GALAS Modeling Methodology Applications In The Prediction Of Drug Metabolism Related Properties

Remigijus Didziapetris[1], Justas Dapkunas[1,2], Andrius Sazonovas[1], Pranas Japertas[1]

[1] ACD/Labs, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania,
[2] Department of Biochemistry and Biophysics, Vilnius University, M.K.Ciurlionio g. 21/27, LT-03101 Vilnius, Lithuania.

INTRODUCTION

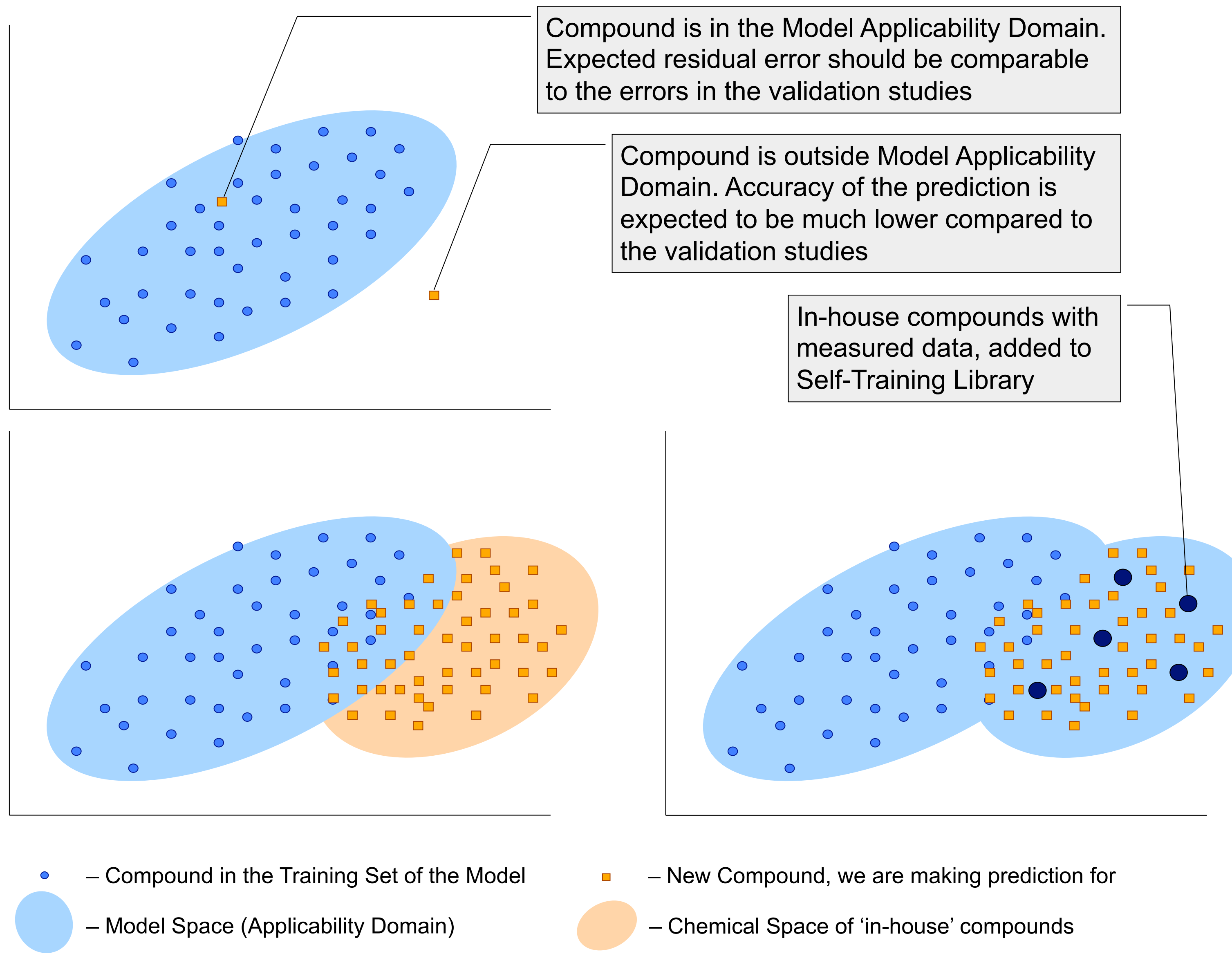
Analytical identification of metabolites for a drug candidate is usually a time consuming and low-throughput task and is performed only at the later phases of drug development. Therefore the possibility to predict possible sites of human liver microsomal (HLM) metabolism using *in silico* techniques would be a very attractive feature for any medicinal chemist. Yet, every model, no matter what 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 the chemical space that 'in-house' projects are focused on. 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 any company to effectively assess the Applicability Domain of any third-party model and to 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 providing a novel solution to this problem. Each GALAS model consists of the following parts:

