Riga Technical University
57th International
Scientific Conference

MSAC 2016

MATERIALS SCIENCE AND APPLIED CHEMISTRY

21st October, 2016, Riga

PROCEEDINGS AND PROGRAMME

RTU Press
Riga 2016

Use of Different Statistical Approaches to Evaluate Performance of Quantum Chemical Computational Methods in Predicting Redox Potentials

Igors MIHAILOVS [1,2], Baiba TUROVSKA [3], Valdis KAMPARS [1], Martins RUTKIS [2]
[1] Riga Technical University, Latvia
[2] University of Latvia, Latvia
[3] Latvian Institute of Organic Synthesis, Latvia

ABSTRACT. The study compares statistical approaches to discriminate significant factors in calculation of redox potentials by the means of quantum chemistry judging from regression and time-related figures of merit. The results are compared with data from effect decomposition/significant effect extraction from a full factorial experiment. Hierarchical cluster analysis is found to perform well, while linear discriminant analysis is somewhat less convenient. Model-based clustering on principal components failed to reproduce factor significance.

KEYWORDS. Classification, Redox potentials, HCA (Hierarchical Cluster Analysis), LDA (Linear Discriminant Analysis)

I. INTRODUCTION

Redox potentials are properties of molecular substance that are of importance in many research areas, such as organic electronics,[1], [2] biochemistry and medical chemistry[3]–[5] and corrosion science. [3], [6] As experimental quantities they, however, require access to appropriate laboratory equipment, skills and, most importantly, need to either synthesize the compounds of interest or extract them from natural materials. Quantum chemical computations are appealing alternative and are widely used to predict redox potentials.[1], [4]–[6] There are, however, not many systematic studies on reliable selection from the wide variety of many computational methods. By reliable selection we mean that based on some statistical approach. Less than ten publications were found during a search in both *Scopus*™ and *Web of Science*™ databases; one of the most reliable was by da Hora, et al.,[7] dealing with selection of best computational method and basis set for optimizing geometry parameters and ionization energy (a quantity close to oxidation potential). This and most other studies focus on representing well absolute values of experimental parameters; however, absolute values are frequently shifted in electrochemical experiment due to various factors.[8] On the other hand, practical applications of redox potentials often require correct trends more than correct absolute values, especially if experimental data is available for some compound in a series studied. Hence, this study focuses on linear regression parameters (slope, intercept and R^2), as well as on computation time per CPU core, which is also of major importance if our goal is computational screening of compounds.

The most reliable way to make the selection was, in our opinion, to perform a full factorial experiment and to decompose resulting values to effects, remove insignificant ones and recalculate back corrected values of all four figures of merit studied.[9] This required us additional coding, as we did not found an *R* package for effect decomposition for the case of multi-level factors. We were, therefore, interested in suggesting some method which would discriminate insignificant factors from just the raw (noisy) data. In our opinion, this could be done via classification of data points in four-dimensional space of slope, intercept, R^2 and time, where each point correspond to a certain

computational parameter combination (a treatment). Widely used tool is principal component analysis (PCA),[7], [10] which is, however, not intended to be used for classification but for reducing data dimensionality. This means all clustering in PCA is merely accidental, on the other hand, some clustering information could be lost. Model-based clustering can be nevertheless performed on all PCA results (not just few first components) to generate clusters in principal components space. Then we can speculate about which cluster contains the point of ideal description. Among methods whose sole purpose is classification, two distinct groups emerge: supervised and unsupervised. Unsupervised classification, or clustering, tries to classify objects without any prior information about them; in this case we again can test to which clusters the point of ideal description is most close. On the other hand, for supervised classification, we should provide training set with certain classes (in our case, certain levels of computational parameters) and then check if clustering obtained separates other data correctly. This also allows to directly check the performance of model by evaluating the relative error of prediction.[10], [11]

