Technical Document

>20 Years of Reliable pK_a Predictions From ACD/Labs

Accuracy Improvements & Customer Collaborations

Executive Summary

ACD/pK_a is an industry leading calculator of acid dissociation constants. It is used globally by computational chemists, medicinal chemists, environmental scientists, and chromatographers, among other scientists to help understand the behavior of molecular structures for complex applications. Through more than 20 years of continual development and data curation ACD/pK_a delivers excellent prediction accuracy with the capability to train the algorithm with in-house experimental data.

This document provides an overview of the ACD/pK_a Classic prediction model and a review of improvements in prediction accuracy including highlights of collaborative projects using proprietary customer data, with excellent results.

Author *Sanjivanjit K. Bhal*

ACD/Labs

Introduction

Version 1 of ACD/Labs' *in silico* pK_a prediction software (ACD/pK_a) was introduced in 1997 and has been used by R&D organizations and researchers for more than 20 years to help inform experiment design, understand the behavior of chemical entities, and build proprietary models for more complex molecular properties.

History

ACD/pK_a is an important part of the Percepta portfolio and is under active development; continually being improved to provide "best-in-class" predictions. First introduced as a stand-alone prediction module (ACD/pK_a DB) the 2009 merger of ACD/Labs and Pharma Algorithms led to the introduction of a second prediction algorithm and evolution to the Percepta platform. As of 2010, ACD/Labs has offered the Classic and GALAS pK_a prediction algorithms. This document relays the continued development and improvements for the pK_a Classic algorithm.

> *The industry standard for pK_a...*
>
> *Dr Tom Carter, University of Oxford, UK*
>
> *The best software for routine pK_a predictions!*
>
> *Renier van der Westhuyzen, Chief Investigator, H3D*

Adoption

As of v2020 almost 1000 R&D organizations have chosen pK_a prediction from ACD/Labs. The software has been cited, evaluated, and peer-reviewed over the years,[1-4] enjoying an excellent reputation in the computational chemistry community.

The Largest Curated Commercial Repository of pK_a Data

The development team for ACD/pK_a seek to add well-curated experimental data to the database to maximize chemical space coverage and prediction accuracy. ACD/pK_a DB represents the largest publicly available compilation of curated experimental pK_a data.

Version	# Compounds in DB
v1–5	8889
v6–2020	15932

The Model Applicability Domain

ACD/pK_a DB is built on a database of curated pK_a values for pharmaceutically relevant compounds. The database is enriched continually to ensure expansion of the applicability domain for our clients. This is achieved through inclusion of proprietary data provided by collaborators discussed in detail later. Model applicability is an important factor to consider when evaluating a prediction model for scientific application.

A Trainable Model for More Than Two Decades

Since R&D organizations work on novel, proprietary chemistry, ACD/Labs' development team understands the importance of a trainable model that can be expanded to include novel structures. Tools for training the model with user experimental data was introduced early in the history of ACD/pK_a. There can be no guarantee that novel chemical structures will be accurately predicted by any commercially available model unless the relevant experimental data can be used to expand the applicability domain of the model. ACD/pK_a enables scientists to create local databases with in-house experimental data that can be shared within the organization and applied to improve prediction accuracy for such novel compound series.

Continual Improvements in Prediction Accuracy

The BioByte Starlist database has long been the gold standard for accurate experimental pK_a data. Including more than 10,000 compounds and >12,000 experimental pK_a values, it also offers the opportunity for a statistically meaningful study of prediction accuracy for commercial prediction models. Summarized here are the findings of studies comparing pK_a values predicted with the pK_a Classic algorithm from ACD/Labs (from version 1 to version 12 which is still a good representation of current version performance) versus experimental pK_a values published in the BioByte database.