- Structure based QSAR/QSPR for the prediction of the property of interest (i.e., baseline model)
- User defined data set with experimental values for the property of interest (i.e., Self-training Library)
- Special similarity-based routine which 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 (i.e., training engine)

The result is a prediction that is corrected according to the 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.



Compound is in the Model Applicability Domain. Expected residual error should be comparable to the errors in the validation studies

Compound is outside Model Applicability Domain. Accuracy of the prediction is expected to be much lower compared to the validation studies

In-house compounds with measured data, added to Self-Training Library

● – Compound in the Training Set of the Model

■ – New Compound, we are making prediction for

● – Model Space (Applicability Domain)

● – Chemical Space of 'in-house' compounds

FIGURE 1. Illustration of the Model Applicability Domain, and its expansion using the GALAS modeling method.

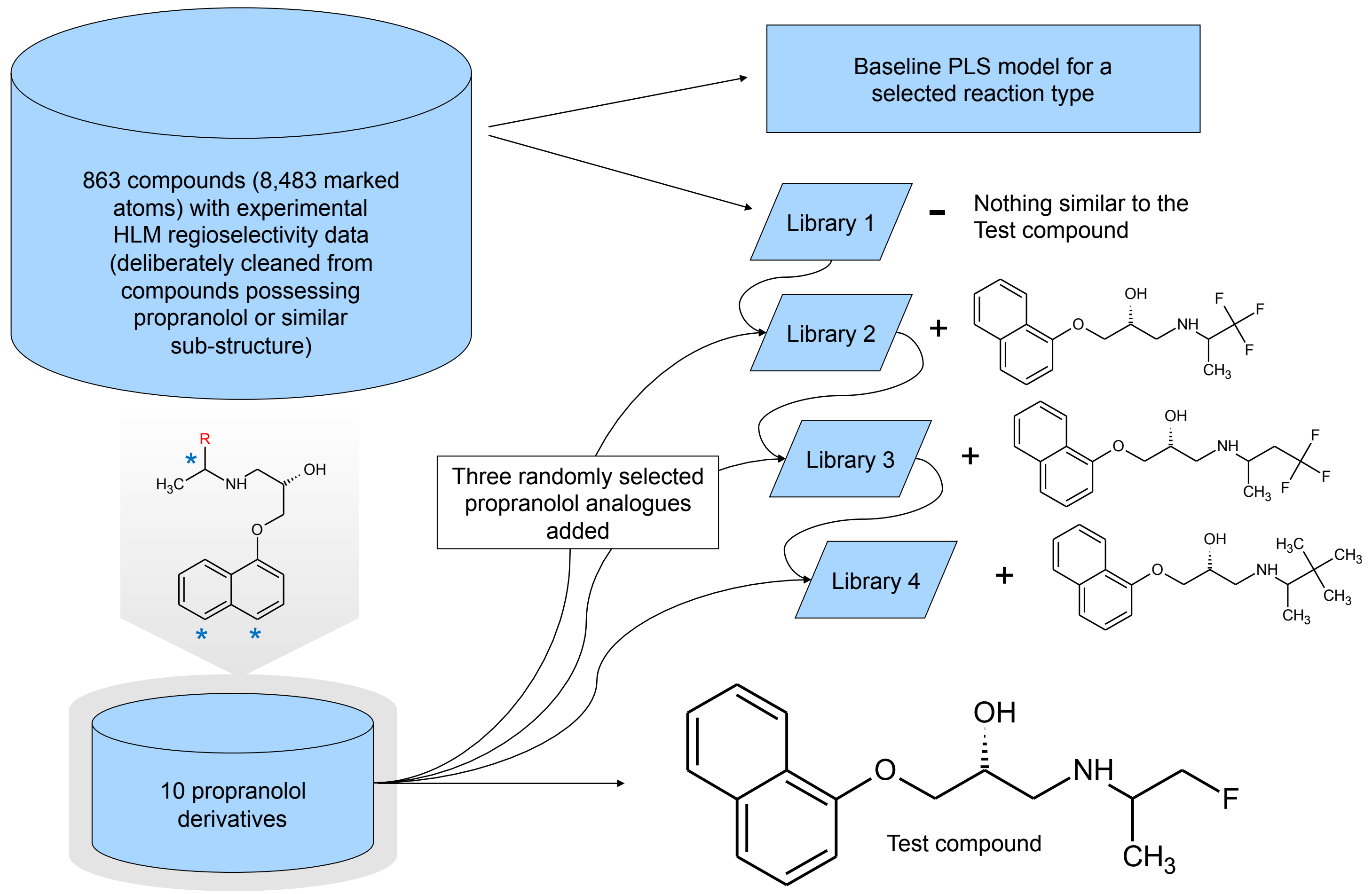
In addition, GALAS modeling methodology allows quantitative assessment of the prediction reliability. This information is contained in the developed Reliability Index (RI) that can provide values in the range [0–1]. Lower values suggest a compound being further from the Model Applicability Domain and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