II. METHODS

The experimental redox potentials are obtained using a computer-controlled electrochemical system PARSAT 2273 using glassy carbon disk (Ø 0.5 cm) as a working electrode. The measurements were carried out using a three-electrode cell configuration. Saturated calomel electrode (SCE) served as a reference electrode and Pt wire – as an auxiliary electrode. The potential scan rate was 100 mV/sec. Electrochemical redox reactions were studied in deaerated 0,1 M tetrabutylammonium tetrafluorophosphate ($TBAPF_6$) solution in acetonitrile (ACN). ACN (Merck, puriss. grade) was distilled over phosphorus pentoxide, then redistilled over potassium carbonate and stored over 0.4 nm molecular sieves. Reduction data of 14 compounds from 2-benzylidene-1,3-indandione series (with different electronically active substituents in 4' position) were obtained.

All quantum chemical calculations were performed using *Gaussian 09*, rev. D.01 program,[12] and statistical analysis was performed in R (v. 3.3.1). A full factorial design was chosen for the study, spanning six factors:

- computational methodology (4 levels – orbital energy approach, vertical ΔSCF, adiabatic ΔSCF and adiabatic ΔSCF with ZPVC correction), termed **m**(0, m_1, m_2,. m_3),
- method (Hamiltonian) of calculation (3 levels – B3LYP with B3LYP/6-31G(d,p) optimized geometry, CAM-B3LYP with B3LYP/6-31G(d,p) optimized geometry and CAM-B3LYP with CAM-B3LYP/6-31G(d,p) optimized geometry), termed **f**(0, f_1, f_2,),
- solvent modelling (2 levels - no modelling and CPCM solvation model), termed **s**(0, s),
- presence of diffuse functions in basis set (3 levels – no functions, functions on heavy atoms and functions on all atoms), termed **d**(0, d_1, d_2,),
- number of functions in the valence region of the basis set (2 levels – double-zeta and triple-zeta), termed **z**(0, z),
- presence of polarization functions in basis set (2 levels – functions on heavy atoms and functions on all atoms), termed **p**(0, p).

The computed values were then subject to linear regression with experimental and subsequently to effect decomposition according to reference [9]. Figure 1 colour scale for both regression coefficients is based on their variance (also recomputed only from significant effects), for R^2 and time it was chosen arbitrarily. All classifications were performed in the four-dimensional space where the slope, the intercept, the R^2 and computation time, centred and scaled, were the initial dimensions. Classification by model-based clustering from principal components,[13] by hierarchical clustering[14] (HCA; unsupervised method) and by linear and quadratic discriminant analyses[15] (LDA and QDA; supervised methods) was then applied. For discriminant analysis, performance tests[11] were performed by collecting relative error of predictions for the test set which was randomly selected from input data as ca. 30 % of all points, the rest being put into the training set. This was performed 100 times, and the mean ration of erroneously predicted factor level was registered. Both methods achieve similarly low performance (errors about 60–70 % of predictions); nevertheless, histogram plots of

LDA show quite high degree of factor separation, so for qualitative judging these were deemed enough precise.

III. Results and discussion

On Figure 1 we provide a matrix produced from effect-decomposition analysis by recalculating factors from significant effects only. This is what we deem 'gold standard' for the study. Clearly, the solvent factor is dominant in the all plots but the time plot, indicating that accounting for solvent in computations is not only crucial but also very cheap. Next significant factor is the computational methodology: here adiabatic ΔSCF performs well (seemingly good performance of vertical ΔSCF can be explained by error cancellation). Method used is also significant: results for CAM-B3LYP are notably better than for B3LYP, although there is almost no difference in which geometry should be used (levels 2 and 3). Diffuse function usage does not strongly affect the performance of methods, but strongly increases computation time; therefore, the effect of this factor is negative. Rest two factors are of little significance.

