J Math Chem (2013) 51:1684–1693
DOI 10.1007/s10910-013-0177-0

BRIEF COMMUNICATION

# CORAL: QSPRs of enthalpies of formation of organometallic compounds

A. P. Toropova · A. A. Toropov · E. Benfenati ·
G. Gini · D. Leszczynska · J. Leszczynski

Received: 11 February 2013 / Accepted: 28 March 2013 / Published online: 10 April 2013

**Abstract** Available on the Internet the CORAL software gives reasonable good prediction for standard enthalpy of formation for selected organometallic compounds (n = 132). The approach is tested using five random splits of the considered data into the sub-training set (n = 32–49), calibration set (n = 36–51), test set (n = 10–29), and the validation set (n = 22–41). Compounds of the validation set are not involved in building up the models. The average statistical quality of prediction is the following: correlation coefficient ($\overline{R^2}$) 0.991 ± 0.005 and standard error of estimation ($\overline{s}$) 22.9 ± 5.6 kJ/mol.

**Keywords** QSPR · Organometallic compound · Standard enthalpy · CORAL software

A. P. Toropova · A. A. Toropov (✉) · E. Benfenati
Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via La Masa 19, 20156 Milan, Italy
e-mail: aatoropov@yahoo.com
e-mail: andrey.toropov@marionegri.it

G. Gini
Department of Electronics and Information, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milan, Italy

D. Leszczynska
Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering,
Jackson State University, 1325 Lynch St, Jackson, MS 39217-0510, USA

J. Leszczynski
Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry,
Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

## 1 Introduction

The standard enthalpy of formation, $\Delta_f H_0$, represents a basic property of any chemical compound. It provides information about compound's thermodynamic stability and facilitates the calculation of enthalpies of reaction [1]. Quantitative structure–property/activity relationships (QSPRs/QSARs) are a tool to estimate the standard enthalpy for substances which were not studied in the experiment [2–5]. Optimal descriptors calculated by the Monte Carlo method with the representation of the molecular structure by simplified molecular input-line entry system (SMILES) [6] and SMiles ARbitrary Target Specification (SMART) [7] also were used as a tool for QSPR prediction of standard enthalpy of organometallic compounds. The CORAL software [8] involves building up of optimal descriptors by the Monte Carlo method. The validation of a model becomes crucial component of the QSPR/QSAR analyses [9]. The aim of the present study is the estimation of statistical quality of QSPRs for the standard enthalpy of formation from elements of organometallic compounds. It is accomplished for few series of random splits of available data into the training set, the calibration set, the test set, and the validation set.

## 2 Method

### 2.1 Data

The values of the gas-phase enthalpies of formation (in kJ/mol) of the organometallic compounds (n = 132) have been taken from the literature [1]. Using five different schemes these compounds were split into the sub-training set, the calibration set, the test set, and the validation set. The details of these splits were selected and executed according to three principles: (i) the majority of molecular features (i.e. SMILES attributes) should be placed in the sub-training set and in the calibration set; (ii) the test set (as well as the validation set) should contain the minimum of rare attributes; and (iii) splits should not be identical. The measure of identity for a pair of splits can be estimated by the formula:

$$Identity_{i,j} = \frac{Nset_{i,j}}{0.5 \times (Nset_i + Nset_j)} \times 100\,\% \tag{1}$$

where set = (sub-training, calibration, test, and validation); $Nset_{i,j}$ is the number of identical substances in the set for ith and jth splits; $Nset_i$ and $Nset_j$ are the total numbers of substances in the set for ith and jth splits, respectively.

Table 1 contains data on the identity of five splits into the sub-training, calibration, test, and validation sets.

### 2.2 Optimal descriptors

The SMILES-based optimal descriptors are calculated as the following:

$$DCW(Threshold, N_{epoch}) = \sum CW(SA_k) \tag{2}$$

**Table 1** The values of identity (%) for all pairs of five splits into sub-training, calibration, test, and validation sets

