Sustainability assessment using local lazy learning: The case of post-combustion CO₂ capture solvents

Gulnara Shavalieva* Athanasios I. Papadopoulosb Sara Baddr Panos Seferlisb
Stavros Papadokonstantakisb

*a Chalmers University of Technology, Department of Space, Earth and Environment, Division of Energy Technology, SE-412 96 Gothenburg, Sweden
b Chemical Process and Energy Resources Institute, Centre for Research and Technology-Hellas, 57001 Thermi, Greece
gulnara.shavalieva@chalmers.se

Abstract

The consideration of sustainability is very important for the assessment of life cycle, environmental, health and safety properties of chemicals used in various applications. The screening of wide ranges of molecular structures, prior to the identification of the optimum and most sustainable options, requires the use of efficient and inclusive predictive models. Group contribution (GC) models are popular for the evaluation of numerous molecular options, but they support the calculation of few properties related to sustainability, while their predictive capabilities are often limited by significant data gaps. To address such challenges, we propose the use of a local learning approach as a means of evaluating sustainability properties for a wide range of molecular structures. Supplementing GC methods with data mining ones, such as local lazy learning approaches and exploiting molecular similarities has a potential to improve the predictive capacity of sustainability indices and offers an alternative when GC methods or empirical models are not available for life cycle assessment (LCA) and environmental, health and safety (EHS) hazard assessment indicators. The proposed approach is applied to predict a set of properties (bioaccumulation, persistence and acute aquatic toxicity) of 101 commercial solvents for post-combustion CO₂ capture.

Keywords: Local-lazy learning, Sustainability, Post-combustion CO₂ capture, EHS, data mining

1. Introduction

GC methods have been used very widely for the evaluation and selection of molecular structures with desired properties out of numerous options. Their exploitation in computer-aided molecular design (CAMD) approaches has resulted in numerous prior implementations in the design of solvents, polymers, biofuel additives etc. (Papadopoulos et al., 2016a).

In such implementations, it was realized that it is important to also account for sustainability in the context of environmental, health and safety aspects. Early evaluation of issues such as hazards associated with toxic releases of chemicals could help prevent incidents and minimize harmful impact to human health and environment. One of the main characteristics shared among studies that address the integration of sustainability assessment to automated selection of chemicals is that they focus on only few sustainability indicators. This is because the available GC and/or empirical models only
support few EHS and LCA endpoints. Recent efforts to incorporate sustainability assessment in GC-based CAMD are reported in Ten et al (2016) who integrates inherent safety and occupational health to the framework. The parameters required for application in CAMD are calculated via property prediction methods available for selected indicators. Another example is the study by Khor et al (2017) who use a CAMD approach to select sustainable solvents for palm oil extraction. The sustainability indices are obtained by GC and empirical models available for the required parameters. Both studies include safety and health indices, whereas potential impact to the environment is either not assessed or considered to a limited extent. However, a holistic sustainability assessment requires diverse sustainability indicators to achieve a reliable differentiation over a certain, typically large, number of molecules. Data mining methods can be used to analyse, interpret and exploit molecular similarities to supplement GC models to predict sustainability properties, to potentially improve their predictive capabilities and to expand the range of sustainability indicators that can be predicted. Local lazy learning is one such method that was proposed in recent years by Lu et al (2014) to predict the chemical toxicity profile of drug compounds, with encouraging results. To this end, we investigate the implementation of this method in the evaluation of the sustainability performance of numerous, diverse molecular structures used as solvents in CO2 capture.

2. Methodology
2.1 Motivation and approach

This work is motivated by Papadopoulos et al. (2016b), where CAMD was used for the design of CO2 capture solvents. The solvent design and selection was based on thermodynamic and reactivity property criteria predicted by GC and empirical models. Sustainability assessment was also performed but only at the post-design stage, where the required information to perform the assessment for few selected solvents was gathered manually from various databases, material data sheets and estimated using available software tools based on Quantitative Structure Activity Relationships. Even though the method provided reliable assessment results, it could not be applied at the design stage hence limiting the evaluated molecular structures.

To estimate safety, health and environmental scores of the metrics, at least one indicator value for dangerous property per aspect should be found for every molecule. An example of dangerous properties and indicators to calculate a score for the environment within EHS is presented in Table 1. To automate the assessment, the dangerous properties should be calculated or estimated by techniques which can be performed computationally and for the desired (wide) range of sustainability indicators. While some of the indicators can be calculated by GC and empirical models, it is troublesome to obtain information for other endpoints which cannot be estimated by such methods. In this work, the local lazy learning (LLL) technique is applied to evaluate its capability in closing data gaps and its potential to be integrated in a CAMD. LLL is a data mining technique which approximates a molecular target function (e.g. a sustainability indicator) locally, based on the known function value of the most similar molecule(s) available in a database (closest neighbours). Closest neighbours can be identified using the Tanimoto coefficient and the extended connectivity fingerprints (ECFPs). ECFPs encode molecular features (e.g. atoms, bonds, stereochemistry) and store them in a bit form. ECFPs are further used to calculate the Tanimoto index to examine similarity between compounds. The most similar compounds have the Tanimoto coefficient close to 1. The aim is to develop LLL.
models and compare their prediction accuracy with GC approaches for three important property indicators mentioned in Table 1 (LC50aquatic, persistence and bioconcentration) using a wide range of molecular structures for which available GC methods can provide predictions.

