GALAS Modeling Methodology Applications in the Prediction of Drug Safety Related Properties

Andrius Sazonovas[a,b], Remigijus Didziapetris[a], Justas Dapkunas[a,c], Liutauras Juska[a,c], and Pranas Japertas[a]

[a] Advanced Chemistry Development, Inc.
[b] Faculty of Chemistry Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania
[c] Faculty of Natural Sciences, Vilnius University, M.K.Ciurlionio g. 21/27, LT-03101 Vilnius, Lithuania

Introduction

Early computational evaluation of a drug candidate's properties, related to pharmaceutical safety (such as hERG inhibition induced cardiotoxicity or CYP3A4 inhibition responsible for various unwanted drug-drug interactions), is becoming increasingly important in the drug discovery process. Yet, every model, no matter the data, descriptors, or modeling techniques used to build it, has a certain applicability domain beyond which the quality of predictions becomes highly questionable. This reality is one of the fundamental issues concerning the effective use of third-party predictive algorithms in industry. The simple reason for this is that literature based training sets rarely cover the specific part of chemical space that 'in-house' projects focus upon. Discrepancies between 'in-house' experimental protocols and methods used to measure properties for compounds in publicly available sources further affect the quality of resulting *in silico* predictions. Therefore, the need has long existed for a method that would allow a company to effectively assess the Applicability Domain of any third-party model, and tailor it to its specific needs using proprietary 'in-house' data.

GALAS Model Methodology and Reliability Index

Addressing the aforementioned issue, a GALAS (Global, Adjusted Locally According to Similarity) model concept has been developed that provides a novel solution to this problem. Each GALAS model consists of the following parts:

- A structure based QSAR/QSPR for the prediction of the property of interest derived from a literature training set—the so called baseline QSAR/QSPR.
- A user defined data set with experimental values for the property of interest—the so called Self-training Library.
- A particular similarity-based routine that identifies the most similar compounds contained in the Self-training Library, and, considering their experimental values, calculates systematic deviations produced by the baseline QSAR/QSPR for each submitted molecule—the so called training engine.

The result is a prediction that is corrected according to experimental values for the most similar compounds present in the user defined Self-training Library, covering the part of the chemical space not initially included in the training set. Both the concept of Model Applicability Domain and its expansion with the help of GALAS modeling methodology are schematically outlined in Figure1.

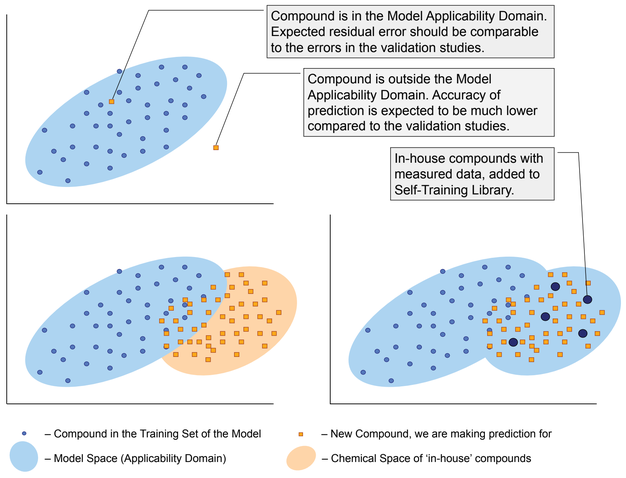


Figure 1: An illustration of the Model Applicability Domain, and its expansion using GALAS modeling methodology.

In addition, GALAS modeling methodology allows quantitative assessment of prediction reliability. This information is contained in the developed Reliability Index (RI) that can provide values in the range [0-1]. Lower values suggest a compound being further from the Model Applicability Domain and the prediction less reliable, while high RI values indicate an increasing confidence about the quality of the prediction. Estimation of the Reliability Index takes into account the following:

- Similarity of the tested compound to the training set—no reliable predictions can be made if no similar compounds exist in the training set.
- Consistency of experimental values for similar compounds—even when similar compounds are present in the dataset, the quality of prediction could be lower if that data is inconsistent. It does not matter what the reason for the inconsistency is (experimental variability or sudden change in mechanism of action due to slight structural changes), it indicates possible problems when trying to give accurate predictions.