| | Set | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | Sub-training | 100 | 26 | 26 | 27 | 26 |
| | Calibration | 100 | 37 | 35 | 41 | 35 |
| | Test | 100 | 20 | 27 | 11 | 30 |
| | Validation | 100 | 17 | 25 | 9 | 33 |
| 2 | Sub-training | | 100 | 33 | 42 | 40 |
| | Calibration | | 100 | 23 | 30 | 29 |
| | Test | | 100 | 12 | 24 | 12 |
| | Validation | | 100 | 8 | 13 | 9 |
| 3 | Sub-training | | | 100 | 22 | 29 |
| | Calibration | | | 100 | 22 | 41 |
| | Test | | | 100 | 13 | 8 |
| | Validation | | | 100 | 8 | 7 |
| 4 | Sub-training | | | | 100 | 33 |
| | Calibration | | | | 100 | 27 |
| | Test | | | | 100 | 21 |
| | Validation | | | | 100 | 16 |
| 5 | Sub-training | | | | | 100 |
| | Calibration | | | | | 100 |
| | Test | | | | | 100 |
| | Validation | | | | | 100 |

where $SA_{\mathrm{k}}$ is a SMILES attribute i.e. a group of symbols which cannot be examined separately, e.g. ‘c’, ‘=’, ‘Cl’, [Si], etc.; $CW(SA_k)$ is correlation weight of $SA_k$, the values of correlation weights are calculated with the Monte Carlo optimization procedure that provides maximum value of the target function calculated as:

$$TF = R + R' - \left|R - R'\right| \times W_R - \left|C_1 - C_1'\right| \times W_C - C_0 - C_0' \qquad (3)$$

$R$ and $R'$ are correlation coefficient for sub-training and calibration sets, respectively; $C_0$, $C_0'$, $C_1$, and $C_1'$ are regression coefficients for sub-training and calibration sets, respectively; $W_R = 0.1$ and $W_C = 0.01$ are empirical constants;

*Threshold* is represented by an integer coefficient for classification of SMILES attributes into two categories rare (noise) and not rare (active). Correlation weights of rare SMILES attributes are fixed to be equal to zero, i.e. they have not influence on a model. Threshold values 1 and 2 were examined in this study (if selected threshold is 1, then attributes which appears in the sub-training at least one times should be involved in the building up of a model);

$N_{epoch}$ is the number of iteration of the Monte Carlo optimization, one iteration represents a variation of all SMILES attributes taken in random sequence [8]. In this study $N_{epoch} = 200$ is used.

**Table 2** List of correlation weights for calculation of *DCW(1,200)* calculated by the Monte Carlo optimization with target function that is calculated with Eq. 3

| $\boldsymbol{SA}_k$ | $\boldsymbol{CW}(\boldsymbol{SA}_k)$ | The number of $\boldsymbol{SA}_k$ in the sub-training set | The number of $\boldsymbol{SA}_k$ in the calibration set | The number of $\boldsymbol{SA}_k$ in the test set |
|---|---|---|---|---|
| # | 4.07900 | 1 | 1 | 0 |
| ( | −0.08425 | 33 | 36 | 18 |
| 1 | 0.10925 | 7 | 11 | 4 |
| 2 | 0.08350 | 6 | 11 | 4 |
| 3 | 0.02700 | 4 | 9 | 2 |
| 4 | 0.0 | 0 | 3 | 0 |
| = | 0.0 | 0 | 3 | 1 |
| B | −1.01975 | 4 | 6 | 4 |
| C ($sp^3$) | −0.37700 | 33 | 37 | 26 |
| Br | −2.83550 | 3 | 2 | 2 |
| I | −1.61025 | 2 | 2 | 3 |
| Cl | −3.44075 | 7 | 4 | 2 |
| O | 0.0 | 0 | 0 | 1 |
| P | −0.87400 | 1 | 1 | 0 |
| [Al] | −0.16775 | 2 | 1 | 1 |
| [Bi] | 5.16150 | 2 | 2 | 0 |
| [As] | 0.0 | 0 | 1 | 0 |
| [Ga] | 0.32600 | 3 | 2 | 0 |
| [Ge] | −0.28650 | 1 | 3 | 2 |
| [Hg] | 2.49175 | 5 | 4 | 9 |
| [Pb] | 4.07825 | 2 | 3 | 1 |
| [Sb] | 2.35850 | 4 | 0 | 1 |
| [Se] | 0.32200 | 2 | 2 | 2 |
| [SiH] | 0.0 | 0 | 1 | 0 |
| [Si] | −1.96150 | 3 | 0 | 0 |
| [Sn] | 1.86675 | 6 | 15 | 4 |
| [Te] | 1.20625 | 1 | 1 | 5 |
| c ($sp^2$) | 0.37175 | 6 | 11 | 4 |

Table 2 contains example of the correlation weights (split1, $Threshold = 1$, $N_{epoch} = 200$). Table 3 contains an example of calculation of DCW(1,200). When the correlation weights which provide maximum values for the target function calculated with Eq. 2 are predicted, one can calculate standard enthalpy of formation applying the model:

$$\Delta_f H_0 = C_0 + C_1 \times DCW(Threshold, N_{epoch}) \tag{4}$$

Apparently, the predictive ability of the model calculated with Eq. 4 should be tested using the external set.

