MACHINE LEARNING OF CORRELATION BETWEEN MOLECULAR STRUCTURE AND QUANTUM MECHANICAL CALCULATIONS OF SOLVATION CHARACTERISTICS

Jie-Jiun Chang¹, Jia-Lin Kang², David Shan-Hill Wong¹, Cheng-Hung Chou¹, Hsuan-Hao Hsu³, Chen-Hsuan Huang¹ and Shang-Tai Lin³

1 Department of Chemical Engineering, National Tsing Hua University, Hsinchu, Taiwan 30013
2 Department of Chemical and Material Engineering, Tam Kang University, New Taipei City, Taiwan
3 Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan

Abstract
In this paper, a supervised machine learning method is proposed to project molecular features as floating-point numbers in a high dimensional space from the language-like description Simplified Molecular Input Line Entry Specification (SMILES) and densing them into a molecular fingerprint known as Molecular ACCess System (MACCS). A neural network model is build using the location of a compound in the high dimensional space as input to predict the “sigma-profile”, the charge distribution of the molecule near a perfect infinite conductor, which is calculated by quantum mechanics. The sigma-profile can be used in the COSMOSAC model for predicting thermodynamic properties such as activity coefficient. Preliminary results showed that an accurate neural work model with generalization ability can be developed. Moreover it was found that the sigma profile prediction accuracy direct use of MACCS as input is much inferior.

Keywords: Machine Learning, Word embedding, MACCS, Sigma profile, COSMO-SAC

Introduction
Soft-sensor, sequence-to-sequence recurrent network.

Computer aided molecular design (CAMD) of products and processes involved following iterative steps (Ng et al 2015):
1. Generation a molecular structure.
2. Extract a molecular features relevant to target properties.
3. Predict target properties using the molecular features using a model.
4. If target properties are satisfactory, stop
5. Else go to step 1.

The prediction step can be performed by semi-empirical methods the group contribution (GC) methods in which molecules were broken down into functional groups. The properties were then predicted by either linear or nonlinear functions of group parameters or interaction parameters between groups. For example, the Joback method (Joback and Reid 1987) for pure component properties is a linear function of group parameters. The UNIFAC method for activity coefficients for binary mixtures (Fredenslund et al 1975) is a nonlinear function of interaction parameters between groups. Alternatively, in quantitative structure properties/activities (QSPR/QSAR), topological indices or molecular descriptors based on chemical graph theory were used to predict properties (Rogers and Hopfinger 1994). In recent years direct quantum mechanical calculations were used to predict thermodynamic properties. For example, the COSMO method, quantum mechanical calculations is used to predict the charge distribution of a molecule near an infinite conductor, known as the sigma-profile (Klamt 1995). The sigma-profile can be used to predict thermodynamic properties of mixtures (e.g. COSMOS-RS Klamt and Eckert 2000, COSMO-SAC Lin and Sandler 2002). Substantial effort needs to be expended in producing the sigma-profile. Database of sigma-profiles of a limited number of compounds was provided (Mullins et al 2006). Yet it is desirable that a fast surrogate generation method be developed to alleviate the load of first principle calculations.

Recently the use machine learning or deep learning models has been used to develop improved QSPR. In some work, e.g. Faber et al 2017, molecular descriptors are still used as the regressors, but machine learning models were used to replace linear or simple functional representations. Alternatively, simple 2-dimensional molecular structure were represented as an image as input a convolutional network for properties prediction (Goh et al. 2017). Yet there are many ways to represent molecular structure ranging from aforementioned group contribution method to text-based description such as Simplified Molecular Input Line Entry Specification (SMILES) (Weininger 1998)
Recent development in machine learning has been able to convert word-based content into vector space (Mikolov et al. 2013). In this paper, a word embedment approach is proposed to map molecular features from SMILES representation into a high-dimensional floating-point vector space. The mapping was trained by projecting positions in this high dimensional space into a simple molecule recognizer, and a molecular fingerprint known as Molecular Access System (MACCS, Durant et al. 2002). The positions in this high dimensional space was then used to develop a neural network to generate sigma-profiles.