Coping with Completely New Chemical Features—an Example Scenario with Prediction of hERG Inhibition

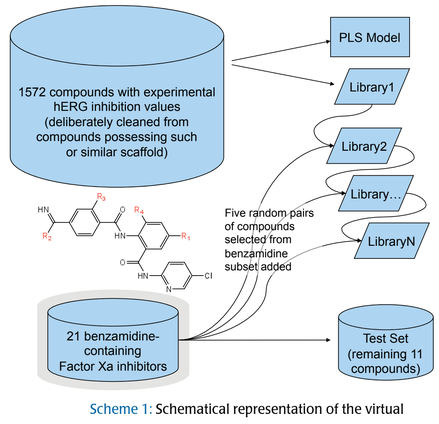
The objectives of this validation study for GALAS modeling methodology were as follows:

- Demonstrate that a GALAS model can be trained to new chemical features absent in the original training set.
- Demonstrate that a small number of compounds with experimental data is sufficient for training.

The first step involved the creation of a set of 1572 compounds with experimental hERG inhibition data, from which compounds possessing the scaffold typical for the benzamidine-containing Factor Xa (thrombokinase inhibitors)[1], or similar, were removed. This set was used for initial training of the baseline PLS model, and as a starting Self-training Library.

The second dataset contained only benzamidine-containing Factor Xa inhibitors, mimicking a project with new chemistry. 10 compounds (chosen at random) from this set were added in five pair-wise additions to the initial Self-training Library, making it more and more aware of benzamidine derivatives. Remaining compounds from the second set were used as a validation set, on which the performance of the methodology was tested. Scheme 1 graphically outlines the procedure.

Table 1 illustrates the model's performance on the validation set of compounds during the course of this virtual experiment. Initially, predictions for all compounds are inconclusive as no similar compounds are present in the library (indicated by low RI values). After just a couple of additions, however, predictions of sufficient reliability start to appear. When 6 compounds of the same class are added to the library, most calculated values become reliable. When all 10 compounds are added, 10 of 11 test set molecules are confidently predicted as either hERG inhibitors or non-inhibitors.



Scheme 1: Schematical representation of the virtual experiment procedures.

	Number of Added Compounds				
Exp.	0	2	4	6	10
Inhibitors	0.46 (0.12)	0.66 (0.49)	**0.80 (0.66)**	**0.87 (0.71)**	**0.91 (0.71)**
	0.68 (0.15)	0.84 (0.49)	**0.89 (0.64)**	**0.90 (0.56)**	**0.94 (0.71)**
	0.42 (0.13)	0.64 (0.38)	0.82 (0.58)	**0.88 (0.70)**	**0.92 (0.72)**
	0.42 (0.11)	0.67 (0.34)	0.81 (0.58)	0.82 (0.34)	**0.90 (0.57)**
	0.59 (0.23)	0.73 (0.33)	**0.88 (0.46)**	**0.89 (0.47)**	**0.94 (0.63)**
	0.53 (0.21)	0.66 (0.29)	0.77 (0.31)	0.81 (0.29)	**0.86 (0.49)**
	0.68 (0.21)	0.73 (0.24)	0.84 (0.36)	**0.91 (0.50)**	**0.88 (0.45)**
Non-inhibitors	0.11 (0.12)	0.18 (0.39)	0.24 (0.63)	0.19 (0.60)	0.19 (0.60)
	0.09 (0.11)	**0.12 (0.45)**	**0.17 (0.63)**	**0.11 (0.79)**	**0.14 (0.62)**
	0.12 (0.14)	0.18 (0.42)	0.24 (0.59)	**0.16 (0.75)**	**0.19 (0.61)**
	0.12 (0.20)	0.17 (0.43)	0.20 (0.57)	**0.17 (0.57)**	**0.10 (0.70)**

– Inconclusive prediction – Predicted Non-inhibitor – Predicted Inhibitor

NOTE: The coloring scheme takes into account both predicted probability and the Reliability Index values

Table 1 : Model performance for test set compounds after different numbers of similar molecules added to the library (numbers in parentheses report prediction Reliability Index values)

Advanced Chemistry Development Inc.
Tel: 1(416) 368-3435
Fax: 1(416) 368-5596
Toll Free: 1-800-304-3988
Email: info@acdlabs.com
www.acdlabs.com

GALAS Model Application on PubChem Data—an Example Scenario with CYP3A4 Inhibition