**Table 3** Example of calculation of the $DCW(Threshold, N_{epoch})$ with the correlation weights calculated by the Monte Carlo method (Table 2) SMILES = CC(C)C[Al](CC(C)C)CC(C)C $DCW(1{,}200) = -5.366$

| $(SA_\mathbf{k})$ | $CW(SA_\mathbf{k})$ |
|---|---|
| C | −0.3770 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |
| [Al] | −0.1677 |
| ( | −0.0843 |
| C | −0.3770 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |
| ( | −0.0843 |
| C | −0.3770 |

## 3 Results and discussion

This study has been carried out according to principle: the analysis of series of the QSPR/QSAR models which are obtained with various splits into the training set (sub-training and calibration sets) and test set is able to provide more reliable data on the predictive potential of an approach than the only one model based on the only one split.

Table 4 contains statistical characteristics of models developed for five random splits. One can see that statistical quality of these models is similar and quite good. Preferable threshold for all models is 1. Table 5 contains the statistical quality of models calculated with *Threshold*=1 and the $N_{epoch} = 200$. The test set was not involved in the building up of the models. However, the test set has been involved in the selection of the optimal threshold value. Consequently, the additional tests of the approach should be done with the external validation set (Table 5). One can see (Table 5) that these models calculated with sub-training set and calibration set are good for both the test set and validation set.

In the case of split1 the model is the following (Fig. 1):

$$\Delta_f H_0[kJ/mol] = 18.0238(\pm 0.6737) + 50.5631(\pm 0.1306) \times DCW(1{,}200) \tag{5}$$

**Table 4** The average statistical quality (three runs of the Monte Carlo method optimization with target function that is calculated with Eq. 3) of models for enthalpy of formation from elements of organometallic compounds for five splits: one can see that the preferable threshold is 1 for all five splits

| Split | Threshold | $N^{*}_{act}$ | $N^{**}_{model}$ | Sub-training set | | | | Calibration set | | | Test set | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | n | $\overline{R^2}$ | $\bar{s}$, kJ/mol | $\overline{F}$ | n | $\overline{R^2}$ | $\overline{s}$, kJ/mol | n | $\overline{R^2}$ | $\overline{s}$, kJ/mol |
| 1 | 1 | 23 | 107 | 36 | 0.9916 | 24.1 | 4,035 | 42 | 0.9870 | 36.3 | 29 | 0.9957 | 22.5 |
| | 2 | 19 | | 36 | 0.9633 | 50.5 | 892 | 42 | 0.9654 | 61.5 | 29 | 0.9540 | 66.6 |
| 2 | 1 | 28 | 110 | 49 | 0.9941 | 20.2 | 7,932 | 51 | 0.9941 | 24.7 | 10 | 0.9985 | 34.0 |
| | 2 | 20 | | 49 | 0.9819 | 35.4 | 2,548 | 51 | 0.9818 | 37.8 | 10 | 0.9749 | 58.7 |
| 3 | 1 | 26 | 102 | 42 | 0.9946 | 22.4 | 7,344 | 37 | 0.9946 | 23.1 | 23 | 0.9680 | 33.0 |
| | 2 | 22 | | 42 | 0.9802 | 42.7 | 1,985 | 37 | 0.9606 | 56.9 | 23 | 0.8669 | 61.4 |
| 4 | 1 | 21 | 91 | 32 | 0.9933 | 22.5 | 4,448 | 36 | 0.9810 | 44.6 | 23 | 0.9873 | 35.6 |
| | 2 | 17 | | 32 | 0.9586 | 56.0 | 694 | 36 | 0.9551 | 77.7 | 23 | 0.9714 | 76.8 |
| 5 | 1 | 24 | 109 | 40 | 0.9945 | 23.3 | 6,861 | 45 | 0.9898 | 26.7 | 24 | 0.9912 | 25.7 |
| | 2 | 19 | | 40 | 0.9832 | 40.6 | 2,223 | 45 | 0.9510 | 59.4 | 24 | 0.9243 | 79.1 |

