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

Jie-Jiun Chang[1], Jia-Lin Kang[2], David Shan-Hill Wong[1], Cheng-Hung Chou[1], Hsuan-Hao Hsu[3], Chen-Hsuan Huang[3] and Shang-Tai Lin[3]

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

Keywords

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

Introduction

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)

or .mol file (Dalby et al 1992) or even three dimensional representation (Humphrey et al 1996).

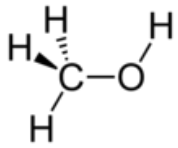
Recent development in machine learning has been able to convert word-based content into vector space (Mikolov et al 2013). In this paper, a word embedment approach is proposed to map molecular features from SMILES representation into a high-dimensional floating-point vector space. The mapping was trained by projecting positions in this high dimensional space into a simple molecule recognizer, and a molecular finger-print 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 (c1ccccc1) easily. The molecular structure of methanol and its SMILES and MACCs representation were given in Table 1.

Table 1: Molecular structure, SMILES and MACCS representation of Methanol

 (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.

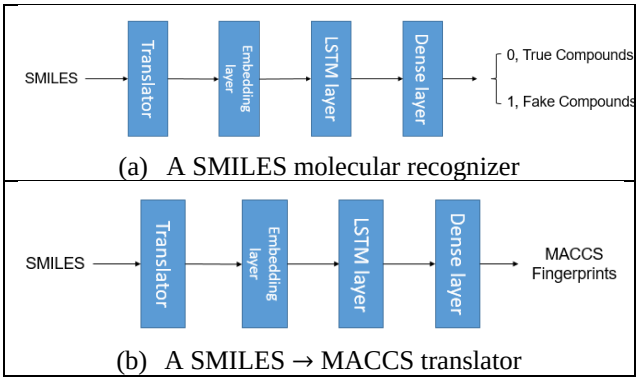
Mapping to a high dimensional space

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 memeory (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.



(a) A SMILES molecular recognizer

(b) A SMILES → MACCS translator

Figure 1: Architecture of a Molecular Recognizer and an MACCS translator

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 2: Test statistics of SMILES_MOLREC and SMILES→MACCS

n	SMILES_MOLREC		SMILES→ MACCS
	False Positive	False Negative	λ
100	1	0	0.0110
200	1	0	0.0105
500	2	0	0.0103

Molecular of the high dimensional representation

Classifications of compounds

Principle 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.

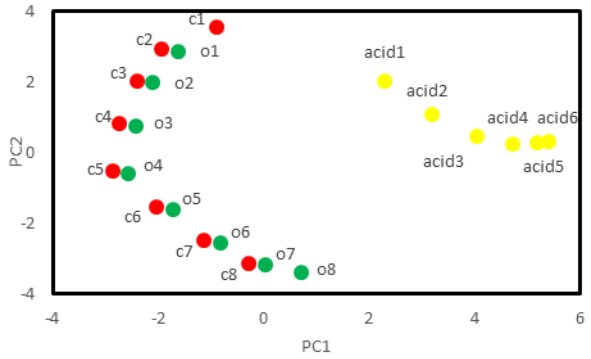


Figure 2: Arrangement of compounds in the PCA subspace of high dimensional map constructed by SMILES_MOLREC

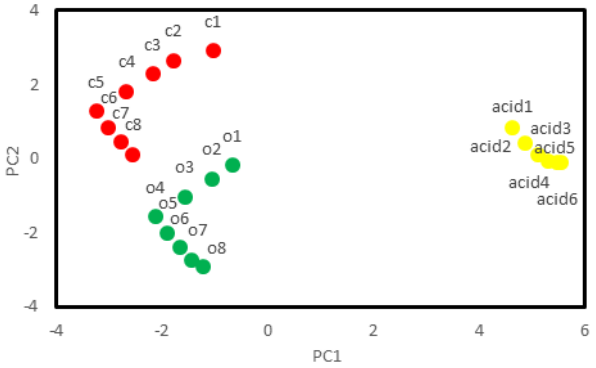


Figure 3: Arrangement of compounds in the PCA subspace of high dimensional map constructed by SMILES→MACCS

Correlation of the Two High Dimensional Map

Sigma profile predictions

Random sampling

To predict the sigma profile, three neural networks were build. One used the MACCS vector directly as input. The other two used the locations of high dimensional maps generated by SMILES_MOLREC and SMILES→MACCS as input.

Table 3: The prediction accuracy (R^2) of test compounds

	R^2 of Test Compounds		
	Mean	Median	Minimum
MACCS	0.7151	0.8368	-1.4565
SMILES_MOLREC	0.8687	0.9510	-1.1238
SMILES→MACCS	0.8962	0.9581	-0.3489

Effect of selective sampling

Table 4: The prediction accuracy (R^2) of test compounds

	R^2 of Test Compounds		
	Mean	Median	Minimum
MACCS			
SMILES_MOLREC	0.9581	0.9732	0.7877
SMILES→MACCS	0.9641	0.9808	0.7835

Infinite Dilution Activity Coefficients

(a)	(b)
(c)	(d)

Figure 4: Comparison infnite 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

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K., Grier, D. L., Leland, B. A., & Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. Journal of chemical information and computer sciences, 32(3), 244-255.

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. Journal of chemical information and computer sciences, 42(6), 1273-1280.

Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., ... & Von Lilienfeld, O. A. (2017). Prediction errors of molecular machine learning models lower than hybrid DFT error. Journal of chemical theory and computation, 13(11), 5255-5264.

Fredenslund, A., Jones, R. L., & Prausnitz, J. M. (1975). Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE Journal, 21(6), 1086-1099.

Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O., & Baker, N. (2017). Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. arXiv preprint arXiv:1706.06689.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. Journal of molecular graphics, 14(1), 33-38.

Jolliffe, I. (2011). Principal component analysis. In International encyclopedia of statistical science (pp. 1094-1096). Springer, Berlin, Heidelberg.

Kier, L. B., & Hall, L. H. (1986). Molecular connectivity in structure-activity analysis. Research Studies.

Klamt, A. (1995). Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. The Journal of Physical Chemistry, 99(7), 2224-2235.

Klamt, A., & Eckert, F. (2000). COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. Fluid Phase Equilibria, 172(1), 43-72.

Joback, K. G., & Reid, R. C. (1987). Estimation of pure-component properties from group-contributions. Chemical Engineering Communications, 57(1-6), 233-243.

Lin, S. T., & Sandler, S. I. (2002). A priori phase equilibrium prediction from a segment contribution solvation model. Industrial & engineering chemistry research, 41(5), 899-913.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mullins, E., Oldland, R., Liu, Y. A., Wang, S., Sandler, S. I., Chen, C. C., & Seavey, K. C. (2006). Sigma-profile database for using COSMO-based thermodynamic methods. Industrial & engineering chemistry research, 45(12), 4389-4415.

Ng, L. Y., Chong, F. K., & Chemmangattuvalappil, N. G. (2015). Challenges and opportunities in computer-aided molecular design. Computers & Chemical Engineering, 81, 115-129.

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Modeling. 1998, 28, 31–36.

Rogers, D., & Hopfinger, A. J. (1994). Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. Journal of Chemical Information and Computer Sciences, 34(4), 854-866.