Word Embedment of SMILES

Encryption of SMILES string

The SMILES format encoded molecular structure clearly with a short ASCII string. It is able to read the structures like chain length, double bond (=), triple bond (#) and aromatic ring (c1cccc1) easily. The molecular structure of methanol and its SMILES and MACCs representation were given in Table 1.

| (a) Molecular structure | CO |
| (b) SMILES | 93: QCH3 |
| | 139: OH |
| | 157: C-O |
| | 160: CH3 |
| | 164: O |
| (c) MACCS |

In order that the SMILES is readable by machine, we need to code the representation into a vector of numbers. We assumed that the maximum length of SMILES is 50, and assigned every elements and symbols used in SMILES to a specific number. Then we can translate the SMILES format data to sequences of numbers.

Mapping to a high dimensional space

However, these number sequences are encrypted codes, they do not have any metric of distance, nor do they “characteristic” or feature of the molecules. This coded sequence is map into a $50 \times n$ matrix using a neural network with 1 layer and $n$ nodes (embedded layer in Figure 1a and b). To ensure that some molecular features were embedded in the high dimensional space, two approaches were tested. The first approach is called a SMILES molecular recognizer (SMILES_MOLREC). In SMILES_MOLREC, the high dimensional vector was projected back into discriminant output which is 0 for a true compound or 1 for false compound. This projection was done by a dense-layer composing of 1 long and short term memory (LSTM) layer and 1 output layer with one node (Figure 1a).

To train SMILES_REC, 1372 true compounds are used and tagged 0, or “true-compound”. Another 1372 compounds with randomly created SMILES file are tagged 1 as “false” or “fake-compound”. Supervised learning are used until the recognizer network can accurately distinguish between the two classes. 274 true compounds and the same amount of fake-compound are used as test set.

In an alternative approach, a SMILES to MACCS (SMILES→MACCS) translator can be build. The high dimensional vector was projected back into a 1x166 vector with entries which is either 1 or 0, which is the MACCS representation of the compound. This projection was done by a dense-layer composing of 1 LSTM layer and 1 output layer with 166 nodes (Figure 1b).

In training the SMILES→MACCS translator, the same 1372 true compounds were used and an additional 274 true compounds were used as test data.

![Figure 1: Architecture of a Molecular Recognizer and an MACCS translator](image)

The test statistics were shown in Table 2. The false positive rate and false negative rate were calculated for SMILES_MOLREC. The average number of erroneous entries in a test compound was given for SMILES→MACCS.

<table>
<thead>
<tr>
<th>n</th>
<th>SMILES_MOLREC</th>
<th>SMILES→MACCS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Molecular of the high dimensional representation

Classifications of compounds

Principal component analysis (PCA, Jolliffe 2011) was performed to help visualize how molecules are distributed in the high dimensional space. The high dimensional subspaces obtained using $n = 100$. If we locate...
homologues of normal paraffins, straight chain alcohols and acids in the PCA subspaces. In both spaces are arranged in an orderly manner, indicating that the embedment is able to transform the text-based SMILES input into a distance-relevant high dimensional space that can be used for molecular recognition and classification.

Effect of selective sampling

Table 4: The prediction accuracy (R²) of test compounds

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACCS</td>
<td>0.7151</td>
<td>0.8368</td>
<td>-1.4565</td>
</tr>
<tr>
<td>SMILES_MOLREC</td>
<td>0.8687</td>
<td>0.9510</td>
<td>-1.1238</td>
</tr>
<tr>
<td>SMILES→MACCS</td>
<td>0.8962</td>
<td>0.9581</td>
<td>-0.3489</td>
</tr>
</tbody>
</table>

Infinite Dilution Activity Coefficients

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMILES_MOLREC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMILES→MACCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Comparison infinite dilution coefficients calculated by COSMOSAC using machine-learned and actual sigma profiles (a) water, (b) n-hexane, (c) DMSO, (d) Nitromethane

Conclusions

The above results serve as a preliminary demonstration that molecular classification and prediction of sigma-profile, results of quantum calculations, using text-based molecular description is possible. A recognizer was trained using word-embedment network, and a LSTM transformation network. The network gave no false negative and very few false positive. PCA analysis showed that the transformed space can be used as for molecular feature representation and classification. Use this space as input, we showed that fairly accurate prediction of sigma-profile can be developed. Optimization of network structure have not yet been considered. The promising results suggest that extension of this approach to a more extensive data base should be a valuable for a priori property prediction and molecular design.

References