$^{*}$) $\mathbf{N_{act}}$ is the number of SMILES attributes which are involved in a model
$^{**}$) $\mathbf{N_{model}}$ is the number of compounds which are involved in the building up of the model (i.e. compounds placed in the sub-training set, the calibration set, and the test set, but not in the validation set)

**Table 5** Statistical quality of models with external validation

| Split | $N_{act}$ | Sub-training set | | | | Calibration set | | | Test set | | | Validation set | | | Prediction for Test settogether with validation set | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | n | $R^2$ | S, kJ/mol | F | n | $R^2$ | S, kJ/mol | n | $R^2$ | S, kJ/mol | n | $R^2$ | S, kJ/mol | n | $R^2$ | S, kJ/mol |
| 1 | 23 | 36 | 0.9916 | 24.1 | 4033 | 42 | 0.9870 | 36.3 | 29 | 0.9958 | 22.5 | 25 | 0.9927 | 23.9 | 54 | 0.9933 | 23.0 |
| 2 | 28 | 49 | 0.9941 | 20.1 | 7959 | 51 | 0.9941 | 24.7 | 10 | 0.9986 | 33.7 | 22 | 0.9972 | 17.2 | 32 | 0.9960 | 22.8 |
| 3 | 26 | 42 | 0.9946 | 22.4 | 7343 | 37 | 0.9946 | 23.0 | 23 | 0.9682 | 33.0 | 30 | 0.9907 | 24.5 | 53 | 0.9846 | 28.1 |
| 4 | 21 | 32 | 0.9933 | 22.5 | 4443 | 36 | 0.9810 | 44.4 | 23 | 0.9873 | 35.5 | 41 | 0.9821 | 32.0 | 64 | 0.9837 | 33.0 |
| 5 | 24 | 40 | 0.9945 | 23.2 | 6881 | 45 | 0.9898 | 26.6 | 24 | 0.9913 | 25.6 | 23 | 0.9929 | 16.9 | 47 | 0.9918 | 21.4 |

Threshold is 1. The number of epochs is 200

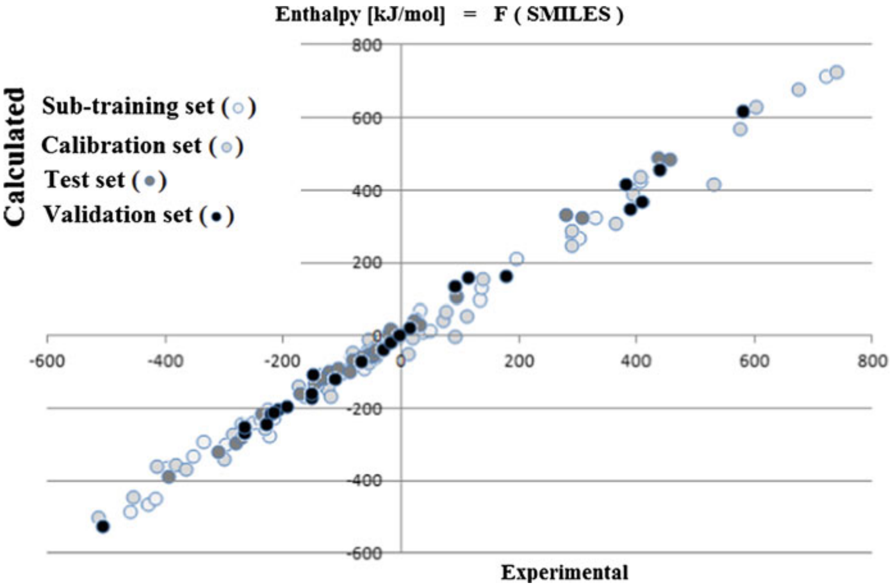

**Fig. 1** The representation of a model that is calculated with Eq. 5

n = 36, $R^2$ = 0.9916, s = 24.1 kJ/mol, F = 4033 (sub-training set);
n = 42, $R^2$ = 0.9870, s = 36.3 kJ/mol (calibration set);
n = 29, $R^2$ = 0.9958, s = 22.5 kJ/mol (test set);
n = 25, $R^2$ = 0.9927, s = 23.9 kJ/mol (validation set);

It is obvious from the data in Table 2 that prevalence of various attributes is different. One can define three groups of attributes. The first group including: bracket that is indicator of the branching in the molecular skeleton, carbon in $sp^3$state, and tin [Sn] has the highest prevalence. The second group (‘1’, ‘2’, and ‘3’ which are indicators of rings; chlorine; ‘[Hg]’; ‘[Ge]’; and carbon in $sp^2$ state) has middle prevalence. The third group (all other attributes) has the lowest prevalence. Thus, the molecular features represented by SMILES attributes of the first and second groups may be qualified as the basis for the definition of the applicability domain. We suggest that SMILES which contains more than 90 % of such attributes should be examined as the applicability domain.

