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

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

- Demonstrate that a GALAS model can be trained on 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 standard Dataset for the hERG family containing Factor Xa (hERG/KCNQ1 inhibition) in a similar manner were removed. This set was used in initial training of the baseline GALAS 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 pairwise additions to the initial training library, making it more and more benzamidine-like. 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: Model performance for test set compounds after different numbers of similar molecules added to the library (numbers in parentheses represent Prediction Reliability Index values)

<table>
<thead>
<tr>
<th>Number of Added Compounds</th>
<th>2</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted IC50 (nM)</td>
<td>0.64 (15)</td>
<td>0.68 (16)</td>
<td>0.70 (18)</td>
</tr>
<tr>
<td>Predicted IC50 (nM)</td>
<td>0.68 (16)</td>
<td>0.70 (18)</td>
<td>0.73 (21)</td>
</tr>
<tr>
<td>Predicted IC50 (nM)</td>
<td>0.69 (16)</td>
<td>0.71 (18)</td>
<td>0.73 (22)</td>
</tr>
<tr>
<td>Predicted IC50 (nM)</td>
<td>0.69 (16)</td>
<td>0.71 (18)</td>
<td>0.74 (23)</td>
</tr>
<tr>
<td>Predicted IC50 (nM)</td>
<td>0.69 (16)</td>
<td>0.71 (18)</td>
<td>0.75 (24)</td>
</tr>
</tbody>
</table>

**COPING WITH COMPLETELY NEW CHEMICAL FEATURES**

**A STRUCTURED APPROACH TO MODELING OF THE PROPERTIES OF INTEREST DERIVED FROM A LITERATURE TRAINING SET**

A series of 12,000 compounds was classified into three different models (CYP3A4, CYP2C9, and CYP2D6), with 100 compounds in each model. The models were trained on a self-training data set, which was composed of 10,000 compounds. The models were then tested on a validation data set, which was composed of 2,000 compounds.

**RESULTS**

The models were able to classify the compounds into the correct model with an accuracy of 95%. The models were then tested on a validation data set, which was composed of 2,000 compounds. The models were able to classify the compounds into the correct model with an accuracy of 95%.