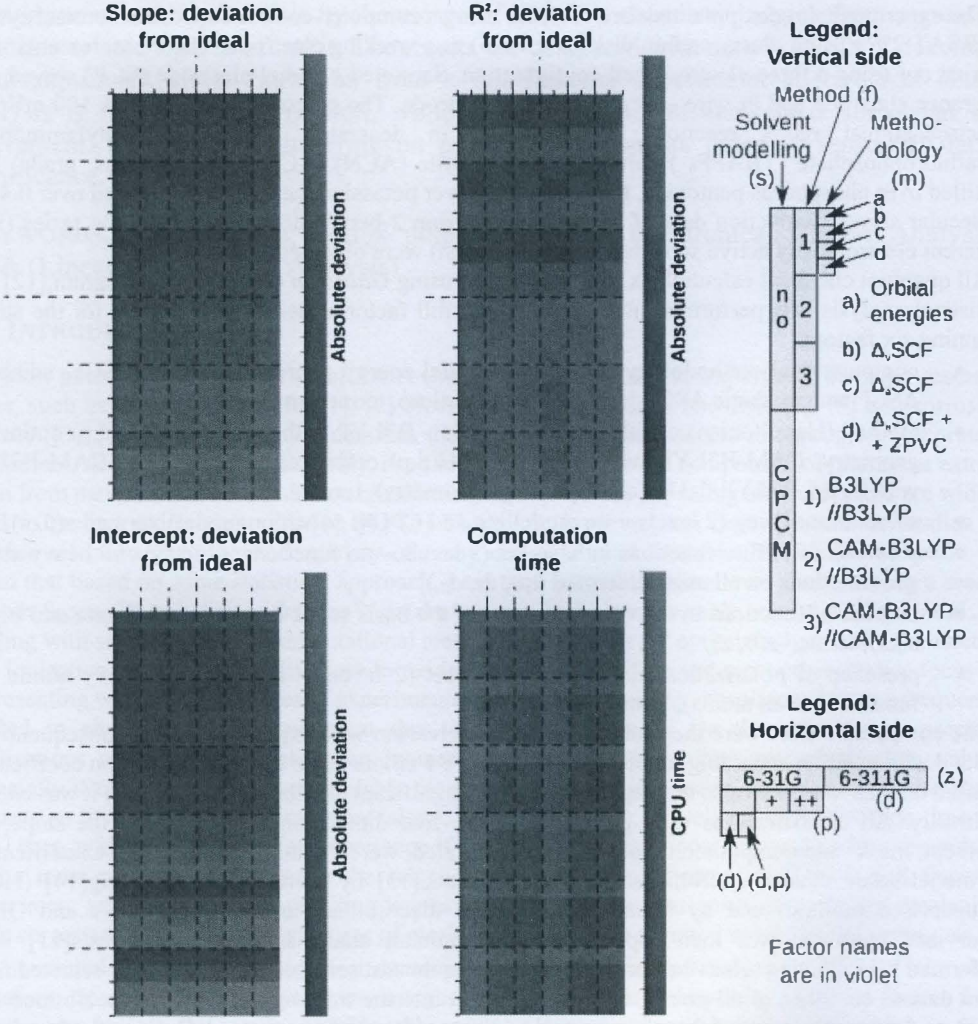

Fig. 1. Results of effect decomposition and treatment recalculation for reduction potential–calculated electron affinity correlations. Reddish areas correspond to less efficient description.

TABLE I

MODEL-BASED CLUSTERING RESULTS

No. of cluster	Level occurrence in cluster (times)	No. of cluster	Level occurrence in cluster (times)
1	24 (p, z), 16 (f_1, f_2), 12 (d_1, d_2, m_1)	6	24 s, 12 (p, z, f_1, f_2), 8 (d_1, d_2)
2	12 (p, z, m_2, m_3)	7	16 s, 8 (p, z, f_1, f_2, m_2, m_3)
3	24 m_1, 12 (p, z, d_1, d_2), 8 (f_1, f_2)	8	24 (s, m_1), 12 (p, z, d_1, d_2), 8 (f_1, f_2)
4	12 (p, d_1, d_2, m_2, m_3), 8 (f_1, f_2)	9	52 s, 28 z, 26 (p, m_2, m_3), 24 (d_1, d_2), 16 (f_1, f_2)
5	Ideal case, 28 s, 14 p, 12 (z, m_1), 4 (d_1, d_2, f_1, f_2), 2 (m_2, m_3)	10	24 z, 12 (p, d_1, d_2, m_2, m_3), 8 (f_1, f_2)

Now we continue with classification analyses. First, we analyze results for model-based clustering. The program selected VVV model (variable volume and shape, ellipsoidal), based on Bayesian information criterion. This resulted in 10 clusters, and assignments are listed in the table 1 below.