2.2 Dataset

The comprehensive database of Strempel (2012) was used to create a dataset to generate the models. This data was sorted and only molecules with the existing value for all three indicators of bioaccumulation, persistence and acute toxicity were left in the database. The database was further reduced to a dataset containing compounds consisting of sixteen functional groups relevant to the CO2 capture case. The selection of the groups was determined by the structure of the molecules identified in the work of Papadopoulos et al. (2016). The influence of functional group selection should be studied further; but it is assumed that such a reduction creates a dataset with more similar compounds that, according to Lu et.al (2014), is needed to enhance the prediction performance of LLL models. This led to a final dataset containing 2106 molecules. Acute toxicity and persistence were converted to logarithmic values.

2.3 LLL Model

For a given compound with an absent indicator value the LLL method identifies n nearest neighbours, i.e. most similar molecules in a dataset, and then uses these molecules to predict a property of the compound as follows (Lu et al, 2014):

\[ y_{\text{pre}} = \frac{1}{n} \sum_{i=1}^{n} S_i \times y_{i,\text{db}} \]  

where \( y_{\text{pre}} \) is the predicted value of the query compound, \( y_{i,\text{db}} \) is the property value of the i-th nearest neighbour from the dataset, \( S_i \) is the Tanimoto coefficient indicating the similarity value between the given compound and the i-th neighbour, and n is the optimized number of nearest neighbours used in the prediction. The Tanimoto coefficient can be calculated as follows:

\[ S = \frac{c}{a + b + c} \]  

where c is the number of atom pairs common for the molecules and a, b-the number of molecules’ unique atom pairs.

Thus, developing an LLL model consists in finding the optimal value of n, over a dataset.
and a target property. This optimal value is found by monitoring the LLL model performance over test compounds which are not included in the dataset used to search for the nearest neighbours. In this study this is achieved by leave-one-out cross-validation (LOO-CV). The procedure was applied as follows (Zheng and Tropsha, 2000):
1) One compound \(i\) of the database of \(N\) compounds was eliminated and the rest of the dataset molecules were used to predict the compound’s property (Eq. 3) using \(k\) (initially set to 1) most similar compounds from the dataset:

\[
y_{i,\text{pre},k} = \sum_{j=1}^{k} \frac{S_k}{\sum_{j=1}^{k} S_j} y_{k,\text{db}}
\]

where \(y_{i,\text{pre},k}\) is the predicted value of the query compound using \(k\) closest neighbours, \(y_{k,\text{db}}\) is the property value of the \(k\)-th nearest neighbour from the database, \(S_i\) the Tanimoto similarity value between the eliminated compound and the \(k\)-th neighbour, \(\sum_{j=1}^{k} S_j\) is the sum of Tanimoto coefficients of the eliminated compound and all its \(k\) nearest neighbours.

The step was repeated for all the compounds in the database \((i=1 \text{ to } N)\).

2) Cross-validated \(Q^2\) (Eq.4) was calculated using all the database molecules and their predicted \(y_{i,\text{pre}}\) and actual values \(y_i\), where \(y_{\text{avr}}\) is the average property value of all molecules in the dataset.

\[
Q^2 = 1 - \frac{\sum(y_i - y_{i,\text{pre}})^2}{\sum(y_i - y_{\text{avr}})^2}
\]

3) Steps 1 and 2 are repeated for \(k\) in range 1-50. The maximum possible \(k\) equals to the number of molecules in the dataset, however further increase of the range was unnecessary since the optimal number of neighbours was found to lay in range from 3 to 16 (see Fig. 1). This range remained the same even when the dataset was reduced to 700 and 1400 molecules to analyze the dependence of a dataset size on \(k\). The value of \(k\) which gave the highest \(Q^2\) was set to \(n\) and used for the final model (Eq. 1).