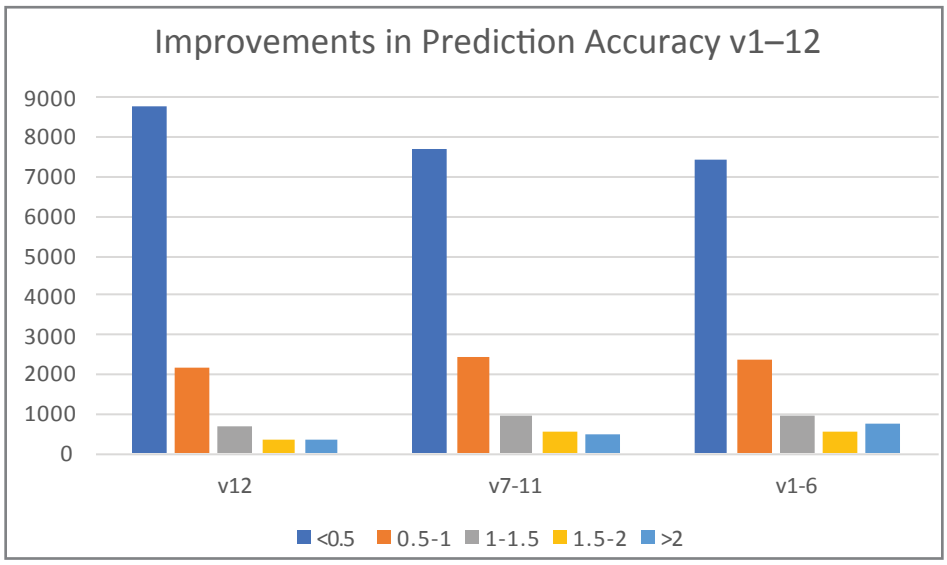

Figure 1: Accuracy of prediction of the ACD/pK_a Classic algorithm. Values are bucketed to show prediction within 0.5 log units, 0.51–1.0 log unit, 1.1–1.50 log units, 1.51–2 log units, and >2 log units.

Software versions were grouped into 3 categories: v1–6, v7–11, and v12. Each category saw significant enhancements to the algorithm, warranting a review of prediction accuracy. Over the years the model has seen a decrease in the average error of prediction (from 0.62 in v1 to 0.42 in v12) and significant improvement in prediction accuracy.

By version 12, 72% of pK_a values are predicted within 0.5 log unit (90% within 1 log unit) compared to version 1 when 62% of pK_a values were predicted within 0.5 log unit (81% within 1 log unit). At the other end of the spectrum the number of pK_a values predicted with an error of greater than 2 log units halved from 6% in version 1, to 3% in version 12.

Third-party References and Peer-Reviewed Publications on ACD/pK_a Prediction Accuracy

1. J. Manchester, G. Walkup, O. Rivin, and Z. You. (**2010**). Evaluation of pK_a Estimation Methods on 211 Druglike Compounds. *J. Chem. Inf. Model.*, *50*: 565–571.
2. C. Liao and Marc C. Nicklaus. (**2009**). Comparison of Nine Programs Predicting pK_a Values of Pharmaceutical Substances. *J. Chem. Inf. Model.*, *49*(12): 2801–2812.
3. Adam C. Lee and Gordon M. Crippen. (**2009**). Predicting pK_a. *J. Chem. Inf. Model.*, *49*: 2013–2033.
4. M. Meloun and S. Bordovská. (**2007**). Benchmarking and validating algorithms that estimate pK_a values of drugs based on their molecular structures, *Anal Bioanal Chem, 389*: 1267–1281.

Collaborative Projects | Enhancing the Model & Providing Value to Partners

Quite often, R&D organizations amass a significant volume of experimental data. While they see the benefit of training predictive models to realize the benefits of expanding chemical space coverage and improved accuracy, resources can be limited for such tasks. ACD/Labs has engaged in projects with leading chemical and pharmaceutical companies to help them leverage in-house knowledge. This saves the organization time in leveraging their data to impact future work, and means that ACD/Labs development team members with deep expertise are able to work with the data and the model to extract maximum value.

