# The Importance of Training Predictors

Sanjivanjit K. Bhal
Advanced Chemistry Development, Inc.
Toronto, ON, Canada
www.acdlabs.com

## Introduction

Prediction accuracy is of prime concern when evaluating a new software predictor. While the quality of the prediction algorithm impacts accuracy, a more important factor is whether the predictor covers the appropriate chemical space for compounds under investigation. Fortunately the applicable chemical space, and therefore accuracy, can be easily improved with training.

### A Model is Only as Good as the Data It's Based On

No matter what statistical methods or descriptors are used in building a model, the model is only reliably applicable within the chemical space covered by the internal dataset, and as good as the associated data upon which it is built. These are the two limiting factors in in-silico modeling.

Datasets of commercially available predictors are generally based on compilations of publicly available data because this provides the broadest chemical space coverage. This often means that data comes from different laboratories that use diverse experimental methods to acquire it. The spread of error distribution for this data is often quite large. Contrast this with measurements made in a single laboratory, and the spread of error distribution will be significantly tighter.

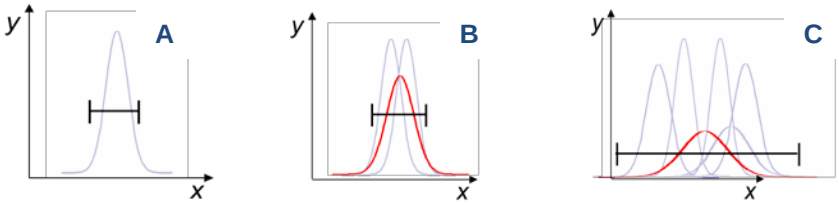


**Figure 1** Plots to illustrate typical error distributions that might be expected for: *A—measurements carried out in a single lab; B—measurements from several laboratories using the same protocol; C—data from a compilation of publicly available sources.*

### Chemical Space Coverage—the Model Applicability Domain

Every predictive model has certain chemical space coverage governed by the compounds in the training set—the model applicability domain (MAD). If the compound for which we wish to make a prediction falls within this, we would expect the error in prediction to be comparable to errors in the model validation. If the compound falls outside of this space, the reliability of the prediction cannot be estimated. Third party/proprietary chemical space will likely overlap the internal training set to some extent, but a significant part will not be represented.

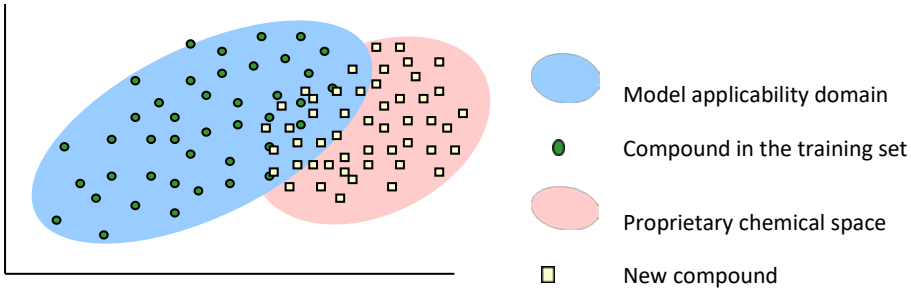


**Figure 2** Model applicability domain vs. proprietary chemical space.

## Training Predictive Models

The two main reasons to train models when the ability is available are:

- It makes use of experimental data to improve prediction accuracy for proprietary chemical space.
- It saves time that might otherwise be spent building and validating a new model.

Any commercially available software predictor is unlikely to sufficiently represent the chemical space of interest for research programs working in novel space, to provide accurate and reliable predictions 'out of the box'. Addition of reliable in-house data to a model not only improves accuracy due to expansion of the model applicability domain to broader chemical space, but the data added is also likely to provide a narrower error distribution for that chemical space. This results in greater prediction accuracy, and a model that is customized for the work carried out in your laboratory/institution.

