

Linear Free Energy Relationships and Latent Variables: Similarity in Modelling Two Environmentally Relevant Properties

Tomas Öberg* and Tao Liu

School of Pure and Applied Natural Sciences, University of Kalmar,
391 82 Kalmar. *E-mail: tomas.oberg@hik.se



Introduction

The equilibrium partition constant between two phases, on a mole fraction basis, is defined as:

$$K_{12} = \exp[-\Delta_{12}G_f/(RT)]$$

The free energy change can be separated into the contributions from van der Waals and polar interactions, assuming that these are additive:

$$\Delta_{12}G_f = \Delta_{12}G_{vdW} + \Delta_{12}G_{polar}$$

Here we will evaluate two approaches to model these interactions and determine the partition constants: linear free energy relationships (LFER) and latent variable models based on partial least squares regression (PLSR).

In LFER-models the factors are aimed to describe different aspects of van der Waals and polar interactions. Abraham et al. developed a general linear solvation energy relationship (LSER) — a LFER for solvent-solute interactions — for the correlation with five solute descriptors [1]:

$$SP = c + e \cdot E + s \cdot S + a \cdot A + b \cdot B + v \cdot V$$

A multidimensional space of theoretical descriptors can be compressed by PLSR to a few informative new “super-descriptors” through linear combinations of the original variables. The solvent property is thus related to a low-dimensional description of the chemical structure and/or properties [2]:

$$SP = SP_{av} + q_1 \cdot t_1 + q_2 \cdot t_2 + q_3 \cdot t_3 + q_4 \cdot t_4 \dots$$

Two environmentally relevant properties, the partitioning between octanol and water (K_{ow}) and water solubility (S_w), were selected as endpoints for comparison and evaluation of the modelling methodology.

Materials and Methods

Abraham solute descriptors were collected from the Absolv Database v.3.5.

Three-dimensional molecular structures were optimised with the software Corina v3.2 and 1664 theoretical descriptors were generated with the software Dragon v.5.4.

Experimentally determined K_{ow} - and S_w -values at 298.15 K were retrieved from the PhysProp Database. The finally selected data sets contained 1505 and 884 compounds, respectively.

Multiple linear regression and partial least squares regression were used for data analysis and modelling.

Results and Discussion

Model performance

The two endpoints, K_{ow} and S_w , were transformed to their logarithms. One third of the data was randomly assigned to an external test set to evaluate predictive performance. Thirty-seven outlying objects were removed from each calibration set, and the model performances are summarized in Table 1.

The PLSR models were based on 92 respectively 388 theoretical descriptors projected down to four respectively five orthogonal factors (same dimensionality as the LSER models).

Table 1: Comparison of model performance

Model		$\log K_{ow}$	$\log S_w$
n	Cal/pred	966/502	552/295
k	Dim	4	5
LSER	R^2_{Cal}	0.979	0.948
	Q^2_{Ext}	0.955	0.879
	RMSEP	0.410	0.937
PLSR	R^2_{Cal}	0.957	0.959
	Q^2_{Ext}	0.933	0.922
	RMSEP	0.498	0.740

The prediction results for $\log K_{ow}$ with the two models are also shown in Figures 1 and 2.

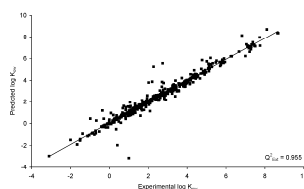


Figure 1: Predicted vs. measured $\log K_{ow}$ for a four-parameter LSER model (external test set).

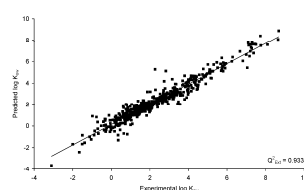


Figure 2: Predicted vs. measured $\log K_{ow}$ for a four-factor PLSR model (external test set).

Interpretability

The LSER models provide a framework to interpret the relationship between the chemical structure and the partitioning of a compound, but also have limitations in the “blending of chemical contributions” [3]. The LSER descriptors are thus not necessarily the realisations of fundamental properties that we search for, but we can use them as provisional conjectures.

The relationships between the individual PLS factors and the solute descriptors then become interesting. The two most important LSER descriptors, in both models, were those encoding for McGowan volume and hydrogen bond basicity, describing:

- van der Waals interactions
- Polar interactions

The two first PLSR factors in the water solubility model correlate strongly with these two descriptors, $R^2=0.87$ and $R^2=0.82$ respectively. In the PLSR model for partitioning between octanol-water a linear transformation was necessary to obtain similar correlations, $R^2=0.88$ and $R^2=0.77$.

Conclusions

This study has shown that a PLSR approach based on theoretical computed descriptors can attain many of the same benefits as a LSER model, with more general applicability and potential for coverage of a larger chemical domain.

- A PLSR model can be interpreted similarly to a LSER model
- The predictive performances seem to be similar with both approaches

The choice of model should therefore primarily be driven by the availability of data and the predictive performance [4].

References

- [1] Abraham, M. H., Ibrahim, A., Zissimos, A. M. 2004. Determination of sets of solute descriptors from chromatographic measurements. *Journal of Chromatography A* **1037**(1-2): 29-47.
- [2] Wold, S., Sjöström, M. 1998. Chemometrics and its roots in physical organic chemistry. *Acta Chemica Scandinavica* **52**(5): 517-523.
- [3] Vitha, M., Peter, W. C. 2006. The chemical interpretation and practice of linear solvation energy relationships in chromatography. *Journal of Chromatography A* **1126**(1-2): 143-194.
- [4] Liu, T., Öberg, T. 2009. Modelling of partition constants: linear solvation energy relationships or PLS regression? *Journal of Chemometrics*, in press.