- Similarity of the tested compound to the known compounds: No reliable predictions can be made if no similar compounds exist in the training set or Self-training Library.
- Consistency of the 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.

COPING WITH COMPLETELY NEW CHEMICAL FEATURES: AN EXAMPLE TRAINING SCENARIO WITH METABOLISM REGIOSELECTIVITY PREDICTION

The objectives of this scenario of the GALAS modeling methodology validation were as follows:

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



863 compounds (8,483 marked atoms) with experimental HLM regioselectivity data (deliberately cleaned from compounds possessing propranolol or similar sub-structure)

Baseline PLS model for a selected reaction type

Library 1 – Nothing similar to the Test compound

Library 2 +

Library 3 +

Library 4 +

Three randomly selected propranolol analogues added

10 propranolol derivatives

Test compound

SCHEME 1. Schematical representation of the virtual experiment procedures.

As illustrated in the scheme of the preparation steps, the compound class mimicking a new drug development project in this virtual experiment was propranolol analogs. These compounds are metabolized by CYP2D6 and CYP1A2 and all of them have two aromatic hydroxylation sites and one N-dealkylation site indicated by asterisks in the above outline [1,2].

The initial model with 0 analogs added predicts only one metabolism site of three. Such prediction would be classified as "satisfactory" according to the prediction quality evaluation criteria utilized in the HLM regioselectivity model evaluation (less than half of experimentally determined metabolism sites obtain probabilities of >0.5). After adding the first similar compound, both aromatic hydroxylation sites are predicted, and after adding another one, these sites are already estimated with high reliability.
The N-dealkylation site is found after addition of the third propranolol analog, resulting in "excellent" prediction (all experimentally determined metabolism sites receive probabilities of >0.5 and the atom ranked 1st is experimentally determined as a metabolism site).

The steady increase of the calculated probabilities for all metabolism sites as well as the growth of Reliability Index values for all atoms, indicate that following the model training the compounds of this novel class are already in the applicability domain of the model.

