**A COMPREHENSIVE EVALUATION OF ACD/LOGD ON A PHARMACEUTICAL COMPOUND SET**

Andrius Sazonovas, Kiril Lanevskij, Remigijus Didziapetris

1. Vilnius University, Lithuania
2. ACD/Labs, Inc., Canada

--

**Introduction**

Lipophilicity, which is often expressed in terms of 1-octanol/water partitioning coefficient logD, or the corresponding pKa-dependent distribution coefficient logP, is one of the key physicochemical characteristics of drug-like molecules. It has a major influence on a variety of the compounds’ properties constituting their ADME, pharmacokinetic, and drug safety profiles. Widely available in silico tools for predicting these properties are mostly based on experimental data for simple organic chemicals and marketed drugs. Consequently, as drug discovery projects are moving to increasingly novel regions of chemical space, utility of existing methods becomes more and more questionable. In several previously published evaluation studies, the mean logD prediction error for in-house compound libraries of pharmaceutical companies was shown to exceed 1 log unit and almost all methods. Prediction of logD is even more challenging, as it requires accurate knowledge of both logD of neutral form and distribution of ionic forms of the compound in the relevant pH range.

**Objectives**

1. Collecting a data set of experimental logD values from recent publications dealing with novel congeneric compound series from drug discovery projects;
2. Evaluating the performance of ACD/logP predictor for the newly collected molecules using different combinations of available logD and pKa calculation algorithms;
3. Investigating the potential for improving prediction accuracy for unknown compound classes by application of automated model training.

**Experimental Data**

The data compilation stage involved careful analysis of recent MEDChem publications dealing with lead optimization in novel series of congeneric compounds, that also reported measured physicochemical property values. The absolute majority of lipophilicity measurements correspond to logD values at pH 7.4, and for simplicity only this pH point was considered in this study. Overall, logD and pKa values were collected for 1659 compounds, and the spread of the data by the types of used assays is shown in Fig. 1.

**Prediction Methods**

Calculation of logD relies on estimating logD of the neutral form and interaction constants (A) for various ionic species i, where i refers to their fractions at given pH:

\[
\log D = \sum_i a_i \log P_A^i
\]

ACD/Percepta includes two different algorithms – Classic and GALAS – for predicting logD and ionization state. A Consensus approach utilizing both methods is also available for logD. The program can be configured to use any combination of the listed algorithms. Therefore, a total of 4 different built-in methods can be used for calculating logD, as outlined in Scheme 1.

**LogD**

Classic: a structural fragment-based approach that calculates logD of the molecule as a sum of increments from functional groups, carbon atoms and intramolecular interactions.

- GALAS: a two-step approach that involves (1) a statistical baseline model based on fragmental descriptors, and (2) a similarity-based routine that introduces corrections to predictions produced by the baseline model according to its performance for most similar compounds from the library. (2)

- Consensus: calculates a weighted average of Classic and GALAS logD predictions with weights of contributions from both methods based on their baseline reliability estimates.

**Initialization**

• Classic: an algorithm predicting apparent pK\textsubscript{a} values on the basis of a database of tautomeric-type and electronic substituent constants for a wide variety of variable centers.

- GALAS: a multi-step procedure involving estimation of pK\textsubscript{a} microconstants for all possible ionization centers in a hypothetical uncharged molecule; corrections according to the surrounding of the reaction center and charge influences of some neighbor ionization centers.

**Automated Model Training**

Although System training feature is available for ACD/logD Classic and ACD/LogP Classic methods, in this work we will focus on self-training functionality offered by ACD/logD GALAS model.

The key features of GALAS methodology that allow for effective training are summarised below:

- GALAS model predictions include Reliability Indices (RI) that serve as a built-in metric indicating whether the compounds belong to the Model Applicability Domain. This concept is highlighted in Scheme 2.

- Training with multiple compounds can be accomplished in a fully automatic manner by simply importing the entire data set as a new self-traning library. The new libraries are immediately available for use in calculations when the import is completed.

- It is possible to use logD values directly in logD training – the imported logD values are converted to intrinsic logD on the fly using one of the available pK\textsubscript{a}/ionization state prediction algorithms to estimate the distribution of ionic forms of indicated pH.

**Training Results**

In training experiments, two best performing logD/pKa method combinations – Consensus/Classic (1), and GALAS/Classic (2) – were used. Since only logD GALAS component was trained, it was anticipated that training would have a stronger effect on method (2). Indeed, as shown in Fig. 2, the predictive power of method (2) quickly surpassed that of method (1) and stayed slightly better throughout the experiment. Notably, in both cases accuracy of predictions for test set molecules was significantly improved after addition of the very first portion of training data. Further improvements were more modest, and eventually statistical parameters stabilized at around R\textsuperscript{2} = 0.7, and RMSE = 0.4.

**Discussion & Conclusions**

The first training/iteration using only 20% of all data accounted for more than 50% of the overall improvement in predictivity, highlighting a key advantage of GALAS model training; performing measurements for a relatively few members of a novel chemical series may suffice to ensure reliable property estimates for the entire class of compounds.

However, R\textsuperscript{2} values never exceeded 0.7, pointing out potential issues with logD determination – there are quite a few discrepant data points for highly similar molecules, different from different laboratories; hence to produce systematic deviations. Even the well-established shake-flask method shows its limits in discriminating highly lipophilic compounds (logD > 3), so plots in the validation stage, calculated on predicted values may contribute valuable information to complement the experimental measurements.

**References**


---

**TABLE 1.** Statistical performance of built-in methods for predicting logD, in ACD/Percepta. The top two methods for each statistical parameter are highlighted green (\(\text{1}^\text{st}\)) and light green (\(\text{2}^\text{nd}\)).

<table>
<thead>
<tr>
<th>Method</th>
<th>Consensus/Classic</th>
<th>GALAS/Classic</th>
</tr>
</thead>
<tbody>
<tr>
<td>R\textsuperscript{2}</td>
<td>0.39</td>
<td>0.42</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>MAE</td>
<td>1.03</td>
<td>1.03</td>
</tr>
</tbody>
</table>

**FIGURE 1.** Composition of logD\textsubscript{i} data set by types of assays.

**FIGURE 2.** Improvement of logD\textsubscript{i} prediction accuracy for validation set compounds by model training.

**SCHEME 1.** Combinations of logD and pKa algorithms that can be used to predict logD in ACD/Percepta.

**SCHEME 2.** Illustration of the Model Applicability Domain, and its expansion by GALAS model training.