The optimal \(n\) number depends on the quality of the database. For instance, a larger database with more similar compounds could possibly result in a lower number of \(n\). The obtained \(n\) number was used to predict properties of solvents for post-combustion \(\text{CO}_2\) capture. The assumption here is that a new optimal number of \(n\) should not be required if this family of molecules has been indirectly included in the initial LLL development (i.e., the initial dataset was obtained by using the same functional groups as those used in the \(\text{CO}_2\) capture molecules). \(Q^2\) was calculated in the same way as stated in Eq. (4), and an

![Figure 1. \(Q^2\) vs number of neighbours \(k\)](image-url)
Sustainability assessment using local lazy learning: The case of post-combustion CO₂ capture solvents

average relative prediction error was determined by Eq. (5):

\[ RE = \frac{1}{N} \sum_{i} \left| \frac{y_i - y_{i,\text{pre}}}{y_i} \right| \times 100 \]  \hspace{1cm} (5)

3. Results

To demonstrate the performance of the model, the results are presented both for molecules from the dataset of 2106 molecules and for a subset previously identified as potential post-combustion CO₂ capture solvents (Papadopoulos et al., 2016b).

3.1. LLL model performance

The results of LOO-CV showed that the number of closest neighbours required to get the highest \( Q^2 \) values are different for every property indicator (Table 2 and Figure 1). After reaching the optimal values, the graphs level up and adding more neighbours into the model does not improve its predictive power. The work carried out by Lu et al. (2014) acquired similar \( Q^2 \) values for predicting \( LD50 \) indicator by LLL and their reference library consisting of 2271 compounds. Table 2 also contains \( Q^2 \) and \( RE \) values for the GC+ method as reported by Hukkerikar et al. (2012) for the case of the stepwise approach. It can be seen that published \( Q^2 \) numbers for the GC+ method are higher than those obtained by the LLL method for the dataset used.

3.2. Post-combustion CO₂ capture case

The best \( n \) values found by LOO-CV were used to predict the property indicators for 101 CO₂ capture commercial solvents from Papadopoulos et al. (2016), which all were also present in the dataset of the 2106 molecules. The results were compared to numbers obtained by the GC+ method, as shown in Table 3. The LLL model has a reasonable performance. Higher accuracy might be achieved by adding more molecules that are relevant for CO₂ capture and more structurally similar. Negative \( Q^2 \) numbers for the GC+ model show that the model fits the data for CO₂ capture solvents poorly.

Table 2. LLL method results compared to results reported for GC+

<table>
<thead>
<tr>
<th>Property indicator</th>
<th>Optimal n</th>
<th>( Q^2 ) for the optimal n value</th>
<th>Average ( RE )</th>
<th>( Q^2 )</th>
<th>Average ( RE )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogBCF</td>
<td>13</td>
<td>0.64</td>
<td>33</td>
<td>0.78</td>
<td>32</td>
</tr>
<tr>
<td>LogLC50</td>
<td>4</td>
<td>0.74</td>
<td>69</td>
<td>0.82</td>
<td>55</td>
</tr>
<tr>
<td>Log P</td>
<td>9</td>
<td>0.64</td>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\*RE of LogBCF is not given and log P is not studied by Hukkerikar, et al (2012)

Table 3. Results for CO₂ capture solvents obtained by LLL and GC+ methods

<table>
<thead>
<tr>
<th>Property indicator</th>
<th>( Q^2 ) for the optimal n</th>
<th>( R^2 )</th>
<th>Average ( RE )</th>
<th>Median ( RE )</th>
<th>( Q^2 )</th>
<th>( R^2 )</th>
<th>Average ( RE )</th>
<th>Median ( RE )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogBCF</td>
<td>0.18</td>
<td>0.55</td>
<td>41</td>
<td>32</td>
<td>-0.71</td>
<td>0.21</td>
<td>71</td>
<td>55</td>
</tr>
<tr>
<td>LogLC50</td>
<td>0.38</td>
<td>0.75</td>
<td>105</td>
<td>32</td>
<td>-0.17</td>
<td>0.51</td>
<td>150</td>
<td>70</td>
</tr>
<tr>
<td>Log P</td>
<td>0.68</td>
<td>0.71</td>
<td>9</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
4. Discussion and Conclusions

LLL gives different $Q^2$ values leading to different prediction capacity for particular indicators, which can probably be influenced by the quality of the dataset or character of molecular structure-dangerous property relationship for every indicator. This aspect needs to be studied further to identify which structural fragments are responsible for the highest contribution to further improve the model. An advantage of the LLL method is that it does not require extensive work prior to implementation and can be applied to any set of molecules, whereas GC methods are restricted to molecules that have groups for which contributions have been calculated. Up to now, there is quite a limited data on GC methods for such an analysis. Depending on the number of missing values and the database size, the LLL method can be computationally expensive. This effort can be reduced with a different $n$ optimization method.

LLL can be integrated into CAMD to predict information when values are missing, providing an alternative when GC methods or empirical models are not available for some EHS and LCA indices.

References

7. Strempel, S., Scheringer, M., Carla A. Ng, Hungerbühler, K., 2012, Screening for PBT chemicals among the “Existing” and “New” chemicals of the EU, Environment Sci. Technol., 46 (11), 5680-5687.