Uniting Cheminformatics and Chemical Theory to Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules

James L. McDonagh, Neetika Nath, Luna De Ferrari, Tanja van Mourik & John B. O. Mitchell

EaStCHEM School of Chemistry, University of St Andrews

North Haugh, St Andrews, Fife, KY16 9ST, UK

jm222@st-andrews.ac.uk & nn223@st-andrews.ac.uk

University of St Andrews

EaStCHEM The Edinburgh and St Andrews Research School of Chemistry

SULSA Scottish Universities Life Sciences Alliance

Scottish Funding Council Promoting further and higher education

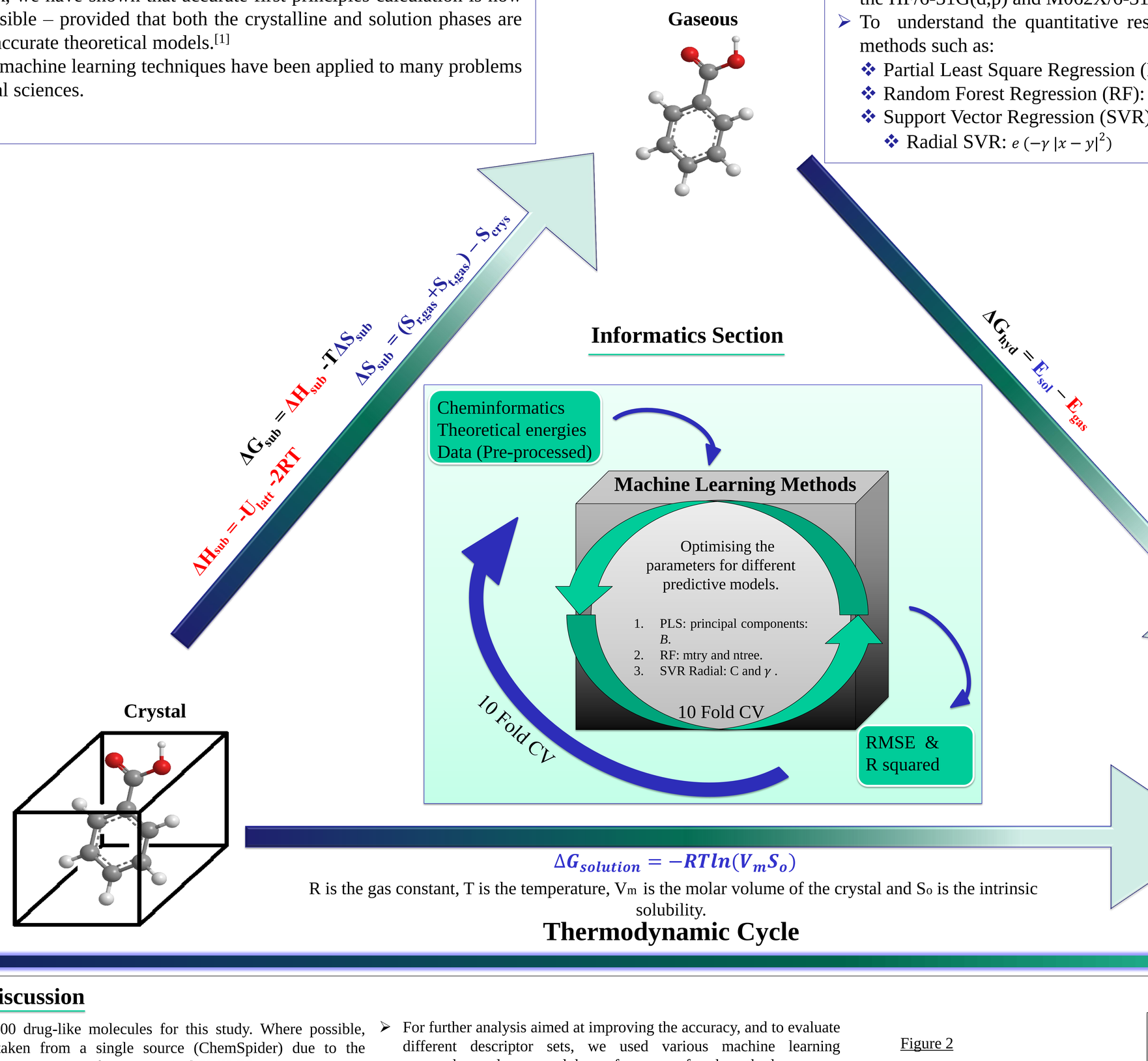
BBSRC bioscience for the future

Introduction

- Poor aqueous solubility remains a major cause of attrition in the drug development process.
- In previous work, energy terms from a computed thermodynamic cycle had been used as descriptors in a multi-linear regression model for intrinsic solubility. Accuracy much better than from direct computation and comparable to leading informatics approaches was achieved.
- In recent work, we have shown that accurate first principles calculation is now becoming possible – provided that both the crystalline and solution phases are described by accurate theoretical models.[1]
- Sophisticated machine learning techniques have been applied to many problems in the chemical sciences.