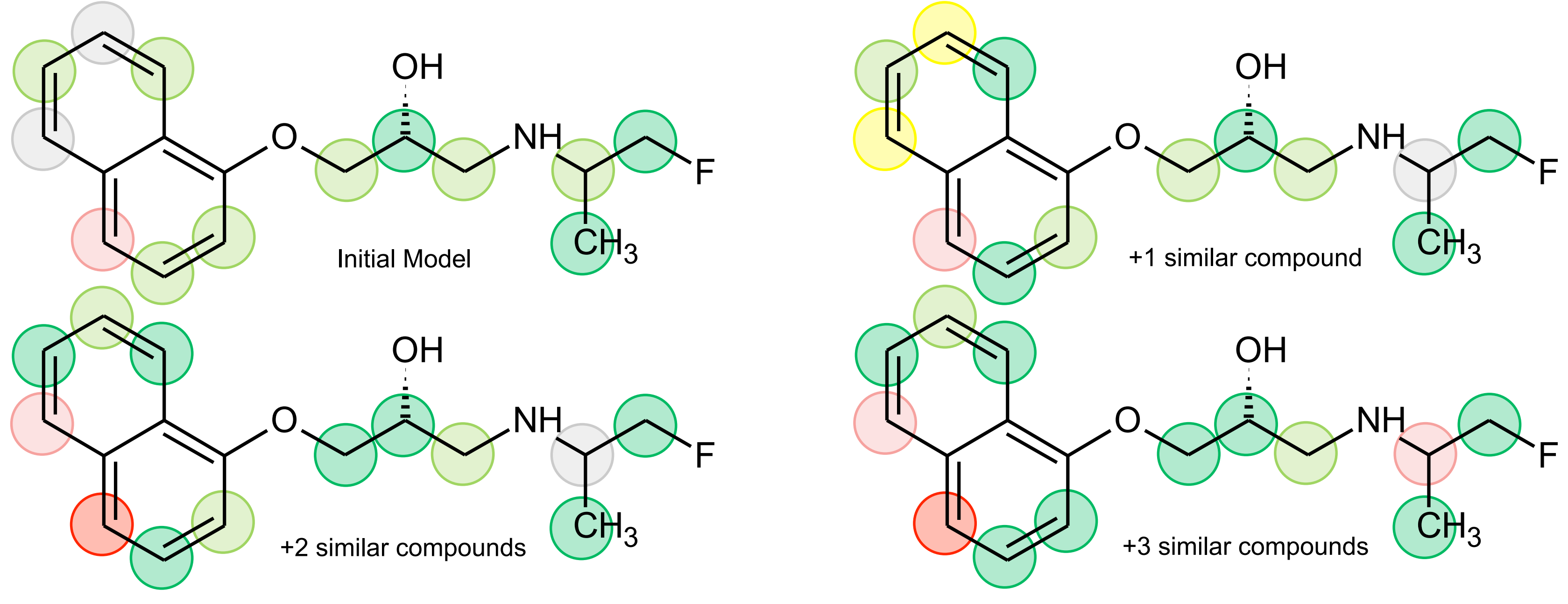


FIGURE 2. Changes in the HLM regioselectivity predictions for a selected testing compound–fluoropropranolol. Colors represent combination of predicted probability and Reliability Index value for every atom: red indicates site of metabolism, yellow–possible site of metabolism, grey–no prediction, green–no metabolism.

GALAS MODEL APPLICATION ON PUBCHEM DATA: AN EXAMPLE TRAINING SCENARIO WITH CYP3A4 INHIBITION

A GALAS model for the prediction of CYP3A4 enzyme inhibition developed at ACD/Labs using a training set of ca. 900 compounds was used as a starting point of this investigation [3]. A recently published PubChem collection [4] containing more than 11,000 individual compounds was chosen as a good representation of an actual 'in-house' project for the external validation of ACD/Labs CYP3A4 inhibition model. For demonstration, the available PubChem data sets (cleaned from salts, mixtures, *etc.*) were classified using different thresholds:

- CYP3A4 inhibition in general (IC_{50} < 50 μM)–8528 compounds
- Effective CYP3A4 inhibition (IC_{50} < 10 μM)–7639 compounds

The first threshold corresponds to the criteria used in classification of the training set data of the ACD/Labs CYP3A4 inhibition model. The second threshold was introduced primarily considering the fact that there is actually no objective definition of what is a CYP3A4 inhibitor, and as a result different classification schemes might exist. Additionally, even with the consistent classification threshold, a simple fact that a certain company is using a property measurement protocol that is different from the ones usually used to measure the publicly reported values of the same property can still result in inconsistent qualitative data. All of these factors introduce additional data variability which is one of the causes contributing to the reduction of prediction quality.

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

Increasing the size of the PubChem-based Self-training Library gives 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), which are correctly classified as positive or negative in terms of the property in all but a few cases.