The ideal case (slope and R^2 equal to 1 and intercept and computation time – to 0) is assigned to the cluster 5. One of the main features of this cluster is relatively frequent occurrence of top level of **s** factor (what is correct); however, there are clusters with even higher frequency for it, so there might be notable interactions. The last conclusion is actually quite wrong, if we return to the standard analysis on Figure 1. Next, cluster 5 also contains less treatments with diffuse functions (d_1 and d_2 levels), which is again true. Nevertheless, there are also quite few appearances of levels with CAM-B3LYP, which is wrong. Among methodologies, adiabatic ΔSCF is clearly depreciated, which is again wrong. Polarization factors are spread quite uniformly over clusters, and that is in accord with the results of effect decomposition analysis. Additionally, the same could be told about **z**, and this is to no surprise, also does not show serious improvement due to this factor. The net conclusion is that model-based clustering is not robust enough for discriminating between significant and insignificant factors, as it judges incorrectly such important factors as methodology and method in use.

Results of hierarchical clustering analysis are available on Figure 3. Among various methods, only Ward's one has managed to produce separation when the 'ideal' case is incorporated in some cluster other than its own one (what results in lamp-like plot). Results for Euclidean and for Mahalanobis distances are quite similar, for the first one somewhat more clear, so presented here. All slices contain results of the same analysis, but data points are classified by colouring according to levels of a specific factor.

In this case, most strongly emerges the solvent factor, and clearly the effect is the same as predicted by effect decomposition. The next optimistic result is that methodology is also correctly classified, with adiabatic ΔSCF situated most closely to the 'ideal' case. It is quite educative to note that in fact the best method is also more or less correctly pointed out, despite that the 'ideal' case seems to lie in the B3LYP-coloured region – the distances between clusters must be measured in vertical direction, and CAM-B3LYP dominates here, with little differences between geometries. Diffuse functions are recommended not to be used (correctly), and for valence and, especially, polarization functions there are no clear dominance of either variant, so the factors are considered insignificant. To sum up, hierarchical clustering performs well in distinguishing valuable and invaluable parameters.

What corresponds to linear and quadratic discriminant analyses, they produced results generally similar to those of HCA. Histogram plots for LDA are shown on Figure 2. Level 0 of **m** factor is distinguished better on plot for second discriminant function (not shown here). This makes the method somewhat less convenient for discriminating factor significance, but its overall performance is also deemed good.

IV. CONCLUSIONS

The study concludes that HCA is well-suited for preliminary analysis of factor importance in redox potential calculations. LDA is also performing well, whereas model-based clustering based on principal components failed to reproduce factor effects on computation–experimental trends figures of merit.

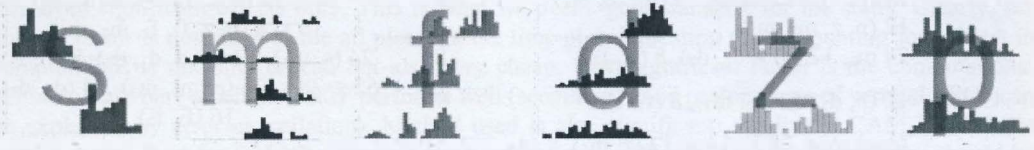

Fig. 2. LDA histogram plots for different levels of factors. Factors with more distant distribution centres are judged to be more significant.