Methods

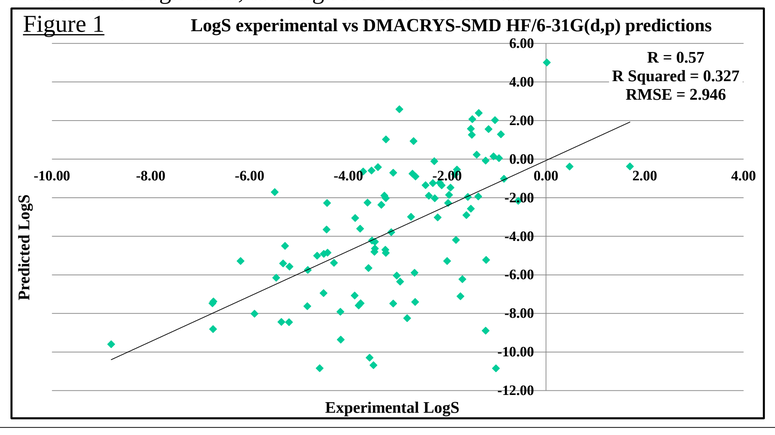
- The free energy terms are partitioned into physically meaningful terms. The relevant solute – solute, solute – solvent and solvent – solvent enthalpy and entropy terms are computed.
- DMACRYS is used to calculate the crystal lattice energies. Gaussian 09 (G09) is used to calculate the gaseous and solution phase energies.
- The SMD solvation model was used; gas and solution phases were calculated at the HF/6-31G(d,p) and M062X/6-31G(d,p) levels of theory.[1]
- To understand the quantitative results, we have performed machine leaning methods such as:
  - Partial Least Square Regression (PLS): $Y=XB+\varepsilon$ where $B=[(X^T X)]^{(-1)} X^T Y$
  - Random Forest Regression (RF): ensemble of many decision trees.
  - Support Vector Regression (SVR): $f(x,\omega) = \sum_{j=1}^{m} \omega_j g_i(x) + b$
    - Radial SVR: $e\,(-\gamma\,|x-y|^2)$



Results and Discussion

- We have used 100 drug-like molecules for this study. Where possible, SMILES were taken from a single source (ChemSpider) due to the variability in the interpretation of non-canonical SMILES strings.
- The descriptors were calculated from SMILES strings in the Chemistry Development Kit (CDK). These descriptors include Molecular Weight, XLogP, Freely rotatable bonds, number of H-bond acceptors etc.
- Figure 1: represents the correlation between the HF/6-31G(d,p) theoretical calculation, which outperformed the M062X/6-31G(d,p) calculation, and the experimental LogS values.
- Table 1 summarises the results of a linear regression analysis of the theoretical prediction against the experimental results.
- Our initial results, a linear regression analysis, suggested an error of almost 3 LogS units, too large for a useful model.

Figure 1



- For further analysis aimed at improving the accuracy, and to evaluate different descriptor sets, we used various machine learning approaches and compared the performance of such methods.
- Figure 2: boxplots represent the distribution of RMSE of various machine learning methods using 10 fold CV for different sets of descriptor. Here, the red dot represents the average performance (RMSE) of different models.
- Table 2 (A and B): reports the average over 10 fold CV of RMSE and R squared scores of different models and descriptor sets. These results suggest that RF performed slightly better than other predictive models when fitted with all sets of descriptors.
- Overall, machine learning methods significantly improved upon the linear relationship, suggesting that the theoretically calculated data could be explained by non-linear QSPR models.
- Our results also suggest that the prediction is significantly improved if theoretical energies are combined with cheminformatics descriptors.
- Machine learning based solely on computed energy terms and a study of variable importance are ongoing.