Proprietary, curated pK_a data shared by the pharma partner has been deconstructed by ACD/Labs to extract Hammet-Taft equations and new ionizable groups with associated pK_a values. This process ensures that important pK_a information can be extracted and used to improve the model without exposing proprietary structures. These projects have delivered significant improvements by expanding the chemical space coverage of the model and improving prediction accuracy. Results of a few such projects are summarized herein.

Project 1

Data Source: A top-ten pharmaceutical organization

About the Data: 2712 curated pK_a values for pharmaceutical compounds

Results: pK_a values for the data were predicted poorly in version 2006, when only 65% were predicted within 1 log unit of the experimental value. Data illustrated in Figure 2 shows the improvement in prediction accuracy for the data after training. Training provided significant improvement in prediction accuracy. After training, 83% of pK_a values were predicted within 1 log unit.

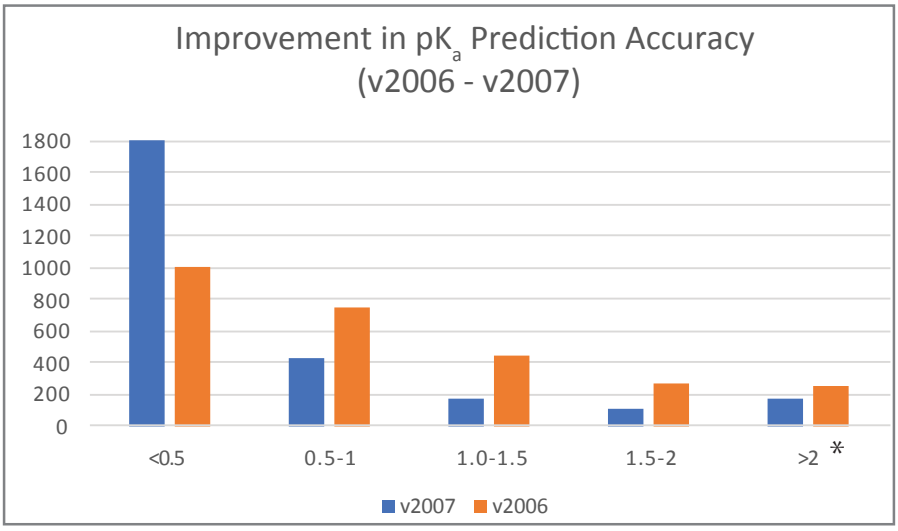

Figure 2: Improvement in prediction accuracy for 2712 pharmaceutically relevant pK_a values in v2007 of ACD/pK_a.

**A significant number of pKa values predicted with a difference of >2 log units in version 2007 are due to structures in irregular tautomeric form, incorrect structures, poor experimental data, and only to small degree due to the limitations of the algorithm*

Project 2

Data Source: A leading chemical organization

About the Data: 500 pK_a values from unique and diverse compounds provided:

- Extraction of 23 new reaction centers
- Augmentation of 10 existing reaction centers
- Addition of 40 new fragments to the internal training set

Results: Linear regression scatter plots (Figure 3a) illustrate the improvement in prediction accuracy before and after training. Inclusion of the new data resulted in a decrease in the average error of prediction from 0.89 before training (v2015) to 0.43 after training (v2015.2). R^2 also saw significant improvement (from 0.74 to 0.93 after training).

In v2015.1 68% of compounds were predicted within 1 log unit whereas by v2015.2 89% of compounds were predicted within 1 log unit.

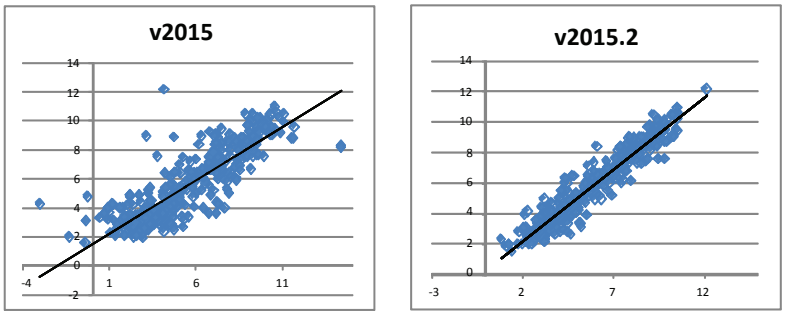

Figure 3a: Regression plots illustrating the improvement in prediction accuracy for 500 chemical compounds from v2015.1 to v2015.2.