This study focuses on the application of models derived using GALAS modeling methodology in situations closely resembling their potential application in a real world environment. The GALAS model for the prediction of CYP3A4 enzyme inhibition developed at ACD/Labs was used as a starting point for this investigation. The model is based on approx. 900 compounds classified as either CYP3A4 inhibitors or non-inhibitors according to experimental data from publicly available publications. With the advent of the PubChem project[2], data for more than 11,000 individual compounds became publicly available. As a result, the PubChem collection has been chosen as a good representation of an actual 'in-house' project for the external validation of ACD/Labs' CYP3A4 inhibition model. For demonstration purposes, the available PubChem data (cleaned of salts, mixtures, *etc.*) classified using different thresholds (with IC_{50} values close to them treated as inconclusive and excluded) was used:

- CYP3A4 inhibition in general (IC_{50} < 50 uM)—8528 compounds
- Effective CYP3A4 inhibition (IC_{50} < 10 uM)—7696 compounds

The first corresponds to the criteria used in classification of the training dataset of ACD/Labs' CYP3A4 inhibition model. The second threshold was introduced primarily considering the fact that there is no objective definition of what a CYP3A4 inhibitor is. As a result of different experimental data interpretations between the algorithm developer and its user, dramatic mis-predictions can be observed even for compounds that are well within the Applicability Domain of the model. Additionally, even with the consistent classification threshold, the fact that one company will use property measurement protocols different from the ones usually used to measure the publicly reported values of the same property, can still result in inconsistent qualitative data due to the differences in quantitative results serving as a basis for the classification. All these factors introduce additional data variability, which is one of the causes contributing to the reduction of prediction quality.

Both PubChem sets were split in half with one part of the compounds intended for gradual addition to the blank Self-training Library, whereas the second was reserved for model performance evaluation.

Figure 2 illustrates a steady rise in the number of test set compounds falling within the Applicability Domain of the model (RI>0.3) and obtaining high quality predictions (RI>0.5) with the increase in size of the PubChem based Self-training Library. As shown in the previous example with hERG inhibition, such compounds receiving sufficiently reliable predictions are correctly classified as positive or negative in terms of the property in all but a few cases.

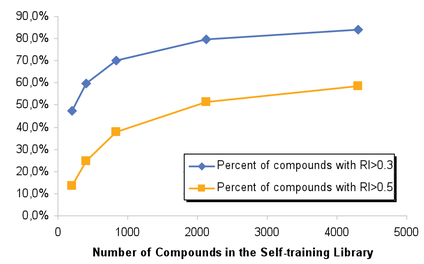


Figure 2: Number of corresponding reliability predictions following each addition of the general inhibition PubChem data to the Self-training Library of the CYP3A4 inhibition GALAS model.

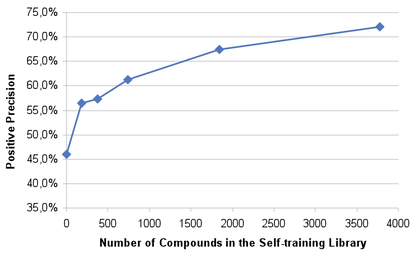


Figure 3: Changes in the positive precision of the GALAS model of CYP3A4 inhibition during its training with the effective inhibition PubChem set.

Note: Positive precision is equal to the fraction of true positives among all positive predictions of the model.

Figure 3 reveals that the precision of the initial ACD/Labs CYP3A4 inhibition model for the effective inhibition test set is ca. 40%. This number remains essentially unchanged while training the model with the general inhibition PubChem set, and is nowhere near the positive precision of the trained model obtained for the general inhibition test set (82%). This is no surprise given the differences in classification thresholds used to obtain general and effective inhibition sets. However, a dramatic impact on positive precision is observed if the first part of the effective inhibition set is used as a Self-training Library.

These observations suggest that the GALAS models can successfully cope with the practical challenges potentially arising during their applications in real life 'in-house' projects.

GALAS Models in ADME and Tox Suites

- CYP3A4 Substrate Specificity
- Genotoxicity (Ames test)
- P450 Inhibition Specificity*
- P-gp Substrate/Inhibitor Specificity
- Plasma protein binding (logK_a and %PPB)
- hERG channel inhibition
- Ionization constants (pK_a)
- Quantitative solubility in pure water (logS_w)
- Quantitative solubility in buffer (log$S$)
- Qualitative solubility in buffer
- Octanol-water or buffer partitioning coefficients (log$P$ and log$D$)

References

1. Zhu et al. *Bioorg Med Chem Lett.*, **16**:5507, 2006.
2. *NCBI PubChem database*, National Center for Biotechnology Information of the National Library of Medicine, Bethesda (MD), United States. Available at http://pubchem.ncbi.nlm.nih.gov/.