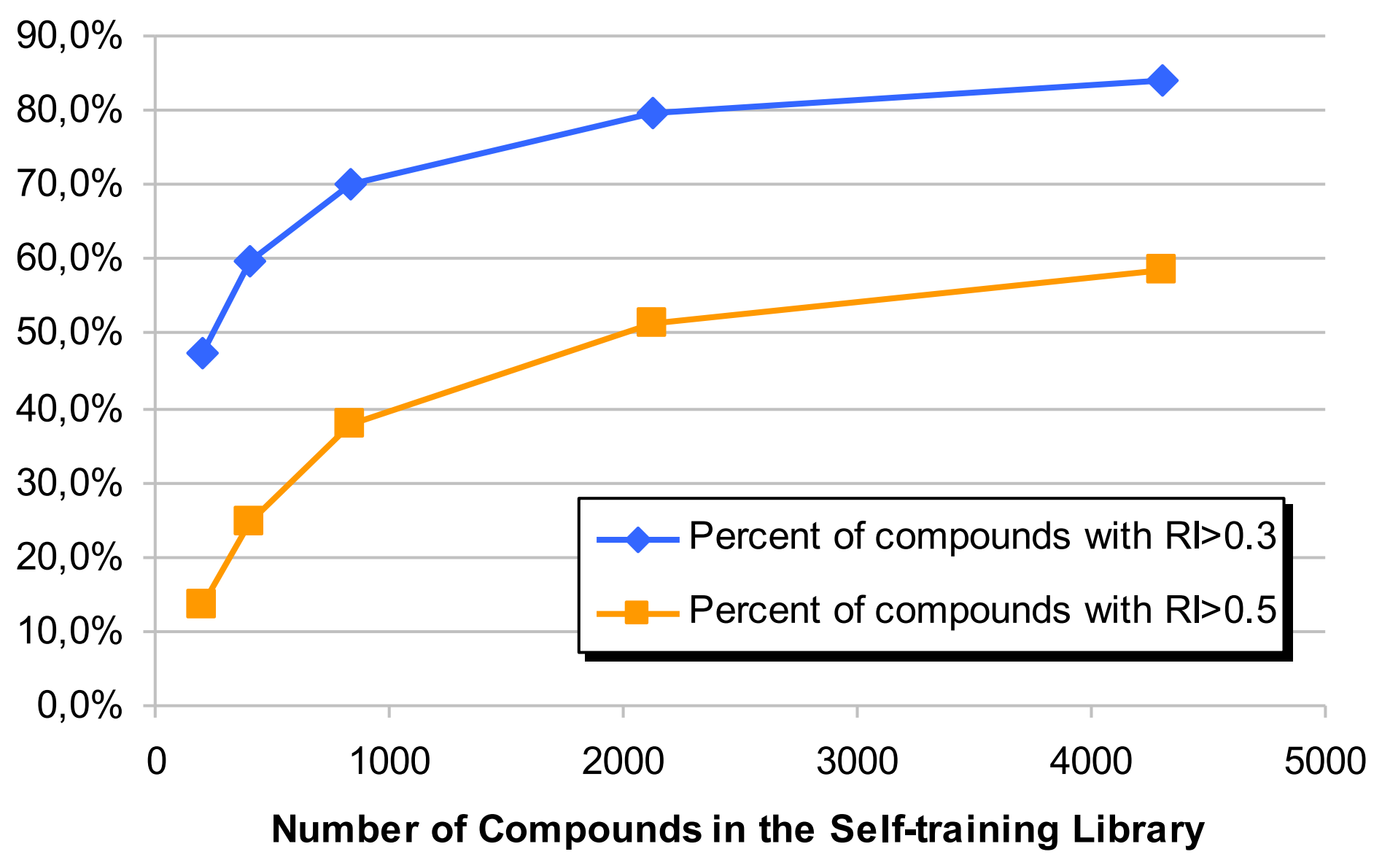


FIGURE 3. Number of corresponding reliability predictions following each addition of the general inhibition PubChem data to the Self-training Library of the CYP3A4 inhibition GALAS model.