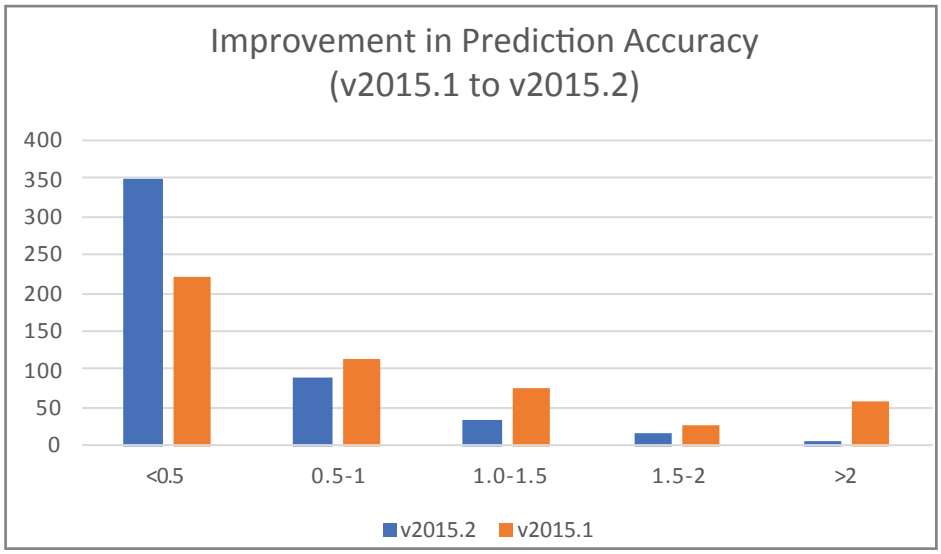

Figure 3b: Improvement in prediction accuracy for a dataset of ~500 compounds included in the v2015 algorithm.

In addition, the collaborating organization tested the performance of the new algorithm with a set of related compounds which were initially poorly predicted by the v2015 algorithm. They reported great satisfaction in the improvement of prediction accuracy for this validation set of compounds consisting of 60 bases and 21 acids (Table 1).

Table 1—Results from testing of the new algorithm (v2015.2) with a collection of 60 bases and 21 acids.

Average Prediction Error		
	v2015	V2015.2
Acids	2.08	0.79
Bases	2.03	0.78

Project 3

Data Source: A leading pharmaceutical organization

About the Data:

- >2300 pK_a values extracted from drug-like compounds
- >100 new Reaction Centers and fragments extracted
- A number of existing reaction centers updated

Results: In version 2015, pK_a values for the data were predicted with an average error of 1.38 log units and R^2 of 0.60; only 55% of compounds were predicted within 1 log unit and 20% were predicted with a difference of more than 2 log units.

After training, the average error of prediction dramatically dropped to 0.55 log units (R^2 0.91), with 82% of compounds predicted within one log unit and only 5% with a difference of 2 log units or more, as illustrated in Figure 4.

