%	46%
with one example	26%	27%
with two examples	8%	6%
with three or more examples	14%	13%

[a]Ionization site identity is based on structural identity of the local environment based on equivalence of rooted fingerprints. The current data set contains 97 ionization sites with 10 or more compounds. Two cut-off values for the average absolute prediction performance were used to discriminate between good and poor performance. The percentage values are not cumulative.

curated assignments. Sometimes, projects synthesize compounds that have structural features that are poorly predicted. In this case, we can give them access to *cutting-edge* models that are trained daily with the latest experimental data.

If we compare the cross-validation study with the retrospective study, it is important to note that only the retrospective analysis could reveal the strong dependence of prediction improvement on ionization site similarity. For the random split used in our cross-validation study, the large data set caused a bias that lead to an overestimate of the predictive performance. The larger the data set, the higher the probability that an example is found in the training set that is structurally similar to a test set compound. With a model that depends strongly on structural features, as we have seen here, this leads to a correlation between training set size and improved predictions caused by a higher probability of test set ionization sites being represented in the training set. To address this issue, leave-cluster-out cross-validation[37] should be used to estimate the impact of structural similarity. As an alternative, it is possible to look at the dependence of predictive performance as a function of similarity to the training set as demonstrated in this publication.

This analysis has shown that prediction performance improves if a similar ionization site is found in the training set. In many cases, training improves with only one or two examples. It will therefore be useful to identify compounds with *novel* ionization sites and determine their experimental pK_a value. This will help strengthen the model.

AUTHOR INFORMATION

Corresponding Author

*E-mail: peter.gedeck@novartis.com.

Present Addresses

⊥F.L.: Chem & Mol Therapeutics, Biogen Idec, 14 Cambridge Center, Cambridge, MA 02142, United States.

○G.D.: Sutro Biopharma, Inc., United States.

Author Contributions

P.G. developed the web application and model-training workflow, carried out the data analysis, and wrote the initial draft of the manuscript. S.S., S.R., and W.J. ran the pK_a experiments. Y.L., S.S., and S.R. contributed to the development of the web application, curated the pK_a assignments, and contributed to the analysis of results. R.V. contributed to the deployment of the model and the interaction with the vendor. G.D., G.B., B.F., and F.L. provided advice and contributed to the discussion during the project. All authors contributed to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We recognize additional experimentalists who contributed to the Novartis pK_a database, including Gina Geraci and Linhong Yang.

ABBREVIATIONS USED

CE, capillary electrophoresis; **LOESS**, locally weighted regression line[38]; **PLS**, partial least-squares; **QSAR**, quantitative

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459

structure−activity/property relationship; **SMARTS**, SMILES arbitrary target specification; **SMILES**, simplified molecular-input line-entry system

REFERENCES

(1) Charifson, P. S.; Walters, W. P. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2014**, *57* (23), 9701−9717.

(2) Manallack, D. T. The pKa Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2007**, *1*, 25−38.

(3) Brown, T. N.; Mora-Diez, N. Computational Determination of Aqueous pKa Values of Protonated Benzimidazoles (Part 1). *J. Phys. Chem. B* **2006**, *110* (18), 9270−9279.

(4) Ding, F.; Smith, J. M.; Wang, H. First-Principles Calculation of pKa Values for Organic Acids in Nonaqueous Solution. *J. Org. Chem.* **2009**, *74* (7), 2679−2691.

(5) Yu, H.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Comparative Analysis of QSAR Models for Predicting pKa of Organic Oxygen Acids and Nitrogen Bases from Molecular Structure. *J. Chem. Inf. Model.* **2010**, *50* (11), 1949−1960.

(6) Yu, H.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Prediction of the Dissociation Constant pKa of Organic Acids from Local Molecular Parameters of Their Electronic Ground State. *J. Chem. Inf. Model.* **2011**, *51* (9), 2336−2344.

(7) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pKa for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **2007**, *47* (2), 450−459.

(8) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47* (6), 2172−2181.

(9) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **2009**, *49* (1), 68−75.

(10) Cruciani, G.; Milletti, F.; Storchi, L.; Sforna, G.; Goracci, L. In Silico pKa Prediction and ADME Profiling. *Chem. Biodivers.* **2009**, *6* (11), 1812−1821.

(11) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pKa by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 870−879.

(12) Meloun, M.; Bordovská, S. Benchmarking and Validating Algorithms That Estimate pK a Values of Drugs Based on Their Molecular Structures. *Anal. Bioanal. Chem.* **2007**, *389* (4), 1267−1281.

(13) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting pKa Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49* (12), 2801−2812.

(14) Settimo, L.; Bellman, K.; Knegtel, R. M. A. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm. Res.* **2014**, *31* (4), 1082−1095.

(15) Manchester, J.; Walkup, G.; Rivin, O.; You, Z. Evaluation of pKa Estimation Methods on 211 Druglike Compounds. *J. Chem. Inf. Model.* **2010**, *50* (4), 565−571.

