Confidence Assessment of Candidate Drug Property Predictions by Subspace Mapping Methods
Axel J. Soto¹, Gustavo E. Vazquez², Marc Strickert³, Ignacio Ponzoni²⁴,
¹Faculty of Computer Science, Dalhousie University - Halifax, Canada
²Laboratory of Research and Development in Scientific Computing (LiDeCC), Department of Computer Science,
Universidad Nacional del Sur - Bahía Blanca, Argentina
³Institute for Vision and Graphics, Siegen University - Siegen, Germany
⁴Planta Piloto de Ingeniería Química, CONICET - Bahía Blanca, Argentina

Background
During the modern drug discovery process it is important to count on cheminformatics methods to improve the pharmacokinetic of the lead compound or modifying its activity. These computational methods are based on structure-property or structure-activity relationships (QSPR or QSAR), which essentially allow predicting target values such as biological properties or activities based on molecular descriptors of a given compound.

Although these methods constitute a rapid and cost-effective way for attaining of otherwise wet-lab empirical data, assessing the confidence in a QSPR-based prediction for every single compound is non-trivial. This problem is also known as the applicability domain estimation [1].

Most applicability domain approaches are based on applying similarity measures to an unseen compound for comparison against the training set of known compounds to evaluate how well represented that unseen compound is. One problem with this kind of approaches is that a given distance measure applied to the original descriptor space does not necessarily measure relatedness to the target values. This is even more relevant in the case of distances measured on a noisy and high-dimensional space, which is in fact the case of molecular descriptors used in QSPR.

Yet another challenge in applicability domain is identifying chemical spaces where the model is not performing reliably, even though that space may be well-represented by the training set of compounds.

Method
One of the objectives of our approach is to obtain an applicability domain method where its parameters and results can be visually understood by a chemical analyst. We aimed at assessing the prediction accuracy by using a probabilistic generative approach. In this way, the probability distributions for the input and the output of the prediction model can be used to determine the applicability domain. However, modeling probability distributions over high-dimensional spaces is computationally intractable. Therefore, a key stage of our approach is to first apply a subspace mapping method, where a low-dimensional representation of the training set of compounds is obtained. This low-dimensional space is sought as a linear transformation where a target-driven (i.e. supervised) criterion is optimized. As a result, compounds with similar target class values will be projected close to each other in an Euclidean distance sense. Three different target-driven projection criteria were applied, where Correlative Matrix Method (CMM) [2] turned out to be the most appropriate one.

On the low-dimensional space then, compounds belonging to the same target class will be modeled as being drawn from a Gaussian distribution. In this way, we can apply a Gaussian Bayes Classifier as the prediction model and, at the same time, make use of the likelihood and posterior probability as a means of providing different quantitative ways to assess prediction confidence. Figure 1 shows a visual representation of the chemical space, and the applicability domain determination for specific threshold values.

Fig. 1: Left: Projection to a 2-dimensional space using CMM (hidrophobicity dataset). Right: Applicability domain with \( \theta_p = 0.85 \) and \( \theta_l = 0.05 \) (posterior and likelihood probability thresholds respectively).

¹ Our QSPR aplicability domain model was tested on a classification setting, i.e. discrete target values are predicted.
Results
We applied this applicability domain estimation method to three different data sets, where the target properties are blood-brain barrier permeation, human-intestinal absorption and hydrophobicity which have target values for 289, 127 and 442 compounds respectively. Results show that by constraining the prediction of compounds to those in the applicability domain, the prediction accuracy increases substantially (Table 1). Although our applicability domain method is specially suited to Bayesian Classifiers, we showed that our approach also works well when other classification methods are used.

In conclusion, this method constitutes a new visually-assisted applicability domain technique for QSAR/QSPR models. A comprehensive explanation of our methodology and a detailed analysis of the experiments can be found in Soto et al. [3].

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood-brain barrier</td>
<td>17.24%</td>
<td>7.45%</td>
<td>35.71%</td>
</tr>
<tr>
<td>Intestinal absorption</td>
<td>36.15%</td>
<td>17.84%</td>
<td>45.34%</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>23.86%</td>
<td>10.88%</td>
<td>33.61%</td>
</tr>
</tbody>
</table>

Table 1: Misclassification averages for three different datasets (in rows) right after the subspace mapping is applied (“Overall” column) and after applying the applicability domain (“Accepted” and “Rejected” columns). Probability thresholds are \( \theta_p = 0.85 \) and \( \theta_l = 0.05 \). A Bayesian classifier was used for computing this table. Results are showed for the test compounds only.

Acknowledgements
This work is kindly supported by NSERC, PGI-UNS (24/ZN16), DFG Graduate School 1564, CONICET-PIP 112-2009-0100322 and MInCyT-BMBF (AL0811-ARG 08/016).

References