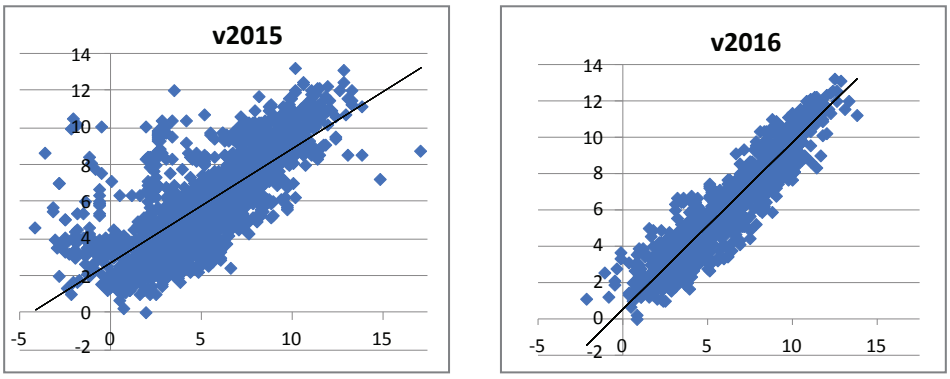

Figure 4a: Regression plots illustrating the improvement in prediction accuracy for druglike compounds from v2015 to v2016.

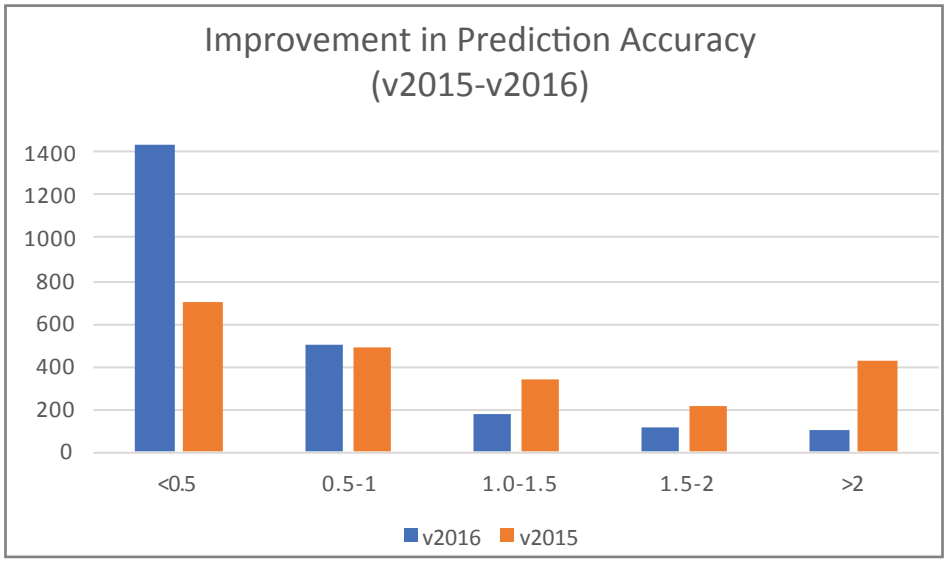

Figure 4b: Improvement in prediction accuracy for >2000 pharmaceutically relevant pK_a values in v2016 of ACD/pK_a.

Project 4

Data Source: A leading pharmaceutical organization

About the Data: Experimental data for 600 pharmaceutically relevant compounds provided:

- 1144 pK_a values
- 50 new reaction centers extracted
- Enrichment of 20 existing reaction centers
- >100 Hammet-Taft equations added

Results: Training of the algorithm resulted in significant improvement in prediction accuracy for these 600 novel pharmaceutical compounds. pK_a for 92% of compounds was predicted within 1 log unit, with 75% predicted within 0.5 log units, after training in v2020.1. This compares to only 64% of compounds being predicted within 1 log unit with the v2019.2 Classic algorithm (Figure 5b).

The linear regression scatter plots (Figure 5a) show the comparison of experimental versus predicted pK_a before and after training of the dataset. The average error of prediction is significantly lower after training in v2020.1 (0.36 versus 0.81) log units in v2019.2, and R^2 is closer to 1 (0.95 versus 0.84 in v2019.2).