Table 6 contains comparison of models developed for the 132 organometallic compounds taken from the literature. Apparently, the realistic comparison of various approaches can be done if each approach is tested for a series of the splits. However, as a rule, such comparison is not available. The statistical quality of models calculated with the CORAL software for splits: #1, #2, #3, and #5 is better than for models published in the literature. In the case of split #4, the statistical quality of the model is poorer, but, in this case, the number of compounds in the test set and validation set is maximal, and consequently, the minimum number of organometallic compounds was involved in the building up of this model.

The details of studied five splits are available on the Internet [8]. Having downloaded CORALSEA.zip, one can repeat computational experiments described in the

**Table 6** Comparison of statistical quality of the prediction for the enthalpy of formation of organometallic compounds

| The number of compounds in the external set | Correlation coefficient, $R^2$ | Standard error of estimation, s, kJ/mol | Reference |
|---|---|---|---|
| 28 | 0.990 | 30.2 | [1] |
| 28 | 0.991 | 29.4 | [6] |
| 28 | 0.991 | 29.4 | [7] |
| Validation set 22–41 | 0.991 ± 0.005 | 22.9 ± 5.6 | In this study |
| Test set and validation set 32–64 | 0.990 ± 0.005 | 25.7 ± 4.3 | |

folder "(7)-Metals-and-Ions", using #Enthalpy-2012(1).txt, #Enthalpy-2012(2).txt, …, #Enthalpy-2012(5).txt for the building up a model and then, using files input.txt, input2.txt, …, input5.txt for the validation of these models [8].

Finally, it should be noted that the CORAL software [8] can be used for the QSAR analysis of various other endpoints [10–13].

## 4 Conclusions

The application of CORAL software allows developing reasonable good models for standard enthalpy of formation from elements for 132 organometallic compounds. The SMILES attributes which are representation of (a) molecular branching; (b) carbon in $sp^3$ state; (c) rings; (d) tin, mercury, and germanium have the highest influence for models of enthalpies of examined organometallic compounds, since they are characterized by maximal prevalence. The statistical quality of models which are calculated with the CORAL software depends on details of a split of available data into the training, calibration, test, and validation sets (Table 5).

**Acknowledgments** We thank CALEIDOS (the project number LIFE11-INV/IT 00295), and the National Science Foundation (NSF/CREST HRD-0833178, and EPSCoR Award #:362492-190200-01\ NSFEPS-090378) for financial support.

## References

1. J. Jover, R. Bosque, J.A. Martinho Simoes, J. Sales, J. Organomet. Chem. **693**, 1261–1268 (2008)
2. K. Roy, I. Mitra, Curr. Comput. Aided Drug Des. **8**, 135–158 (2012)
3. E.A. Castro, F.M. Fernández, P.R. Duchowicz, J. Math. Chem. **37**, 433–441 (2005)
4. O. Ivanciuc, T. Ivanciuc, D.J. Klein, W.A. Seitz, A.T. Balaban, J. Chem. Inf. Comput. Sci. **41**, 536–549 (2001)
5. E. Estrada, L. Torres, L. Rodríguez, I. Gutman, Indian J. Chem. Sect. A **37**, 849–855 (1998)
6. A.A. Toropov, A.P. Toropova, E. Benfenati, Chem. Phys. Lett. **461**, 343–347 (2008)
7. A.A. Toropov, A.P. Toropova, E. Benfenati, A. Manganaro, J. Comput. Chem. **30**, 2576–2582 (2009)
8. CORAL, http://www.insilico.eu/coral/
9. A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, O. Igglessi-Markopoulou, G. Kollias, Mol. Divers. **14**, 225–235 (2010)

10. J.C. Garro Martinez, P.R. Duchowicz, M.R. Estrada, G.N. Zamarbide, E.A. Castro, Int. J. Mol. Sci. **12**, 9354–9368 (2011)
11. J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, Graph. Model. **31**, 10–19 (2011)
12. L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, Chemometr. Intell. Lab. **107**, 269–275 (2011)
13. E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, Chemometr. Intell. Lab. **110**, 81–88 (2012)