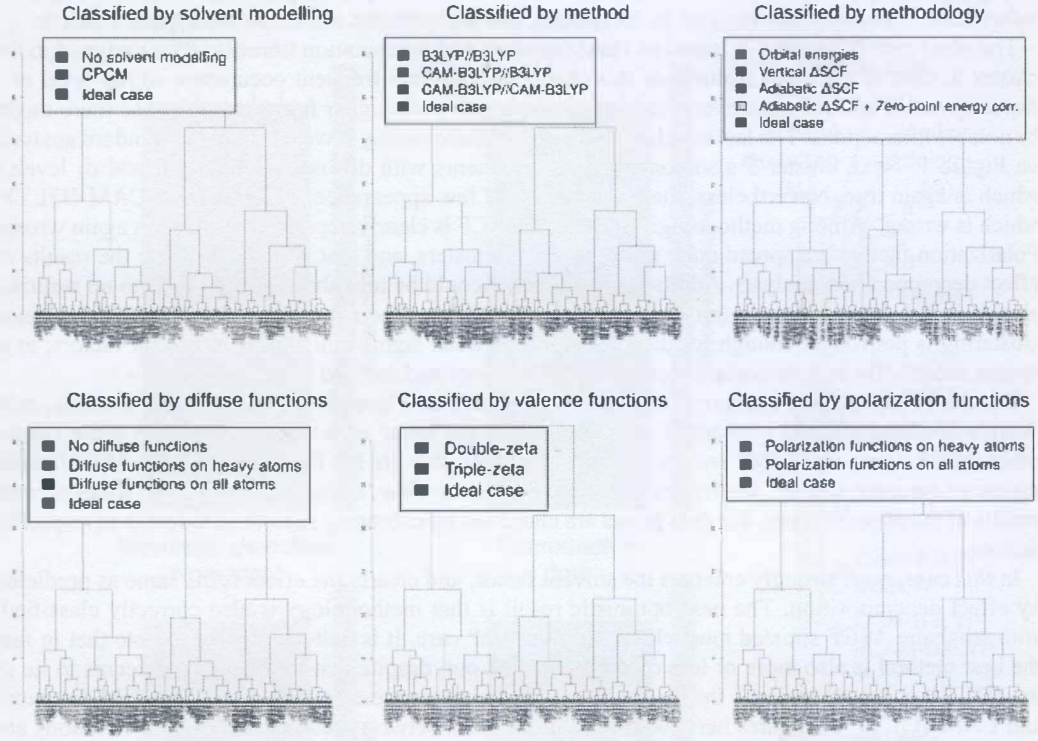

Fig. 3. Results of hierarchical clustering analysis, coloured by different factors. Depending on the position of 'Ideal' case point, factors from the top row, as well as the leftmost factor in the bottom row are considered important.

ACKNOWLEDGEMENT

This work has been supported by the National Research Program of Latvia in Materials Science "MultIfunctional Materials and composItes, photonicS and nanotechnology (IMIS2)".

REFERENCES

[1] P. E. Schwenn, P. L. Burn, and B. J. Powell, "Calculation of solid state molecular ionisation energies and electron affinities for organic semiconductors," *Org. Electron.*, vol. 12, no. 2, pp. 394–403, 2011 [Online]. Available: http://dx.doi.org/10.1016/j.orgel.2010.11.025

[2] J.-L. Bredas, "Mind the gap!," *Mater. Horizons*, vol. 1, no. 1, pp. 17–19, 2014 [Online]. Available: http://dx.doi.org/10.1039/C3MH00098B\nhttp://xlink.rsc.org/?DOI=C3MH00098B

[3] M. D. Archer, "Fundamentals and applications in electron-transfer reactions," in *Nanostructured and Photoelectrochemical Systems for Solar Photon Conversion*, M. D. Archer and A. J. Nozik, Eds. Singapore: World Scientific, 2008, pp. 209–274. [Online]. Available: http://dx.doi.org/10.1142/p217

[4] E. D. Raczyńska, "Quantum-chemical studies of the consequences of one-electron oxidation and one-electron reduction for imidazole in the gas phase and water," *Comput. Theor. Chem.*, vol. 993, pp. 73–79, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.comptc.2012.05.036