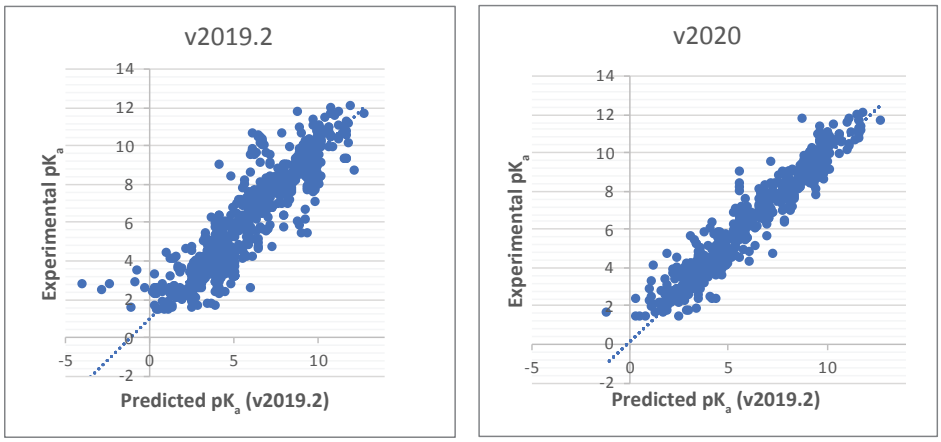

Figure 5a: Regression plots illustrating the improvement in pK_a prediction accuracy for 600 pharmaceutically relevant compounds from v2019.2 to v2020.

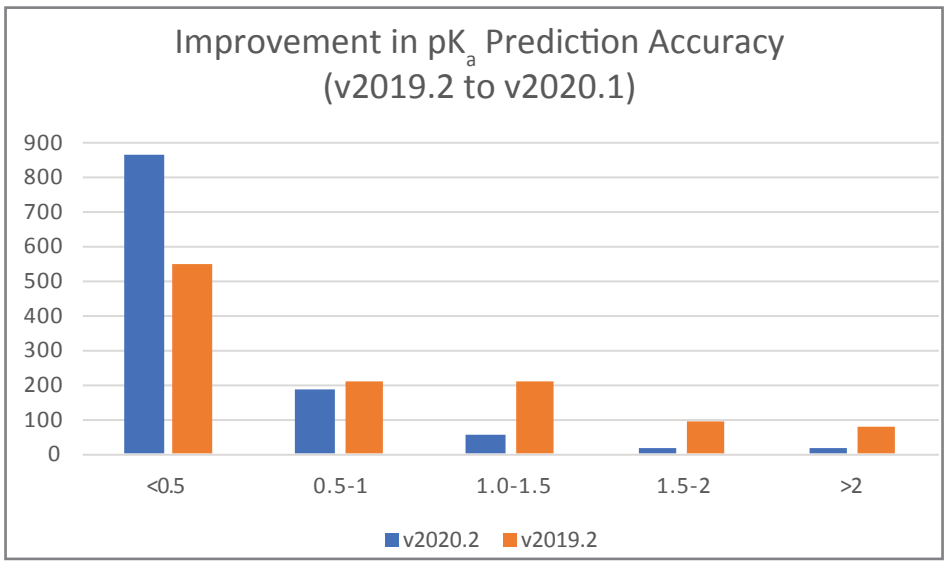

Figure 5b: Improvement in prediction accuracy (pK_a Classic algorithm) for 1144 ionization centers in 600 pharmaceutical compounds.

The Value of Collaborative Projects

Through all of these projects the ACD/Labs' team were delighted to satisfy the needs of collaborators by providing expert model training with their own experimental data. In addition, the broader R&D community have access to a commercially available predictor with improved chemical space coverage and prediction accuracy. Specifically, for the Pharma/Biotech community, 6659 pK_a values from modern drug-like compounds were added to the internal training set, equivalent to a 30% expansion of the original database. Curation of the data to remove structures allowed the data to be used to expand the applicability domain of the model while protecting valuable intellectual property.

Please enquire to engage in similar projects with in-house experimental data.

ACD/Labs®

info@acdlabs.com

www.acdlabs.com
@ACDLabs

1 800 304 3988 (US and Canada)
+44 (0) 1344 668030 (Europe)