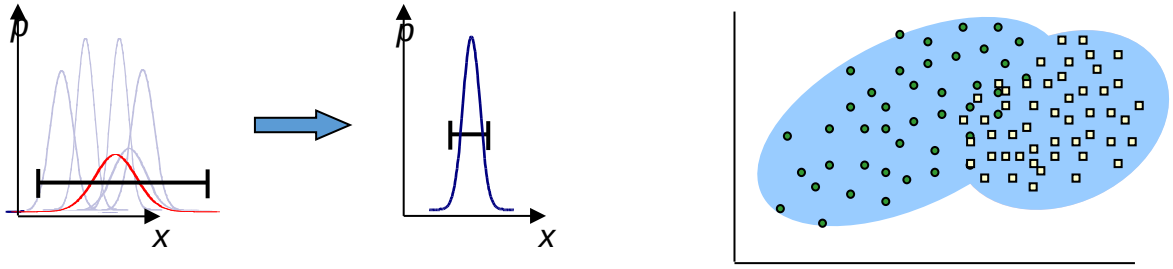


**Figure 3** Narrowing of error distribution and expansion of model applicability domain with addition of in-house experimental data.

# Who Should Be Training?

Everyone! Regardless of whether you have access to data for thousands of compounds from a variety of projects, or limited data for one project, it all adds value to your predictive model. While the addition of data for thousands of compounds will make the model more widely applicable, the addition of data for a particular compound series can be equally as valuable for a particular research project where changes around a common scaffold are being investigated.

With ACD/Labs' PhysChem and ADME/Tox predictive tools you don't have to be a statistical or software expert to apply training. The majority of our predictive modules offer the ability to train, and our training tools have been specifically designed to be easy to use, and allow facile addition of data for many thousands of compounds, or a handful of four or five.

So, whether you are starting a software evaluation, or getting started installing and using our software, be sure to ask about training and apply it to get the best from your evaluation/purchase.

To get value-for-money from any predictive model, it is imperative that you add reliable in-house data when it is available.

## Appendix: The Effects of Training on Four ACD/Percepta Predictive Models

Use your own experimental data to expand applicability domain of predictions and to adapt them to your screening protocol. Often adding data of just few compounds can increase the accuracy for entire chemical class however in the case of diverse compounds it is advisable to use much larger data set.

In the table below experimental data was randomly subdivided into training and test sets. The Training sets were gradually added to the algorithm to assess the effect on the test set.

Table. Algorithm training with User's Data

| Algorithm | Training Set size | MAE[a)] | Algorithm | Training Set size | MAE[a)] |
|---|---|---|---|---|---|
| Log $P_{OW}$, (Test set: N = 1,000) [1] | 0 | 0.59 | Intrinsic Log $S_W$ (Test set: N = 400) [2] | 0 | 0.84 |
| | 1,000 | 0.53 | | 250 | 0.84 |
| | 2,500 | 0.48 | | 750 | 0.66 |
| | 4,000 | 0.46 | | 913 | 0.62 |
| Log $D_{OW}$, (Test set: N = 11) [2] | 0 | 0.67 | CYP3A4 inhibition (Test set: N = 4,228) [3] | 0 | 76% [c)] |
| | 100 | 0.66 [b)] | | 400 | 80% [c)] |
| | 149 | 0.60 [b)] | | 2,000 | 84% [c)] |
| | 198 | 0.51 [b)] | | 4,300 | 86% [c)] |
| | 265 | 0.48 [b)] | | | |

[a)] Mean Absolute Error for RI > 0.3.

[b)] Log D of the training set was obtained under different exp. conditions than log D of the built-in set.

[c)] MAE replaced with Accuracy (%) of qualitative predictions

[1] Japertas Pet al. Evaluation of Structure Based Methods for the Prediction of LogP for Agrochemicals, 234th ACS National Meeting, Boston, MA, August 19-23, 2007.

[2] Japertas Pet al. Similarity Based Correction for the Predictions of Compounds Physicochemical Properties, 235th ACS National Meeting, New Orleans, LA, April 6-10, 2008.

[3] Didziapetris R et al. Trainable structure-activity relationship model for virtual screening of CYP3A4 inhibition. J Computer Aided Mol Des. 2010, prepared for publication.