Figure 2

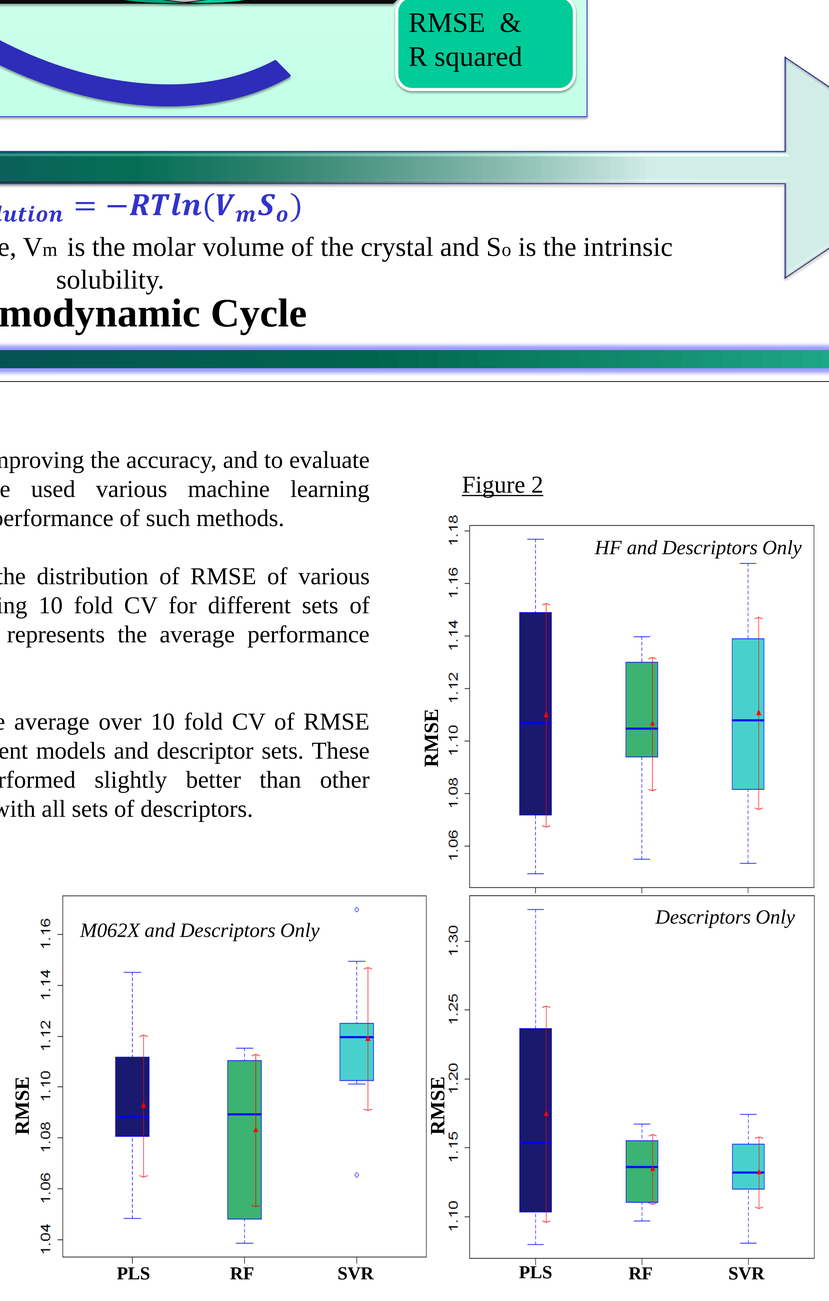


Table 1

Theoretical Methods	DMACRYS + SMD M062X	DMACRYS + SMD HF
RMSE	4.045	2.946
R Squared	0.252	0.327

Table 2(A)

RMSE	M062X & Descriptors	HF & Descriptors	Descriptors Only
PLS	1.093	1.110	1.174
RF	1.086	1.107	1.134
SVR	1.119	1.111	1.132

Table 2(B)

R Squared	M062X & Descriptors	HF & Descriptors	Descriptors Only
PLS	0.595	0.594	0.559
RF	0.602	0.583	0.559
SVR	0.575	0.576	0.559

Acknowledgements

JMcD, TvM and JBOM are grateful for access to the EaStCHEM Research Computing Facility and to Dr Herbert Früchtl for its maintenance. We are all grateful to Dr Graeme Day (University of Southampton) for providing a script to calculate entropies of crystal structures in DMACRYS.

The University of St Andrews is a charity registered in Scotland No: SC013532

References

1. Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V., *Journal of Chemical Theory and Computation*, **8**, 3322 (2012).