[5] K. Tang, H.-T. Sun, Z.-Y. Zhou, and Z.-Z. Wang, "Density Functional Theory Study on the Ionization Potentials and Electron Affinities of Adenine–Formamide Complexes," *J. Theor. Comput. Chem.*, vol. 8, no. 2, pp. 187–201, 2009.[Online]. Available: http://dx.doi.org/10.1142/S0219633609004733

[6] L. T. Sein, Y. Wei, and S. A. Jansen, "Corrosion inhibition by aniline oligomers through charge transfer: a DFT approach," *Synth. Met.*, vol. 143, no. 1, pp. 1–12, May 2004 [Online]. Available: http://dx.doi.org/10.1016/j.synthmet.2002.06.002

[7] G. C. A. da Hora, R. L. Longo, and J. B. P. da Silva, "Calculations of structures and reaction energy profiles of As2O3 and As4O6 species by quantum chemical methods," *Int. J. Quantum Chem.*, vol. 112, no. 20, pp. 3320–3324, 2012 [Online]. Available: http://doi.wiley.com/10.1002/qua.24196

[8] R. Rybakiewicz, P. Gawrys, D. Tsikritzis, K. Emmanouil, S. Kennou, M. Zagorska, and A. Pron, "Electronic properties of semiconducting naphthalene bisimide derivatives - Ultraviolet photoelectron spectroscopy versus electrochemistry," *Electrochim. Acta*, vol. 96, pp. 13–17, 2013 [Online]. Available: http://dx.doi.org/10.1016/j.electacta.2013.02.041

[9] R. Mead, S. G. Gilmour, and A. Mead, *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. New York: Cambridge University Press, 2012.

[10] T. Næs, T. Isaksson, T. Fearn, and T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester: NIR Publications, 2002.

[11] G. M. Venturini, "Discriminant Analysis in R," *RPubs*, 2014. [Online]. Available: http://www.rpubs.com/gabrielmartos/discriminantR . [Accessed: 21-Aug-2016]

[12] M. J. Frisch, G. W. Trucks, H. B. Schlegel, et al., "Gaussian 09, Revision D.01." Gaussian, Inc., Wallingford, CT, 2009.

[13] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, "mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation," Washington, D.C., 2012.

[14] F. Murtagh and P. Legendre, "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," *J. Classif.*, vol. 31, no. 3, pp. 274–295, Oct. 2014 [Online]. Available: http://link.springer.com/10.1007/s00357-014-9161-z

[15] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. Springer, 2002.

Igors Mihailovs (Mg. chem., 2015) is a PhD student at the Faculty of Materials Science and Applied Chemistry of Riga Technical University and engineer at the Laboratory of Organic Materials of Institute of Solid State Physics, University of Latvia. He performs quantum chemical computations of molecules at the Laboratory and is interested in applying statistical methods to result analysis. E-mail: igorsm@cfi.lu.lv

Baiba Turovska (Dr. chem., 1989) is a researcher at Latvian Institute of Organic Synthesis. She is member of New York Academy of Sciences, Electrochemical Society and International Society of Electrochemistry. Her main interests are physical organic chemistry and electrochemistry. E-mail: turovska@osi.lv

Valdis Kampars (Dr. habil. chem., 1983) is the director of the Institute of Applied Chemistry and leading researcher at the Faculty of Materials Science and Applied Chemistry of Riga Technical University, as well as Secretary General of the Latvian Academy of Sciences. He has broad scientific interests spanning from fuel science to chemistry of organic materials for electronics and photonics. E-mail: kampars@ktf.rtu.lv, ORCID ID: 0000-0001-5490-8928.

Martins Rutkis (Dr. phys., 1992) is the director and leading researcher at the Institute of Solid State Physics, University of Latvia. He is member of Optical Society of America, American Chemical Society, Royal Society of Chemistry and a corresponding member of Latvian Academy of Sciences. His main research interests are optics, photonics and organic electronics. E-mail: martins.rutkis@cfi.lu.lv, ORCID ID: 0000-0002-5929-2031.