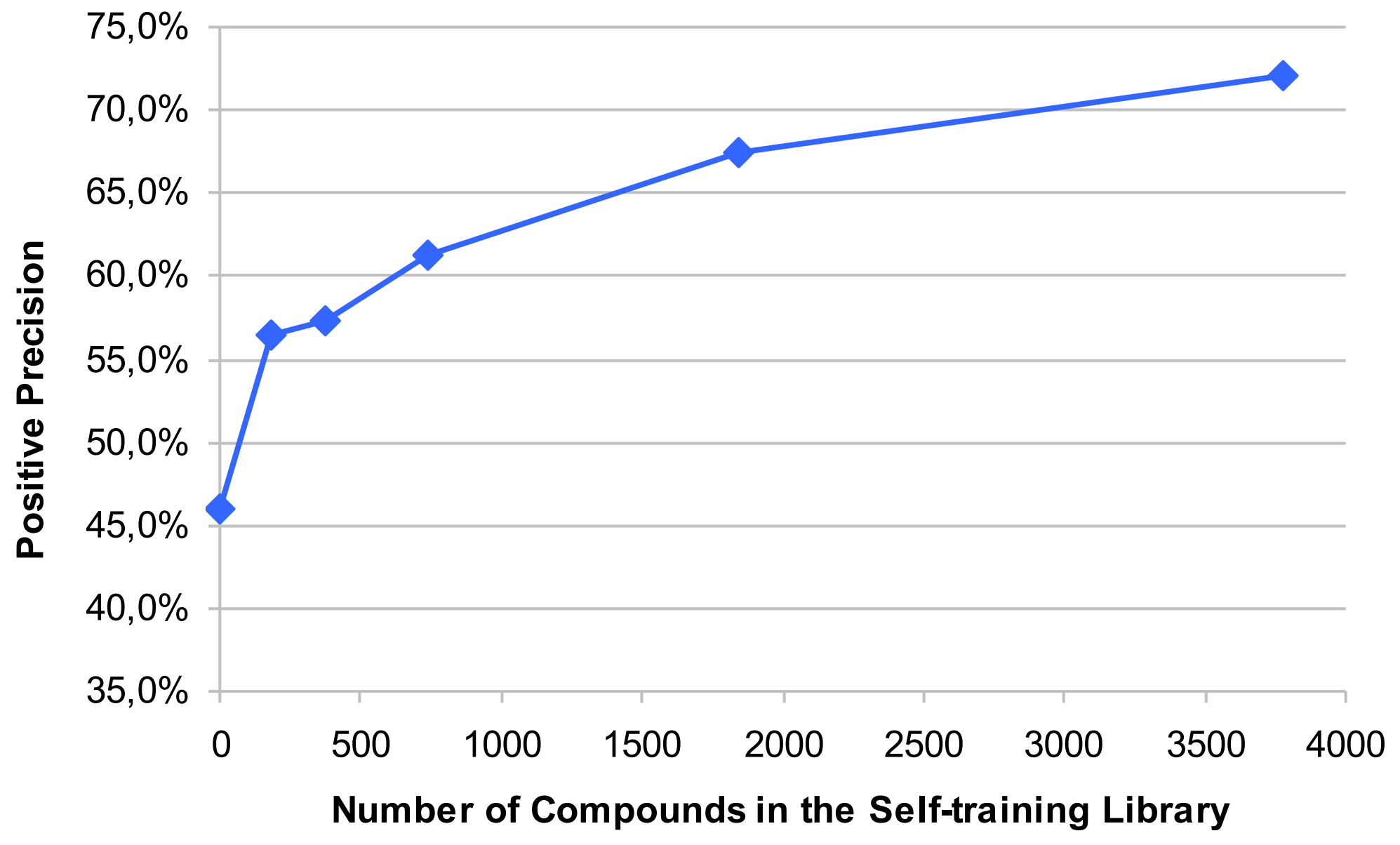


FIGURE 4. Changes in the positive precision of the GALAS model of CYP3A4 inhibition during its training with the effective inhibition PubChem set.

The positive precision (i.e., the fraction of true positives among all positive predictions of the model) of the initial ACD/Labs CYP3A4 inhibition model for the effective inhibition test set is ca. 40%. 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 ACD/ADME SUITE AND ACD/TOX SUITE

Currently ACD/Labs software products contain trainable GALAS models for the following properties:

- Genotoxicity (Ames test)
- Acute rodent toxicity (LD_{50})*
- Aquatic toxicity (LC_{50})**
- P450 Substrate Specificity***
- P450 Inhibition Specificity***
- P-gp Substrate/Inhibitor Specificity
- hERG channel inhibition
- Plasma protein binding (logK_a and *%PPB*)
- 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*)

* - Mouse OR, IP, IV, SC, and Rat OR, IP systems

** - fathead minnow (Pimephales promelas), and water flea (Daphnia magna) species

*** - CYP3A4, CYP2D6, CYP2C9, CYP2C19, and CYP1A2 isoforms

REFERENCES

[1] Upthagrove AL et al. *Drug Metab Dispos* **2001**, 29, 1377.

[2] Upthagrove AL et al. *Drug Metab Dispos* **2001**, 29, 1389.

[3] Didziapetris R et al. *J Comput Aided Mol Des* **2010,** 24, 891.

[4] *NCBI PubChem database*. Available at http://pubchem.ncbi.nlm.nih.gov/.

ACD/Labs

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