(16) Shelley, J. C.; Calkins, D.; Sullivan, A. P. Comments on the Article "Evaluation of pKa Estimation Methods on 211 Druglike Compounds. *J. Chem. Inf. Model.* **2011**, *51* (1), 102−104.

(17) Milletti, F.; Storchi, L.; Goracci, L.; Bendels, S.; Wagner, B.; Kansy, M.; Cruciani, G. Extending pKa Prediction Accuracy: High-throughput pKa Measurements to Understand pKa Modulation of New Chemical Series. *Eur. J. Med. Chem.* **2010**, *45* (9), 4270−4279.

(18) Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico pKa Prediction. *J. Chem. Inf. Model.* **2015**, *55* (2), 389−397.

(19) Rodgers, S. L.; Davis, A. M.; van de Waterbeemd, H. Time-Series QSAR Analysis of Human Plasma Protein Binding Data. *QSAR Comb. Sci.* **2007**, *26* (4), 511−521.

(20) Rodgers, S. L.; Davis, A. M.; Tomkinson, N. P.; van de Waterbeemd, H. Predictivity of Simulated ADME AutoQSAR Models over Time. *Mol. Informatics* **2011**, *30* (2−3), 256−266.

(21) Wood, D. J.; Buttar, D.; Cumming, J. G.; Davis, A. M.; Norinder, U.; Rodgers, S. L. Automated QSAR with a Hierarchy of Global and Local Models. *Mol. Informatics* **2011**, *30* (11−12), 960−972.

(22) Gavaghan, C.; Arnby, C.; Blomberg, N.; Strandlund, G.; Boyer, S. Development, Interpretation and Temporal Evaluation of a Global QSAR of hERG Electrophysiology Screening Data. *J. Comput. Aided Mol. Des.* **2007**, *21* (4), 189−206.

(23) Sherer, E. C.; Verras, A.; Madeira, M.; Hagmann, W. K.; Sheridan, R. P.; Roberts, D.; Bleasby, K.; Cornell, W. D. QSAR Prediction of Passive Permeability in the LLC-PK1 Cell Line: Trends in Molecular Properties and Cross-Prediction of Caco-2 Permeabilities. *Mol. Informatics* **2012**, *31* (3−4), 231−245.

(24) Sirius analytical. http://www.sirius-analytical.com/ (accessed Nov 14, 2014).

(25) Allen, R. I.; Box, K. J.; Comer, J. E.; Peake, C.; Tam, K. Y. Multiwavelength Spectrophotometric Determination of Acid Dissociation Constants of Ionizable Drugs. *J. Pharm. Biomed. Anal.* **1998**, *17* (4−5), 699−712.

(26) Takács-Novák, K.; Box, K. J.; Avdeef, A. Potentiometric pKa Determination of Water-insoluble Compounds: Validation Study in Methanol/water Mixtures. *Int. J. Pharm.* **1997**, *151* (2), 235−248.

(27) Avdeef, A. pH-Metric Log P. Part 1. Difference Plots for Determining Ion-Pair Octanol-Water Partition Coefficients of Multiprotic Substances. *Quant. Struct.-Act. Relationships* **1992**, *11* (4), 510−517.

(28) Avdeef, A. pH-metric Log P. II: Refinement of Partition Coefficients and Ionization Constants of Multiprotic Substances. *J. Pharm. Sci.* **1993**, *82* (2), 183−190.

(29) Shalaeva, M.; Kenseth, J.; Lombardo, F.; Bastin, A. Measurement of Dissociation Constants (pKa Values) of Organic Compounds by Multiplexed Capillary Electrophoresis Using Aqueous and Cosolvent Buffers. *J. Pharm. Sci.* **2008**, *97* (7), 2581−2606.

(30) *MoKa 2.5.4*; Molecular Discovery, 2014.

(31) *Python Language Reference*, Version 2.7; Python Software Foundation, 2014.

(32) *Django*, Version 1.5; Django Software Foundation, 2013.

(33) *CORINA - Fast Generation of High-Quality 3D Molecular Models*; Molecular Networks GmbH, 2014.

(34) Pagliara, A.; Carrupt, P.-A.; Caron, G.; Gaillard, P.; Testa, B. Lipophilicity Profiles of Ampholytes. *Chem. Rev.* **1997**, *97* (8), 3385−3400.

(35) Gedeck, P.; Lewis, R. A. Exploiting QSAR Models in Lead Optimization. *Curr. Opin. Drug Discovery Devel.* **2008**, *11* (4), 569−575.

(36) Vulpetti, A.; Hommel, U.; Landrum, G.; Lewis, R.; Dalvit, C. Design and NMR-Based Screening of LEF, a Library of Chemical Fragments with Different Local Environment of Fluorine. *J. Am. Chem. Soc.* **2009**, *131* (36), 12949−12959.

(37) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (11), 1961−1969.

(38) Cleveland, W. S.; Devlin, S. J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Am. Stat. Assoc.* **1988**, *83* (403), 596−610.

DOI: 10.1021/acs.jcim.5b00172
J. Chem. Inf. Model. 2015, 55, 1449−1459